Coevolutionary history of ERVs and Perissodactyls inferred from the retroviral fossil record

Submitted in fulfillment of the requirements for the Degree of Doctor of Philosophy in Infection and Immunity

Institute of Infection, Immunity & Inflammation

College of Medical, Veterinary & Life Sciences University of Glasgow

01 April 2018

Abstract

The horse (*Equus caballus*) is an economically and scientifically important species of mammal. The horse genome (and that of other mammals) contains thousands of sequences derived from retroviruses, called endogenous retroviruses (ERVs). These sequences are highly informative about the long-term interactions of retroviruses and hosts. They are also interesting because they have influenced the evolution of mammalian genomes in various ways.

Horses belong to the family *Equidae* in the order *Perissodactyla* - comprising 16 extant species of strict herbivores adapted for running and dietary specialisation. This PhD thesis describes my work developing and applying a novel bioinformatics approach for characterising ERVs. I used this approach to characterise ERVs in genomes of *Hippomorpha* species in relation to those found in a representative of the *Ceratomorpha* - the white rhinoceros (*Ceratotherium simum*).

Through comparative analysis of these three genomes, I derive a calibrated timeline describing the process through which ERV diversity has been generated in the equine germline. My project has provided an overview of retrotranspositional activity in perissodactyl ERV lineages and identified individual ERV loci that show evidence of involvement in physiological processes and/or pathological conditions. The dataset generated in this project will be of great utility for future studies aiming to investigate the potential functional roles of equine ERVs and their impact on equine evolution.

Table of Contents

List of Tables

List of Figures

Acknowledgement

I would like to express my greatest gratitude and appreciation to both my supervisors - Dr Pablo R. Murcia and Dr Robert J. Gifford for giving me the opportunity to work between your labs, for your continuous support of my PhD study, for your patience, motivation, and immense knowledge. Without your help, I would never make it.

Besides my supervisors, I would like to thank Dr Joseph Hughes and Dr Quan Gu. Thank you so much for your guidance, encouragement and advice during my study. And with sincere thanks to my colleges Dr Caroline Chauché, Dr Joanna Crispell, Dr Tristan Dennis, and Dr Yi Jin, for their insightful comments and encouragement. Also, I would like to give special thanks to Dr Gaelle Gross for taking care of us. In the end, I am enormously delighted to everyone in Murcia group, Gifford group and Bioinformatics group.

To the people without whom I would not be here: to mon and dad, I love you both so much; to Lingling Chen, my lovely fiancée, I finally made it and it time to begin our new life together.

Author's Declaration

I, Henan Zhu, declare that, except where explicit reference is made to the contribution of others, that this dissertation is the result of my own work and has not been submitted for any other degree at the University of Glasgow or any other institution.

Printed Name: Henan Zhu

Signature: _____

Abbreviations

AIDS	Immunodeficiency syndrome
ALV	Avian leucosis virus
ALV-J	Avian leukaemia virus type J
BERV	Bovine endogenous retrovirus
BFV	Bovine foamy virus
BIV	Bovine immunodeficiency virus
BLAST	Basic local alignment search tool
BLV	Bovine leukaemia virus
CA	Capsid
CERV	Chimpanzee endogenous retrovirus
ChiRV	Chicken retrovirus
CoeEFV	Coelacanth enogenous foamy virus
CSV	Comma-separated values
DIGS	Database-integrated genome screening
DNA	Deoxyribonucleic acid
DU	Dutpase
E.caballus	*Equus caballus*
E.derm	Equine dermis cell line
EFV	Equine foamy virus
EIAV	Equine infectious anemia virus
EqERV.b1	Equine endogenous retrovirus beta1
EMBL	European Molecular Biology Laboratory
EMBOSS	European Molecular Biology Open Software Suite
ENA	European nucleotide archive
ENTV	Enzootic nasal tumor virus
ERVAP	Endogenous retrovirus annotation pipeline
ERVs	Endogenous retroviruses
EVEs	Endogenous viral elements
FeLV	Feline leukaemia virus
FFV	Feline foamy virus
FIV	Feline immunodeficiency virus
GaLV	Gibbon ape leukaemia virus
GLUE	Genes Linked by Underlying Evolution
GtRNAdb	Genomic trna Database
HERV	Human endogenous retrovirus
HIV-1	Human immunodeficiency virus 1
HIV-2	Human immunodeficiency virus 2
HMM	Hidden Markov models
HTLV-1	Human T-lymphotropic virus 1
ICTV	International Committee on Taxonomy of Viruses
IN	Integrase

JSRV	Jaagsiekte sheep retroviruses
KERV	Kangaroo endogenous retrovirus
KoRV	Koala retrovirus
KwERV	Killer whale endogenous retrovirus
LPDV	Lymphoproliferative disease virus
LTRs	Long terminal repeats
MA	Matrix
MDEV	Mus dunni endogenous retrovirus.
MLV	Murine leukaemia virus
MMTV	Mouse mammary tumour virus
MPMV	Mason-Pfizer monkey virus
MSA	Multiple sequence alignment
MuLV	Murine leukaemia virus
MUSCLE	Multiple Sequence Comparison by Log-Expectation
Mya	Million years ago
Myr	Million years
NC	Nucleocapsid
NCBI	National Center for Biotechnology Information
NGS	Next-Generation Sequencing
ORF	Open reading frame
PBS	Primer binding site
PERV	Porcine endogenous retrovirus
pol	Polymerase
PPT	Purine-rich sequence
PR	Protease
pSIVgml	Prosimian endogenous immunodeficiency virus
PyERV	Python endogenous retrovirus
R	Repeat
RELIK	Rabbit endogenous lentivirus K
RERV	Rabbit endogenous retrovirus
REV	Reticuloendotheliosis virus
RNA	Ribonucleic acid
RNase H	Ribonuclease H
RNA-Seq	RNA sequencing
RSV	Rous sarcoma virus
RT	Reverse transcriptase
SA	Splice acceptor site
SD	Splice donor site
SFVspi	Spider monkey foamy virus
SMRV	Squirrel monkey retrovirus.
SnRV	Snakehead retrovirus
SRA	Sequence read archive
SRLV	Small ruminant lentivirus

SRV	simian retrovirus
SSSV	Salmon swimbladder sarcoma virus
SU	Surface
TgERV-2	Taeniopygia guttata endogenous retrovirus 2
TM	Transmembrane
TMRCA	Time of the most common ancestor
tRNAs	Transfer RNAs
U3	Unique 3’ sequence
U5	Unique 5’ sequence
WDSV	Walleye dermal sarcoma virus
WEHV I	Walleye epidermal hyperplasia viruses type I
WEHV II	Walleye epidermal hyperplasia viruses type II
WGS	Whole Genome Shotgun
XMRV	Xenotropic MLV-related retrovirus

1 Introduction

1.1 Retroviruses (exogenous and endogenous)

Retroviruses (family *Retroviridae*) are enveloped viruses that infect vertebrates. The retroviral infection causes a variety of disease including immunosuppressive disease syndromes (Sepkowitz, 2001), leukaemias (Hayward, Neel and Astrin, 1981; Payne *et al.*, 1981, 1991) lymphomas (Storch *et al.*, 1985), sarcomas (Mayer, Hamaguchi and Hanafusa, 1988) other tumors of mesodermal origin; mammary carcinomas (Salmons and Günzburg, 1987)and carcinomas of liver, lung and kidney (Palmarini *et al.*, 1999; Cherkasova, Weisman and Childs, 2013; Hashimoto *et al.*, 2015) autoimmune diseases (Nexø *et al.*, 2016) lower motor neuron diseases (Jolicoeur, 1991) and several acute diseases involving tissue damage.

The *Retroviridae* are divided into two subfamilies: *Orthoretrovirinae* and *Spumaretrovirinae* (King *et al.*, 2011). All retroviruses are characterised by a replication strategy in which the viral RNA genome is converted to DNA and stably integrated into the genome of the host cell (a form referred to as 'provirus') (Coffin, 1990). Retroviral infection of germline cells (i.e. sperm, eggs or early embryo) can lead to vertical inheritance of proviral loci as host alleles termed endogenous retroviruses (ERVs) (Vogt, 1997). Mammalian genomes typically contain thousands of ERV loci, reflecting a long-term co-evolutionary relationship with retroviruses (Holmes, 2011).

ERV sequences in mammalian genomes typically group into phylogenetically distinct lineages (sometimes referred to as 'families') that are thought to have arisen from a small number of 'germline colonisation' events in which integration of proviral sequences into the germline has been followed by copy number expansion, either through reinfection of germline cells, or retrotransposition (Wilkinson, Mager and Leong, 1994; Sverdlov, 1998; Tristem, 2000). A subset of ERV insertions have been genetically fixed in the host germline, and these sequences constitute a genomic 'fossil record' from which the long-term evolutionary history of retroviruses can be inferred. In addition, recent studies have demonstrated that ERVs sequences have often been co-opted or exapted by host genomes, and this has exerted a profound impact on mammalian evolution and biology (Best *et al.*, 1996; Arnaud *et al.*, 2008; Dupressoir, Lavialle and

Heidmann, 2012; Babaian and Mager, 2016; Blanco-Melo, Gifford and Bieniasz, 2017).

1.1.1 Retrovirus genome structure

Virus particles of the subfamily *Orthoretrovirinae* carry two copies of the linear, single-stranded, positive-sense RNA genome, while those of the subfamily *Spumaretrovirinae* are dsDNA (Coffin, Hughes and Varmus, 1997). In general, the retroviral genome is around 7-12 kb in length, and the coding region is approximately 5-10kb (Coffin, Hughes and Varmus, 1997). Infectious viruses encode four major coding domains for virion proteins including *gag*, *pro*, *pol* and *env* (Figure 1-1).

A short repeat (15-250 nt) attaches to both ends of genomic RNA, and this region is termed as 'R' (Repeat). A unique 5' sequence (U5) positions between R and the primer binding site (PBS) (Damgaard et al., 2004). Moreover, the PBS is usually 18 nt in length and complementary to the 3'end of a specific host tRNA (Goldschmidt et al., 2002). At the 3'end of viral RNA there is a unique 3' sequence (U3) between 7-18 nt long, a purine-rich sequence (PPT) and R. The unintegrated viral DNA and provirus comprises two identical long terminal repeats (LTRs). Long terminal repeats consisted of U3, U5 and R in the form of 5'U3-R-U5-3'. Before reverse transcription, genomic RNA is organised in the form 5'R-U5-*gag-pro-pol-env*-U3-3'R. After the reverse transcription, the viral DNA is organised in the following order: 5′LTR-*gag-pro-pol-env*-3′LTR (Coffin, Hughes and Varmus, 1997; Gifford and Tristem, 2003).

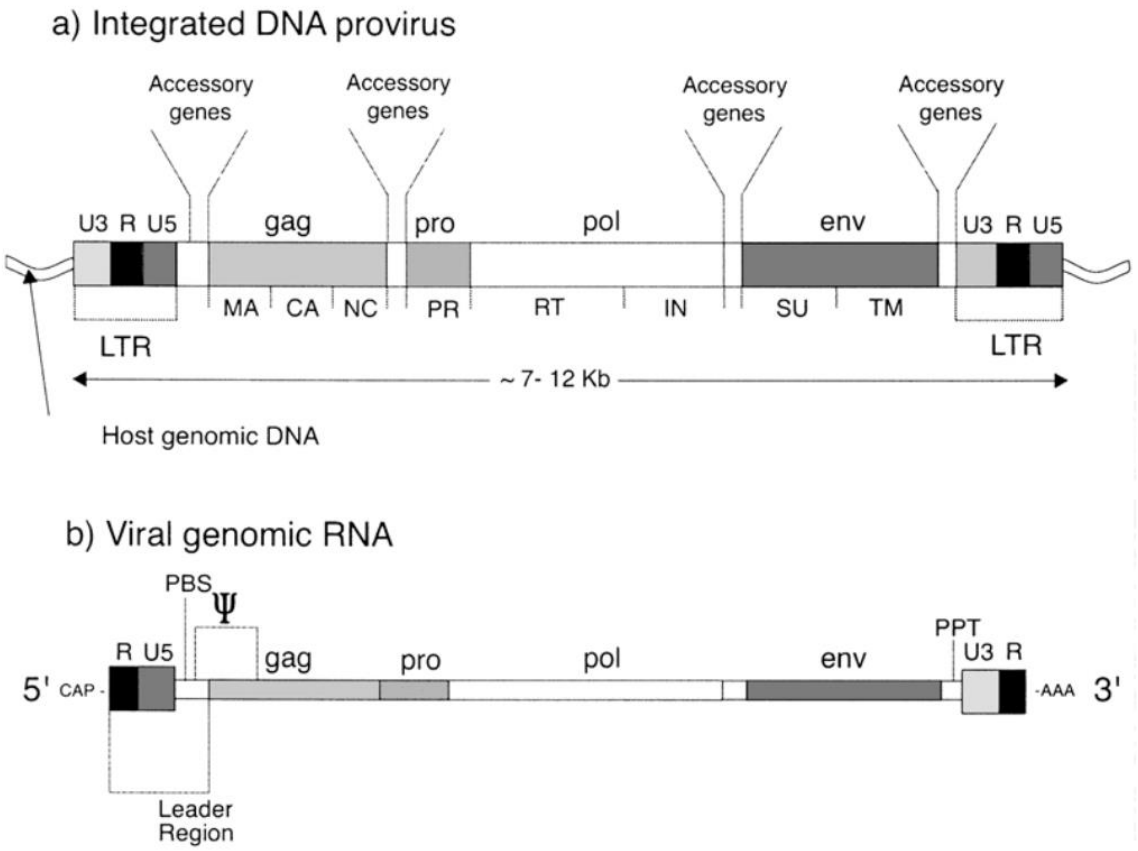


Figure 1-1 Main genome structures of a retrovirus. The genome structure of viral genomic RNA and integrated DNA provirus are generalised to show the common structure for all retroviruses: a) integrated DNA provirus has two long terminal flanking repeats (LTRs composed of U3-R-U5) flanking the internal coding region. Genomic DNA is organised in order: 5'LTR-*gag* (MA, CA, NC)-*pro* (PR)-*pol* (RT, IN)-*env* (SU, TM)-5'LTR; b) viral RNA only has a repeat (R) flanking the internal coding region. The organisation of viral genomic RNA is in order of 5'R-U5-*gag*-*pro*-*pol*-*env*-U3-3'R (Gifford and Tristem, 2003). Permission to reproduce this figure has been granted by the Copyright Clearance Center (License Number: 4354250433044).

Starting from the 5' end, the first coding sequence is *gag* (Vogt, 1997). It is found in all known replication-competent retroviruses. The *gag* gene encodes the polyprotein that controls the assembly and release of the virion. Its cleavage products are the structural components of the viral core (Vogt, 1997). For the *Orthoretrovirinae*, it can be cleaved into three subunits including matrix (MA), capsid (CA), and nucleocapsid (NC) (Swanstrom and Wills, 1997). However, for *Spumaretrovirinae*, it can only be cleaved into large ($p68^{Gag}$) and small ($p71^{Gag}$) products (Swanstrom and Wills, 1997; Cartellieri *et al.*, 2005).

The second coding sequence is *pro* (Vogt, 1997). The *pro* gene is a small coding domain that is essential for viral propagation. It always encodes protease (PR) which is initially synthesised with *gag* and *pol* as polyprotein precursors (Swanstrom and Wills, 1997). The protease embedded within polyprotein

precursors can cleave itself out and subsequently cleave the reminding bonds within polyproteins (Dunn *et al.*, 2002; Goodenow *et al.*, 2002).

Pol is the third coding domain presenting in all replication-competent retroviruses (Swanstrom and Wills, 1997). It encodes part of the Gag-Pro-Pol polyprotein, and it can always be cleaved into reverse transcriptase (RT) and integrase (IN) (Telesnitsky and Goff, 1997). The reverse transcriptase, also known as RNA-directed DNA polymerase, is a critical enzyme for generation of retroviral DNA (Telesnitsky and Goff, 1997). Another essential enzyme encoded by *pol* gene is integrase (IN) which is responsible for the processing and joining steps of integration (Andrake and Skalka, 1996; Brown, 1997; Hindmarsh and Leis, 1999).

The last coding domain is *env*. Virions are non-infectious without envelope glycoproteins. The *env* gene encodes two polypeptides - surface (SU) and transmembrane (TM) (Hunter and Swanstrom, 1990; Vogt, 1997). These polypeptides are responsible for viral adsorption by binding specific cell surface receptors. SU and TM together form an oligomeric knob or knobbed spike on the surface of the viral particle (Hunter, 1997).

Additional, some retroviruses encode dUTPase (DU) in various locations. DU can be encoded between the 3'end of *gag* and 5'end of *pol* in betaretroviruses, or at the 3'end of *pol* in some lentiviruses (Hizi and Herzig, 2015). Furthermore, retroviruses with complex genome organisation also encode up to six non-structural regulator proteins, for example, Tat, Rev, Nef, Vpr, Vpu, Vif, Vps of lentiviruses, Tax and Rex of gammaretroviruses, Tas and Bet of spumaviruses. Moreover, there are some other structural features, such as Cap site, TAR, splice donor site (SD), splice acceptor site (SA), Poly(A) tract (Vogt, 1997).

1.1.2 Retrovirus replication

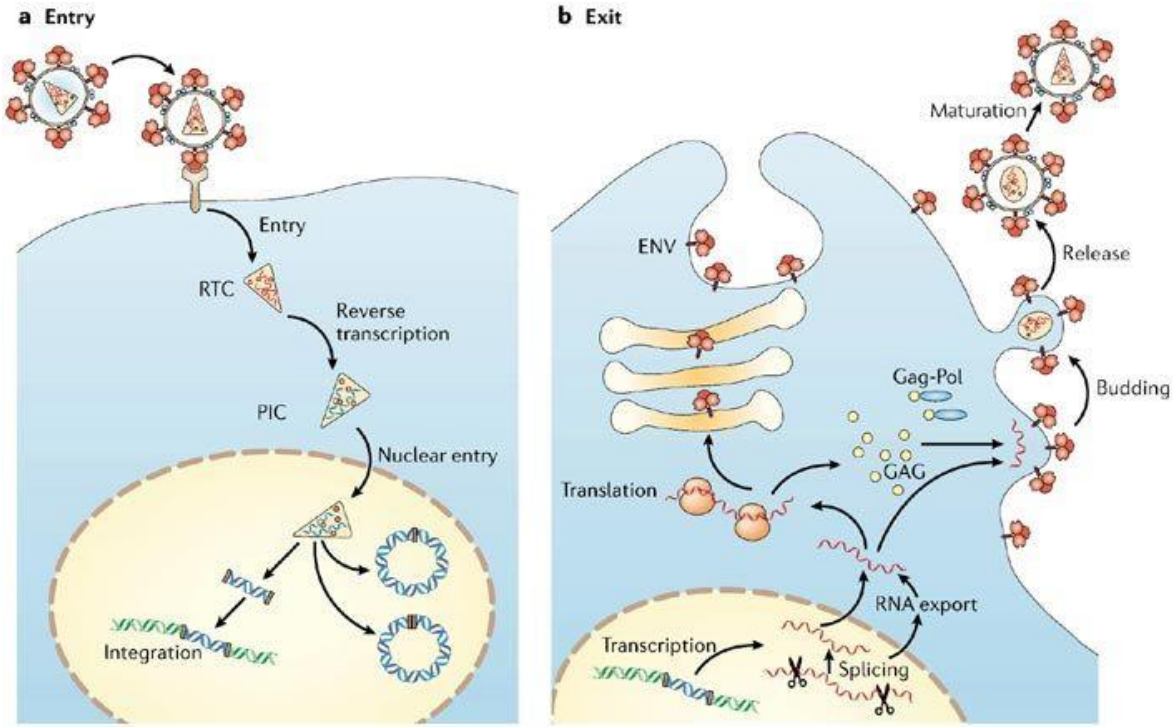


Figure 1-2 Retrovirus replication cycle. Generalised steps in the replication cycle of retroviruses are illustrated: a) viral entry into the host cell including following steps: binding to receptor of cell surface, form membrane fusion, interlocution and uncoat vial core, reverse transcript to synthesis dsDNA, viral dsDNA entry into nucleus, integration; b) viral exit involves the following steps: transcript provirus, nuclear export of viral mRNA with splicing or without splicing, translation of viral proteins and virion assembly; RNA packing; budding through the cell membrane; release infectious virion from cell surface (Goff, 2007). Permission to reproduce this figure has been granted by the Copyright Clearance Center (License Number: 4354250433044).

Receptor binding, internalisation and uncoating

Retroviral entry processes are mediated by interactions between receptors on the cell surface and envelope proteins on the virion surface. (Hunter, 1997; Goff, 2013). SU plays a critical role in the virus replication cycle via binding to a specific receptor molecule on the host cell (Miller, 1996). Transmembrane (TM) mediates the fusion of the virion with the host-cell membrane. After virion cores are delivered into the cytoplasm of the infected cell, they uncoat and reverse transcription is initiated (see below).

Reverse transcription

Soon after the virion core is released into the cytoplasm, the reverse transcription begins in the cytoplasm (Hunter, 1997). Reverse transcription is the defining characteristic and why retroviruses got their names (Telesnitsky and Goff, 1997). In this step, single-stranded viral RNA is used as a template and converted into double-stranded DNA that can be integrated into the host cellar DNA. The entire process of reverse transcription relies on two enzymatic activities of reverse transcriptase: DNA polymerase and ribonuclease H (RNase H) (Telesnitsky and Goff, 1997; Goff, 2013).

Nuclear entry and integration

The linear double-stranded viral DNA needs to be integrated into the cellular DNAs (Brown, 1997; Goff, 2013). Such process is called 'integration' which is a crucial step and a defining characteristic of retroviruses (Brown, 1997). The Integration process is mediated by the viral integrase enzyme. Viral DNA is transmitted through the cytoplasm and then enters the nucleus. In the nucleus, the ends of the linear viral DNA are joined to the cellar DNA (Brown, 1997).

Following integration, the location of provirus in the host DNA is permanent (Brown, 1997). Although proviruses can lose the internal region via the homologous recombination between flanking LTRs (Varmus, Quintrell and Ortiz, 1981), there is no direct mechanism to accurately excise provirus from the host genome. The preference of integration site varies across different retroviruses (Kitamura, Lee and Coffin, 1992; Withers-Ward et al., 1994; Kim et al., 2008; McCallin, Maertens and Bangham, 2015). For example, lentiviruses preferentially insert into transcriptional units (Schröder et al., 2002), whereas gammaretroviruses tend to insert nearby to promoter sequences (Wu, 2003).

Transcription of the provirus

To produce a new infectious virion, the integrated provirus is transcribed and packaged into the virion (Rabson and Graves, 1997). The full-length transcripts have several usages. Some transcripts are used to form the virion core. These transcripts are exported to the cytoplasm directly and packaged into the virion

particle. A portion of transcripts comprising the whole genome is used for the translation of Gag and Gag-Pol polyproteins. A smaller portion of transcripts is spliced to generate the precursor of the envelope proteins. Moreover, for the complex retroviruses, multiply spliced transcripts are used for the translation of accessory regulatory genes (Rabson and Graves, 1997).

Translation of the RNAs

These spliced transcripts shared a common sequence at their 5'ends. Most translation products are polyproteins (Swanstrom and Wills, 1997; Goff, 2013). The *gag*, *pro* and *pol* genes are expressed by complex mechanisms to form precursor proteins and then cleaved to become mature.

In type-C mammalian gammaretroviruses (e.g., MuLV) and epsilonretroviruses (e.g. MDSV), Gag and Pro-Pol are in the same ORF. Translation of *pro* and *pol* involves bypassing translational termination signals by translation readthrough - that is the UAG stop codon at the boundary between Gag and Pro-Pol is suppressed (Yoshinaka et al., 1985). However, for alpharetroviruses (e.g., ALV) and lentiviruses (e.g., HIV-1), the Gag and Pol are encoded in different reading frames. The formation of large precursor protein is via translational frameshifting (Jacks and Varmus, 1985). The ribosome can slip back one nucleotide when translation reaches a specific site near the termination signals. In the betaretroviruses (e.g., MMTV) and deltaretroviruses (e.g., BLV, HTLV-1), the *pro* gene is present at the ORF differed from that of gag and pol. Translation of the long Gag-Pro-Pol fusion protein requires two successive fameshifts - the ribosome can slip back one nucleotide twice near the 3' end of the gag ORF and near the 3'end of the pro ORF. For spumaviruses, *pol* is translated individually instead of forming a Gag-Pol fusion protein (Enssle et al., 1996; Löchelt and Flügel, 1996; Holzschu et al., 1998).

Assembly of the virion

Once the Gag, Gag-Pro-pol and Env polyproteins are synthesised, they come together with two copies of viral RNA and tRNA primers to form progeny virions. The assembly happens at a common site on the plasma membrane (Henderson,

Krutzsch and Oroszlan, 1983) or in the cytoplasm (Rhee, Hui and Hunter, 1990). The uncleaved Gag precursors are responsible for virion assembly.

Packaging of the viral RNA genome

The viral genome harbours an RNA packaging signal located at the 5'end between U3 and *gag* of the viral RNA (Mann, Mulligan and Baltimore, 1983; Kaye, Richardson and Lever, 1995; McCann and Lever, 1997; Zaitseva, Myers and Fassati, 2006). This specific RNA sequence is termed as 'Psi' or 'Ψ'. The RNA packaging signal can interact with specific residues in the NC domain of Gag precursor for the viral genome to incorporate into the virion (Mann, Mulligan and Baltimore, 1983; Kaye, Richardson and Lever, 1995; McCann and Lever, 1997; Zaitseva, Myers and Fassati, 2006).

Budding and release of the virions

After the virion assembly and RNA packaging, virions are released from the cell by the process of budding, which occurs preferentially at lipid rafts (Coffin, Hughes and Varmus, 1997).

1.2 Retrovirus diversity

1.2.1 Taxonomy of exogenous retroviruses

The retroviral subfamily *Spumaretrovirinae* only has one genus: *Spumavirus*. In contrast, there are six officially recognised genera in the subfamily *Orthoretrovirinae* are *Alpharetrovirus*, *Betaretrovirus*, *Deltaretrovirus*, *Epsilonretrovirus*, *Gammaretrovirus* and *Lentivirus*. This classification is based on the virus taxonomy (2017 release) of International Committee on Taxonomy of Viruses (ICTV).

Alpharetrovirus has widespread distribution in chickens and some other birds. The prototype virus is Avian leucosis virus (ALV). Based on their receptor usage, ALV isolates are classified into ten subgroups (Petropoulos, 1997). All known ALV subgroups are all exogenously acquired infections.

Betaretrovirus includes only viruses isolated from mammals, (Gifford and Tristem, 2003; Baillie *et al.*, 2004; Hayward *et al.*, 2013). Liquid hybridisation data suggested betaretroviruses are widely distributed in mammals (Hecht et al., 1996). Betaretroviruses consist of mammalian type-B and type-D retroviruses (Weiss, 1996). The viral particles of MMTV are assigned to type-B morphology, while all other members of *Betaretrovirus* exhibit a type-D morphology (King *et al.*, 2011). The prototype species of type-B virus is the Mouse mammary tumour virus (MMTV), while the type-D prototype virus is Mason-Pfizer monkey virus (MPMV, also known as SRV-3).

Gammaretrovirus was first described as aetiological agents of leukaemias and sarcomas within mice (Gross, 1951; Levy, 1973). Gammaretrovirus exhibits as type C morphology for their virion structure. Gammaretroviruses are widely spread in several vertebrates including mammalian, reptilian, avian and amphibians (Tristem et al., 1996; Martin et al., 1999), e.g. murine leukaemia virus (MuLV) (Shinnick, Lerner and Sutcliffe, 1981), Reticuloendotheliosis viruses (REVs) (Purchase et al., 1973; Payne, 1992).

Epsilonretrovirus is comprised of fish retroviruses. Infection with exogenous viruses is associated with tumours in fish (Lepa and Siwicki, 2011; Coffee, Casey

and Bowser, 2013). There are several well-known epsilonretroviruses including Walleye dermal sarcoma virus (WDSV) (Walker, 1969), Walleye epidermal hyperplasia viruses type I and II (WEHV I and II) (LaPierre et al., 1998), Snakehead retrovirus (SnRV) (Frerichs et al., 1991), salmon swimbladder sarcoma virus (SSSV) (Paul et al., 2006). Although these viruses are classified into the same genus, both SnRV and SSSV may provide the basis for additional genera (Lepa and Siwicki, 2011; Naville and Volff, 2016).

Deltaretrovirus is restricted to mammalian species. All exogenous members are found in primates and cattle, e.g. human T-lymphotropic virus 1 (HTLV-1) (Verdonck et al., 2007) and Bovine leukaemia virus (BLV) (Miller and Van Der Maaten, 1977).

Lentivirus is the most well-known and well-studied retrovirus genus of the subfamily. The most famous examples are Human immunodeficiency virus 1 and 2 (HIV-1 and 2) which causes acquired immunodeficiency syndrome (AIDS) (Barre-Sinoussi et al., 1983; Gallo et al., 1983; Weiss, 1993; Douek, Roederer and Koup, 2009). Except for HIV-1 and 2, lentiviruses were also discovered to infect a variety of primates and ungulates, e.g. goats, sheep, cattle and horses (Barboni *et al.*, 2001; Leroux, Cador and Montelaro, 2004; Bhatia, Patil and Sood, 2013; Larruskain and Jugo, 2013).

Spumavirus is the only genus of *Spumaretrovirinae* subfamily. Unlike viruses of *Orthoretrovirinae*, the Gag protein of spumaviruses is not cleaved into subunits in infectious virions (Flügel and Pfrepper, 2003). Exogenous spumaviruses are broadly found in mammals. However, infection with spumaviruses has no association with disease (Santillana-Hayat et al., 1996; Heneine et al., 2003).

1.2.2 Taxonomy of endogenous retroviruses

Unfortunately, the nomenclature of endogenous retroviruses classification and exogenous retroviruses taxonomy are developing separately and thus is hard to integrate. There is no systematic way to incorporate ERVs into the existing retroviral taxonomy (Blomberg *et al.*, 2009). This situation has become more complicated with increasing development of ERV classifications in a variety hosts since there is no consensus method to describe what they found. Also, current

studies frequently assign different ERV lineages to ‘family’ and ‘class’, though ICTV groups the whole *Retroviridae* as one ‘family’ (Fauquet and Fargette, 2005). Thus, it is essential to develop a retroviral taxonomy, which incorporates both endogenous and exogenous viruses.

Throughout this thesis and to describe ERVs identified from the genomes of interests I will use a combined approach that brings together the nomenclature of HERVs classification and the ICTV retroviral taxonomy was used to describe ERVs identified from the genomes of interests (Chapter IV). The HERVs classification is based on the review of Gifford and Tristem (2003). This classification was investigated based on the phylogenetic comparison and the identification of the PBS for higher resolution within ERV lineages. The phylogenetic comparison was performed based on sequences of RTs. Since it was the fact that the retroviral *pol* gene is well conserved across different endogenous and exogenous retroviruses (Williams and Loeb, 1992; Sala and Wain-Hobson, 2000). Thus, retroviral RT sequences can be used to infer the retroviral phylogenies (Doolittle et al., 1989; Xiong and Eickbush, 1990; Tristem, 2000; Song et al., 2013; Naville and Volff, 2016).

HERVs thus are generally divided into three major ‘classes’ (Figure 1-3). ‘Class I’ includes ERVs that are phylogenetically clustered with *Gammaretroviruses* and *Epsilonretroviruses*. HERVs that showed relatively close relation to the *Betaretroviruses* were termed as ‘Class II’. HERVs closely related to *Spumaviruses* are termed as ‘Class III’. In this thesis, these groups are referred as ‘clades’ to avoid confusion with the taxonomic meaning of the word ‘class’ (Tristem, 2000).

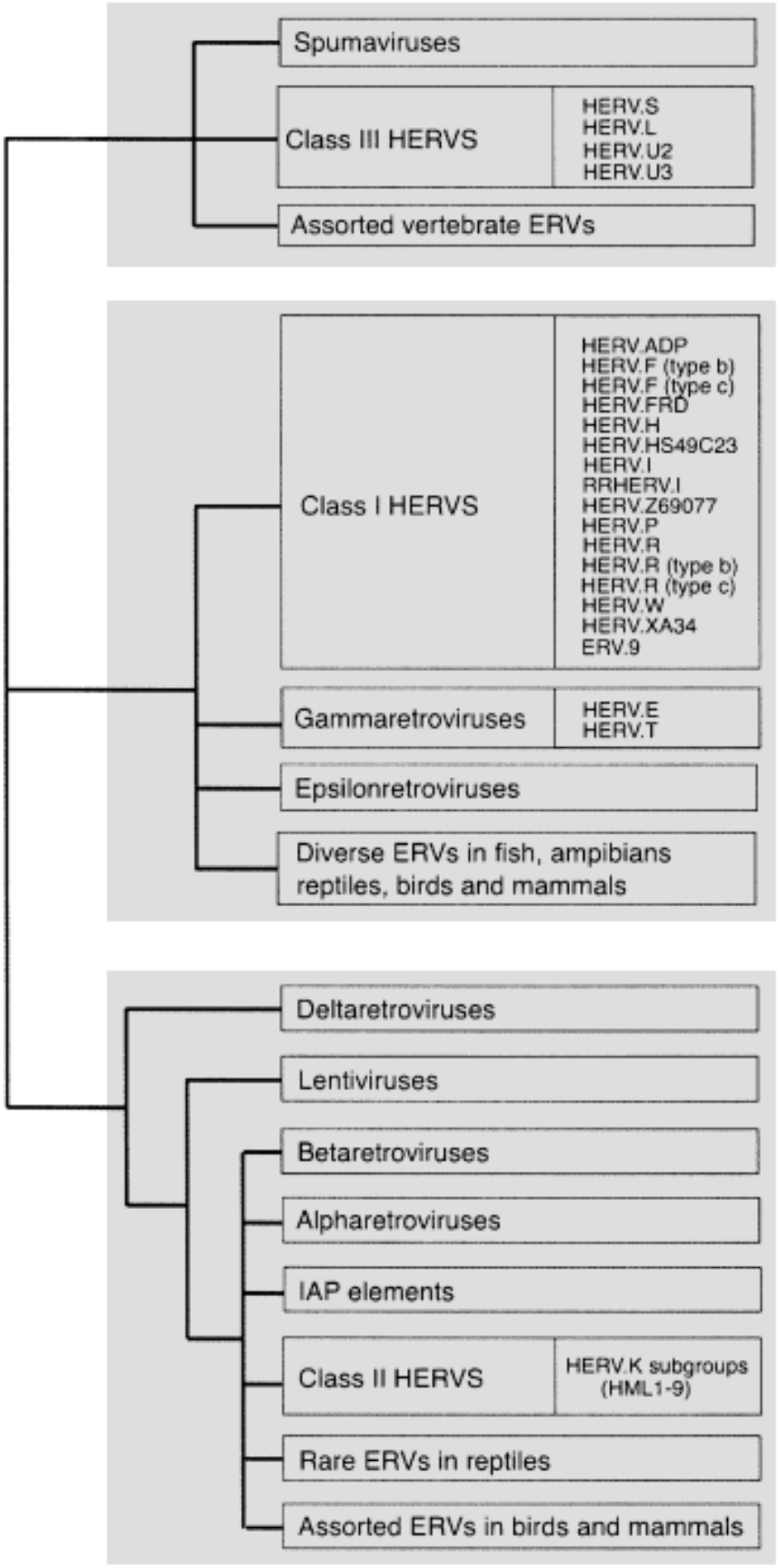


Figure 1-3 Association between HERV classification and ICTV taxonomy. Illustration of retrovirus evolutionary relationships is based on the phylogenetic reconstruction of retroviral RT genes. Major classes are frame coloured by grey. Branches within each major group are summarised as boxes with group names. (Gifford and Tristem, 2003).

1.3 Detecting and characterising ERVs

1.3.1 Early studies of ERVs using laboratory approaches

The early discovery of ERVs was based on a combination of virological and immunological techniques with Mendelian genetics. Simultaneously, crucial evidence of three ERVs was found for the endogenous avian leucosis virus (ALV) in *Gallus gallus* (domestic fowl), and murine leukaemia virus and murine mammary tumour virus in *Mus musculus* (laboratory mouse) in the late 1960s (Subramanian *et al.*, 2011). Nucleic acid hybridisation then confirmed the existence of a retroviral genome. Since then, numerous ERVs were identified in the human genome using wet-lab techniques, e.g. low-stringency hybridisation (Martin et al., 1981), PCR strategies (Medstrand and Blomberg, 1993).

1.3.2 Bioinformatics approaches for detection of ERVs

The development of sequencing technology has enabled researchers to efficiently sequence the whole genome of a species at a lower cost. Based on these sequencing data, researchers can apply *in silico* screening methods to identify and characterise ERVs at the nucleotide level.

Bioinformatics tools are now the most common methods to mine and annotate ERVs in the genome. Owing to the advances in the genome sequencing and *in silico* screening approaches, numerous ERVs families have been identified in genomes of various organisms to date, e.g. human (Lander et al., 2001), mouse (Mouse Genome Sequencing Consortium et al., 2002), chicken (Hillier et al., 2004), dog (Jo et al., 2012), sheep (Klymiuk *et al.*, 2003) and sharks (Han, 2015).

ERV detection methods can operate on two categories of genome data: assembled genomes and WGS reads. In principle, detection tools using WGS data aim to identify reads counting junction of ERVs and host DNA sequence (Li et al., 2005). In addition, comparative genomics methods can apply for detecting ERVs (e.g. the UCSC and Ensembl genome browsers) (Caspi, 2005). Herein, I reviewed the detection tools using assembled genomes.

Computational tools developed for detection in assembled genomes can be categorised into two major groups: homology-based and *de novo*. The homology-

based approaches require prior information of ERVs (e.g. Repbase) and utilise similarity to identify known ERVs. Whereas, *de novo* approaches rely on the nature of ERVs including repetitiveness and structural signatures (i.e. long terminal repeats). As results, *de novo* detection tools can identify novel ERVs that have not been described or lose the features for homology-based search.

Table 1-1 Current available tools for ERV detection

Name	References	Comments
General homology search tools		
BLAST		BLAST is a suite of programs, provided by NCBI, which can be used to quickly search a sequence database for matches to a query sequence.
BLAT	Kent, 2002	BLAT is a very fast sequence alignment tool similar to BLAST typically used for searching similar sequences within the same or closely related species.
HMMER	Eddy, 2001	HMMER is based on profile hidden Markov models (HMMs), it finds evolutionarily related proteins and/or domains, close and remote homologs.
DIGS		Systematic screening using BLAST and a relational database
TE homology search tools		
RepeatMasker	Smit *et al.*, 2013	Screens DNA sequences for interspersed repeats and low complexity DNA sequences
CENSOR	Jurka *et al.*, 1996	A software tool which screens query sequences against a reference collection of repeats and "censors" (masks) homologous portions with masking symbols.
TE de novo search tools		
RECON	Levitsky, 2004	Designed for constructing profiles of nucleosome potential, characterising the probability of nucleosome formation along DNA sequences.
PILER	Edgar *et al.*, 2005	An approach to de novo repeat annotation that exploits characteristic patterns of local alignments induced by certain classes of repeats.
LTR_par	Kalyanaraman *et al.*, 2006	LTR_par identifies regions in a genomic sequence that show structural characteristics of LTR retrotransposons
LTR_STRUC	Eugene *et al.*, 2003	Identifies and automatically analyses LTR retrotransposons in genome databases by searching for structural features characteristic of such elements.
Hybrid search tools/strategies		
Retrotector	Sperber *et al.*, 2009	Specific detection of ERVs using combined de novo and homology-based approaches
GenomeTools	Gremme *et al.*, 2013	A bioinformatics environment that includes several tools relevant to ERV detection
LTR_FINDER	Xu *et al.*, 2007	A tool for the prediction of full-length LTR retrotransposons

Homology-based detection

From many aspects, the most straightforward method of identification is the direct searching of sequences that are similar to the query database, if an ERV reference library is available. Such detection can be simply and efficiently achieved using any sequencing alignment tools, for examples, BLAST (Camacho et al., 2009) and BLAT (Kent, 2002). These tools can report any sequences with homology to the reference sequence in the query database. Among all sequencing alignment tools, the RepeatMasker (Smit, AFA, Hubley, R & Green, 2013) is the most popular programs for this task. RepeatMasker uses RMBlast (RepeatMasker compatible version of the standard NCBI BLAST) or cross_match (Tempel, 2012) as the search engine to screen DNA sequences for interspersed repeats. Then RepeatMasker will mask repeats in sequence with ambiguous characters (i.e. Ns) for further analysis like gene prediction.

The sensitivity of homology-based detection tools greatly relies on the prior knowledge, and in particular, on a reference library. To date, Repbase is the most widely used database of repetitive DNA elements (Jurka et al., 2005). Repbase contains a wide collection of consensus sequences of repetitive DNA elements from a wide range of eukaryotic species.

Also, if researchers apply screening methods using probabilistic inference methods based on hidden Markov models, e.g. nhmmer (Wheeler and Eddy, 2013), the Dfam database can provide the hidden Markov models (HMM) of repetitive DNA element sequence alignments for eukaryote genomes. Also for human-specific ERVs detection, the Human Endogenous Retroviruses Database (Paces, Pavlícek and Paces, 2002a), a lineage-specific database of human ERVs, is available.

An alternative method is to detect protein-coding sequences using known protein domains. The advantages to detecting protein-coding sequences are that the discovery of protein-coding sequences is more likely to be *bona fide*. However, it also means that this method cannot detect any ERVs that have lost all coding regions.

The common program for protein-coding detection is the HMMer package (Finn, Clements and Eddy, 2011). Some programs implement HMMer as a search engine

and achieve an output similar to HMMer but with their constraints for different purposes, e.g. LTRdigest (Steinbiss, Willhoeft, *et al.*, 2009). Furthermore, tBLASTn of the NCBI BLAST+ package (Camacho et al., 2009) is also an efficient choice. For using HMMer and HMMer-based programs, the most widely used library is Pfam (Finn et al., 2016). Pfam provides a collection of protein families in the HMMs format. It is also the common choice for HMMer screening.

***De novo* detection**

The major motivation for the development of *de novo* detection methods is to detect ERVs without prior knowledge of sequences. This is particularly useful for the screening performed on species for which ERVs have not been fully characterised.

Since *de novo* detection utilises the repetitive features of ERVs (the paired LTR sequences that flank integrated proviruses), it does not require any references to identify novel ERVs. Rather, these approaches are based on detecting pairs of identical or near identical sequences that are of reasonable length and distance apart that they could potentially represent ERV proviruses. *de novo* strategies usually entail a 'self-comparison' following a clustering step as described below. For the initial self-comparison, most programs initially align the query sequence with itself and then find all multiple possible matches caused by repeats. Some programs use standard similarity search tools like BLAST and BLAT for this purpose; others use custom tools.

Numerous popular programs for *de* novo detection are currently available: e.g. REPuter (Kurtz *et al.*, 2001), RECON (Bao and Eddy, 2002) and PILER (Bao and Eddy, 2002). RECON is one example of a program using a self-comparison strategy. The initial alignment of RECON program is generated by implementing WU-BLAST and then clustering the local pair-wise alignments.

However, the detection tools mentioned above are designed for more general purposes than simply detecting ERVs – they are designed to detect all reparative elements. In most cases, the clustering function of these tools cannot distinguish ERVs from the other repeats. Thus, even after clustering, an additional step of identification is still needed to filter ERVs from the results. To further automate

the identification step, LTR retrotransposons detection tools have been developed. ERVs share many structural features with other types of LTR retrotransposons. Thus, LTR retrotransposons detection tools can be used as ERV-specific detection tools.

Instead of searching any similar sequence pairs, LTR retrotransposons detection tools aim to find the LTRs initially. Full length and nearly-full length proviruses are ideal targets for the detection. Many programs have been designed for the *de novo* LTR retrotransposons detection. LTR_STRUC (McCarthy and McDonald, 2003) is one of the most popular detection tools used for LTR detection. It has been applied to a variety of organisms including fruit fly (Franchini, Ganko and McDonald, 2004), rice (McCarthy et al., 2002) and mouse (McCarthy and McDonald, 2004).

Hybrid approaches

To further improve the accuracy of prediction, some programs consider internal structural features, e.g. *gag*, *pol*, and *env*. These tools are no longer a typical *de novo* detection tools. They are more likely to be a hybrid of homology-based and *de novo* detection. They initially screen the query sequences for flanking LTRs using the *de novo* method and then annotate the internal region of flanking LTRs for internal structure features. These tools usually inherit prior information of LTR retrotransposons features including PBS, PPT, ORFs and other genetic features.

Some tools also accept a custom library for a flexible detection. RetroTector also applies a ‘fragment threading’ process to convert detected LTRs and conserved retroviral motifs into chains which represent more or less full-length ERVs (Sperber et al., 2007). The well-known tools include LTR_FINDER (Xu and Wang, 2007), as well as LTRharvest (Ellinghaus, Kurtz and Willhoeft, 2008) and LTRdigest (Steinbiss, Willhoeft, *et al.*, 2009) of GenomeTools packages (Gremme, Steinbiss and Kurtz, 2013).

1.4 Analysis of equine ERVs

1.4.1 Why analyse ERVs in the horse genome

So far, studies of mammalian ERVs have tended to focus on primates and rodents, reflecting the importance of these mammalian groups in biomedical research. However, whole genome sequences are now available for a much broader range of mammalian groups, making more wide-ranging investigations possible. Currently published studies focused on the modern horse, but not in the wider context of related species. Characterising ERVs across a wider context will enable comparative investigations that can shed light on the biology of ancient retroviruses and reveal insights into the co-evolutionary processes through which ERVs have shaped host genomes.

ERVs have been shown to be involved in controlling gene expression and pluripotency in mammals. (Kamat et al., 1998; Mi et al., 2000; Conley and Hinshelwood, 2001; van de Lagemaat et al., 2003; Dupressoir et al., 2009). Several previous studies have observed similar biological phenomena (Moreton *et al.*, 2014). Multiple ERVs insertions seem to have transcript activities in the horse tissue. 79 ERV loci were found to have expression level of RPKM >1 in the RNA transcriptome of kidney, jejunum, liver, spleen and mesenteric lymph nodes of horses (Brown *et al.*, 2012). Also, another study suggested that an equine ERV *env* is expressed in multiple horse tissues, with expression in the equine fetal part of the placenta being significantly higher than the others (liver, spleen, lung and kidney) (Stefanetti et al., 2016). Moreover, in this study, I found some *pol* genes have different expression in the cerebellum of two different horse breeds via reverse transcription quantitative real-time PCV (RT-qPCR) (Gim and Kim, 2017). Understanding how ERVs influenced gene expression in equids may facilitate the development of stemcell based therapeutics for horses. It also provides insight into the ERV studies of other organisms.

1.4.2 Evolution of the horse

Evolution of Perissodactyls

The *Perissodactyla* are also known as ‘odd-toed ungulates’. Members of the order *Perissodactyla* are strict herbivores with an odd number of toes and adapted for

running and dietary specialisation (Radinsky, 1966). The *Perissodactyla* can be divided into two suborders: *Hippomorpha* and *Ceratomorpha* (Prothero and Schoch, 1989). The *Hippomorpha* only has one family: *Equidae*. The *Ceratomorpha* contains families of *Tapiridae* and *Rhinocerotidae* (Radinsky, 1966; Prothero and Schoch, 1989; Wilson and Reeder, 2005). The *Equidae* comprises all living species of horses, asses, and zebras in the genus Equus and many other species only known from fossils. The *Ceratomorpha* includes four tapirs of the family *Tapiridae*. Moreover, five rhinoceroses in four genera belong to family *Rhinocerotidae*. Living perissodactyls represent a small remnant of a diverse group of mammals widespread on all continents apart from Australia and Antarctica (Radinsky, 1966; Prothero and Schoch, 1989; McKenna and Bell, 1997).

The common ancestor of the *Perissodactyla* diverged from the *Laurasiatheria* around 77 Mya (Murphy et al., 2007; Meredith et al., 2011; dos Reis et al., 2012; Waku et al., 2016). The common ancestors of the *Equidae* diverged from other species of the *Perissodactyla* around 55 Mya (CI: 53-56 Mya) (MacFadden, 2005; Franzen, 2011; Steiner and Ryder, 2011). Moreover, the divergence of *Tapiridae* and *Rhinocerotidae* was around 50 Mya (CI: 46-53 Mya) (Steiner and Ryder, 2011).

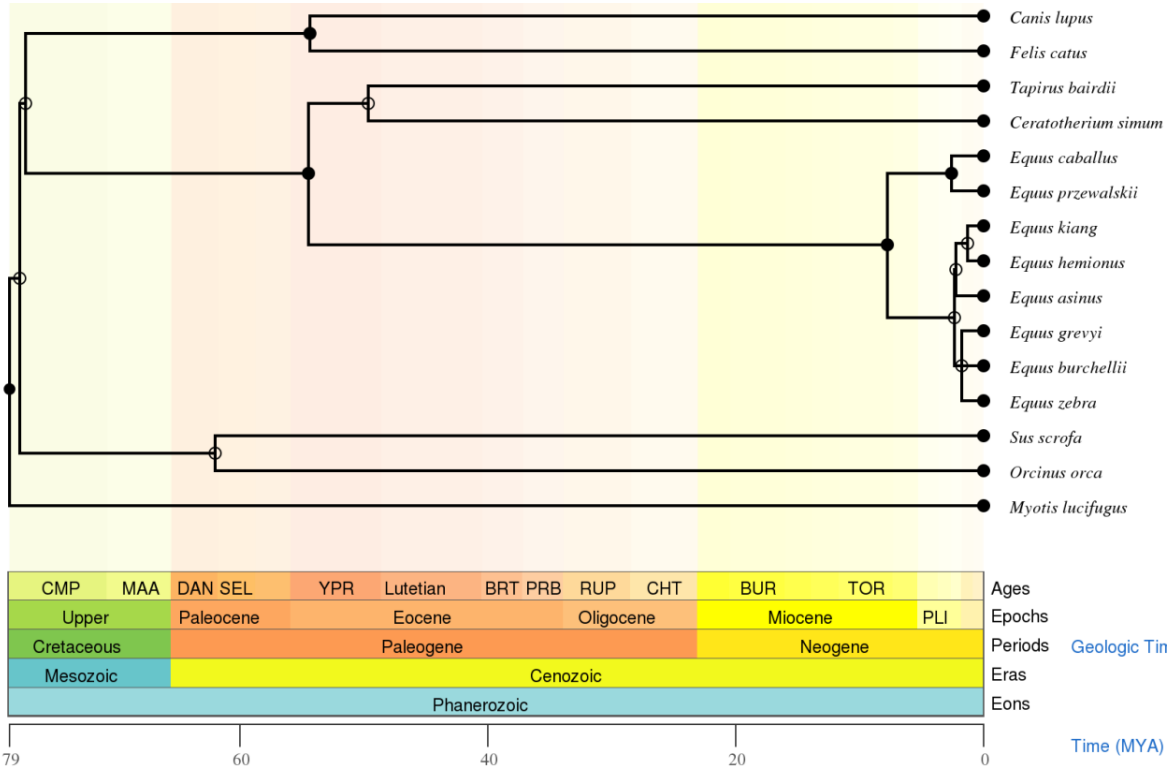


Figure 1-4 The timetree for the *Laurasiatheria* and geographic timescale.The topology of timetree was obtained from the TimeTree resource (Kumar *et al.*, 2017). It was summarised based on the published studies.

The divergence of Equus genus

The earliest equid was a fox size, multi-toed forest-dwelling animal. After 50 million years evolution, however, equids have transformed into the modern, large species adapted to run and the steppe (Franzen, 2011). Currently, all living species of *Equus* genus, including horse, donkey, half ass and zebra, were suggested (Macfadden, 1997) to evolve from the same ancestor, *Dinohippus* (B J MacFadden, 1986; Quinn, 1955), which is an early horse living in North America approximately 3.6-10.3 million years ago (B J. MacFadden, 2000). These estimates were originally based on fossil evidence, and are now also supported by molecular data. Phylogenetic reconstructions based on the whole genome (Orlando et al., 2013) and mitochondrial DNA (Vilstrup et al., 2013) of ancient and modern equids dated the time of most recent common ancestor (TMRCA) of the *Equus* genus to 4.25 Mya.

Migration of extended equids

The ancestor of all extended equids (i.e. including the wild donkey, Asian wild ass and zebra) was suggested to have first diverged from an ancestral population in America, and later to have migrated to Asia. Mitochondrial phylogenomic studies (Vilstrup et al., 2013) pushed the divergence time back to around 2.87 Mya. The ancestors of zebra diverged from other equids at 2.78 Mya (Vilstrup et al., 2013) and moved to Africa (Franzen, 2011). The wild donkey and half-ass diverged from each other at around 2.62 Mya (Vilstrup et al., 2013). The wild donkey migrated to Africa, while half-ass remained in Asia.

Migration of equines

The ancestor of the wild horse was the last lineage to leave North America through the Bering Sea Bridge. They first migrated to Asian and spread to the whole Eurasian (Franzen, 2011). There is no direct evidence showing that the ancestor of the horse reached Africa. After that, the Pleistocene to Holocene extinction wiped out all horse ancestors in North and South America, presumably due to climatic and vegetational changes. These changes also impacted the European horse species (Bendrey, 2012; Sommer et al., 2011) driving surviving populations

to refuges in the Eurasian steppe and the Iberian Peninsula (Warmuth et al., 2011). Horses and donkeys were reintroduced to America by European colonists.

Currently, the only true wild horse left is the Przewalski's horse, which is endangered. All current Przewalski's horses were descended from 13-14 individuals due to a reintroduction project (Ryder, 1993). This species was once considered as one of the domestic horses (Cai et al., 2009) but changed to be sister species based on phylogeny later (Goto et al., 2011).

1.5 Thesis aims

The aims of this PhD project were as follows:

1. To develop an enhanced mechanism for identifying and annotating ERVs in assembled genomes

2. To comprehensively and systematically classify ERVs in the equine genome using a phylogenetic approach

3. To investigate the long-term co-evolutionary relationships between retroviruses and equids using genomic data.

In the following chapters, I describe the work performed during my PhD in pursuit of these three aims.

2 Materials and Methods

2.1 Materials

2.1.1 Whole genome and transcriptome sequences

This project used a number of different NGS resources using different sequencing technologies. A detailed description of these follows. All NGS data are publicly available in the NCBI Sequence Read Archive (https://www.ncbi.nlm.nih.gov/sra) and European Nucleotide Archive (https://www.ebi.ac.uk/ena).

Whole genome sequences

The reference genomes of thoroughbred horse (*Equus caballus*) (Wade *et al.*, 2009), Przewalski's horse (*Equus przewalskii*) (Huang *et al.*, 2014), Mongolian horse (Huang *et al.*, 2014) and southern white rhinoceros (*Ceratotherium simum simum*) were obtained from the NCBI Genome database (NCBI Resource Coordinators, 2018).

All the other genomes used in the study were only available in raw read format (via the European Nucleotide Archive database), as detailed in Table 2-1. There are two versions of domestic donkey (*Equus asinus africanus*) genome assembly. GCF_001305755.1 is publicly available on the NCBI Genome database. DNA from a male Guanzhong donkey was sequenced to 42.4-fold coverage (~2.36Gb), resulting in a *de novo* assembly (Huang *et al.*, 2015). The second version was published by the Orlando group in 2013 (Orlando *et al.*, 2013), and is also a *de novo* assembly. Samples were collected from a domestic donkey, called 'Willy'. Samples have been sequenced to 12.04-fold coverage and approximately 2.35Gb. The 'Willy' donkey assembly was used as the reference due to non-availability of GCF_001305755.1 (available at 2015/10/02) at the beginning of this study (2014/10). Another reason is that the 'Willy' assembly was used as a reference for assembly of the half-ass and zebra genomes used in this study (Jónsson *et al.*, 2014). To be consistent with previous research, the 'Willy' donkey assembly was utilised in preference to the NCBI version.

Table 2-1 Whole genome sequence assemblies used in this study

Taxonomy			Assembly			
Organism	**Common Name**	**TaxaID**	**Accession**	**Synonyms**	**Level**	**Coverage**
Rhinocerotidae						
Ceratotherium simum	Southern white rhinoceros	73337	GCF_000283155.1	cerSim1	Scaffold	91x
Equidae						
Equus asinus africanus	Donkey	582580	PRJNA205517	N/A	Scaffold	12.04x
Equus asinus somalicus	Somali wild ass	73336	PRJEB7446	N/A	Scaffold	21.43x
Equus burchellii boehmi	Plains zebra	89250	PRJEB7446	N/A	Chromosome	20.6x
Equus burchellii quagga	Burchell's zebra	89252	PRJEB7446	N/A	Chromosome	7.92x
Equus caballus	Horse (thoroughbred)	9796	GCF_000002305.2	equCab2	Chromosome	6.8x
Equus caballus	Horse (Arabian)	9796	PRJNA205517	N/A	Chromosome	11.03x
Equus caballus	Horse (Icelandic)	9796	PRJNA205517	N/A	Chromosome	8.43x
Equus caballus	Horse (Norwegin Fjord)	9796	PRJNA205517	N/A	Chromosome	7.86x
Equus caballus	Horse (Standardbred)	9796	PRJNA205517	N/A	Chromosome	12.16x
Equus caballus	Horse (Connemara Pony)	9796	PRJNA205517	N/A	Chromosome	N/A
Equus caballus	Horse (Mongolian)	9796	GCA_000696655.1	Ajinai1.0	Scaffold	90.57x
Equus ferus przewalskii	Przewalski's Horse	9798	GCA_000696695.1	Burgud	Scaffold	85.63x
Equus grevyi	Grevy's zebra	9792	PRJEB7446	N/A	Chromosome	17.05x
Equus hemionus	Onager	9794	PRJEB7446	N/A	Scaffold	18.65x
Equus kiang	Kiang	94398	PRJEB7446	N/A	Scaffold	13.26x
Equus zebra hartmannae	Hartmann's mountain zebra	73335	PRJEB7446	N/A	Chromosome	17.33x

N/A: non -available

The newest version of horse reference genome is EquCab2.0 (GCF_000002305.2) and was sequenced and assembled by the Broad Institute (Wade *et al.*, 2009). Excluding gaps in scaffolds, the total size of the whole genome is 2.43 Gb (2.68 Gb with gaps). Because the animal sequenced was a female thoroughbred horse (named "Twilight"), the horse Y chromosome is missing in the assembly. Although many studies have sequenced or cloned the partial horse Y chromosome (Raudsepp *et al.*, 2004; Wallner *et al.*, 2013), there is still complete Y chromosome reference sequence available for *E.caballus*.

Transcriptomes of 17 tissues and E.derm cell line

Table 2-2 Transcriptome dataset

Tissues & Cell Lines	BioProject	Reference
Cell line		
E.derm	Unpublish	Unpublish
Tissues		
Bone Marrow	PRJNA266428	Tallmadge *et al.* (2015)
Brain	PRJNA184055	Fushan *et al.* (2015)
BrainStem	PRJNA318917	Unpublish
Inner Cell Mass	PRJNA223157	Iqbal *et al.* (2014)
Kidney	PRJNA184055	Fushan *et al.* (2015)
Lamellar	PRJEB6100	Holl *et al.* (2015)
Skin	PRJEB6101	Holl *et al.* (2016)
Liver	PRJNA184055	Fushan *et al.* (2015)
Oviduct	PRJNA297894	Smits *et al.* (2016)
Peripheral blood mononuclear cell	PRJEB7497	Pacholewska *et al.* (2015)
Placental (donkey)	PRJNA153313	Wang *et al.* (2012)
Placental (hinny)	PRJNA153313	Wang *et al.* (2012)
Placental (horse)	PRJNA153313	Wang *et al.* (2012)
Placental (mute)	PRJNA153313	Wang *et al.* (2012)
SpinalCord	PRJNA318917	Unpublish
Trophectoderm	PRJNA223157	Iqbal *et al.* (2014)
Uterus	PRJNA270116	Marth *et al.* (2015)

18 RNA-Seq raw reads dataset were used to examine patterns of equine ERV expression (Table 2-2). The RNA-Seq dataset of the equine dermis cell line (E.derm) was prepared and sequenced by Dr Joanna Crispell. The E.derm cell line dataset is not available to download at the time of writing. All other RNA-Seq data were obtained from the SRA database or ENA database. These data were downloaded at 2016/07. RNA-Seq data published after that are not included in this study.

2.1.2 Software and tools

Read processing: quality control and trimming

FastQC is a quality control tool for NGS reads. It implements a set of modules to analyses the read quality and then visualises the quality via multiple plots and statistical reports (Andrews, 2010). FastQC v0.11.6 was used to check the raw read quality and determine the length cut-off for discarding reads.

Trim Galore is a Perl script for automated adapter trimming and quality control (Krueger, 2015). Trim Galore v0.4.4 was used to trim adapters from all raw reads and reads whose length is shorter than a user-defined threshold.

Whole genome assembly

Bowtie2 is an alignment program which uses an extended full-text minute index-based approach. It permits the gapped alignment of NGS reads to long reference sequences (Langmead and Salzberg, 2012). Bowtie2 v2.3.3.1 was used to align trimmed reads to the reference sequences.

SAMtools (Li *et al.*, 2009) and BCFtools are utility toolset for interacting with and post-processing NGS read alignment in SAM, BAM and CRAM formats. The combination of SAMtools (v1.3) and BCFtools (v1.3) was used to generate the consensus sequences.

Transcriptomics

I used TopHat (version 2.1.1) a splice junction mapping program designed for RNA-Seq reads, to identify splice junctions (Trapnell, Pachter and Salzberg, 2009). I used the Cuffquant and Cuffnorm utilities, both included in the Cufflinks package (version 2.2.1), to measure and normalise RNA expression levels (Trapnell *et al.*, 2012).

Genome-wide screening for RT loci

The database-integrated genome screening (DIGS) tool (version 1.1) is open source (https://giffordlabcvr.github.io/DIGS-tool/). All programs used in the framework

of the DIGS tool are freely available for non-commercial use. The DIGS tools was used to perform systematic screening of whole genome sequence assemblies (Zhu *et al.*, 2018).

Annotation of ERV internal coding region

LTRharvest and LTRdigest are implemented utilities of the GenomeTools package. GenomeTools v1.5.8 was applied in this study. LTRharvest is a *de novo* detection tool designed specifically for LTR retrotransposons (Ellinghaus, Kurtz and Willhoeft, 2008). LTRdigest is the annotation tool for characterising the internal coding region defined by LTRharvest (Steinbiss, Willhoeft, *et al.*, 2009). The domain detection function of LTRdigest is performed by using phmmer, a program of the HMMER package.

AnnotationSketch is a C-based drawing library for visualised GFF3-compatible genomic annotations. It was one of the tools included in Genometools package (Steinbiss, Gremme, *et al.*, 2009; Gremme, Steinbiss and Kurtz, 2013). AnnotationSketch was applied to visualise the proviral genome structure.

The tRNAscan-SE a program aiming to detect transfer RNA genes in genomic sequence. The tRNAscan-SE performs prediction via RNA covariance models based on stochastic context-free grammars (Lowe and Eddy, 1997). The tRNAscan-SE v2.0 was applied.

EMBOSS Transeq is a program for translating nucleic acid sequences to peptide sequences. It can translate all six reading frames. EMBOSS Transeq is part of the **E**uropean **M**olecular **B**iology **O**pen **S**oftware **S**uite (EMBOSS) (Rice, Longden and Bleasby, 2000).

HMMER (Eddy, 2001) is a package of a program designed for searching sequence databases for sequence homologs using probabilistic models - profile hidden Markov models (profile HMMs). HMMER applied in this study was version 3.1b2.

Exonerate is a pairwise sequence aligner (Slater and Birney, 2005). The version 2.2.0 of exonerate program was applied to quickly determine the relative coordinate of RT locus in the extracted sequences.

Phylogeny and alignment

MUSCLE is multiple sequence aligner for both nucleotide sequences and protein sequences, which stands for **MU**ltiple **S**equence **C**omparison by **L**og-**E**xpectation (Edgar, 2004). MUSCLE v3.8.31 created all multiple sequence alignment (MSA) used in this study.

All substitution model selections for phylogenetic analysis were performed using ModelFinder, a function of IQ-TREE (Kalyaanamoorthy *et al.*, 2017). Phylogenetic reconstructions were performed using RAxML v8.0.20 and IQ-TREE v1.4.4. RAxML stands for **R**andomized **A**ccelerated **M**aximum **L**ikelihood, and it is a program for phylogenetic analysis using maximum likelihood method (Stamatakis, 2014). IQ-TREE is a software package for phylogenomic inference with several key features including tree reconstruction, ModelFinder for model selection and UFBoot for bootstrap approximation (Nguyen *et al.*, 2015).

Detection of solo LTRs

RepeatMasker is a program for screening interspersed repeats and low complexity on a genome-wide scale (Smit, AFA, Hubley, R & Green, 2013). RepeatMasker v4.0.7 was used for identifying solo LTRs. The RMBlast, the NCBI BLAST modified for RepeatMasker, was used as sequence search engine (Tempel, 2012). The RMBlast was build based on the NCBI BLAST v2.6.0 and the isb package 2.6.0.

Collation of ERV sequences and auxiliary data

I used GLUE - an open, data-centric software environment specialised in capturing and processing virus genome sequence datasets, which collated the sequences, alignments and associated data used in this investigation (Singer *et al.*, 2018).

Other software and computational tools

I used ORF-FINDER, available on the NCBI website (Rombel *et al.*, 2002), to identify all putative protein coding regions in the DNA sequences.

JalView (Clamp *et al.*, 2004), SeaView (Gouy, Guindon and Gascuel, 2010) and AliView (Larsson, 2014) are graphical multiple sequence alignment editors. They were applied to convert sequence format to fit the input requirement of different programs. Also, they were used to edit sequences manually.

Bedtools is a set of utilities that are used for a wide-range of genomics analysis task (Quinlan and Hall, 2010). Bedtools allows the user to intersect, merge, count, complement and shuffle genomic intervals in various formats, e.g. BAM, BED, GFF/GTR/VCF.

Perl is a family of high-level programming languages. All pipelines and scripts described in this study are based on Perl 5.

R is a system consisting of a programming language and run-time environment with graphics. It is designed for statistical computation and graphics. R version 3.4.2 was applied for any applications based on R.

A set of R packages were used in this study. The ggplot2 (v2.2.1) was used to draw statistics plots (Wickham, 2016), the karyoploteR package (v1.4.1) was used to estimate and visualise the gene density (Gel and Serra, 2017). The IWTomics package (v1.2.0) is an R package that used to investigate discrimination of the given set of genomic features on different groups of genomic regions (Cremona *et al.*, 2017).

2.1.3 Annotation profiles and reference libraries

RT reference library

An RT reference library (Appendix I) was used for screening with the DIGS tool. The library was obtained from Dr R.J. Gifford who collated it from previous studies. The reference library contains 63 reference sequences, including exogenous retroviral sequences from the RefSeq database (Pruitt *et al.*, 2014), previously characterized ERV sequences (Sverdlov, 2000; Tristem, 2000; Bénit, Dessen and Heidmann, 2001; Villesen *et al.*, 2004), and previously inferred consensus sequences (Jern *et al.*, 2005; Lee and Bieniasz, 2007).

Equine genome annotations

Analysis of transcriptome data requires a genomic annotation profile. The genomic annotation profile is a genome-wide prediction of transcripts. A genomic annotation profile for the domestic horse was obtained by Ensembl (Paces, Pavlícek and Paces, 2002b). This annotation profile is the product of the Ensembl mammalian annotation pipeline (Aken *et al.*, 2016) using the EquCab2.0 assembly for the domestic horse genome. Annotations include available data from EMBL, UniProtKB ('UniProt: the universal protein knowledgebase', 2017) and NCBI RefSeq and predictions (Ensembl release 88.2, March 2017). The gene-set contained 29,196 gene transcripts. It is composed of 20,449 coding genes, 2,142 non-coding genes and 4,400 pseudogenes.

Repeatmasker libraries

To annotate the long terminal repeats (LTRs) and detect solo LTRs, I used a RepeatMasker library from Repbase website (Jurka *et al.*, 2005). Repbase provides a repeat reference collection of prototypic sequences from different eukaryotic species. The RepeatMasker library is a special edition of Repbase library. However, RepeatMasker library is not the same as Repbase library (Tempel, 2012). Sequences of RepeatMasker library has been optimised for RepeatMasker program, and labels of RepeatMasker library may not include in Repbase. Also, Repbase references may match multiple RepeatMasker library references, as Repbase breaks long consensus sequence into several fragments for improving search sensitivity. To improve both the search time and selectivity I extracted all *Equus caballus* repeats, as well as ancestral (shared) repeats (repeats that are classified at a higher taxonomic rank) instead of the whole RepeatMasker library. The extracted library had 218 records (edition 2017/01/27).

Protein profile-HMM (hidden Markov model)

HMMER performs sequence similarity searches based on profile hidden Markov models (profile HMMs). The profile HMM is a position-specific scoring system that is generated from a multiple sequence alignment. The profile HMM is usually used for searching databases for homologous sequences (Eddy, 1998). Pfam is a database which collates multiple sequence alignment and profile HMMs for protein

domain families. The data presented in Pfam is based on the UniProt Reference Proteomes (Finn *et al.*, 2016). The profile HMMs related to retrotransposons were obtained from Pfam. In total, 110 domain records are downloaded.

To identify the primer binding site of putative ERVs, the prediction of tRNA sequences was downloaded from Genomic tRNA Database (GtRNAdb). GtRNAdb has a collection of predicted transfer RNAs (tRNAs) from different species (Chan and Lowe, 2009). GtRNAdb uses tRNAscan-SE (Lowe and Eddy, 1997) to search complete or nearly complete genomes and predicted tRNA sequences. In total, 494 and 519 tRNA sequences of equine and white rhinoceros are obtained. Donkey tRNA sequences are not available in GtRNAdb. To obtain a set of donkey tRNAs, tRNAscan-SE was used to scan 'Willy' donkey assembly. In total, 504 tRNA sequences were predicted and passed the threshold (Score $\geq$ 40).

2.2 Methods

2.2.1 Whole genome assembly for data mining

Quality control was first analysed by FastQC and then performed by Trim Galore. Adapters were removed from short reads. Reads were discarded if read length were shorter than 20 bp before or after trimming process. All short reads were aligned using Bowtie2 with a very-sensitive–local option (equal to –D 20 –R 3 –N 0 –L 20 –i S,1,0.50).

Following read mapping, a single SAM file was created for each species. Each SAM file was then converted to a sorted BAM file using SAMtools, and consensus genomes were generated using a combination of SAMtools and BCFtools.

2.2.2 Homology-based screening using the DIGS tool

The DIGS tool links similarity searches (as implemented in the Basic Local Alignment Search Tool (BLAST) (Camacho *et al.*, 2009)) to a MySQL database. Minimal requirements for performing DIGS are (i) 'target' sequences (i.e. whole genome sequences) for screening; (ii) 'probe' sequences to use as queries in similarity searches; (iii) a reference sequence library for classification of sequences identified via screening. For each DIGS project, the screening is defined by control file that specifies parameters for screening (e.g. file paths and cut-offs).

Before performing a project, DIGS tool creates a distinct MySQL with four table (shown in Appendix II). As illustrated in Figure 2-1, DIGS tool performs each project in two steps. First, the implemented BLAST functions are applied to search sequences of targets ('target' for BLAST) with probe sequences ('query' for BLAST). Depended on the type of probe sequence, DIGS can use BLASTn or tBLASTn for the nucleic acid or protein sequences. Sequences exhibiting similarity to probe sequences are recorded as 'hits'.

Second, stored hits are compared to the reference library by BLAST. This comparison allows hits to be assigned to a board classification of sequences. It is important because query sequences may not be the closest reference sequence to the hits, and hits can be adjusted to the other reference sequences if a better

alternative sequence exists. This step provides an adequate approach for the first-pass description of sequence diversity (Gifford, et al., 2006). Then all assigned hits are captured in a MySQL database.

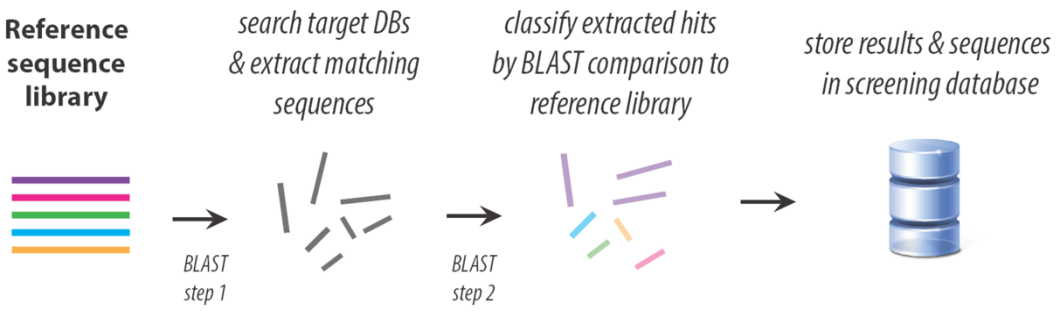


Figure 2-1 Genome screening using the DIGS tool. Sequences from the reference are selected as probes and used to screen target sequence databases (e.g. genome assemblies), with all matches being extracted and classified by comparison to the reference library.

The DIGS tool has functions for dealing with contingencies associated with fragmented or overlapping hits, picking the longest hit if one locus matches multiple distinct probes. If several hits matched to the same probe occur within a given range, the DIGS tool will extract the entire region spanned by these hits as one hit.

2.2.3 ERV detection using Genometools

Identification of full-length provirus candidates by ERVAP

LTRharvest was used to identify LTR pairs within the extracted DNA sequences based on the following parameters: MINLENLTR = 200; MAXLENLTR = 1500; MINDISTLTR = 1000; MAXDISTLTR = 15000; SIMILAR = 80; MINTSD = 5; MAXTSD = 20. Two LTRs that meet constraints are considered as an LTR pair. LTRharvest only reports LTR pairs with E-value below 10^{-6}. All detected LTR pairs will be further analysed separately.

The Exonerate aligner was used calculate the relative position between identified RT and LTR pairs. The ERVAP only considered LTR pairs flanking identified RT as candidates.

Annotation of internal coding regions

LTRdigest was used to annotate putative internal coding region of candidates. All six reading frames were screened. A loose constraint was applied: PPTLEN_MIN =10; PPTLEN_MAX = 30; PBSOFFSET_MIN = 0; PBSOFFSET_MAX = 100; PBSRADIUS = 100.

HMMER was used to annotate the extracted sequences without clear LTR boundaries. The whole extracted sequence was first translated in six frames by EMBOSS Transeq. Then HMMER was performed to search for potential protein domains. Only hits with E-value of ≤ 5e-5 were reported.

2.2.4 Detecting solo LTRs using RepeatMasker

Solo LTRs were detected using RepeatMasker and a custom library. LTRs identified by *LTRharvest* program was first assigned to the RepeatMasker library by BLAST. Unassigned LTRs were considered as 'novel'. The custom library consisted of selected references from RepeatMasker library and novel LTRs. RepeatMasker used RMBlast as a search engine to search for solo LTRs. The screening was performed with the default setting.

2.2.5 Summary of all information for annotation profile

In the final stage, the ERVAP summarised all information generated by each previous stage. ERVAP returned an annotation profile in comma-separated values format (CSV). Also, ERVAP visualised genomes of all identified ERVs using AnnotationSketch.

2.2.6 Sequence alignments and phylogenetic analysis

All multiple sequences alignments (MSA) were generated using MUSCLE. All MSA were manually edited using AliView or SeaView based on the input sequence format.

All phylogenetic reconstructions were performed using maximum likelihood approach. For a tree with less than 200 taxa, RAxML was applied. Others were inferred using IQ-TREE. The best-fit substitution model was selected by the

ModelFinder of IQ-TREE. Support for any phylogenies was assessed via 1000 non-parametric bootstrap replicates.

2.2.7 Calculating the integration time

For the dating of solo LTRs, the maximum likelihood distance (ML distance) was estimated by the distance between solo LTR and the consensus sequence of its relative LTR group. JalView was used to generate consensus sequences based on solo LTR alignments and the majority rule (majority ≥ 60%). The ML distance of paired LTRs was the calculation of divergence of 5' between 3' LTR. LTR pairs were confirmed by ERVAP.

RAxML was applied to compute pairwise maximum likelihood distance for both solo and paired LTRs. GTR+G model was applied as RAxML only allowed this model for pairwise distance function. The rate of neutral substitution for the equine genome has been estimated to be 2.2×10^{-9} substitutions per site per year (Kumar and Subramanian, 2002). The integration time of ERVs is calculated as follows:

$$Date = divergence \div substitution\ rate \div 2$$

2.2.8 Visualising the integration time

The number of integration happened in the evolutionary history was assumed to be a continuous random variable whose values is underlaying an unobserved probability density function. Thus, the probability of the integration falling within a particular interval (or time period) can be visualised using the density plot. The Density plot is a variation of a histogram which use kemel smoothing to plot values over a continuous interval (Hazewinkel, 1994). Density plots were used to display where integration happened concentratedly over the interval (density of integration vs. Mya). The density of number of intention events of each LTR group is estimated based on the estimated integration time of solo LTRs and paired LTRs. The empirical cumulative density function plot is used to visualise the distribution fiction associated with the empirical measure of total number of integration. The ECDF plot displays the fraction of observations of insertion that happened earlier than the specified time point (fraction of integration vs. Mya). For each LTR group, the density plot and ECDF plot are generated using ggplot2.

2.2.9 Orthologue dating

To detect potential orthologs of ERV sequences identified in this study, sequences representing pairs of ERV loci combined with 100 bp flanking DNA were pairwise aligned. If either or both flanking regions could be aligned along with the expected ERV sequence (cut-off 95% identity and 95% query coverage), the loci were considered to be orthologues. If both flanking regions were identified (using the same cut-off) but no ERVs were present the matching site was assumed to represent the pre-integration locus.

Transcriptome of ERVs in equine tissues

Raw reads of 17 tissue samples were downloaded from ENA and NCBI SRA (Table 2-2). Dr Joanna Crispell provided the raw reads of E.derm. The read quality was first visualised by FastQC. Moreover, then Trim Galore was applied to remove adapters and quality control. Reads shorter than 20nt were discarded. The trimmed reads were aligned to the horse reference genome (EquCab2) using TopHat (Trapnell *et al.*, 2012). The Cuffquant program was used to measure the expression of ERV loci, and Cuffnorm was used to normalise expression of the different dataset to the same scale.

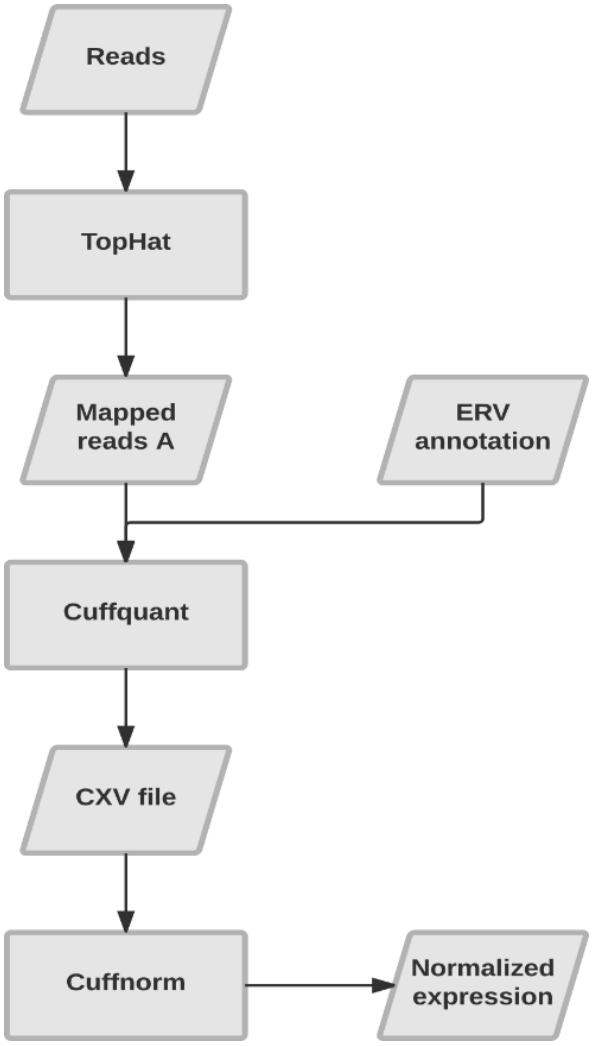


Figure 2-2 Flowchart of transcriptomic analysis. Reads are mapped to the genome using TopHat. Mapped reads are provided as input to Cuffquant directly for estimating expression. The output of CXV is provided as input to Cuffnorm and normalised to the same scale.

3 Development of a novel ERV detection pipeline

3.1 Introduction

In this chapter, I describe the development of a novel bioinformatics pipeline for identification and annotation of ERV proviruses. This pipeline combines phylogenetic screening using the DIGS tools with other software tools for ERV identification and annotation.

3.1.1 Limitations of existing ERV detection tools

Homology-based detection tools for ERV detection required a preconceived notion of the target sequences. Also, homology-based approaches struggle to differentiate between ERVs and other related retroelements, in part because many ERVs are extremely ancient and highly mutated (e.g. HERV-L and MALR which are estimated to have integrated into the genome of early vertebrates over 100 Mya (Smit, 1999)). The internal coding regions of many such ancient proviruses are barely recognisable, and most of them have become solo LTRs. Also, In-frame stop codons and indels cause particular problems for recovering pol sequences. The recovered sequences were usually fragmented or truncated. Furthermore, long pol protein sequences from different ERV classes were relatively divergent, which leads to uncertainties in alignment and phylogenetic inference.

De novo detection tools have been designed for both identification and characterisation of ERVs, but most have been developed to identify full-length ERVs. Thus, a limitation of the *de novo* approaches is that they fail to identify a large number of ERV sequences that are degraded and fragmented.

3.1.2 Phylogenetic screening using the DIGS tool

In this chapter, RT amino acid sequences were used as queries. Because all retroviruses encode RT protein, RT proteins, therefore, can be used to reconstruct evolutionary relationships across the entire *Retroviridae* (Xiong and Eickbush, 1990). Thus, phylogenetic approaches can be used to classify RT loci that are identified by homology-based screening (Tristem, 2000).

In practice, phylogenetic screening can be performed using the DIGS tool. The DIGS tool returns extract sequences of match region in the target sequences. If the probes can be used to infer the phylogeny, the inference of phylogeny based on the results of DIGS tool can be used to improve the DIGS results further when the screening has completed (Figure 3-1).

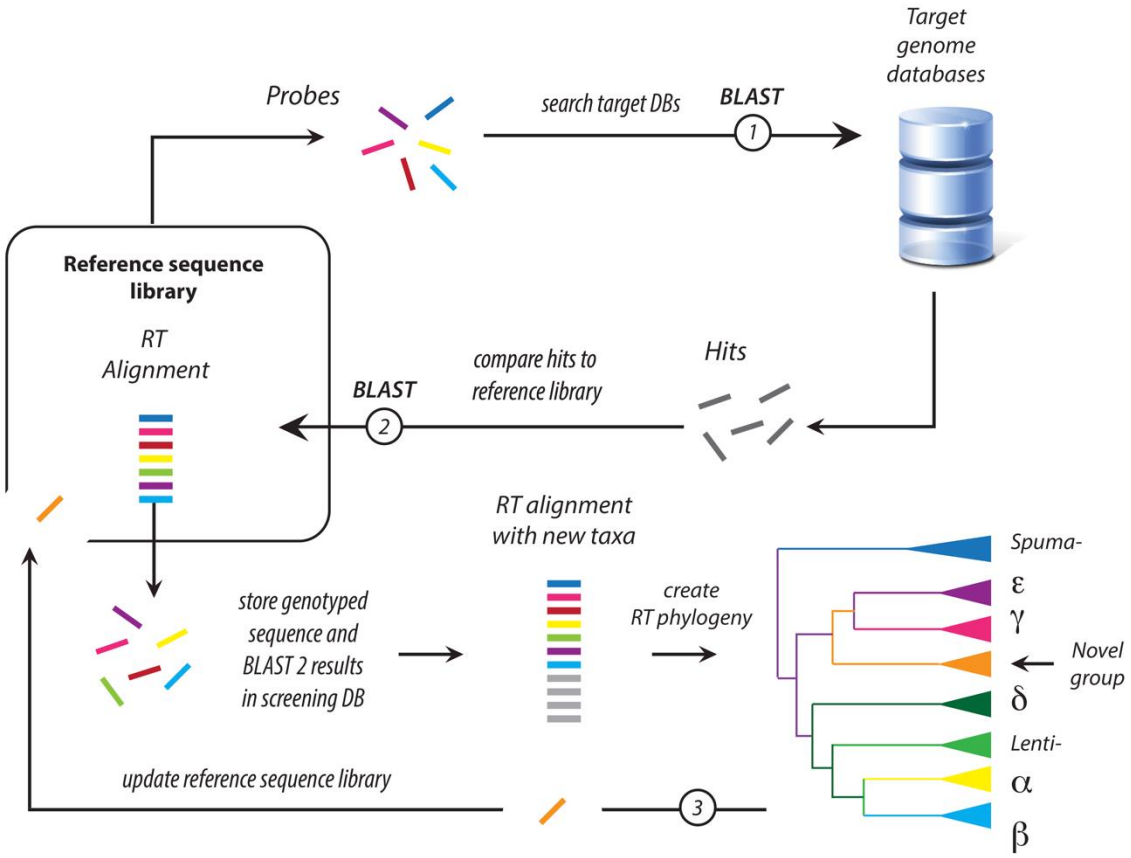


Figure 3-1 Principle of phylogenetic screening using DIGS tool. The general progress of phylogenetic screening using has three steps as marks by number. Probes are extracted from reference library and used by DIGS to screen the target genome. Returned hits are compared to the reference library and then used to infer phylogeny with references. Representatives revealed from the phylogeny are added back to the reference library. Then the whole progress starts again until no more new clades can be found in the phylogeny.

Phylogenetic screening using RT sequences has some limitations. First, some ERV loci might have lost their RT-coding region. In this case, such ERV loci were missed during the identification and classification. Second, the size of RT protein sequence is relatively short; sometimes it will reduce the confidence of phylogeny, e.g. bootstrap value and posterior probability.

3.1.3 The vision for a combined pipeline

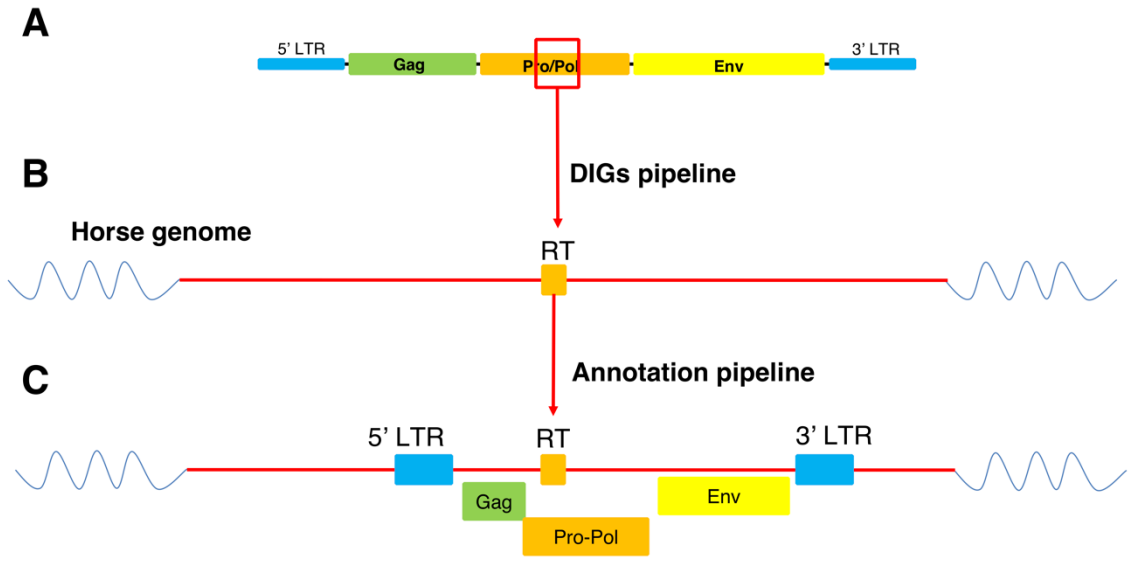


Figure 3-2 Principle of the combined pipeline. A) Preparation of RT reference library and probes; B) DIGS identify RT locus in the horse genome; C) ERVAP annotates the flanking region. DIGS tool and ERVAP are used to predict and annotate the ERVs in the genome. In here, the horse genome is used as an example. RT segments are extracted from the reference sequences. DIGS tool uses RT references to identify RT locus in the horse genome. ERVAP is used to annotate the flanking region of identified RT locus.

In this chapter, I describe the development of an ERV identification and annotation pipeline that integrates a phylogenetic screening approach with other homology-based and *de novo* tools for annotating ERVs.

To allow this, the DIGS tool was used to identify the RT loci in the genome via phylogenetic screening. Then a further set of tools was used to investigate RT loci identified via DIGS screening.

3.2 Results

3.2.1 Validation of the DIGS tool using EVE data

Before building the DIGS tool into my pipeline, I performed a validation of this tool. I used the DIGS tools to detect endogenous retroviral elements (EVEs) from the vertebrate genome. In general, the reference sequence library comprised 53,610 polypeptide gene products of 4,927 viruses obtained from the NCBI virus genomes database. Probes were selected from this library to represent five virus families (*Bornaviridae*, *Filoviridae*, *Circoviridae*, *Parvoviridae* and *Hepadnaviridae*) that have been shown to occur as EVEs in vertebrate genomes (Katzourakis and Gifford, 2010).

Table 3-1 Summary of vertebrate EVEs identified using the DIGS tool

Virus family	Vertebrate lineage											
	Fishes				*Squamates*				*Mammals*			
	S		***NS***		***S***		***NS***		***S***		***NS***	
ssRNA												
Arenaviridae	**0**	(7)	**-**		**0**	(27)	**-**		**-**		**-**	
Flaviviridae	**-**		**-**		**-**		**-**		**-**		**0**	*(1)*
Bornaviridae	**1**		**2**		**8**		**1**		**265**		**98**	*(98)*
Filoviridae	**2**	(30)	**-**		**2**	(88)	**-**		**37**	*(51)*	**17**	*(17)*
Paramyxoviridae	**-**		**2**		**-**		**-**		**-**		**-**	
Nyamiviridae	**-**		**1**		**-**		**-**		**-**		**-**	
ssDNA												
Circoviridae	**-**		**3**		**-**		**4**		**-**		**58**	
Parvoviridae	**3**		**7**		**14**		**11**		**152**		**182**	
Retro-transcribing												
Hepadnaviridae	**-**		**0**	*(2)*	**48**		**193**	(195)	**-**		**0**	*(1)*
Caulimoviridae	**-**		**0**	*(88)*	**-**		**0**	(628)	**-**		**0**	*(1)*
Totals	**6**		**15**		**72**		**209**		**454**		**355**	

S: structural proteins; NS: non-structural proteins; Numbers in brackets to the right show the number of hits obtained in the initial DIGs screen; Bold numbers to the left show the final count, following updates to the reference library. Hyphen represents that no relative hits were found.

Initial screening identified a proportion of hits that were spurious to be derived from non-retroviral EVEs including endogenous retroviruses, retrotransposons and some other genomic sequences. Therefore, I selected representatives of these

sequences and incorporated representative into the reference sequence library. As shown in Table 3-1, these non-retroviral hits were removed from the final output. In sum, 187 vertebrate genomes were screened. All previously reported EVEs for the five virus families were identified. In addition, I identified 744 novel EVEs that have not been described, including 341 novel filovirus and bornavirus-derived EVEs, as well as 328 novel parvovirus-derived EVEs (Katzourakis and Gifford, 2010; Cui *et al.*, 2014).

3.2.2 Development of the ERV Annotation Pipeline (ERVAP)

The **ERV A**nnotation **P**ipeline (ERVAP) uses a combination of homology-based and *de novo* methods to annotate RT-encoding proviral loci that have classified via phylogenetic screening using the DIGS tool. ERVAP uses PERL to negotiate the information flow between different annotation tools and summarise output in a final annotation table.

The ERVAP pipeline has four main stages, illustrated in Figure 3-3:

1. Extract RT loci together with a 20kb flanking region (10Kb each side of the target locus) and run the *LTRharvest* program to identify the boundary of ERV. *LTRharvest* attempts to identify paired LTRs adjacent to the locus.

2. Split loci into those that have putative paired LTRs, and those that do not

    a. RT loci with putative paired LTRs: use *LTRdigest* program to annotate features within the internal coding regions defined by flanking LTRs. These include protein domains (identified using *HMMER*) and several non-coding features (PBS, PPT).

    b. All RT loci: scan for protein-coding domains using *HMMER*

3. For RT loci with paired LTRs – assign LTRs to Repbase groups by blastn, then *RepeatMasker* use identified LTRs as queries to detect solo LTRs from the host genome.

4. Summarise information generated by the pipeline and return an annotation profile.

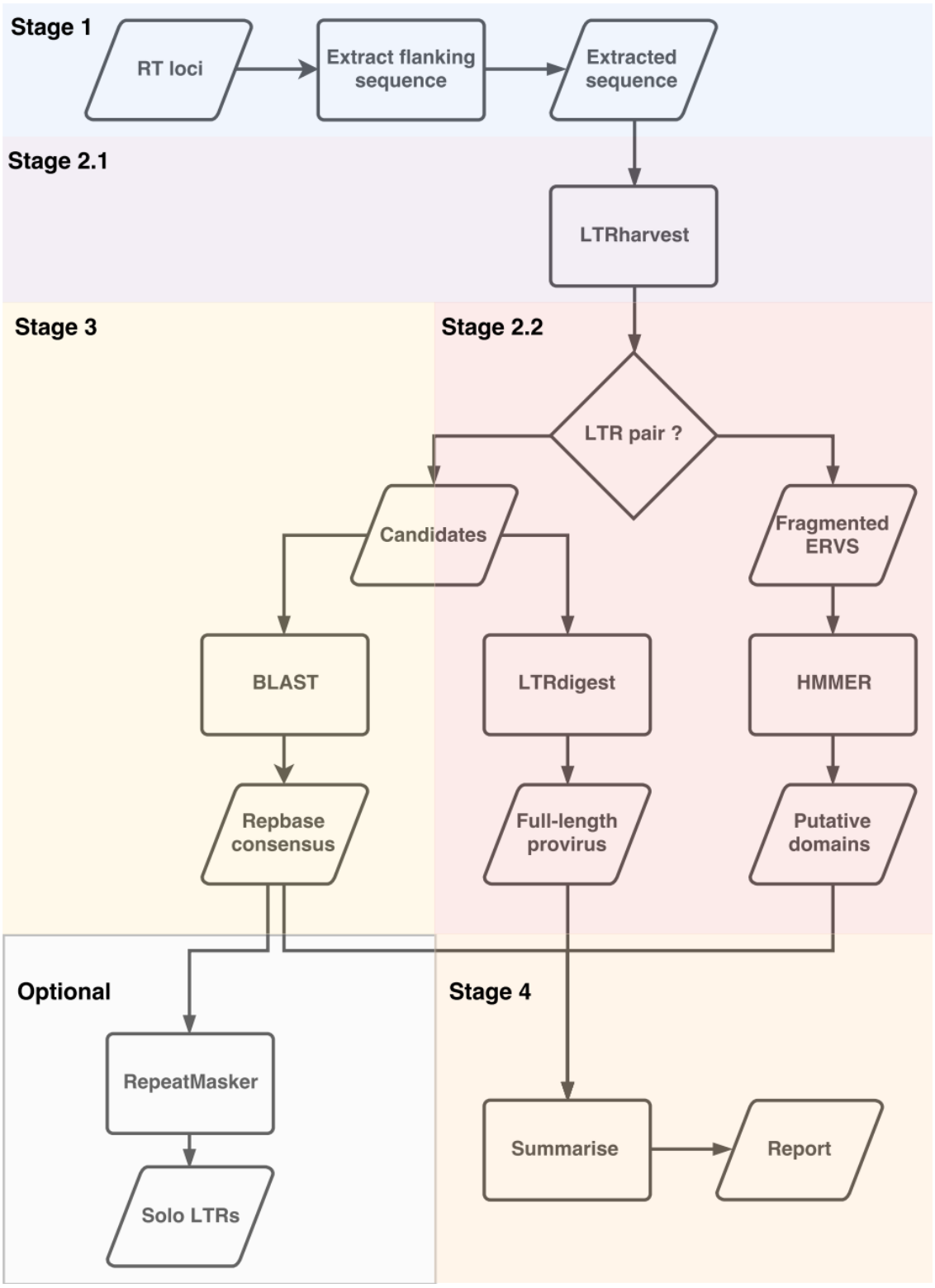


Figure 3-3 Flowchart of ERVAP. The ERVAP consisted of four essential stages and optional stage, framed with Stage 1 (blue), extract the flanking regions of identified RT loci; Stage 2.1 (purple), detect putative LTRs using LTRharvest; Stage 2.2 (red), detect structural features using LTRdigest and HMMER; Stage 3 (yellow), detect and classify LTRs; Stage 4 (orange), summarise and generate final report; Optional stage (grey), detect solo LTRs.

Stage 1. Extraction of RT loci sequences

ERVAP only checks the flanking regions of a candidate RT locus detected by DIGS for LTRs. For each identified RT locus, 10kb sequences are extracted from the upstream and downstream flanking regions (i.e. 5'-10kb + RT locus + 3'-10kb).

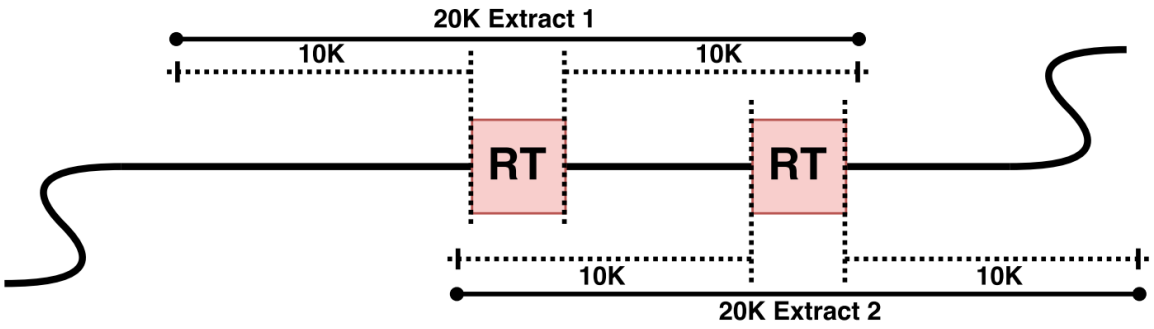


Figure 3-4 The principle of 'fragment' procedure. The genome sequence is shown as the black line. Three identified RT hits are shown as red frames and marked with a number. Dash cross bars show the range of extracted 10k flanking regions from each side of RT hits. Black lines with circle heads represent the final region of the extracted sequence.

The length of an extracted sequence is limited by the maximum length of known ERVs and exogenous retroviruses (~12kb). The relative location of RT queries using by DIGS is around 5k~6k in a 10kb provirus. Moreover, for a full-length provirus, the distance from the start of potential LTRs to identified RT locus is around 5kb. Also, potential provirus region may contain long insertions. Thus, the length of the extracted flanking region is set to 10kb for each side to cover the potential provirus region completely. The extracted sequence of each RT locus and its flanking regions is stored into an individual file. As flanking regions of identified RT loci are considered as potential ERVs loci, even these flanking regions may overlap each other (Figure 3-5). Such settings are to avoid missing mutilated or previously unknown ERVs.

Stage 2.1 Detection of putative long terminal repeat (LTR) pairs in the extracted sequence

The LTRharvest program is run on each extracted sequence individually. For each extracted sequence, LTRharvest is utilised to identify two nearly identical LTRs matching the similarity and length constraints (Ellinghaus, Kurtz and Willhoeft, 2008). Any pair of matching sequences found by LTRharvest that meet the criteria for being LTRs is considered 'candidates'. As multiple invasion and

retrotransposition events can happen in the same or near locations, and tandem repeats are abundant in the mammalian genomes, LTRharvest sometimes can detect multiple candidates at the same locus. In ERVAP, all detected candidates are considered separately for further analysis (see Figure 3-5).

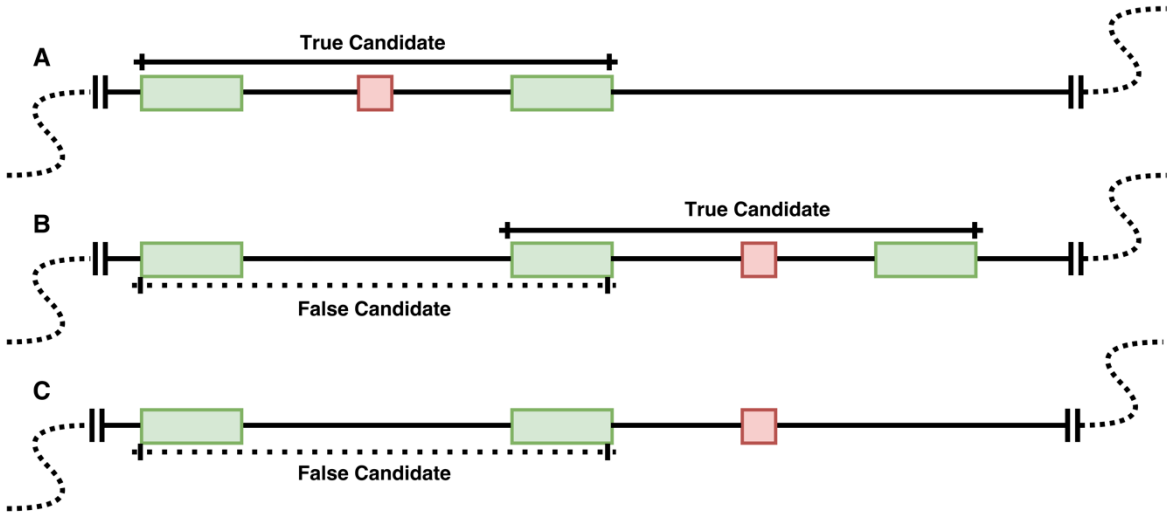


Figure 3-5 The 'candidates' chosen by ERVAP for analysis with LTRdigest.A) True candidate; B) Candidates of tandem repeat; C) False candidate in the flanking region. The genome sequence is shown as black lines. LTRs identified by *LTRharvest* are shown as green frames, RT hits detected by DIGS are shown as red frames. Black crossbars represent candidates that ERVAP accepts, while dash crossbars show the false candidates that ERVS discards.

Stage 2.2 Detection of conserved protein domains within the retrovirus internal region

ERVAP uses both LTRdigest and HMMER to detect structural features in the internal regions of putative proviruses. For RT loci that are flanked by paired LTRs, LTRdigest is used (Figure 3-6). LTRdigest is a downstream analysis tool for LTRharvest; it requires a prediction of LTR boundaries to define the internal region. The LTRdigest tool includes the functionality of the HMMR package. One function of HMMR (called pHMMER) is used to identify putative retroviral coding domains. LTRdigest also contains custom-built algorithms for detection of the PBS and PPT in sequences in candidate proviruses. PBS detection requires a tRNA library that contains tRNA sequences for the species under comparison. Importantly, LTRdigest considers the orientation of genome features within candidate proviruses in the annotation.

Some RT loci are not flanked by paired LTRs – either due to truncation or very high divergence. Entire proviruses can be truncated due to the poor-quality sequencing or long deletions. In these cases, HMMer is used to search for protein-coding domains adjacent to the RT hit directly.

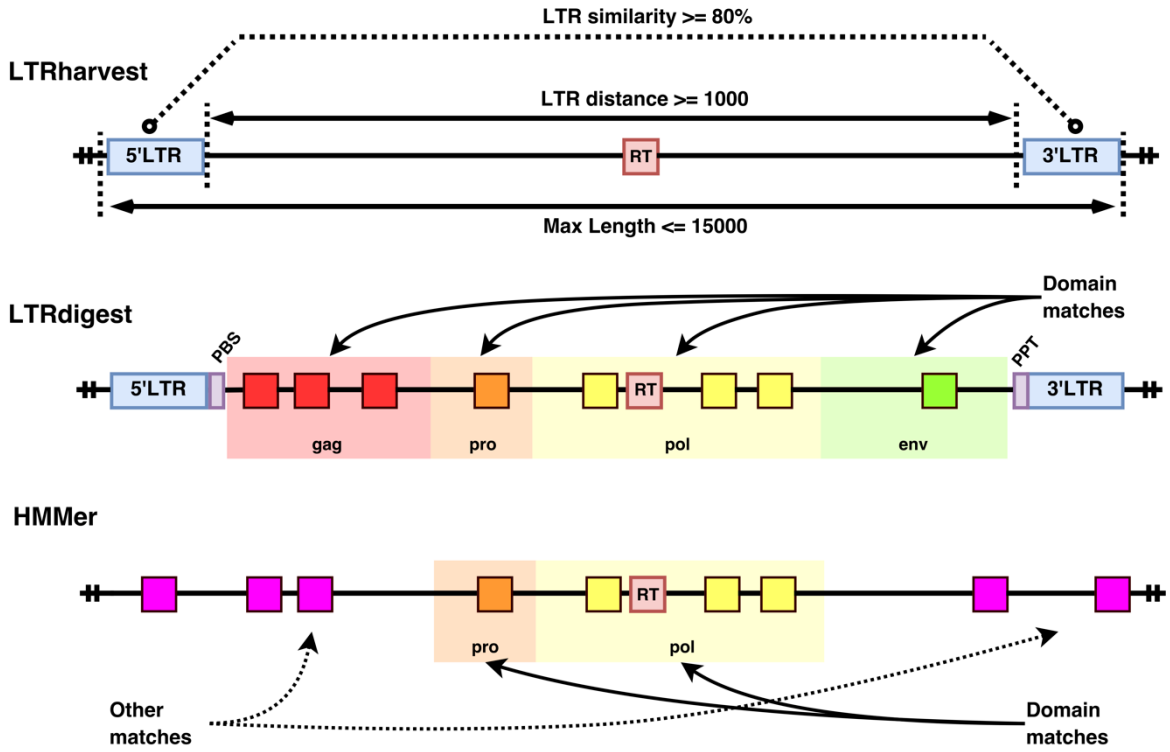


Figure 3-6 Example of annotation processes of the ERVAP pipeline.LTRharvest is used to predict the paired LTR. LTRdigest is used to annotate the internal region. If no paired LTR can be found, HMMer is used to screen the whole extract region. Identified LTRs are shown as blue frames, and RT hits detected by DIGS are shown as red frames.

Stage 3. Detection and classification of solo LTRs

The ERVAP pipeline uses RepeatMasker to perform the detection of solo LTRs. RepeatMasker is a general repeat detection program; it can efficiently and pervasively detect repeat sequences from a large genome. However, the processes of RepeatMasker is time-consuming. Therefore, ERVAP does not run RepeatMasker directly. Alternatively, it provides the query library for RepeatMasker. More specifically, ERVAP compared the identified paired LTR with consensus sequences of Repbase and assigned the identified LTRs with Repbase labels by BLAST. Only paired LTRs go through this process because the process of DIGS and LTRdigest has linked these LTRs with specific ERVs. If BLAST cannot assign identified LTRs with consensus sequences, identified LTRs are considered as ‘novel’.

Stage 4. Detection and classification of solo LTRs

As the last step of the ERVAP, all information generated by the stage 1~3 was summarised to generate the final report. The final report included two major sections. The first section is a report in CSV format which contains RT coordinates estimated by the DIGS tools (Figure 3-7), the RT classification inferred by phylogenetic reconstruction, LTRs found by LTRharvest and LTR classification assigned to Repbase, as well as structural features annotated by LTRdigest and HMMER (Figure 3-8).

The second section is the visualisation of records in the final report. The schematic representation is generated using the *AnnotationSketch* function of the Genometools package. This section is still in the early stage which can only generate a rough layout without explicit annotations. Thus, the second section has not been included in the final report yet.

	K	L	M	N	O	P	Q	R	S	T	U	V	W	X
1	RT_FILE	FLANKING_FILE	REPEAT_S	REPEAT_E	LLTR_S	LLTR_E	RLTR_S	RLTR_E	LLTR_REP_ID	LLTR_REP_SIM	LLTR_REP_SOCRE	RLTR_REP_ID	RLTR_REP_SIM	RLTR_REP_SOCRE
2	chr1_110164	chr1_110164521_110	1	20399	0	0	0	0	NONE	0	0	NONE	0	0
3	chr1_26249(	chr1_26249090_2624	1	20399	0	0	0	0	NONE	0	0	NONE	0	0
4	chr1_156536	chr1_156536706_156	1	20399	0	0	0	0	NONE	0	0	NONE	0	0
5	chr1_111612	chr1_111612360_111	1	20399	0	0	0	0	NONE	0	0	NONE	0	0
6	chr1_533921	chr1_53392193_5339	1	20399	0	0	0	0	NONE	0	0	NONE	0	0
7	chrUn_8822(	chrUn_88220_88528_	1	20399	0	0	0	0	NONE	0	0	NONE	0	0
8	chrUn_3770:	chrUn_37702446_377	1	20399	0	0	0	0	NONE	0	0	NONE	0	0
9	chr1_408669	chr1_40866919_4086	5080	14612	5085	5494	14177	14607	ERV1-4-EC_LTR	85.1	255	ERV1-4-EC_LTR	86.72	241
10	chr1_111787	chr1_111787203_111	8887	14644	8892	9343	14187	14639	ERV2-1-EC_LTR	92.49	453	ERV2-1-EC_LTR	91.61	453
11	chr17_25166	chr17_25166971_251	1	20399	0	0	0	0	NONE	0	0	NONE	0	0
12	chr1_135972	chr1_13597213_1359	1	20399	0	0	0	0	NONE	0	0	NONE	0	0
13	chr1_583696	chr1_58369630_5836	1	20399	0	0	0	0	NONE	0	0	NONE	0	0
14	chr1_759863	chr1_75986333_7598	1	20399	0	0	0	0	NONE	0	0	NONE	0	0
15	chr1_783235	chr1_78323554_7832	1	20399	0	0	0	0	NONE	0	0	NONE	0	0
16	chr1_825696	chr1_82569604_8257	5846	13953	5851	6309	13484	13948	ERV2-1-EC_LTR	92.16	459	ERV2-1-EC_LTR	92.32	456
17	chr1_920956	chr1_92095659_9209	1	20399	0	0	0	0	NONE	0	0	NONE	0	0
18	chr7_796407	chr7_79640797_7964	1	20399	0	0	0	0	NONE	0	0	NONE	0	0
19	chr1_838346	chr1_83834658_8383	6225	18589	6230	6808	17984	18584	NONE	0	0	NONE	0	0
20	chr1_101592	chr1_101592236_101	1	20399	0	0	0	0	NONE	0	0	NONE	0	0
21	chr1_408527	chr1_40852749_4085	1	20399	0	0	0	0	NONE	0	0	NONE	0	0
22	chrX_675890	chrX_67589055_6758	1	20399	0	0	0	0	NONE	0	0	NONE	0	0
23	chr1_110447	chr1_110447504_110	1	20399	0	0	0	0	NONE	0	0	NONE	0	0
24	chr1_110559	chr1_110559230_110	1	20399	0	0	0	0	NONE	0	0	NONE	0	0
25	chr1_294821	chr1_29482117_2948	1	20399	0	0	0	0	NONE	0	0	NONE	0	0
26	chr1_183836	chr1_183836165_183	1	20399	0	0	0	0	NONE	0	0	NONE	0	0
27	chr1_177343	chr1_177343939_177	1	20399	0	0	0	0	NONE	0	0	NONE	0	0
28	chr1_225753	chr1_22575320_2257	6239	13982	6244	6701	13520	13977	ERV2-1-EC_LTR	92.37	459	ERV2-1-EC_LTR	92.37	459
29	chr10_14383	chr10_14383167_143	1	20399	0	0	0	0	NONE	0	0	NONE	0	0
30	chr1_447891	chr1_44789199_4478	1	20399	0	0	0	0	NONE	0	0	NONE	0	0
31	chr4_260408	chr4_26040870_2604	1	20399	0	0	0	0	NONE	0	0	NONE	0	0
32	chr7_342883	chr7_34288383_3428	1	20399	0	0	0	0	NONE	0	0	NONE	0	0
33	chr1_126883	chr1_126883949_126	1	20399	0	0	0	0	NONE	0	0	NONE	0	0
34	chr1_126884	chr1_126884714_126	1	20399	0	0	0	0	NONE	0	0	NONE	0	0
35	chr1_130484	chr1_130484671_130	1	20399	0	0	0	0	NONE	0	0	NONE	0	0

Figure 3-7 Example of DIGS and ERVAP report (part 1). By summarising results of *LTRharvest*, *LTRdigest* and HMMer, ERVAP generates a CSV file for all screened regions. Colour frames covered the major information section. Blue squares circle the example predications of *LTRharvest* and *LTRdigest* for the full-length ERVs. Red square example predications of HMMer for potential regions without paired LTRs.

	Y	Z	AA	AB	AC	AD	AE	AF	AG	AH	AI	AJ	AK	AL	AM	AN	AO	AP	AQ	AR	AS	AT
1	PBS_S	PBS_E	PBS_TYPE	RT_S	RT_E	RT_INSIDE	PROT_HITS_NUM	GENES														
2	NONE	NONE	NONE	10000	10393	NO	18	RNase_H	RVT_1	RVT_1	RVT_1	RVT_1	RVT_1	RVT_1	RVT_1	rve	rve	Gag_p30	Gag_p30	Gag_p30	Gag_p30	Exo_endo
3	NONE	NONE	NONE	10000	10366	NO	2	RNase_H	RVP													
4	NONE	NONE	NONE	10000	10192	NO	17	RVT_1	zf-CCHC	zf-CCHC	zf-CCHC	zf-CCHC	GP41	Gag_p24	Gag_p24	Gag_p24	rve	dUTPase	Gag_p10	RVT_thun	DUF1725	DUF1725
5	NONE	NONE	NONE	10000	10402	NO	10	RNase_H	RVP	RVT_1	zf-CCHC	zf-CCHC	zf-CCHC	rve	Gag_p30	Gag_p30	DUF1725					
6	NONE	NONE	NONE	10000	10396	NO	17	RVT_1	RVT_1	RVT_1	RVT_1	zf-CCHC	zf-CCHC	GP41	Gag_p24	rve	dUTPase	Gag_p10	Exo_endo	RVT_thun	DUF1725	Exo_endo
7	NONE	NONE	NONE	10000	10309	NO	12	RVT_1	RVT_1	RVT_1	RVT_1	RVT_1	RVT_1	rve	Exo_endo	Exo_endo	DUF1725	Exo_endo	Exo_endo	_phos_2		
8	NONE	NONE	NONE	10000	10396	NO	12	RVP	RVP	RVP	RVP	RVT_1	RVT_1	RVT_1	rve	Exo_endo	DUF1725	DUF1725	Exo_endo	_phos_2		
9	14151	14161	trna=Equus_caballu	10000	10150	YES	4	TLV_coat	RNase_H	RVT_1	Gag_p30											
10	NONE	NONE	NONE	10000	10396	YES	7	RVT_1	RVT_thun	RVT_thun	Integrase_	rve	rve	GP41								
11	NONE	NONE	NONE	10000	10264	NO	11	RNase_H	RNase_H	RVP	RVT_1	RVT_1	TLV_coat	TLV_coat	Gag_p30	Gag_p30	BLVR	BLVR				
12	NONE	NONE	NONE	10193	10213	NO	0															
13	NONE	NONE	NONE	10000	10309	NO	6	RVT_1	rve	Exo_endo	Exo_endo	DUF1725	DUF1725									
14	NONE	NONE	NONE	10000	10189	NO	4	RNase_H	RVT_1	Exo_endo	DUF1725											
15	NONE	NONE	NONE	10000	10189	NO	3	RNase_H	RVT_1	RVT_1												
16	NONE	NONE	NONE	10000	10405	YES	9	GP41	rve	Integrase_	RVT_thun	RVT_1	zf-CCHC	Gag_p24	Gag_p10	dUTPase						
17	NONE	NONE	NONE	10000	10387	NO	5	RNase_H	RVT_1	RVT_1	RVT_1	zf-H3C2										
18	NONE	NONE	NONE	10000	10228	NO	11	RNase_H	RVP	RVT_1	RVT_1	RVT_1	RVT_1	RVT_1	rve	Gag_p30	Gag_p30	DUF1725				
19	NONE	NONE	NONE	10000	10342	YES	1	DUF1725														
20	NONE	NONE	NONE	10000	10396	NO	1	RNase_H														
21	NONE	NONE	NONE	10000	10261	NO	3	zf-CCHC	TLV_coat	DUF1725												
22	NONE	NONE	NONE	10000	10246	NO	11	RNase_H	RVP	RVT_1	RVT_1	zf-CCHC	zf-CCHC	zf-CCHC	rve	Gag_p30	Gag_p30	DUF1725				
23	NONE	NONE	NONE	10000	10240	NO	4	RNase_H	RVT_1	RVT_1	rve											
24	NONE	NONE	NONE	10000	10258	NO	2	RVT_1	RVT_1													
25	NONE	NONE	NONE	10000	10405	NO	14	RVP	RVT_1	zf-CCHC	zf-CCHC	GP41	Gag_p24	Gag_p24	rve	Gag_p10	RVT_thun	RVT_thun	DUF1725	zf-CCHC_5	zf-CCHC_5	
26	NONE	NONE	NONE	10000	10411	NO	12	RNase_H	RVP	RVT_1	zf-CCHC	zf-CCHC	zf-CCHC	TLV_coat	rve	rve	Gag_p30	Gag_p30	zf-H2C2			
27	NONE	NONE	NONE	10000	10351	NO	10	RNase_H	RVP	zf-CCHC	rve	rve	Gag_p30	Gag_p30	Gag_p30	DUF1725	Asp_protease_2					
28	NONE	NONE	NONE	10000	10405	YES	7	GP41	rve	RVT_thun	RVT_1	zf-CCHC	Gag_p24	dUTPase								
29	NONE	NONE	NONE	10000	10213	NO	10	RNase_H	RVT_1	RVT_1	RVT_1	RVT_1	rve	Gag_p30	Gag_p30	Gag_p30	zf-H2C2					
30	NONE	NONE	NONE	10000	10393	NO	1	DUF1725														
31	NONE	NONE	NONE	10000	10345	NO	10	RNase_H	RNase_H	RVT_1	RVT_1	RVT_1	Gag_p30	Gag_p30	Gag_p30	Gag_p30	DUF1725					
32	NONE	NONE	NONE	10000	10330	NO	10	RNase_H	RVT_1	RVT_1	RVT_1	RVT_1	rve	rve	Gag_p30	Gag_p30	Gag_p30					
33	NONE	NONE	NONE	10000	10117	NO	10	RNase_H	RVP	RVT_1	RVT_1	zf-CCHC	TLV_coat	rve	Exo_endo	DUF1725	Asp_protease_2					
34	NONE	NONE	NONE	10000	10096	NO	10	RNase_H	RVP	RVT_1	RVT_1	zf-CCHC	TLV_coat	rve	Exo_endo	DUF1725	Asp_protease_2					
35	NONE	NONE	NONE	10000	10249	NO	4	RNase_H	RVT_1	zf-CCHC	rve											

Figure 3-8 Example of DIGS and ERVAP report (part 2). By summarising results of *LTRharvest*, *LTRdigest* and HMMer, ERVAP generates a CSV file for all screened regions. Colour frames covered the major information section. Blue squares circle the example predications of *LTRharvest* and *LTRdigest* for the full-length ERVs. Red square example predications of HMMer for potential regions without paired LTRs.

3.2.3 Demonstration of the ERVAP pipeline

In this section, I present an example to demonstrating how ERVAP was used for the identification and annotation of EqERV.Beta1 in the horse genome. Results were compared to that of the previous study. EqERV.Beta1 was the first ERV to be identified in the horse genome by *in silico* screening. A full-length provirus of EqERV.Beta1 was identified on the chromosome 5:1998769-2009202(-) (NW_001867417.1, Feb 2011). The length of provirus was 10434 nt, and it has two LTRs around 1361 nt in length and four nearly complete genes (van der Kuyl, 2011).

The DIGS tool was used to identify EqERV.Beta1 RT loci in the horse genome (EquCab2). The query for similarity searches consisted of the EqERV.Beta1 RT sequence. Three EqERV.Beta1 RT loci were identified on chromosome 5 (NC_009148.3, Jan 2018), seven loci located on chromosome "unknown". LTRharvest found that paired LTRs flanked only the RT locus on chromosome 5: 16,132,369-16,132,776(-). The paired LTRs identified by LTRharvest suggested that potential provirus located at chromosome 5: 16,136,909-16,147,356 (-) (Figure 3-9).

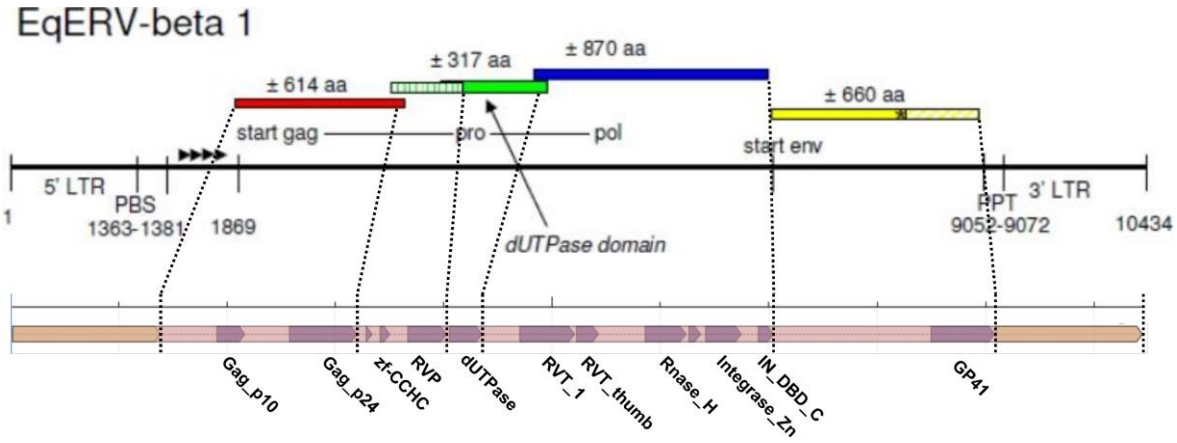


Figure 3-9 Comparison of previous study and ERVAP annotation. (A) Schematic representation of the EqERV.Beta1 genome organization (van der Kuyl, 2011); (B) Schematic representation of the EqERV.Beta1 generated by ERVAP.

The coordinates of EqERV.Beta1 provirus reported by van der Kuyl was based on the NW_001867417.1. This record has been removed, and the new reference sequence of horse chromosome 5 was NC_009148.3. The DIGS tool for this example used NC_009148.3. BLAST was performed to adjust coordinates between NW_001867417.1 and NC_009148.3. After adjusting coordinates, the provirus

identified by ERVAP was located at the same location of the provirus reported EqERV.Beta1.

The identified LTRs were 1365 nt in length. LTR sequence was extracted and compared to the EqERV.Beta1 LTR sequence by BLAST. The identity was 100%. The provirus was 10437 nt in length. Protein domains, Gag_p10, Gag_p24 of *gag*, rev, integrase, RNase_H, RVT_1, dUTPase of *pol*, PRV of pro and GP41 of *env*, were identified within the internal coding region defined by LTRharvest (Figure 3-9). Additional RT loci were identified by DIGS located at 11,522 bp downstream of 3'LTR identified by LTRharvest. Protein domains of all four retroviral genes were found to cluster in this region. This result suggested the presence of tandem EqERV.Beta1 insertion. This finding corresponds precisely to the previous report.

The identified LTRs were then used as a custom library for RepeatMasker to detect solo LTRs of EqERV.Beta1. In total, RepeatMasker identified 350 solo LTRs, while the previous report suggested 227 loci.

3.3 Conclusions

In this chapter, I developed ERVAP - a novel pipeline for performing efficient, comprehensive genome-wide screening of ERVs that integrates a phylogenetic screening approach (implemented using the DIGS tool) with other software tools for detecting and characterising ERVs (GenomeTools and HMMR). This pipeline combines homology-based, and *de novo* approaches to ERV detection, providing added power for detecting and characterising ERVs. An example based on the horse genome was used to demonstrate the application of this pipeline.

The screening strategy implemented in ERVAP has two important advantages. First, it exploits the sensitivity of homology-based screening using RT to identify divergent sequences. Such insertions are easily missed by more stringent ERV-specific detection tools optimised for full-length elements. Secondly, it combines the classification power of phylogenetic screening with a high throughput approach for annotating ERV sequences, including both full-length proviruses, truncated ERVs and even highly degenerated fragments such as solo LTRs.

4 Identification, phylogenetic classification and characterisation of ERVs in perissodactyl genomes.

4.1 Introduction

The first assembled horse genome (EquCab1) was released by the Broad Institute in January 2007 and updated to the current version (EquCab2) in September of the same year (Wade *et al.*, 2009). Since then four separate investigations of equine ERVs have been performed (van der Kuyl, 2011; Brown *et al.*, 2012; Garcia-Etxebarria and Jugo, 2012; Gim and Kim, 2017).

The first published study of an equine ERV focused on a *Betaretrovirus* lineage. A full-length provirus belonging to this lineage was identified on chromosome 5. The pol gene showed a very close relationship to MMTV and was named 'EqERV-beta1' (van der Kuyl, 2011). In the previous chapter, this ERV was used as an example to test the ERV annotation pipeline I have developed.

Two further studies were published in 2012. The first of these identified 1947 putative ERV insertions. These insertions were then grouped into 15 families and three major classes. ERV families were termed as 'EqERV1-15'. (Garcia-Etxebarria and Jugo, 2012). The second reconstructed phylogenetic trees based on the alignment of *gag*, *pol* and *env* with known viruses, respectively. In total, 978 ERV insertions were identified and categorised as gamma-, epsilon- and betaretroviruses (Brown *et al.*, 2012).

The fourth and most recent study was published last year (2017) and identified 22 different ERV types in the horse genome. ERV types were defined based on the tRNA used by the PBS. All 22 ERVs types are categorised into six families in ERV classes I and II. This study used the RetroTector program to generate representative genome structures of ERV families (Gim and Kim, 2017).

In all studies, ERVs belonging to both class I and II were identified. Brown *et al.* (2012) and Gim and kim (2017) suggested that the class I ERVs included Gamma- and Epsilon-like ERVs. However, Garcia-Etxebarria *et al* (2014) did not present similar evolutionary relationships within class I ERVs. For the class II, all studies

found four distinct ERV lineages and one of which is EqERV.Beta1. Moreover, the other three lineages were suggested as Beta-like elements. For the class III, only Garcia-Etxebarria et al., 2012 showed the presence of two distinct families of class III.

In the previous chapter, I developed and tested a novel pipeline for ERV identification and characterisation (ERVAP) that combines homology-based phylogenetic screening with other approaches for ERV identification and annotation. In this chapter, I describe the use of this pipeline to characterise ERVs in the *E.caballus* genome and those of other perissodactyls: the donkey (*Equus asinus*), the white rhinoceros (*Ceratotherium simum*), as well as several half-asses and zebras.

4.2 Results

4.2.1 Collation and preparation of perissodactyl genome sequences

At the time this work was initiated, whole genome assemblies were available for four perissodactyl species: the domestic horse (*Equus caballus*); the domestic donkey (*Equus asinus africanus*); Przewalski's horse (*Equus ferus przewalskii*), and the southern white rhinoceros (*Ceratotherium simum*). The domestic horse has been assembled to chromosome level, while the genomes of the donkey, Mongolian horse, Przewalski's horse and white rhinoceros are assembled to scaffold level via *de novo* assembly (Huang *et al.*, 2014).

Also, several equine genomes were available in raw read format. These included several species: the Somali wild ass (*Equus asinus somalicus*), onager (*Equus hemionus*), kiang (*Equus kiang*), plains zebra (*Equus burchellii boehmi*), Burchell's zebra (*Equus burchellii quagga*), Hartmann's mountain zebra (*Equus zebra hartmannae*) and Grevy's zebra (*Equus grevyi*) as well as genome sequences of one *E.caballus* breeds (see Table 2-1 in chapter II). Genome sequences that were only available in raw read format were assembled (Table 4-1) so that they could be screened for ERVs using the DIGS tool and ERVAP.

Table 4-1 Assembly summary

Organism	Reference	Trimmed Reads	Alignment rate
Equus asinus somalicus	Willy	447,96,530	94.62%
Equus burchellii boehmi	Willy	53,869,589	95.67%
Equus burchellii quagga	Willy	54,396,006	77.25%
Connemara pony	EquCab2	315,753,535	98.25%
Equus grevyi	Willy	11,833,549	95.62%
Equus hemionus	Willy	44,708,162	97.46%
Equus kiang	Willy	15,737,212	97.31%
Equus zebra hartmannae	Willy	7,801,162	95.81%

De novo assembly is the most common method for assembling short reads without the knowledge of reference sequences. I used comparative genome assembly to assemble these genomes, using the thoroughbred horse (EquCab2) and donkey (Willy) as reference genomes for other unassembled species. First, the horse

reference genome (EquCab2) was used to build the assembly of all domestic horse breeds. Second, all short reads of Somalian wild ass, half asses and zebras were aligned against both the horse reference genome (EquCab2) and the donkey scaffolds (Willy).

4.2.2 Identification of RT sequences via phylogenetic screening

A total 18,290 RT hits were identified in the 17 genomes screened here (Table 4-2). In the first iteration of the screening pipeline, all RT hits were 'genotyped' by BLAST-based comparison to a reference library comprised of RT sequences from previously characterised ERVs and exogenous retroviruses (Appendix I). All identified ERVs were found to belong to the *Orthoretrovirinae* subfamily.

Using the initial library, 73.67% of the orthoretrovirus RT sequences I identified were more similar to those of ERVs and exogenous retroviruses from non-perissodactyl species than to perissodactyl ERVs. This finding suggested that representatives of one or more perissodactyl ERV lineages were missing from the initial reference library. This was expected since relatively few equine ERVs and retroviruses were included at the outset. I, therefore, decided to identify these unrepresented RT lineages and included representatives of them in the reference library used with DIGS.

To do this, I created maximum likelihood phylogenies of the RT sequences recovered from three representative perissodactyl species: the horse, donkey and white rhinoceros (Figure 4-1, 4-2, 4-3). For each species, all ERVs that are longer than 300 bp with bit-score > 50 were aligned with the 66 RT sequences in the initial reference library (Appendix 1). 175, 370 and 288 identified RT sequences were used for the phylogenetic reconstruction of donkey, horse and rhinoceros, respectively. All three trees had a highly similar topology, exhibiting multiple robustly-supported clades that were comprised entirely of sequences recovered from perissodactyl genomes. I, therefore, selected 3~4 representative RT sequences for each of those nine clades and included them into the reference library. When sequences recovered by DIGS were classified using this updated reference library, 99% of 18,290 RT sequences were assigned to perissodactyl ERV lineages.

Table 4-2 Summary of RT hits identified by DIGS in perissodactyls

Organism	Common Name	Count
Rhinocerotidae		
Ceratotherium simum	Southern white rhinoceros	1506
Equidae		
Equus asinus africanus	Donkey	1127
Equus asinus somalicus	Somali wild ass	903
Equus burchellii boehmi	Plains zebra	893
Equus burchellii quagga	Burchell's zebra	841
Equus grevyi	Grevy's zebra	903
Equus zebra hartmannae	Hartmann's mountain zebra	902
Equus caballus	Horse (thoroughbred)	1384
Equus caballus	Horse (Arabian)	1254
Equus caballus	Horse (Icelandic)	1213
Equus caballus	Horse (Norwegin Fjord)	1236
Equus caballus	Horse (Standardbred)	1246
Equus caballus	Horse (Connemara Pony)	1366
Equus caballus	Horse (Mongolian)	569
Equus ferus przewalskii	Przewalski's Horse	1165
Equus hemionus	Onager	912
Equus kiang	Kiang	870
Total		18290

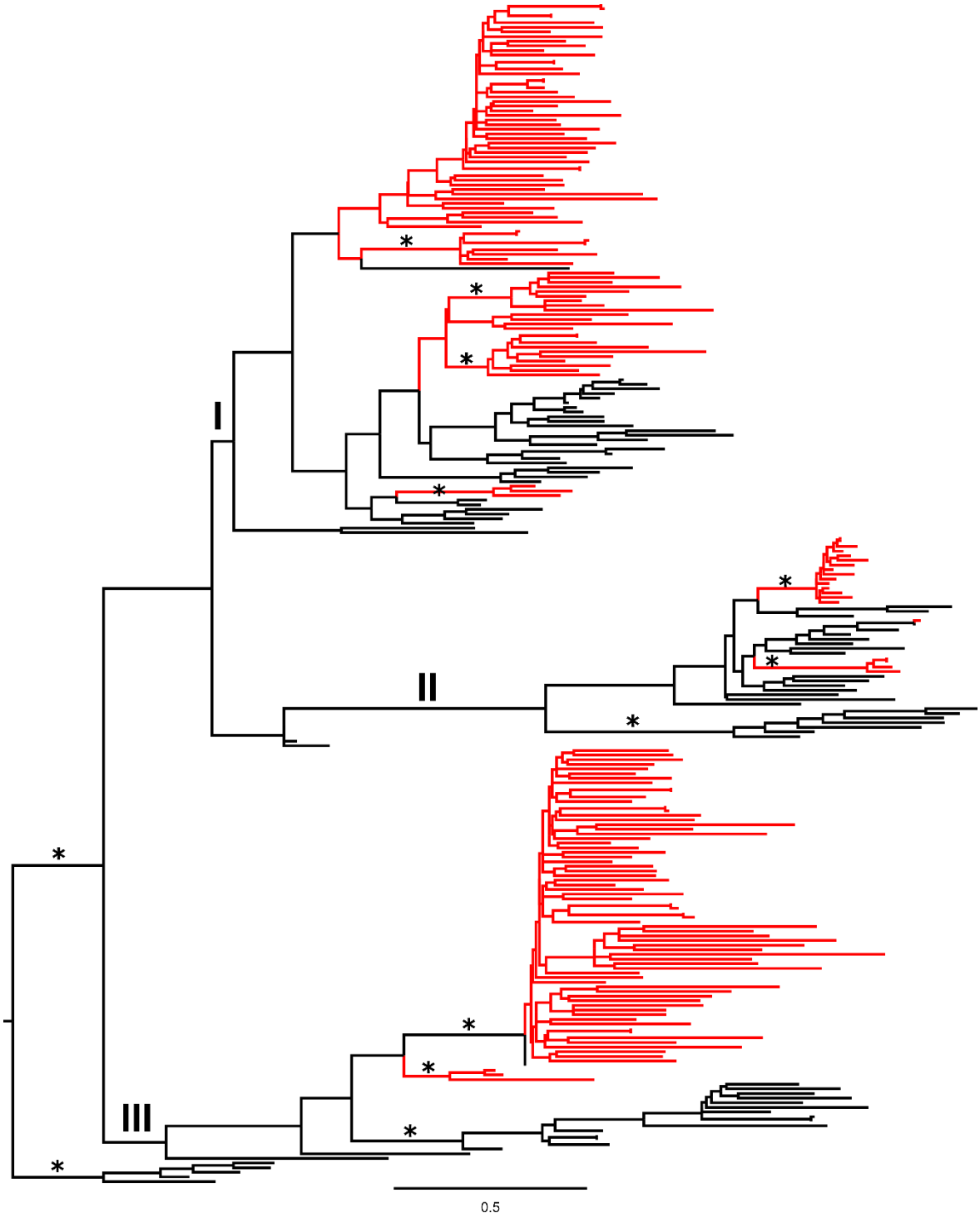


Figure 4-1 Phylogenetic screening of RTs in the donkey genome. Phylogeny of reference RT sequences and 175 RT sequences detected from the donkey genome by DIGS. Main branches that lead to Class I, II and III ERVs are marked. RT references are shown in black and detected RT sequences in red. The asterisk marks the main branches with a bootstrap value over 80.

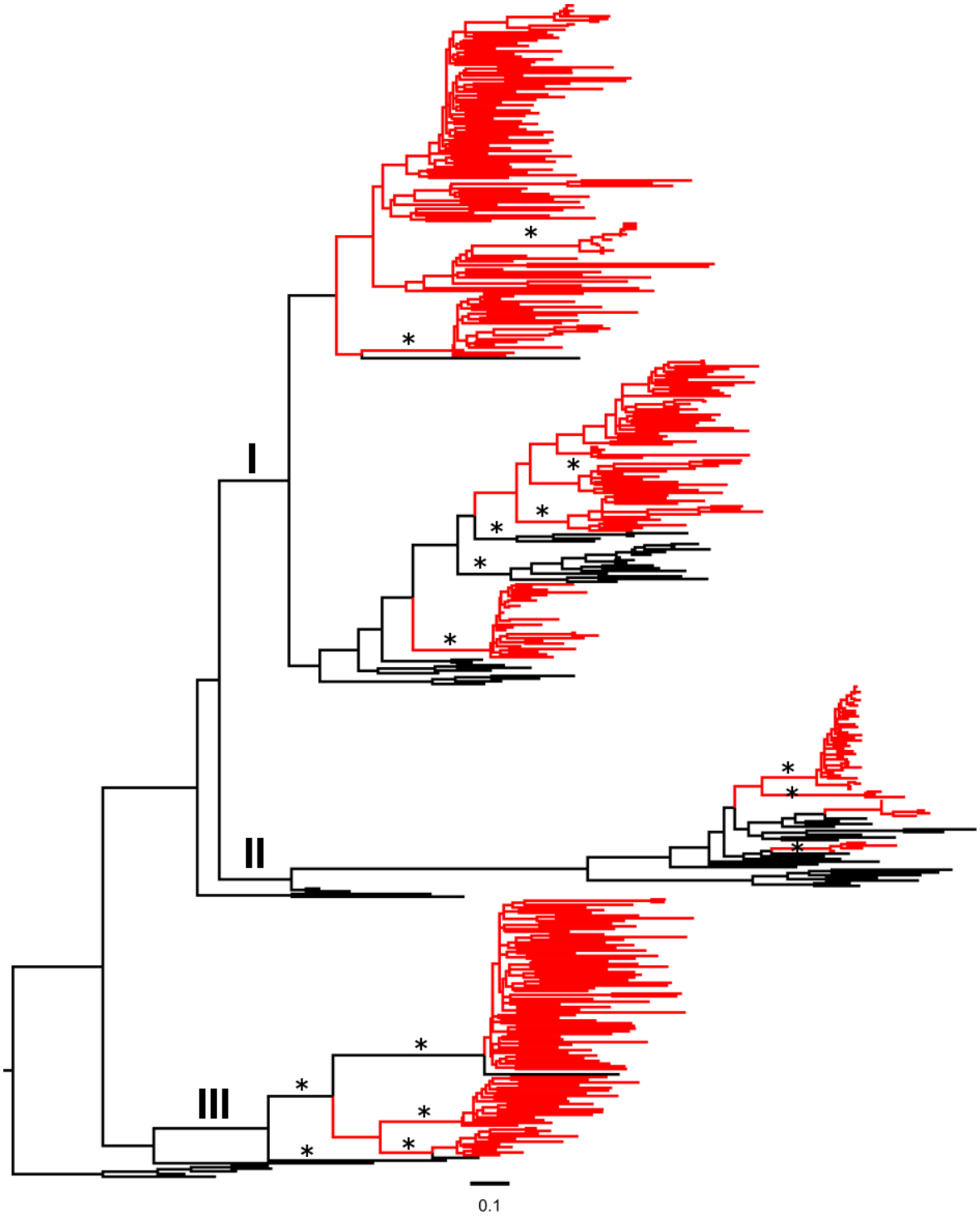


Figure 4-2 Phylogenetic screening of RTs in the horse genome. Phylogeny of reference RT sequences and 370 RT sequences detected from the horse genome by DIGS. Main branches that lead to Class I, II and III ERVs are marked with Roman numbers. RT references are shown in black and detected RT sequences are shown in red. The asterisk marks the main branches with a bootstrap value over 80.

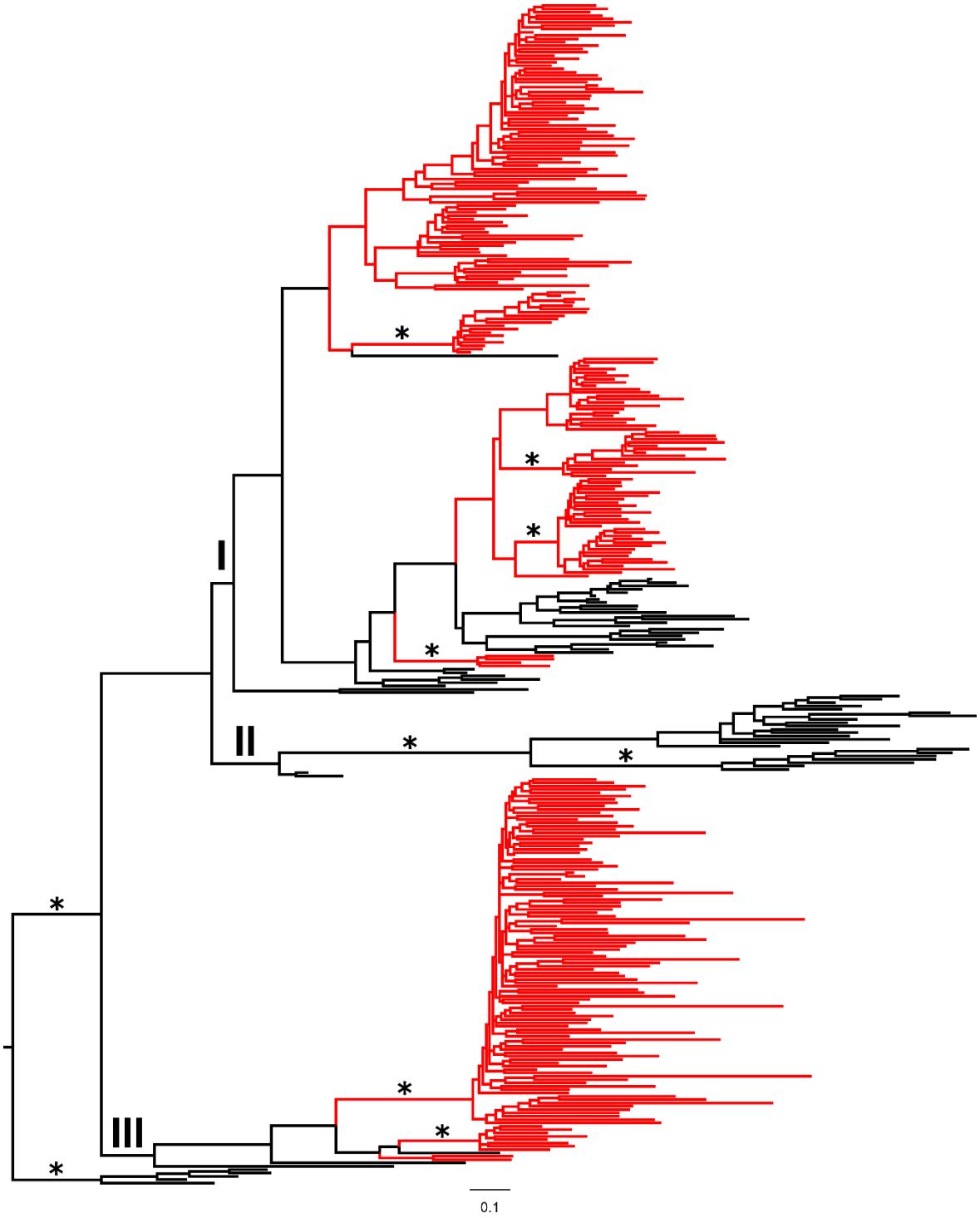


Figure 4-3 Phylogenetic screening of RTs in the rhinoceros genome. Phylogeny of reference RT sequences and 288 RT sequences detected from the rhinoceros genome by DIGS. Main branches that lead to Class I, II and III ERVs are marked by number, respectively. RT references are shown in black and detected RT sequences are shown in red. An asterisk marks the main branches with a bootstrap value over 80.

4.2.3 Classification of perissodactyl ERVs

The phylogenetic screening provided an overview of the evolutionary relationships between major perissodactyl ERV lineages, previously characterised ERVs, and exogenous retroviruses (Figure 4-4). The phylogeny was rooted on the spumaviruses (subfamily *Spumavirinae*), as these constitute a well-established outgroup to the orthoretroviruses (subfamily *Orthoretrovirinae*).

The RT phylogeny revealed three major clades corresponding to ERV classes I, II and III. Each major clade was further divided into multiple sub-lineages. I considered a clade to be a distinct perissodactyl ERV lineage if it was; i) comprised entirely of perissodactyl ERVs; ii) had bootstrap support ≥ 80%; and iii) was robustly separated from other lineages of perissodactyl ERVs by ERVs or exogenous retroviruses from non-perissodactyl hosts. On this basis, I established that there are at least nine distinct ERV lineages in the perissodactyl germline, each generated by an independent genome invasion event.

Of these nine ERV lineages, five were present in both rhinoceroses and equids; four lineages were only present in equids. There were no ERV lineages unique to the rhinoceros.

Table 4-3 Nomenclature comparisons with previous studies.

Group	Clade	Prototype	Name	Garcia-Etxebarria and Jugo, 2012	Brown *et al.*, 2012	Gim and Kim, 2017
Rho	I	HERV.R(b)	Rho.1	EqERV1-3	Gamma	EqERV-Y1~3
Zeta	I	HERV.W	Zeta.1	EqERV4	Gamma	EqERV-E1/I1~7/M1/P1~4/S2
Theta	I	HERV.L(b)	Theta.1	EqERV6-9	Gamma/epsilon	N/A
	I		Theta.2	EqERV5	Gamma	EqERV-S3
Beta	II	MMTV	Beta.1	EqERV12	EqERV.Beta1	EqERV-M2
Kappa	II	HERV.K(HML2)	Kappa.1	EqERV14	Beta	N/A
			Kappa.2	EqERV13	Beta	N/A
U1	II	N/A	U1	EqERV15	Beta	EqERV-Y4
U2	III	N/A	U2	N/A	N/A	N/A
Lambda	III	HERV.L	Lambda	EqERV10	N/A	N/A
Sigma	III	HERV.S	Sigma	EqERV11	N/A	N/A

N/A: non-available

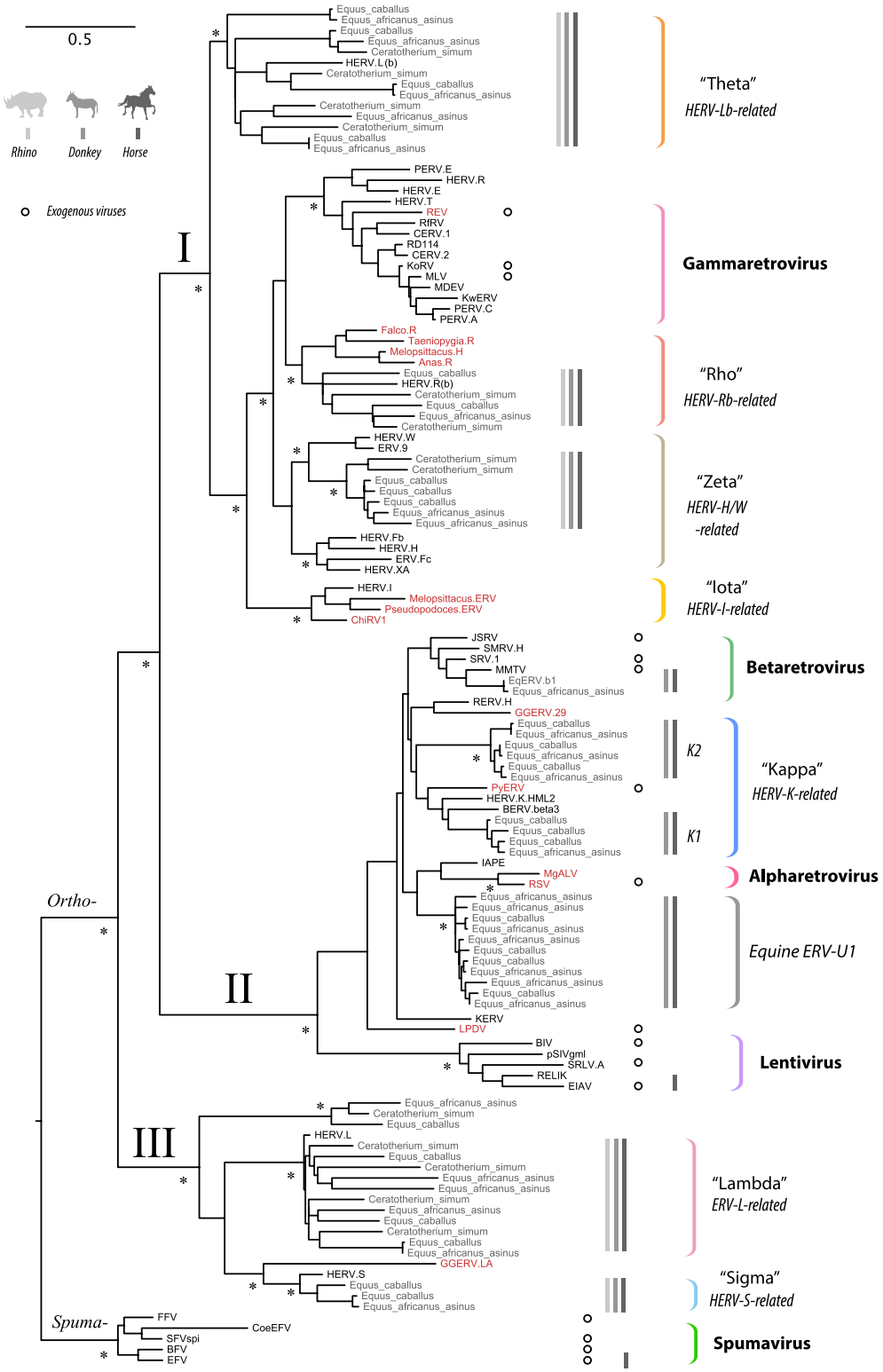


Figure 4-4 ERV diversity in the Perissodactyl germline. Maximum likelihood phylogeny showing the estimated evolutionary relationships of perissodactyl ERV RT sequences to those of previously characterised ERVs and exogenous retroviruses. Taxa labels for RT sequences detected in this study indicate the species in which they were identified. Other taxa labels show the abbreviated name of the virus or ERV. Sequences identified in non-mammalian hosts are indicated in red. RT sequences derived from exogenous virus references are marked with open circles. Retrovirus subfamilies and orthoretrovirus clades (clades I, II and III) are indicated on basal branches, while retroviral genera and ERV lineages defined in this study are indicated by coloured brackets on the right. For each of these groups, the presence of sequences in the rhinoceros, donkey and horse in each genus is indicated using grey bars as indicated in the key (top left). Asterisks indicate nodes with bootstrap support above 70%. The scale bar shows evolutionary distance in substitutions per site. Names of references can be found in Abbreviations.

Clade I: Rho, Theta, and Zeta

Clade I ERVs comprises viruses that cluster with the gamma- and epsilon-genera. As shown in Figure 4-4, three well-supported, monophyletic lineages fell immediately basal to the one that contains exogenous gammaretroviruses and could be included within a broader definition of the Gammaretrovirus genus. However, as shown in Figure 4-4, there was no RT sequence that could be clustered with known endogenous or exogenous gammaretroviruses such as murine leukaemia virus (MuLVs) and reticuloendotheliosis virus (REV).

The Rho lineage is closely related to HERV.R (type b) based on the phylogeny in Figure 4-4. In the phylogenies of all identified RTs (Figure 4-5), Rho lineages could be divided into at least three sublineages. This finding is consistent with Garcia-Etxebarria and Jugo (2012), who termed Rho sublineages as 'EqERV1, 'EqERV2, 'EqERV3' (Table 4-3). The observed relationship of the Zeta lineage was close to HERV-H and ERV.9. This is consistent with Garcia-Etxebarria and Jugo (2012) who termed Zeta as EqERV4 (Table 4-3).

The third clade I lineage was named Theta. This lineage was closely related to HERV.L b type based on the phylogeny shown in Figure 4-4. The Theta lineage could be divided into at least two sublineages: One RT was close to HERV.L(b) found in human, the other RT was different from any known ERVs of class I. Phylogenetic reconstruction of RT sequences showed that two different Theta RT could still form a monophyletic clade together, which suggested both of RT had the same origin. Therefore, Theta was further divided into two sublineages. ERVs with HERV.L(b)-like RT was termed as 'Theta.1', and the other Theta ERVs were termed as 'Theta.2'.

Brown *et al.*, (2012) suggested a large group of sequences consistently with HERV.E. Such cluster was not observed in the phylogenies based on RT sequences. By comparing with published ERV annotation, HERV.E-like sequences suggested by Brown *et al.*, (2012) were distinct from HERV.E and formed a subdivision of Theta lineage (HERV.Lb-related). Also, the perissodactyl germline appeared to lack any RT-encoding ERVs that groups with HERV-I, despite such ERVs being very broadly distributed throughout vertebrates (Martin *et al.*, 1997).

To investigate further, the DIGS screening was performed using RT sequences of known endogenous and exogenous gammaretroviruses. Phylogenetic reconstruction was performed based on the multiple sequence alignment of recovered RT sequences. Still, no sequences were found to cluster with known gammaretroviruses. Instead, all obtained sequences clustered with the HERV.R(b), HERV.H and HERV.L(b) as the phylogeny is shown in Figure 4-4. Thus, 'true' endogenous gammaretroviruses seem to be absent in the perissodactyl germline.

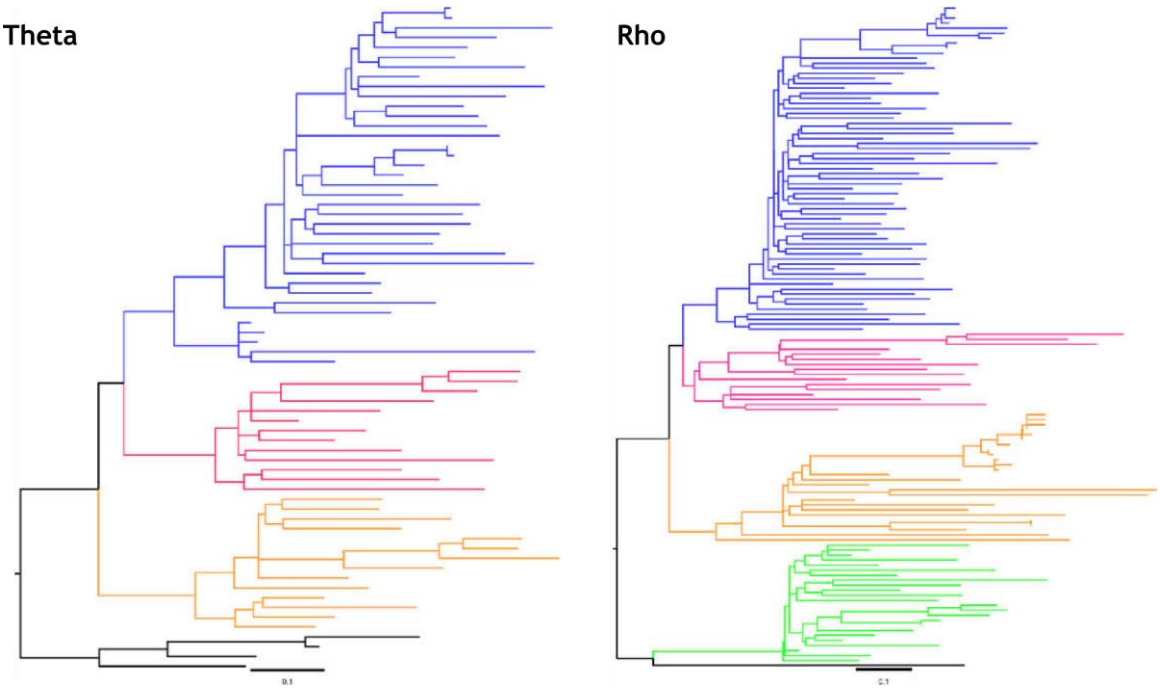


Figure 4-5 Phylogeny of identified Rho and Theta RTs from the horse genome. Phylogenies were rooted using RT references as outgroups. RT references are coloured as black; potential subclades are coloured as blue, pink and orange. Bootstrap values are not shown.

The phylogenetic screening was performed on the genome of Mongolian horse, which is a native horse breed of Mongolia. Sequencing samples of Mongolian horse were collected from a stallion, and de novo assembly was performed (Huang *et al.*, 2014), and phylogenetic screening still suggested that the gammaretrovirus lineage was absent on the Y chromosome. Screening of genomes of white rhinoceros and donkey also suggested the same conclusion. This finding indicated the even the most recent common ancestor of all perissodactyls did not have gammaretroviruses lineage, which suggested that perissodactyls have not been invaded by gammaretroviruses in the last 54 Myr.

Clade II: Beta1, Kappa, and U1

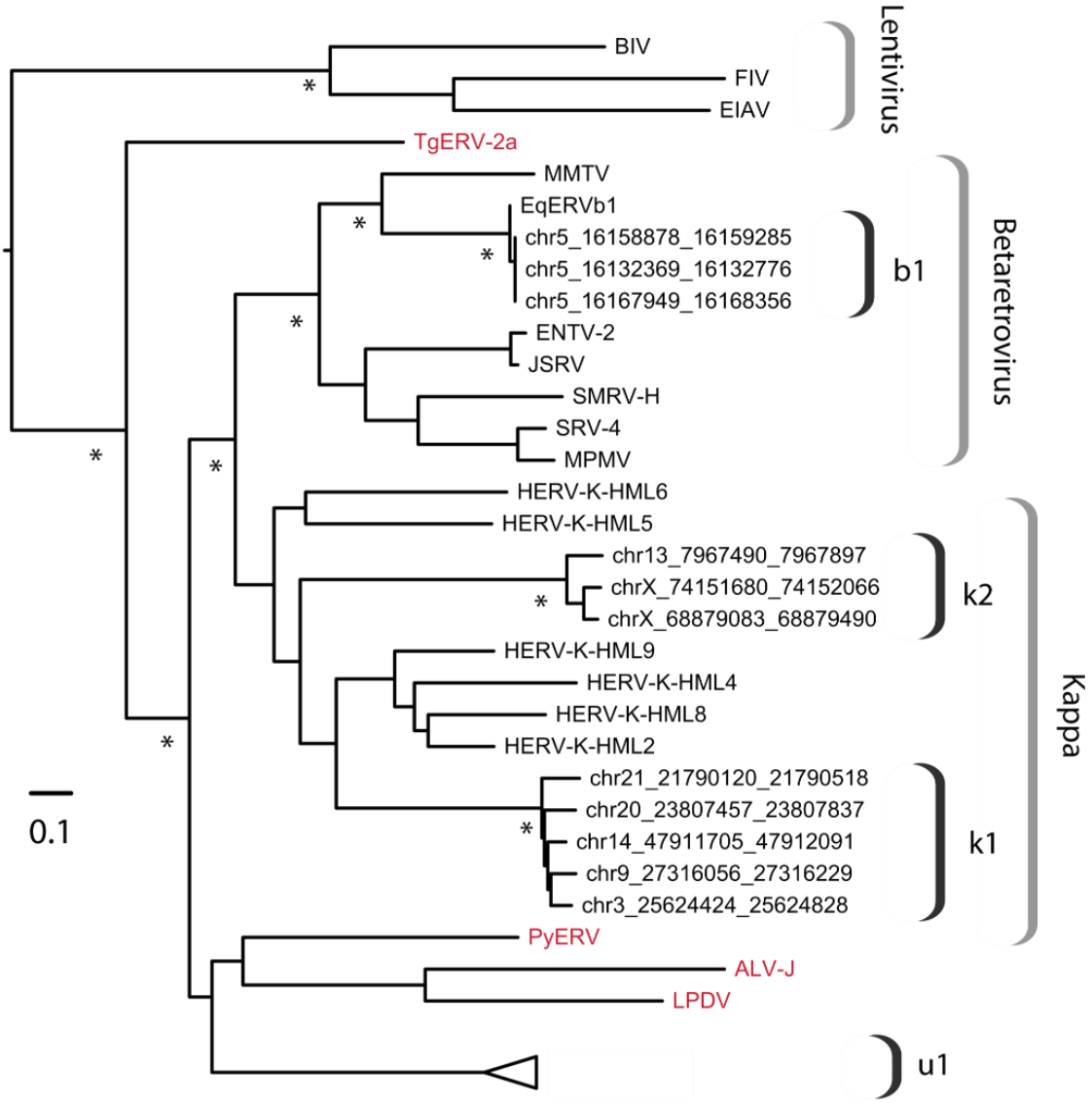


Figure 4-6 Phylogeny of Clade II polymerases from the horse genome The maximum likelihood phylogeny representing the estimated evolutionary relationships between Pol sequences derived from clade II ERVs in perissodactyl genomes, and those of previously characterised ERVs and exogenous retroviruses. Taxa labels for RT sequences detected in this study indicate the species in which they were identified. Other taxa labels show the abbreviated name of the virus or ERV. Sequences identified in non-mammalian hosts are indicated in red. Brackets on the right indicate ERV lineages and retroviral genera. Asterisks indicate nodes with bootstrap support above 70%. Names of references can be found in Abbreviations.

Clade II ERVs are related to the *Alpharetrovirus*, *Betaretrovirus*, *Deltaretrovirus*, and *Lentivirus* genera. The phylogeny in the Figure 4-4 has distinguished four lineages of clade II from the other major clades. However, bootstrap values were not high enough to support the relationship within the clade. This might be due

to the short sequence length of RT sequence. To overcome this issue, I inferred the phylogenetic tree of clade II based on the entire Pol protein sequences. The phylogeny shown in Figure 4-6 indicates that the relationship of four lineages of clade II is consistent with the phylogeny shown in Figure 4-4.

The phylogeny is shown in Figure 4-4 and 4-6 placed Beta1 closed to mouse mammary tumour virus (MMTV). This was consistent with previous work (van der Kuyl, 2011). Interestingly, Beta1 was absent in the genome of the white rhinoceros, but it was present in all equid genomes. Two lineages, referred to here as Kappa, grouped with HERV.K as part of a well-supported sister clade to the Betaretroviruses. Within two Kappa lineages, one lineage group together with HERV.K (HML2), whereas another was distinct from any known HERV.K viruses. The fourth lineage of modern ERVs in the horse genome, U1, is not closely related to any previously characterised retrovirus or ERV. Phylogenetic inference using Pol proteins indicated the distinctiveness of this lineage, grouping it as a robustly supported sister clade to all ERVs and exogenous betaretroviruses found in birds and reptiles.

The phylogenetic reconstruction of RTs suggested that clade II ERVs were found to be completely absent from the rhinoceros genome. This finding also suggests that the integration of clade II ERVs happened after the divergence of *Hippomorpha* and *Ceratomorpha*, estimated to be 54 Mya. To further investigate this situation, I performed a DIGS screening was performed on 181 Eukaryotic species genomes using all Pol proteins of clade II ERVs as queries. Recovered sequences were aligned with the same clade II references and are shown in Figure 4-4 together with horse clade II Pol proteins. Phylogenetic analyses of the recovered *pol* sequences from Eukaryotic species indicated that the clade II ERVs found in equids were only present in equids. Any detected RT hits from non-equids species were proved to be false-positive according to the phylogenies.

Clade III: Lambda and Sigma

Clade III ERVs have a distant relationship with the *Spumaretrovirus* genus. Three lineages were detected within this clade. One lineage grouped with ERV.L and one lineage was placed as a sister clade to the HERV.S according to the RT phylogeny.

I referred these two lineages as Lambda (ERV.L-related) and Sigma (HERV.S-related), respectively.

The relationship of the third lineage to any other previously characterised ERVs or exogenous retroviruses was not evident. The copy number of the third lineage was low (n=3), and all sequences were highly degraded. Even though, this lineage may originate from a distinct invasion, I did not have sufficient information to determine whether this lineage was genuinely distinct from Lambda or Sigma or it oringated from an individual gemer-line invasion. Thus I did not analyse any further.

4.2.4 *In silico* characterisation of perissodactyl ERV lineages

In this section, the ERVAP pipeline was used to investigate equid RT loci identified via DIGS, in an effort to recover representative proviruses for each of the nine perissodactyl ERV lineages identified via phylogenetic screening.

The ERVAP pipeline was used to annotate 1381 RT loci in the horse genome. I found that 146 of 1381 RT loci were flanked by putative paired LTR sequences (similarity threshold for LTR identification ≥ 80%), whereas a further 798 RT loci contained additional retroviral genes but lacked paired LTRs.

A total of 3475 retrovirus-related domains were identified within the 1381 loci (360 *gag*, 180 *pro*, 1615 *pol* and 117 *env*). Any locus that contained at least one retroviral gene flanked by paired LTRs was considered a “provirus”. In sum, 134 proviruses were detected. 92 of 134 paired identical sequences were assigned to 17 LTR consensus sequences in Repbase.

RepeatMasker was performed on the horse reference genome. In total, 479,592 solo LTRs were identified by RepeatMasker, but only 3.84% (n=18422) could be assigned to 17 LTR groups previously described. The detection summary of major lineages and LTRs are shown in Tables 4-4 and 4-5, respectively.

Table 4-4 Profile of perissodactyl ERV lineages in the horse genome

Genus /Group	Clade	Prototype	Name	PBS	RepBase LTR subgroups	Copy #			
						RT	Provirus	*env*(+)	Solo LTRs
Rho	I	HERV.R(b)	Rho.1	Arg(CCG)	1-2, 1-3, 15, 45, 72A, 72B, 8B, 8E, 8F	151	20	6	4062
Zeta	I	HERV.W	Zeta.1	Leu(TAA)	1, 14, 1420	37	13	5	3953
Theta	I	HERV.L(b)	Theta.1	*ND*	1-4, 27_FC	251	11	2	295
	I		Theta.2	*ND*	1-4B, 1-6, 13A, 19, 23B, 6, 6B, MER34A_CF, MER34A1	67	9	6	8540
Beta	II	MMTV	Beta.1	Lys(TTT)	N/A	10	1	2	350
Kappa	II	HERV.K(HML2)	Kappa.1	Lys(CTT)	2-2	5	4	4	80
	II		Kappa.2	Lys(CTT)	N/A	3	1	1	35
U1	II	N/A	U1	Trp(CCA)	2-1	45	32	32	703
U2	III	N/A	U2	*ND*	*ND*	54	NA	NA	NA
Lambda	III	HERV.L	Lambda	*ND*	None identified	691	NA	0	NA
Sigma	III	HERV.S	Sigma	Ser(AGA) Ser(CGA)	3-1C, 74	67	1	2	293
Totals						1381	92	57	18410

Table 4-5 Long terminal repeats detected by RepeatMasker

Clade/Group	Repbase ID	Count
Clade I		
Rho	LTR15_EC	130
	ERV1-2-EC_LTR	312
	LTR8E_EC	79
	ERV1-3-EC_LTR	229
	LTR45_EC	147
	LTR8B_EC	1671
	LTR72A_EC	1096
	LTR72B_EC	345
	LTR8F_EC	60
Zeta	ERV1-LTR_EC	978
	LTR14_EC	1251
	LTR1420_EC	1633
theta.1	ERV1-4-EC_LTR	351
	ERV1-4B-EC_LTR	1859
	LTR27_FC	0
theta.2	ERV1-6-EC_LTR	895
	LTR13A_EC	966
	LTR19_EC	346
	LTR23B_EC	96
	LTR6_EC	220
	LTR6B_EC	97
	MER34A1_EC	4196
	MER34A_CF	0
Clade II		
Beta.1	Own label	351
Kappa.1	ERV2-2-EC_LTR	79
Kappa.2	Own label	34
U1	ERV2-1-EC_LTR	705
Clade III		
Sigma	ERV3-1C-EC_LTR	218
	LTR74_EC	78

4.2.5 Representative genome structures of perissodactyl ERVs

While some recent ERV insertions are relatively intact, most are millions of years old and have accumulated numerous mutations, deletions, and insertions. However, the multicopy nature of many ERV lineages makes it possible to infer the functional sequences of ancient ancestral retroviruses directly – indeed, the consensus sequence of an ERV can approximately represent the original sequence at the time of integration if selection is neutral (Mayer and Meese, 2002; Jern, Sperber and Blomberg, 2004; Lavie *et al.*, 2004; Flockerzi *et al.*, 2005; Jern *et al.*, 2005). By examining an alignment of ERV loci, it is possible to infer the approximate sequences of the retroviral proviruses that founded the ERV lineage. Since it is unlikely that deletions or insertions will occur in the same precise position in different proviral copies, most insertions and deletions that have occurred subsequent to integration are evident.

Prototypic members of each of ERV lineages were investigated to provide further information about these elements. Although it was difficult to identify the exact 5' and 3' ends of the *gag*, *pol*, and *env* genes due to insertions or deletions, and in-frame stop codons, the presence or absence of these genes could still be established by the identification of certain motifs conserved among different retroviruses. In the following section, the consensus structures determined for each ERV lineages are described.

Clade I: Rho

At least 23 proviruses of the Rho lineage were identified by ERVAP, only 6 of which exhibited *env*. 66 loci contained one to three viral coding regions but lacked LTRs. A 7,325 bp region was identified on the sense strand of chromosome 5 (77,379,247-77,386,572) with the expected retroviral structure of LTR-*gag*-*pro*-*pol*-*env*-LTR. Four additional loci were found on chromosome 5, 18 and X.

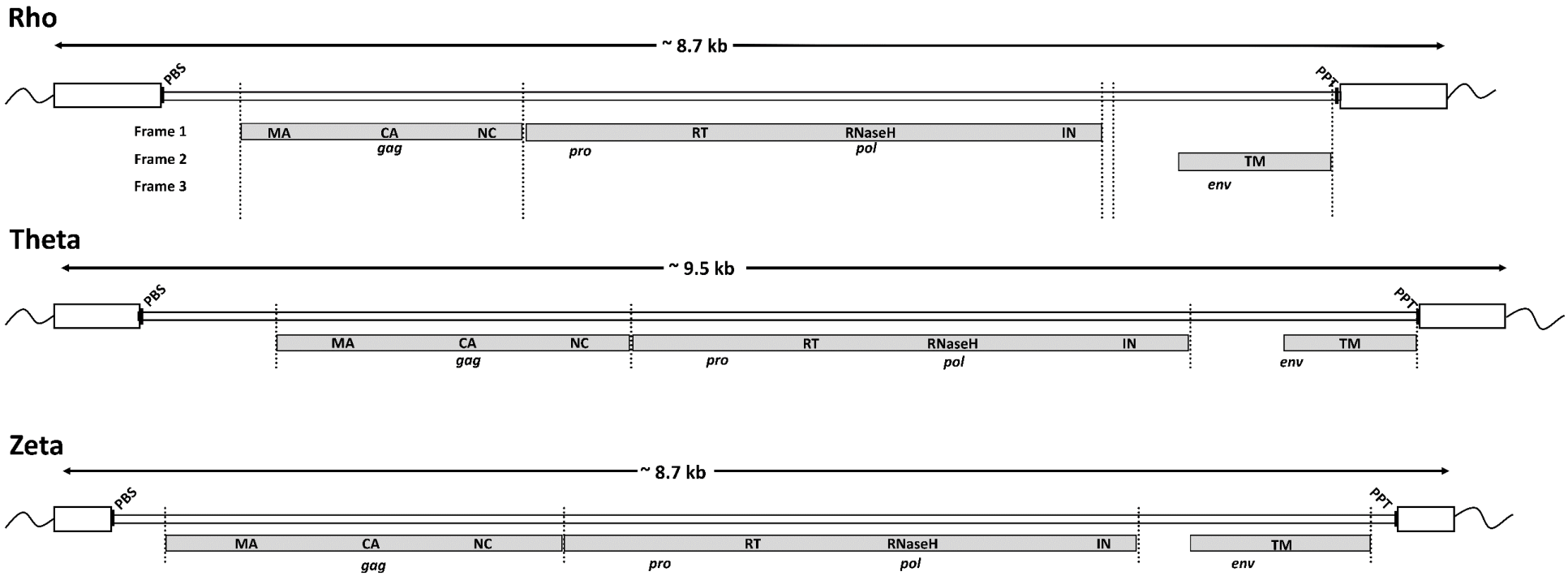


Figure 4-7 Schematic representation of Rho, Theta and Zeta proviruses. The Gag protein encodes the MA, CA, and NC. The *pro* gene is located between *gag* and *pol*. The *pol* encodes the RT, RNase H, and IN. The *env* coding domains encode SU and TM. The ORF of *env* is uncertain. The estimated positions of PBS and PPT are marked with black bars. The long terminal repeats are shown as white boxes, and the host genome is shown as wavy lines. Grey boxes range the coding regions. The scale is shown at the top of each genome structure.

The consensus genome structure was inferred from five Rho proviruses (Figure 4-7). The Rho lineage *pol* gene was found to have a typical retroviral organisation, encoding domains associated with the *pro*, RT, IN, and RNase H. The border between *pro* and *pol* could not be distinguished. At least nine LTR groups were identified according to the Repbase consensus sequences. The average length of Rho LTRs is around 533 bp (from 461 bp to 664 bp), except in one LTR group, which is 1301 bp long. Of the 23 proviruses, five were primed by $tRNA^{Arg}$, while in the others the PBS sequence was not detected.

Clade I: Zeta

A second clade I lineage termed Zeta was represented by at least 41 RT sequences. Of these 41 loci, 17 were determined to be proviruses that contained at least one retroviral gene flanked by paired LTRs, whereas 17 loci showed the presence of retroviral genes but lacked paired LTRs. Interestingly, five proviruses exhibited the LTR-*gag*-*pro*-*pol*-*env*-LTR structure.

The consensus proviral sequence of the Zeta lineage was inferred based on eight proviruses (Figure 4-7). The consensus provirus was approximately 8.57 kb in length. The 17 proviruses all utilised one or the other of two LTR groups (ERV1-LTR_EC and LTR1420_EC). The lengths of two LTR groups differed (454 bp vs 696 bp), but two LTR groups show high identity at their 3'ends (similarity = 95%).

Clade I: Theta

A third lineage clade I lineage was termed as Theta and contained 251 RT loci. However, ERVAP only found a few proviral loci for this lineage. Three proviruses exhibited a complete genome, whereas 19 RT loci were found to contain a least one retroviral gene. The consensus sequence of the Theta lineage is ~8.5 kb in length and has the structure LTR-*gag*-*pro*-*pol*-*env*-LTR (Figure 4-7).

A total of 19 distinct LTR pairs were identified for Theta. These LTRs were assigned to 11 Repbase LTR groups. Two of 11 LTR groups were identified as *Felis catus* LTR and *Canis familiaris* LTR, but no solo LTRs belonging to these LTR groups were detected in the horse genome by Repbase. Thus, these LTRs were probably misassigned. The average length of Theta LTRs is around 494 bp.

Clade II: Beta1

The Beta1 was the first equine ERV lineage reported (van der Kuyl, 2011). The intact Beta1 provirus is ~10k long with a relatively intact genome structure. The LTR of Beta1 was 1350 nt in length, RepeatMasker detected 350 solo LTRs in the horse genome. There were no additional ERV lineages found in the other equids. Due to the unusual length of the LTRs in this ERV lineage, further investigation was performed to find potential ORFs in the Beta1 LTR.

The Beta1 lineage groups closely with MMTV in phylogenies, and it is known that MMTV encodes an extra gene – the superantigen gene (*sag*) - in its LTR. I did not detect an open reading frame in the Beta1 LTR. This could potentially be due to neutral mutations having disrupted the frame subsequent to integration, but in this case, I would still expect HMMR to detect some homology, as there is an HMM for the Sag protein.

Clade II: Kappa1 and Kappa2

The human genome contains a range of ERV lineages that are related to betaretroviruses, but cluster outside the main *Betaretrovirus* clade. These lineages are referred to as the 'HERV-K superfamily' by some authors and are here given the name 'Kappa'. Two Kappa-related lineages were identified in the horse genome (Kappa1 and Kappa2). Of five Kappa1 loci, four were identified as proviruses due to the presence of flanking paired LTR and internal coding regions. All four proviruses were relatively intact with a typical retroviral genome structure of LTR-*gag*-*pro*-*pol*-*env*-LTR. The LTRs were assigned to 'ERV2-2-EC_LTR' in Repbase and were 522 bp in length.

The consensus sequence of Kappa1 was generated based on four identified proviruses (Figure 4-9). The consensus sequence suggested that coding sequences of *gag*, *pro* and *pol* were present in three different-frames, as common for betaretroviruses. A dUTPase was encoded between *pro* and *pol*.

A fragment of the Rec protein (109 aa) was found at the 3'end of the *env* gene, which shared the same reading frame with *env*. The product was identified as the orthologous of Rec protein of HERV-K(HML2). The *Rec* coding region was observed

at the same position in three of four Kappa1 proviruses. The presence of Rec suggested that Kappa1 utilised a homolog of *Rec* for complex regulation of viral gene expression.

The only full-length Kappa2 provirus was retrieved from the chromosome 13 in the horse genome (Figure 4-9). The other two loci were not flanked by paired LTRs and presented like tandem repeats that were adjacent to LINE1. These two copies could not be used to generate consensus sequences.

The provirus was 7,295 bp in length with the usual retroviral genome LTR-*gag*-*pro*-*pol*-*env*-LTR. The length of LTRs was approximately 354 bp and paired LTRs differed around 5% from each other, suggestive of a relatively recent integration. The full-length provirus indicated that the *gag* and *pro* of Kappa2 shared the same reading frame but differed from *pol*.

A long non-coding region was present between *pol* and *env*, and the length of the identified *env* coding region was 336 aa. This finding suggests the *env* gene was incomplete. Searching for ORFs in the non-coding region failed to identify any potential matches. This suggested that unlike Kappa1, Kappa2 might not encode a Rec protein.

Clade II: U1

The U1 lineage had the largest number of proviruses overall (N = 45) and abundant solo LTRs (N = 705). Intriguingly, this lineage also shows indications of relatively recent activity. Alignment of full-length proviruses was used to infer a consensus genome structure (Figure 4-9). This revealed that there were, in fact, two distinct genomic organisations of U1 proviruses. In the first (type I), the *pro* encodes a dUTPase domain at the 3' end, as observed in other betaretroviruses. However, the majority of U1 insertions exhibited a more unusual genome structure (type II) in which the dUTPase encode within *gag*. This second type of genome structure has not previously been reported in any retrovirus.

Beata.1: chr 5: 16,154,742 – 16,168,965 (-)

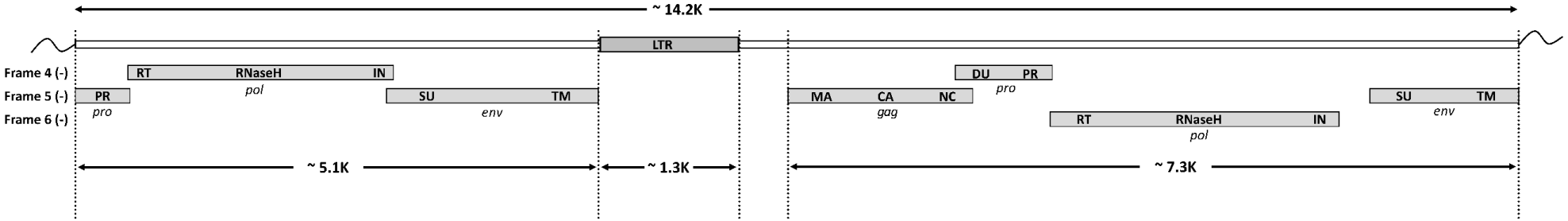


Figure 4-8 A tandem repeat of Beta1. The structure of Beta1 tandem repeat in chr5: 16,154,742-16,168,965(+).. The genome structure of two Beta1 proviruses is the same as provirus described by van der Kuyls (2011).

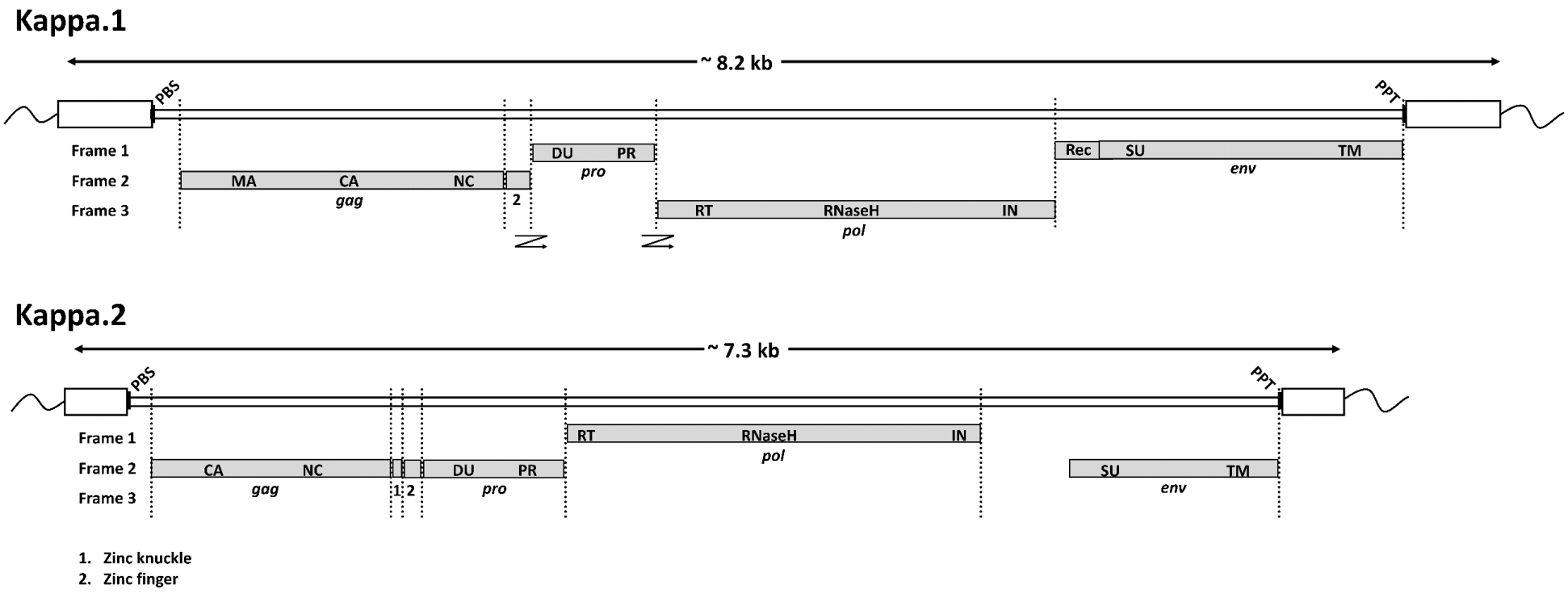


Figure 4-9 Schematic representation of Kappa proviruses. The *gag* encodes the MA, CA, and NC. The *pro* locates between *gag* and *pol*. The *pol* encodes the RT, RNase H, and IN. The *env* encodes SU and TM. The estimated positions of PBS and PPT are marked with black bars. The LTRs are shown as white boxes, and the host genome is shown as wavy lines. Grey boxes range the coding regions. Sites of translation frameshifting at the *gag-pro* ORF junctions and *pro-pol* junctions are shown as fold lines. The scale is shown at the top of each genome structure. Abbreviations: DU (dUTPase).

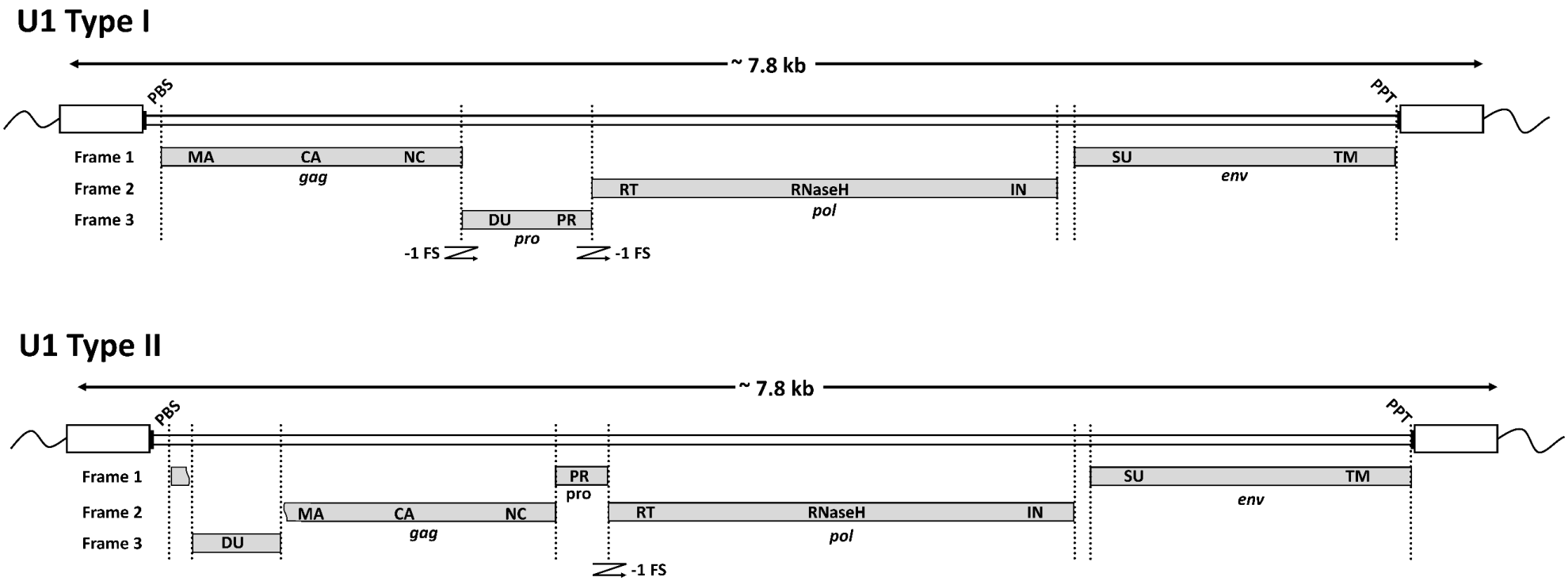


Figure 4-10 Schematic representation of U1 proviruses. The *gag* encodes the MA, CA, and NC. The *pro* locates between *gag* and *pol*. The *pol* encodes the RT, RNase H, and IN. The *env* encodes SU and TM. The estimated positions of PBS and PPT are marked with black bars. The LTRs are shown as white boxes, and the host genome is shown as wavy lines. Grey boxes range the coding regions. Sites of translation frameshifting at the *gag-pro* ORF junctions and *pro-pol* junctions are shown as fold lines. The scale is shown at the top of each genome structure. Abbreviations: DU (dUTPase).

Clade III: Lambda

ERVAP identified 361 loci containing *Lambdaretrovirus* (ERV.L-related) elements. However, none of these loci contained an identifiable *gag* or *env*. Importantly, however, this might be due to the lack of knowledge of ERV-L *gag* and *env* in the Pfam database. Indeed, among all nine lineages, the Lambda lineage was the most degraded, and no intact coding regions were found in any Lambda provirus loci. Also, LTRharvest did not identify any paired LTRs flanking lambda RTs. Nevertheless, I could identify most of the pol gene, and a dUTPase encoded after *pol* - a feature of the lambda lineages such as MuERV-L and HERV-L. Because the equine lambda lineage was so highly degraded (and also because the lineage has more in common with LTR-retrotransposons than retroviruses) I did not generate a consensus genome for Lambda.

Clade III: Sigma

All elements in the Sigma lineage were defective. I detected 76 copies of the Sigma lineage in the horse genome. Only six of these 76 copies contained flanking paired LTRs. No *gag* could be identified in any copies. Two LTR groups were identified, and they were 449 bp and 312 bp in length, respectively. One locus was identified as provirus locus at chr9:55,409,357-55,415,972(+) with structure LTR-*pol*-*env*-LTR. It was 6.62 kb in length. Although DNA sequence between 5'LTR and *pol* was longer than 1500 bp in length, there was no evidence of the presence of *gag*. Indels and in-frame stop codons were observed in *pol*. The *env* was nearly intact with only one in-frame stop codon. A consensus sequence was generated based on six provirus sequences (Figure 4-10).

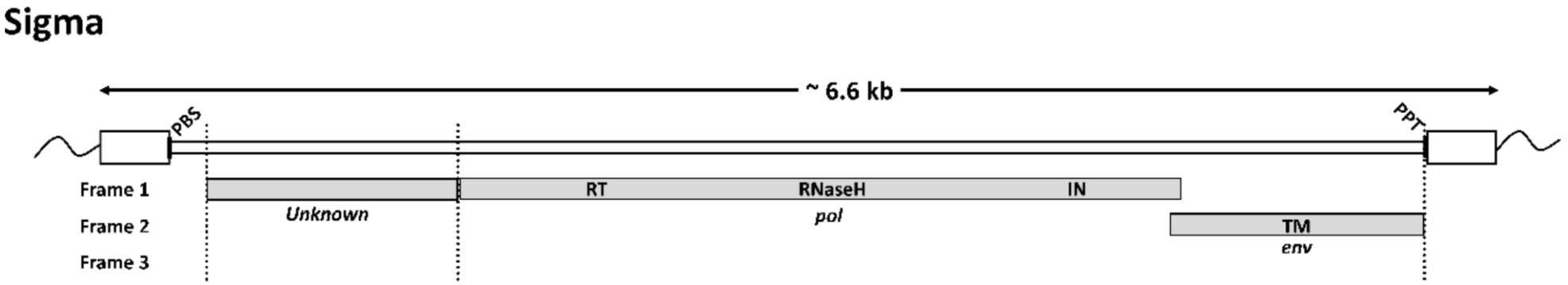


Figure 4-11 Schematic representation of Sigma. The *gag* encodes the MA, CA, and NC. The *pro* locates between *gag* and pol. The *pol* encodes the RT, RNase H, and IN. The *env* encode SU and TM. The estimated positions of PBS and PPT are marked with black bars. The LTRs are shown as white boxes, and the host genome is shown as wavy lines. Grey boxes range the coding regions. The scale is shown at the top of each genome structure.

4.3 Discussion

4.3.1 ERV diversity in the equine genome

Via phylogenetic screening, I determined that there are at least nine distinct ERV lineages in the perissodactyl ERV germline. Interestingly, no *bona fide* Gammaretroviruses were identified in perissodactyl genomes, despite these ERVs being very common in other mammalian genomes. I show that the three lineages of gamma-related ERVs are more closely related to ancient HERVs than to any known exogenous gammaretroviruses, and group outside the main *Gammaretrovirus* clade as defined by exogenous isolates.

Similarly, I find no evidence that the horse genome contains epsilonretrovirus-derived ERVs, as has been reported previously. There are some ERVs in perissodactyl genomes that are distantly related to epsilonretroviruses, but they group far outside the *Epsilonretrovirus* clade as defined by exogenous isolates. This finding is consistent with previous results on the ERV diversity in fish (Basta *et al.*, 2009; Han, 2015; Naville and Volff, 2016).

While I did not identify any true gammaretroviruses in perissodactyl genomes, I did identify several distinct lineages of clade I (gammas-related) ERVs. Here, I refer to these three lineages as Rho (HERV.Rb-related), Zeta (HERV.H/HERV.W-related) and Theta (HERV.Lb-related). It is important to know that HERV.L(b) belongs to class I, and HERVL(b) is not a subtype of HERV.L. HERV.L(b) was named due to its PBS ($tRNA^{Leu}$) which is homologous to PBS of HERV.L (Katzourakis and Tristem, 2005). However, both phylogenetic reconstruction based on domain 1 to 7 of RT of HERV families indicated that HERV.L(b) belongs to class I (Katzourakis and Tristem, 2005). Thus, Theta lineage is a clade I lineage instead of clade III lineage.

Both Rho and theta lineages can be further divided into multiple sublineages (figure 4-5). This finding consists of previous reports (Brown *et al.*, 2012). However, based on the different standard and method (e.g. LTR, tRNA or RT), the number of sublineages can be various. In this chapter, the sublineages were assumed based on the phylogeny. Each monophyletic clade can be counted as one sublineage, and each sublineage can be obtained from an individual germline invasion. However,

there is no direct method to count the exact number of invasion happened in the evolutionary history. Thus, to avoid the uncertainty, the total number of ERV lineage excludes all sublineages, which narrows the originates of ERV lineages in perissodactyl to nine major germ-line invasion events.

Strikingly, clade II ERVs were found to be completely absent from the rhinoceros genome. In equids, by contrast, four clade II (Beta-related) lineages are present, one of which (Beta1) represents a *bona fide* betaretrovirus, and has previously been described in detail. I identified two additional clade II lineages that grouped with representatives of the HERV-K 'supergroup', which I refer to here as 'Kappa'. I named these two lineages as Kappa.1 and Kappa.2. The fourth lineage of clade II ERVs was found to be distinct from all previously characterised retroviruses and ERVs and was named unclassified equine ERV 1 (U1).

I identified numerous RT sequences belonging to the clade III lineage ERV.L lineage (referred to here as Lambda) (Bénit et al., 1999). As expected, none of these RT hits was in proviruses containing *env* genes. However, I did identify additional lineages of clade III ERVs, one of which disclosed relatedness to the primate HERV.S lineage (referred to here as Sigma), and did encode an *env* gene.

4.3.2 Consensus proviral genome structures of ERV lineages

Modern equine ERV lineages have been present in the germline for a relatively short period of time. Single provirus that acquired deletions and insertions have not got a chance to retrotranspose in a retroviral fashion to new genomic sites, giving rise to new proviruses carrying the deletion.

A consensus sequence containing major retroviral proteins was generated for each ERV lineage identified here. Although a previous paper (van der Kuyl, 2011) has described proviruses of the Beta1 lineage in detail, little is known about most other equine ERV sequences. Therefore, this represents the most detailed characterisation of ERVs in perissodactyl species to date.

Deletions, insertions and in-frame stop codons were frequently observed in most of the proviral gene coding regions of all nine ERV lineages. So, it is clear that many retroviral genes were unable to be translated. However, some of them are

still able to retrotranspose within the genome. It will be interesting to see how a proviral genome and the corresponding RNA maintained retrotransposition activity. One way is reinfection. ERVAP identified a few *env* genes, which are necessary for movement between cells. The existence of *env* genes suggested that some ERV lineages might be able to increase their copy number via germline reinfection. Another possible way is via retrotransposition in *cis*. When an ERV integrated into the LINE1 elements or attached to the end of LINE1, it may be able to retrotranspose together with LINE1. In this study, LINE1-related domains were found by ERVAP in the flanking region of some provirus loci, which suggested these loci could be consequences of retrotransposition rather than reinfection. Further investigations were performed to determine how equine ERV increased their copy number (see next chapter).

4.3.3 Approach limitations

The Y chromosome is not available in the current horse genome assembly

Overall, 18,290 RT sequences were revealed from the host genome using DIGS. However, the true copy number should be larger than 18,920. As the horse reference genome was generated from a mare, the Y chromosome was not included. So, the exact location and number of ERVs on the Y chromosome were uncertain, and there is no whole Y chromosome sequence available yet.

However, the classification of ERVs is still trustful. The phylogenetic screening was performed based on the Mongolian horse genome. The *de novo* assembly of Mongolian horse was based on the sequencing sample collected from a stallion. Phylogenetic reconstruction using the RT sequences identified from the Mongolian horse assembly showed a highly similar topology as that of horse reference genome. Thus, there were no putative ERV lineages lost due to the unavailability of the Y chromosome reference.

Underestimation of ERV counts due to the *de novo* assembly

De novo assembly methods have issues regarding the assembly of repeat regions. *De novo* assembly often can map reads to paralogous loci, which will reduce the length of repeat region or even break the contig into two parts. As a result, ERV loci might be lost during genome assembly.

In general, the number of RTs identified from the horse reference genome, and genomes of the other horse breeds were more likely closer to the true copy number of RTs. This is due to the fact that the horse reference genome has been assembled to the chromosome level, with sequences of repeat regions being more likely to be true.

Limitations of reference-based genome assembly

However, there is a certain limit to this comparative approach. The unique ERV lineages of half asses and zebras might be lost. For example, high rates of chromosomal loss were observed during the *caballine/noncaballine* divergence.

Mountain zebra experienced almost four times more chromosome losses than gains, resulting in the smallest number of chromosomes in the entire genus (2n=32). Using the donkey or the horse genome as references may not reflect the true situation of mountain zebra. Also, NGS reads that cannot be mapped to the reference were not included in the screening. Some ERVs may not be observed due to the mapping process.

Phylogenetic reconstruction using truncated RT sequences

To obtain a better phylogeny, RT sequences were edited manually to avoid large indels, and only the most conservative region was maintained. This strategy reduced the evolutionary distance between sequences, especially the distance between equine RT sequences and RT reference obtained from other species (e.g. HERV). In phylogeny, RT reference obtained from other species will cluster in the centre of the monophyletic clade instead of being basal to the clade. Thus, the evolutionary relationship shown in figure 4-4 slightly differed from the relationship shown in figure 4-1, 4-2 and 4-3.

4.4 Conclusion

This chapter has described the use of DIGS and my ERVAP pipeline to detect and annotate ERVs in 17 perissodactyl genomes. A total of 18,290 RT loci were identified. The phylogeny of detected RT sequences was reconstructed together with the RT reference sequences from the previous characterised ERVs and exogenous retroviruses. At least nine major ERV lineages were detected. Interestingly, comparison of the diversity of ERVs in the perissodactyl species suggested that gammaretroviruses and epsilonretroviruses are absent in all perissodactyls, and class II (referred to clade II in the chapter) ERVs are absent from rhinoceroses.

Next, I characterised the genome structure for each identified perissodactyl ERV lineage. The ERVAP pipeline was used to investigate the genomic regions flanking each RT locus identified by DIGS. Representative genome structures and consensus sequences were generated based on the recovered proviral sequences of each major ERV lineage. Except for Lambda, representative proviruses were generated for all other ERV lineages. The U1 lineage even showed two different genome structures (Type I and II).

5. Characteristic of ancestral and modern ERV lineages in the horse

5.1 Introduction

In the previous chapter, nine major endogenous retrovirus lineages were identified in the perissodactyl germline. Here, I investigate the evolutionary history of these ERV lineages, examining their retrotranspositional activity over time, and their properties in relation to potential exaptation or co-option by host genomes.

5.1.1 Calibrating the timescale of ERV evolution

The integration times of individual ERV loci can be estimated to calibrate an evolutionary timeline for specific ERV lineages. The most straightforward method is based on the detection of orthologous insertion – since it can be assumed for orthologous pairs of ERVs that integration occurred prior to the divergence of the host genomes in which they occur, the time of most recent common ancestor (tMRCA) of these two species provides a minimum age of integration. The oldest ERV ortholog that has been detected belongs to the ERV-L lineages and predates the divergence of placental mammals ~ 104-110 Myr (Lee *et al.*, 2013).

The age of ERVs can also be estimated by using the assumption of a neutral molecular clock (i.e. after duplication, two duplicated sequences that are under neutral selection accumulate mutations independently in a clock-like manner). The genetic divergence between duplicated ERV sequences is calculated, and a neutral rate calibration (i.e. the estimated neutral rate in the host species being examined) is applied.

This approach can be used to date individual proviral loci – since the LTRs flanking proviruses are known to be identical at the time of integration, the divergence between these two sequences provides one way of estimating provirus age (Tristem, 2000; Lavie *et al.*, 2004; Sinzelle *et al.*, 2011; Brown, Emes and Tarlinton, 2014).

In addition, ERV loci can be dated using a clock-based approach by comparing against an estimated ancestral virus sequence. Since the number of solo LTR

sequences in most ERV lineages is relatively high, ancestral LTR sequences can be estimated for many ERV lineages. The age of individual solo LTR loci can thus be estimated by measuring their divergence from this ancestor and applying a molecular clock (Subramanian *et al.*, 2011).

5.1.2 Co-option of ERV sequences by host genomes

Recent studies have demonstrated that ERVs sequences have often been co-opted or exapted by host genomes, and this has exerted a profound impact on mammalian evolution and biology (Rowe *et al.*, 2010; Dupressoir, Lavialle and Heidmann, 2012; Redelsperger *et al.*, 2016).

Some ERVs benefit the host by rendering it resistant to the infection by exogenous viruses (Goff, 2013). Perhaps the most famous examples are *Fv1* and *Fv4*. *Fv1* is thought to be derived from the *gag* gene of an ERV-L provirus and can block MLV infection (Pincus, Rowe and Lilly, 1971; Lilly and Pincus, 1973; Best *et al.*, 1996). *Fv4*, on the other hand, originated from an *env* gene fragment. It can render mice resistant to the exogenous viral infection by down-regulating the receptors (Kozak *et al.*, 1984; Ikeda and Sugimura, 1989).

Surprisingly, ERVs sequences also play a crucial role in vertebrate development (Sugimoto and Schust, 2009). Many vertebrates contain genes called *syncytins* that are derived from a retroviral *env*. Interestingly, acquirement of a retroviral *env* gene for placenta development occurred independently in three different order of mammals involving different groups of ERVs (Heidmann *et al.*, 2009). For example, human (*syncytin-1* and *syncytin-2*) and mouse (*syncytin-A* and *syncytin-B*) are acquired independently, and all of them express specifically in the placenta and contribute to the formation of giant syncytia (Mi *et al.*, 2000; Dupressoir *et al.*, 2009).

Also, some specific sequences carried by retroviral proviruses have been co-opted into regulatory networks that control the synthesis and processing of viral RNA. This is thought to have occurred through ERV sequences being targeted for repression – initially to suppress their activity. However, repression of ERV loci can have modulatory effects on expression of host genes in close physical proximity to the repressed locus, and these can be selected so that new gene

regulatory networks emerge (Imbeault, Helleboid and Trono, 2017). Also, LTRs of proviruses naturally carry transcriptional regulatory signals for viral replication. Thus, these signals allow LTRs to work as alternative promoters for the adjacent host gene (van de Lagemaat *et al.*, 2003).

ERV insertions can also modulate patterns of splicing and expression in host genomes. Integration frequently occurs within introns, and when this occurs, the splice acceptor site of proviruses can interfere with the splicing of host mRNA and form a host-virus hybrid (Maksakova *et al.*, 2006).

5.1.3 Aims of this chapter

In this chapter, I will investigate the activity of distinct ERV lineages over time and discriminate those that are 'ancestral' (shared by all perissodactyls) from those that are 'modern' (unique to horses and/or other equids).

Ancestral ERV lineages that predate the divergence of rhinoceroses and horses are unlikely to express replication-competent viruses. However, the long residence of these lineages in the germline may reflect a role in one or more physiological processes. Therefore, I will look for loci within these lineages that show evidence of having been co-opted or exapted.

By contrast, ERV lineages that are unique to equids might potentially be capable of retrotransposition activity. I will look for evidence of recent activity among modern ERV lineages found in the horse genome. I will also look for evidence of co-option or exaptation in these younger lineages.

5.2 Categorising perissodactyl ERVs

To aid investigation of perissodactyl ERVs, I created a distinction between 'ancestral' lineages that entered the perissodactyl germline prior to the divergence of the two major sublineages (*Hippomorpha* and *Ceratomorpha*) and 'modern' ERV lineages that entered after this point.

From the investigation in chapter IV, it was evident which ERV lineages belonged to each category. Ancestral ERV lineages are expected to be present in all perissodactyl lineages where they have not been lost, and exhibit signs of their age, as they tend to be relatively degraded. By contrast, modern ERV lineages are likely to be found in a more restricted range of species, and more frequently have nearly intact open reading frames.

Nevertheless, I sought to demonstrate the ancestral origin of particular ERV lineages by identifying within them clear and unambiguous examples of loci that were orthologous in rhinos and equids. Using a BLAST-based approach, I identified several loci in the Lambda, Sigma, Theta and Rho lineages that were orthologous between the donkey, horse and rhinoceros.

By contrast, I could not identify any orthologous loci in the Zeta lineage, despite this lineage being present in both rhinos and equids. Furthermore, I identified examples of empty Zeta integration sites in the rhinoceros genome – indicating that this lineage, is likely to have entered the perissodactyl germline prior to the divergence of *Hippomorpha* and *Ceratomorpha*, but did not generate fixed copies before this, and remained active subsequently.

Since there were no clade II ERVs (Beta1, Kappa1, kappa2 and U1) identified in the rhinoceros genome, these lineages are categorised as modern, along with Zeta. This was in accordance with findings in the previous chapter, which suggested all four of these lineages have a lower degree of degradation than others.

The divergence of *Hippomorpha* and *Ceratomorpha* is estimated to have occurred 54 million years ago. Since the rhinoceros does not harbour any unique ERV lineages (i.e. lineages that are present in the rhino, but not in the horse), I concluded that no exogenous retrovirus has successfully invaded the rhinoceros

germline subsequent to this time. As far I can determine, this is the longest time any mammal lineage has existed without acquiring a new lineage of ERVs that left some fixed copies in its germline.

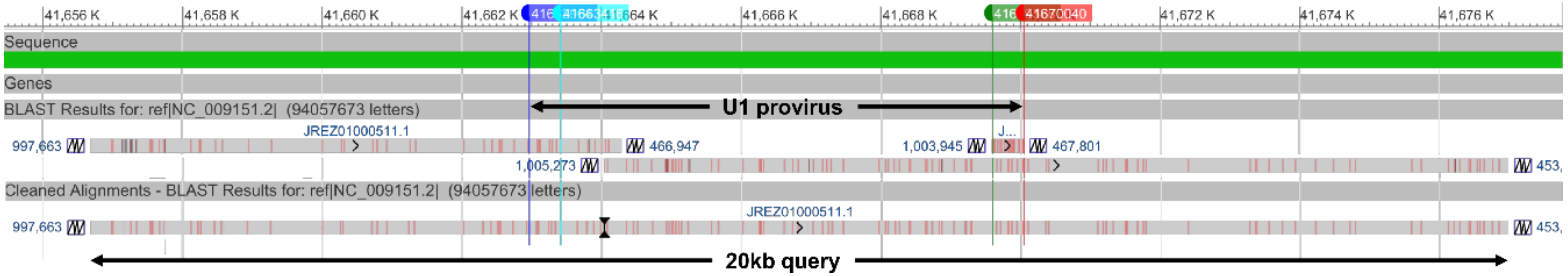


Figure 5-1 The example of U1 orthologous. BLASTn is used to align the horse genome (green) and the rhinoceros scaffold (short red and grey bars). The donkey scaffold JREZ01000511 is aligned to the horse chromosome 8: 41,656,684-41,676,977. Colour lines and black arrows are used to show borders of aligned regions.

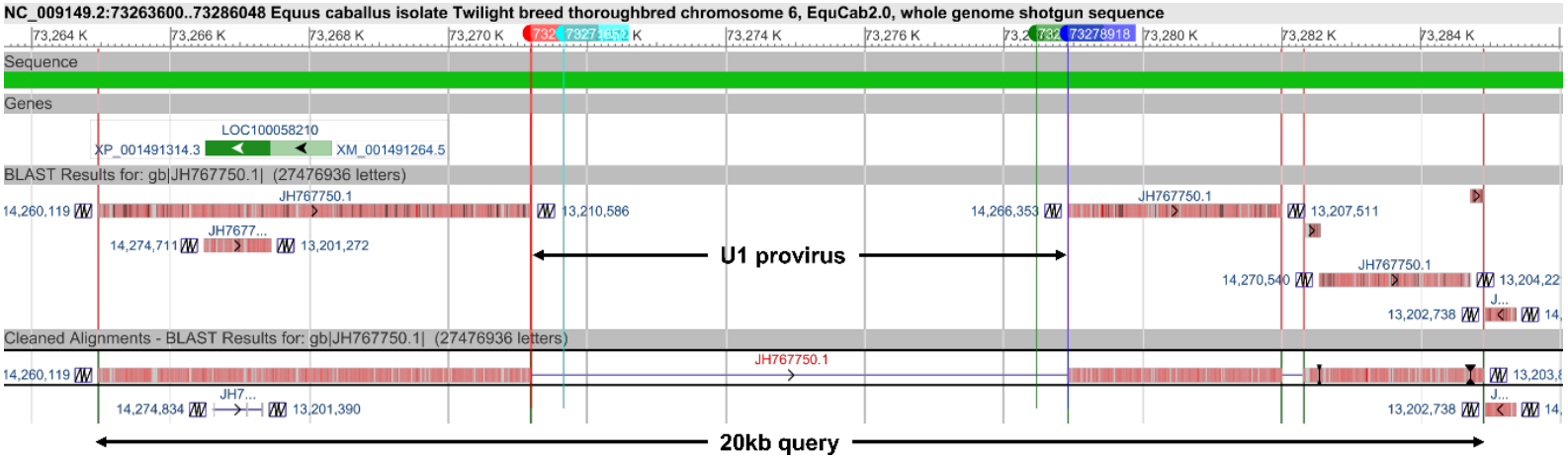


Figure 5-2 The example of U1 empty insertion site. BLASTn is used to align the horse genome (green) and the rhinoceros scaffold (short red and grey bars). The rhinoceros scaffold JH767750.1 is aligned to the horse chromosome 6: 73,264,962-73,285,366. Colour lines and black arrows are used to show borders of aligned regions.

5.3 Ancestral ERV lineages in the horse genome

5.3.1 Clade I: Rho

The copy number of Rho ERV insertions was much larger than any modern ERV lineages (Table 4-4 and 4-5). In total, 151 potential provirus loci and 4062 solo LTR loci were identified, as well as five proviruses with complete genome structures. The relatively large number of loci indicates that Rho expanded massively during perissodactyl evolution. Furthermore, multiple LTR groups were identified in this lineage, suggesting that several distinct germline invasions events may have occurred for this lineage.

Rho proviruses that retained internal coding regions were degraded. Nevertheless, some reasonably long regions of the intact coding sequence (i.e. >300 aa) were identified – mostly derived from *pol* and *gag* coding domains. I found, however, that the longest intact regions among these were derived from fusions of *pol* and LINE1 coding domains. All other coding regions were less than 600 aa – i.e. shorter than the normal length of the major retroviral coding domains.

By annotated the flanking regions of identified RT loci without flanking LTRs, I found a large proportion (n=48) of Rho ERV loci were adjacent to LINE1 elements, and others (n=53) still kept *gag and* pol genes.

Estimates based on the paired LTRs indicated that the age of proviral Rho insertions ranged between 3.18 Mya and 34.77 Mya (Table 5-1). Most were estimated to be ~30 million years old, but one provirus of Rho was estimated to be only 3.18 million years old. As this was inconsistent with its presence as an ortholog in the rhinoceros genome, it might reflect an artefact generated by gene conversion.

Table 5-1 Integration time of Rho proviruses using paired LTR dating

CHR	RT START	RT END	LTR ID	DISTANCE	MYA
chr20	24646310	24655093	ERV1-3-EC_LTR	0.014	3.18
chr5	77382062	77382466	ERV1-3-EC_LTR	0.026	5.91
chr18	1692239	1692421	ERV1-3-EC_LTR	0.027	6.14
chr1	27396955	27397359	ERV1-3-EC_LTR	0.028	6.36
chr10	15563937	15564281	LTR8B_EC	0.066	15.00
chr1	116006166	116006486	LTR8E_EC	0.07	15.91
chr5	44221225	44221536	LTR15_EC	0.073	16.59
chr3	10459139	10459495	ERV1-2-EC_LTR	0.08	18.18
chr5	78218517	78218912	ERV1-2-EC_LTR	0.092	20.91
chr16	31131256	31131672	LTR15_EC	0.095	21.59
chr9	28678508	28678918	LTR8B_EC	0.096	21.82
chr20	32668167	32668487	LTR15_EC	0.105	23.86
chr20	30844305	30844538	ERV1-2-EC_LTR	0.108	24.55
chr1	34479199	34479612	LTR8F_EC	0.113	25.68
chr24	9457937	9458128	ERV1-2-EC_LTR	0.124	28.18
chrX	53127434	53127835	LTR45_EC	0.127	28.86
chr22	24094857	24095075	LTR15_EC	0.129	29.32
chr7	63949233	63949505	LTR8E_EC	0.129	29.32
chr5	32012982	32013122	LTR72A_EC	0.145	32.95
chr7	42537733	42538146	ERV1-2-EC_LTR	0.149	33.86
chr1	161956638	161956748	ERV1-2-EC_LTR	0.152	34.55
chr25	17884308	17884712	ERV1-2-EC_LTR	0.153	34.77

CHR: chromosome; LTR ID: LTR ID used by Repbase; Distance: pair-wise maximum likelihood distance between 5' and 3' LTRs; MYA: million years ago

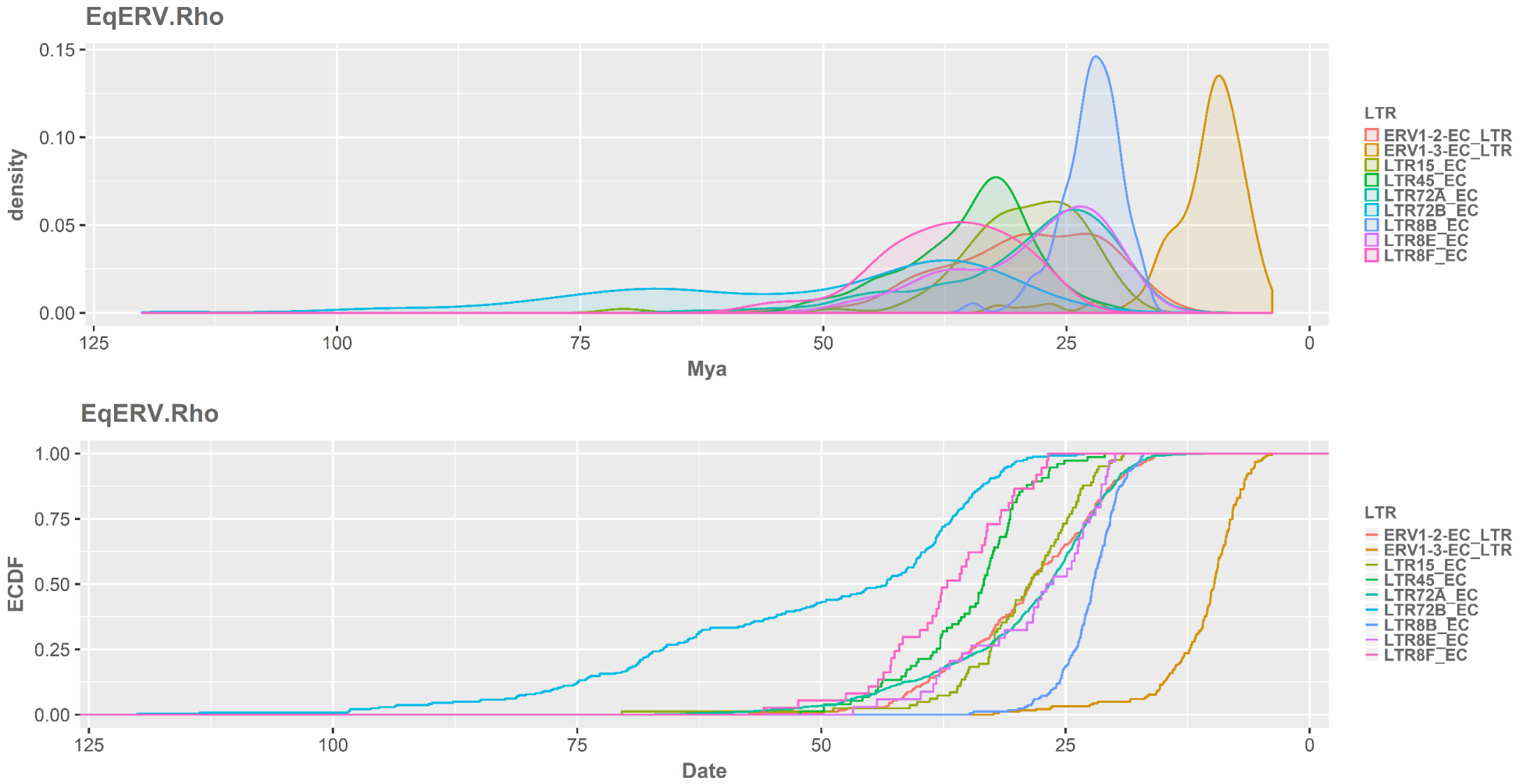


Figure 5-3 Density plot and ECDF plots of Rho solo LTRs. (i) Density plot for the distribution along the time scale; (ii) ECDF plots for the cumulative proportion of observed LTRs versus time scale. The x-axis shows time in millions of years before present, and the y-axis shows the cumulative proportion. LTRs from the same ERV lineage are shown in the same plot with different colours. All X axes are adjusted to the same scale.

To gain an overview of Rho integration history, further age estimations were obtained from solo LTRs, based on their similarity to consensuses derived from Repbase. Using this approach, the maximum age of solo LTRs was 120 Myr, and minimum age was 3.84 Myr (Figure 5-3). Analysis of density plot of solo LTRs suggested that the activity of Rho continued at a low level for over 100 Myr, and the massive expension began around the speciation of the *Equus* genus (~54 Mya) until 25 Mya. Although the distributions of the integration time of solo LTRs between 25 Mya and 54 Mya, the majority of solo LTRs appeared in the same period. Also, the increasing speed of copy number increase in each LTR groups was similar, as shown by the ECDF plot. After 25 Mya, two LTR groups – LTR8B_EC and ERV1-3-EC_LTR - contributed most additional Rho insertions. After 10 Mya, only LTR15_EC was still active. These results indicate that the Rho lineage expended in the horse genome via multiple events and reminded to active until relatively recent.

5.3.2 Clade I: Theta

Similar to the Rho lineage, the Theta lineage was highly abundant in the horse genome and contained two major sublineages (Table 4-4 and 4-5). Of these, Theta.1 contained more RT loci (251 vs 67). By contrast, however, the number of Theta.2 solo LTRs was 40 times larger than the number of Theta.1 LTRs (8540 vs 295). This indicates that, for some reason, more loci in the Theta.1 lineage have been retained as proviruses.

Only 20 loci were flanked by paired LTRs including 11 Theta.1 and 9 Theta.2. Of 20 loci, five loci contain a provirus with complete genome structure – 5'LTR-*gag-pro-pol-env*-3'LTR. However, none of these loci has intact genes. The longest coding region (~790 aa) was identified in the *pol* domain of a Theta.1 insertion on chromosome 1. All other coding regions were shorter than 600 aa. There were no *gag* domains >300 aa in length, but I did find many ORFs over 300 aa that were fused with LINE1 elements. I also found several Theta proviruses that were associated with amino acid permease genes.

The overall degradation of loci suggested that Theta has resided in the perissodactyl germline for a very long time. This inference was supported by the integration dates estimated from paired LTRs (Table 5-2). The majority of proviruses with paired LTRs were estimated to be ~9 Myr, with one provirus on

chromosome 11 was estimated be ~4.3 Myr. Thus, the observed proviruses of Theta were all established before the divergence between horse and donkey.

Nine major LTR groups were observed from the Theta provirus loci. Interestingly, all Theta.1 proviruses have the same LTRs, whereas Theta.2 has eight different LTR groups. All Theta.1 LTRs were assigned to ERV1-4-EC_LTR of Repbase. The density plot of solo LTR dating showed two peaks of Theta.1 activity. One occurred around the speciation of the *Equus* genus ~54 Mya, and another occurred from 40 Mya until relatively recently. This second expansion was greater in extent and contributed the majority of Theta.1 insertions.

In contrast to Theta.1, most Theta.2 insertions were established in the host genome around 30 Mya, but the maximum date of integration of Theta.2 lineage was much bigger than the maximum integration time of Theta.1 lineage. Furthermore, the growth speed of copy number increases for of all LTR groups was similar to each other according to the ECDF plot (Figure 5-4).

Table 5-2 Integration time of Theta proviruses using paired LTR dating

CHR	RT START	RT END	LTR ID	DISTANCE	MYA
chr11	60594390	60588146	ERV1-4-EC_LTR	0.019	4.31
chr4	55652267	55652524	ERV1-6-EC_LTR	0.042	9.54
chr28	2591726	2592133	ERV1-4-EC_LTR	0.059	13.40
chrX	69213526	69213921	LTR13A_EC	0.065	14.77
chr29	8718582	8718791	ERV1-4-EC_LTR	0.07	15.90
chr2	15924940	15925239	LTR27_FC	0.073	16.59
chr9	34319072	34319254	MER34A1_EC	0.078	17.72
chrX	13238144	13238416	ERV1-4-EC_LTR	0.085	19.31
chr14	18648767	18648946	ERV1-4-EC_LTR	0.088	20.00
chr5	8822660	8823055	LTR23B_EC	0.092	20.90
chr10	12947010	12947282	ERV1-4-EC_LTR	0.096	21.81
chr25	28959648	28959860	LTR19_EC	0.105	23.86
chr10	28918333	28918521	ERV1-4-EC_LTR	0.106	24.09
chr1	40866919	40867068	ERV1-4-EC_LTR	0.109	24.77
chr7	5153007	5153249	MER34A1_EC	0.114	25.90
chr2	119104373	119104747	LTR6_EC	0.122	27.72
chr1	10888231	10888512	MER34A1_EC	0.123	27.95
chr15	46321810	46322058	LTR6B_EC	0.156	35.45
chr18	52452681	52452851	ERV1-6-EC_LTR	0.163	37.04
chrX	120359468	120359668	ERV1-6-EC_LTR	0.17	38.63

CHR: chromosome; LTR ID: LTR ID used by Repbase; Distance: pair-wise maximum likelihood distance between 5' and 3' LTRs; MYA: million years ago

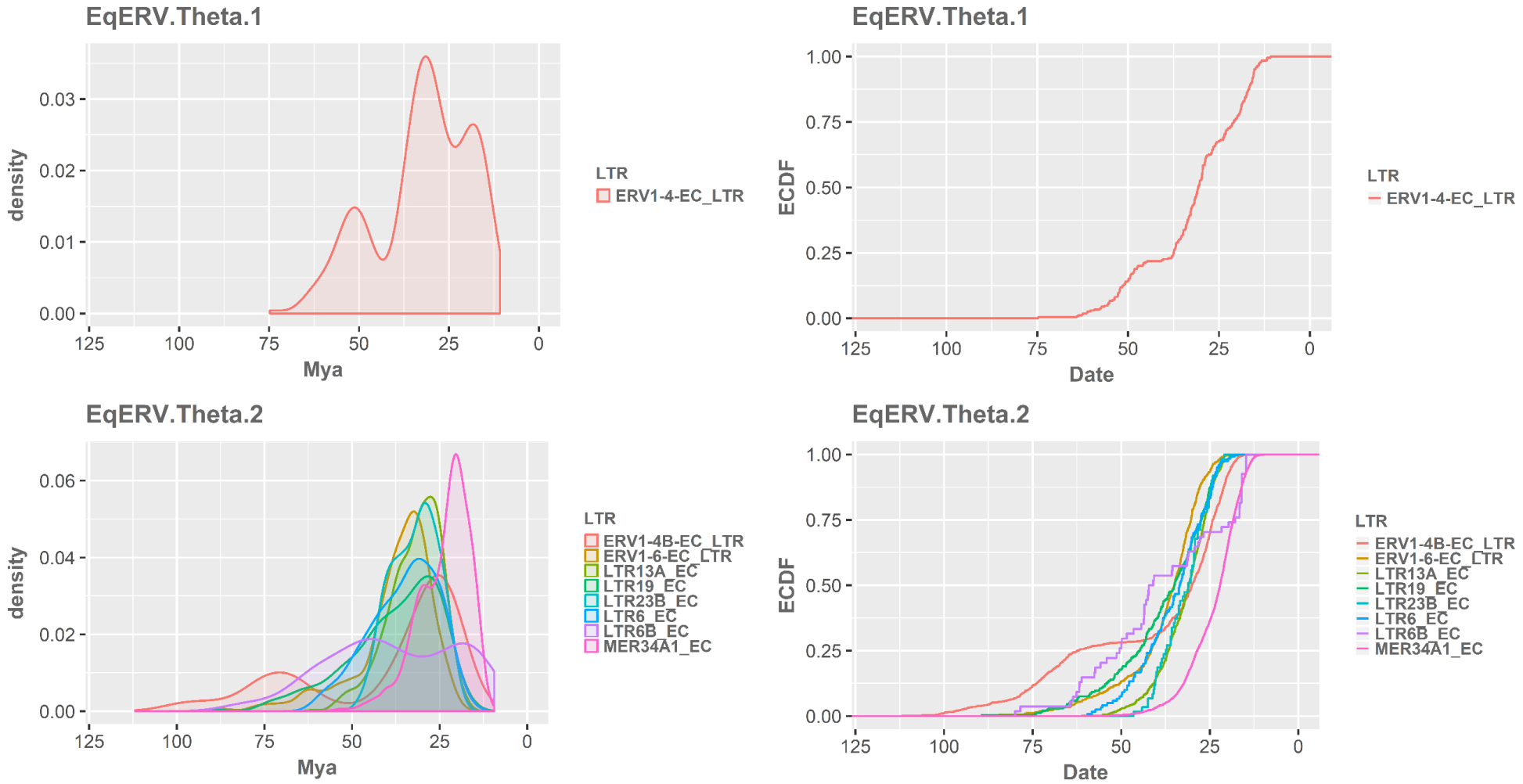


Figure 5-4 Density and ECDF plots of Theta solo LTRs. (Left) Density plot for the distribution along the time scale; (Right) ECDF plots for the cumulative proportion of observed LTRs versus time scale. The x-axis shows time in millions of years before present, and the y-axis shows the cumulative proportion. LTRs from the same ERV lineage are shown in the same plot with different colours. All X axes are adjusted to the same scale.

5.3.3 Clade III: Lambda

Based on the phylogenetic reconstruction of RTs, Lambda lineage was suggested to be one of the spuma-like virus clade III lineages. Lambda lineage was the most abundant ERV lineage in the horse genome. 723 RT loci were found in the horse genome, and orthologous loci were found in the donkey and rhinoceros genomes. However, none of these RT loci was flanked by paired LTRs. The LTRharvest program identified nine loci that were flanked by two similar sequences (similarity > 80%), but all these sequences were assigned to the LINE1 consensus sequences of Repbase. Also, HMMER failed to identify any *gag* or *env* genes in the flanking regions of Lambda RT loci. All coding domains found in Lambda RT loci were interrupted by indels and stop codons.

The longest ORFs was *pol* with 842 aa in length. Another five ORFs were found to have a length >600 aa, but all of them were fusions of LINE1 and partial pol genes. Annotation of flanking regions of Lambda RT loci suggested that Lambda RT were frequently adjacent to the LINE1 elements. Thus, I inferred that Lambda is the most ancient origins of any perissodactyl ERV lineage, and its expansion was likely mediated via non-LTR retrotransposition.

5.3.4 Clade III: Sigma

Sigma is the second spuma-like ERV lineage identified in the perissodactyl germline. Phylogenetic reconstruction suggested that Sigma was closely related to HERV.S and distinct from Lambda.

Table 5-3 Integration time of Sigma proviruses using paired LTR dating

CHR	RT START	RT END	LTR ID	DISTANCE	MYA
chrX	51971865	51972254	ERV3-1C-EC_LTR	0.049	11.14
chr26	34336721	34336885	LTR74_EC	0.1	22.73
chr9	38452893	38453228	LTR74_EC	0.14	31.82

CHR: chromosome; LTR ID: LTR ID used by Repbase; Distance: pair-wise maximum likelihood distance between 5' and 3' LTRs; MYA: million years ago

In contrast to the Lambda lineage, the Sigma copy number was quite low. 75 RT loci were identified from the horse genome; six loci were defined as potential proviruses loci due to the existence of paired LTRs. Interestingly, HMMER and LTRdigest failed to identify any *gag* genes from the RT loci with or without flanking LTRs. However, three *env* genes were found among the proviral loci. Dates obtained from paired LTRs were consistent with the ancestral origin of the Sigma lineage (Table 5-3).

Comparison of the paired LTR sequences identified in proviral loci to RepBase consensuses sequences indicated that the Sigma lineage contained two distinct LTR groups. Estimation of integration time using solo LTRs suggested that the integration activity of Sigma was ancient and continuous up until 15 Mya (Figure 5-5). Both two LTR groups could be dated back to 60 Mya. The maximum integration age of LTR group 'LTR74_EC' was larger than 'ERV3-1C-EC_LTR', and LTR74_EC had more copies over 60 Mya. Similar to the other ERV lineages, the most active period of Sigma was around 30 Mya. However, the copy number increased gently and reached the peak at 30 Mya. Also, as showed in the ECDF plot, the copy number of the Sigma insertion expended gently, in contrast, other ERV lineage usually expanded rapidly.

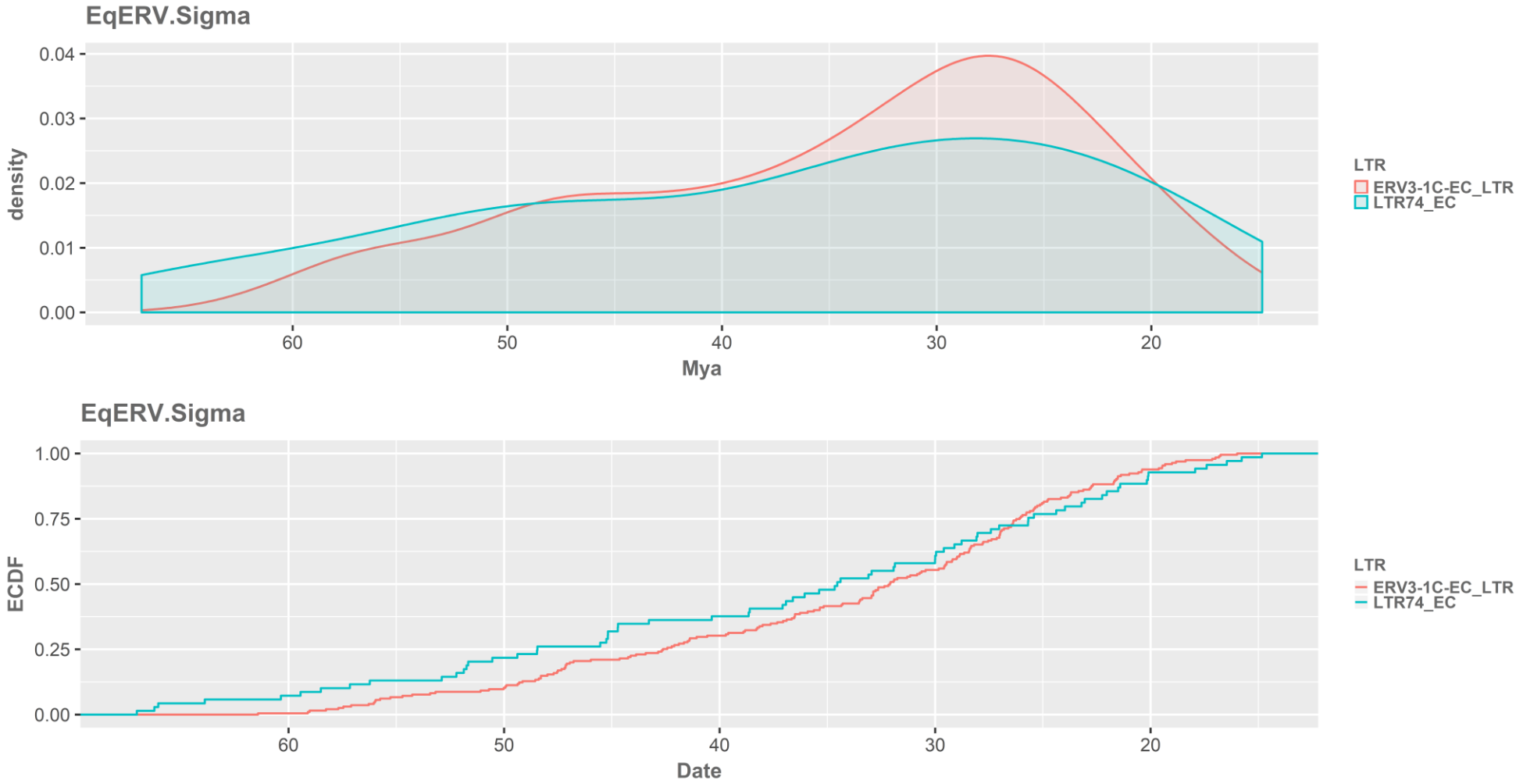


Figure 5-5 Density ECDF plots of Sigma solo LTRs. (Up) Density plot for the distribution along the time scale; (Down) ECDF plots for the cumulative proportion of observed LTRs versus time scale. The x-axis shows time in millions of years before present, and the y-axis shows the cumulative proportion. LTRs from the same ERV lineage are shown in the same plot with different colours. All X axes are adjusted to the same scale.

5.4 Modern ERV lineages in the horse genome

5.4.1 Clade I: Zeta

Zeta insertions were present in both rhinoceros and horses. However, I did not identify any orthologous insertions in those species. Therefore, the Zeta lineage is the only group of perissodactyl ERVs in clade I that was put into the 'modern' category.

A total of 17 Zeta proviruses loci and 3953 associated solo LTRs were identified in the horse genome. This suggested that the horse and its ancestors experienced a massive expansion of Zeta ERVs during their evolution. Also, there were least three different LTR groups present within the lineage, corresponding to three LTR consensuses present in Repbase. These LTR groups are clearly distinct, yet are associated with proviruses that are closely related. These data indicate that there may have been multiple episodes of germline colonisation by related viruses in this lineage.

It was surprising to find that Zeta proviruses also had the most intact coding regions found among all nine lineages. 34 coding domains >300 aa were detected from the 15 Zeta provirus loci. The longest coding domains were found from a full-length provirus on the chromosome 2 (1,716,256-11,716,660). It was 1211 aa in length encoding an intact pro-pol protein. Of 34 domains, 19 domains were *pol*-related, and nine domains were *gag*-related. Two long partial *env* coding regions were also found on chromosome 5 and 11. One coding regions contained both LINE1 and *pol* sequences. The relatively large number of long coding domains suggested that Zeta proviruses could have been active quite recently.

By comparison with Repbase database, three LTR consensus sequences – ERV1-LTR_EC, LTR14_EC and LTR1420_EC - could be assigned to flanking LTRs of Zeta lineages. The uncorrected genetic distance between three different LTRs was 0.176 base substitutions per site. It was interesting that three LTRs shared the conserved R and U5 regions, but the U3 regions were highly variable (Figure 5-6).

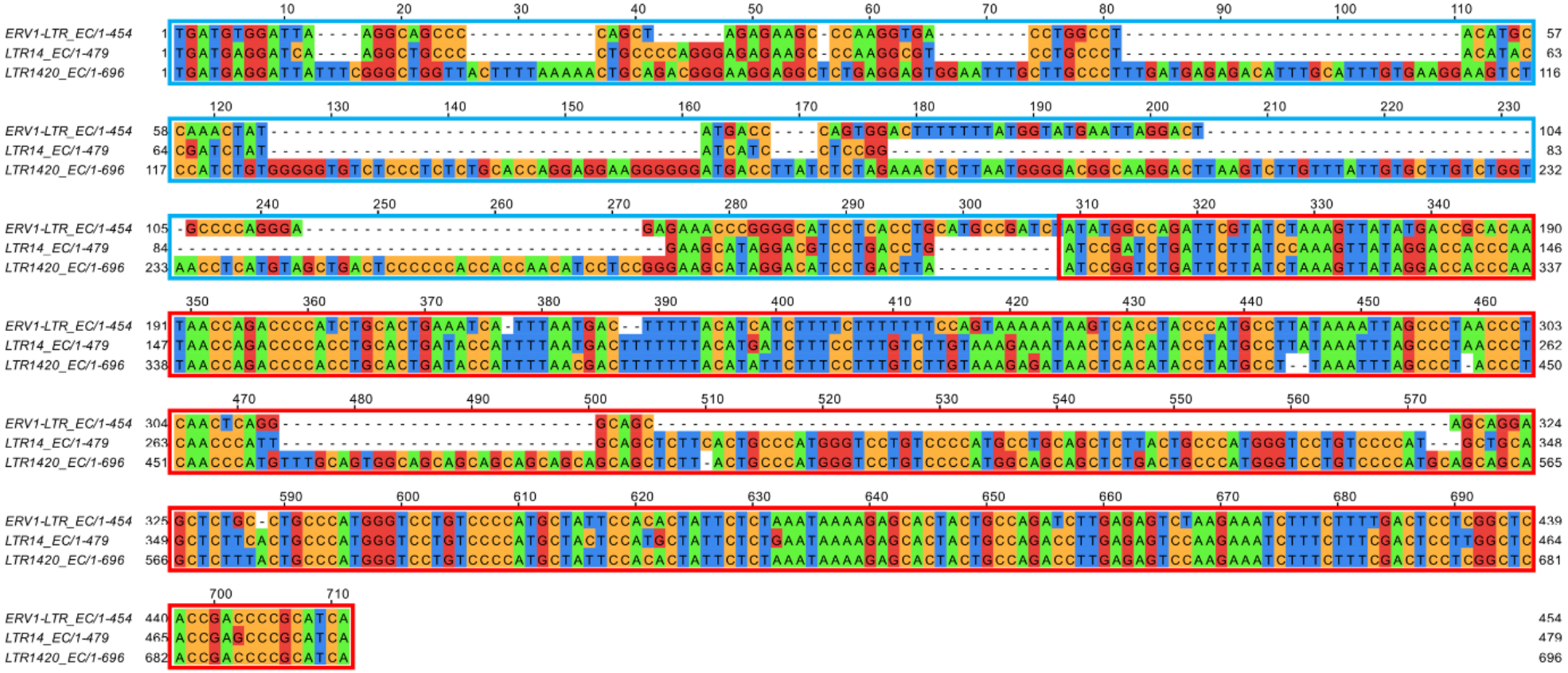


Figure 5-6 Alignment of three Zeta LTR consensus of Repbase. The alignment of consensus sequences was generated by MUSCLE. Blue frame shows the region of U3 while red frame shows the regions of R and U5.

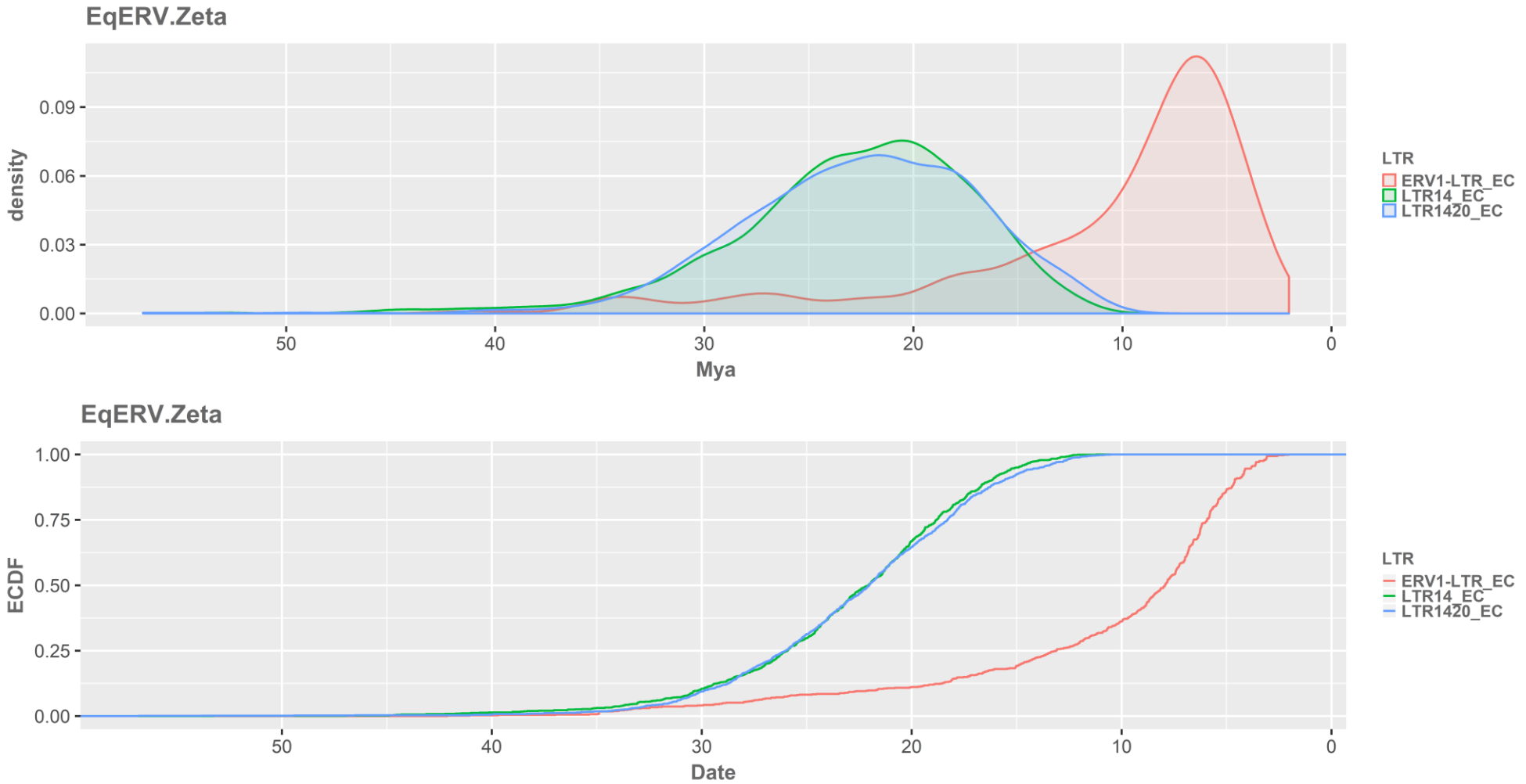


Figure 5-7 Density ECDF plots of Zeta solo LTRs. (Left) Density plot for the distribution along the time scale; (Right) ECDF plots for the cumulative proportion of observed LTRs versus time scale. The x-axis shows time in millions of years before present, and the y-axis shows the cumulative proportion. LTRs from the same ERV lineage are shown in the same plot with different colours. All X axes are adjusted to the same scale.

In general, the estimations of integration age were conducted using 685 ERV1-LTR_EC, 924 LTR14_EC and 1137 LTR1420_EC, respectively. As the density plot shown in Figure 5-7, LTR14_EC and LTR1420_EC had a similar distribution of integration times. Most integration occurred between 10 Mya and 40 Mya. Due to the high similarity of density distribution, it was not surprising to find that ECDF plots of LTR14_EC and LTR1420_EC overlapped each other. Instead, the copy number of ERV1-LTR_EC remained at a low level early on when the copy numbers of LTR14_EC and LTR1420_EC were expanding rapidly. However, ERV1-LTR_EC was more active from ~20 Mya and expanded ~15 Mya rapidly, during a period in which the other groups were active only at low levels.

Estimations based on the flanking paired LTRs indicated a recent activity of Zeta ERVs (Table 5-4). Two provirus loci were dated to 1.59 Mya and 3.86 Mya, indicating they were generated after the divergence of donkey and horse. Consistent with this, I identified the orthologous empty insertion site in the donkey for a Zeta provirus on horse chromosome 5 (27,326,038-27,333,559). However, the provirus on chromosome 4 (58,024,286-58,032,024), which was dated to 3.8 Mya, was present as an ortholog on the donkey scaffold, which suggested that the integration time was supported to over 4.5 Mya.

Table 5-4 Integration time of Zeta proviruses using paired LTR dating

CHR	RT START	RT END	LTR ID	DISTANCE	MYA
chr5	27326038	27333559	ERV1-LTR_EC	0.007	1.59
chr4	58024286	58032024	LTR14_EC	0.017	3.86
chr2	11706256	11726660	ERV1-LTR_EC	0.036	8.18
chrX	44321574	44341975	ERV1-LTR_EC	0.044	10.00
chr11	17081079	17101477	LTR14_EC	0.052	11.82
chr1	105932965	105953060	LTR14_EC	0.082	18.64
chr4	47574607	47594774	LTR14_EC	0.086	19.55
chr7	47460934	47481287	LTR1420_EC	0.094	21.36
chr21	44206723	44227127	LTR14_EC	0.104	23.64
chr27	16314830	16335225	LTR14_EC	0.13	29.55

CHR: chromosome; LTR ID: LTR ID used by Repbase; Distance: pair-wise maximum likelihood distance between 5' and 3' LTRs; MYA: million years ago

Furthermore, the most recent integration time of LTR14_EC solo LTRs was 11.21 Mya. The existence of proviruses with LTR14_EC LTRs indicated that the activity of Zeta lineage with LTR14_EC was much longer than solo LTRs dating suggested. Also, it was interesting that the copy number of proviruses with LTR14_EC was higher than those with LTR1420_EC. Considering the distribution of integration time of LTR14_EC and LTR1420_EC, it seems that more proviruses with LTR14_EC were retained in the horse genome during its evolution.

5.4.2 Clade II: Beta1

Beta1 is a *Betaretrovirus* lineage found in the horse genome, which has previously been described in detail (van der Kuyl, 2011). A full-length Beta1 provirus was found on the positive strand of chromosome 5. It has intact *gag*, *pro* and *pol* coding domains, and an *env* gene interrupted by a single stop codon. This was the most intact ERV provirus identified among in a perissodactyl ERV lineage.

The full-length Beta1 provirus suggested the integration occurred recently. When paired LTRs are used to estimate the age of this provirus, estimates between 0.3-2.27 Mya were obtained depending on the substitution rate used (i.e. before the divergence of donkeys and horses). However, it was clear from the presence of this sequence as an ortholog in the donkey genome, that it predated this event. Indeed, genome screening demonstrated that the Beta1 lineage was present in all Equus species. Thus, the available evidence indicates that the initial germline colonisation event for the Beta1 lineage took place at least 4.5 Mya.

Estimation of integration time using 195 solo LTR sequences showed that the range of Beta1 integration was between 3.62 Mya and 19.61 Mya (Figure 5-8). The overwhelming presence of solo LTRs suggested that the ancestor of *Equus* species experienced a massive expansion of this lineage. According to the density plot, the period of massive integration was more likely to be around 3 to 10 Mya. Also, the ECDF plot suggested that the copy number of Beta1 lineage had the highest increase rate. Together, these results established a minimum age for the Beta1 that is considerably more ancient than the 0.5 Mya (Assuming a nucleotide substitution rate of 10^{-8} substitution/base pair/generation) suggested previously and indicated that Beta1 was still active after the speciation of horse and donkey (van der Kuyl, 2011).

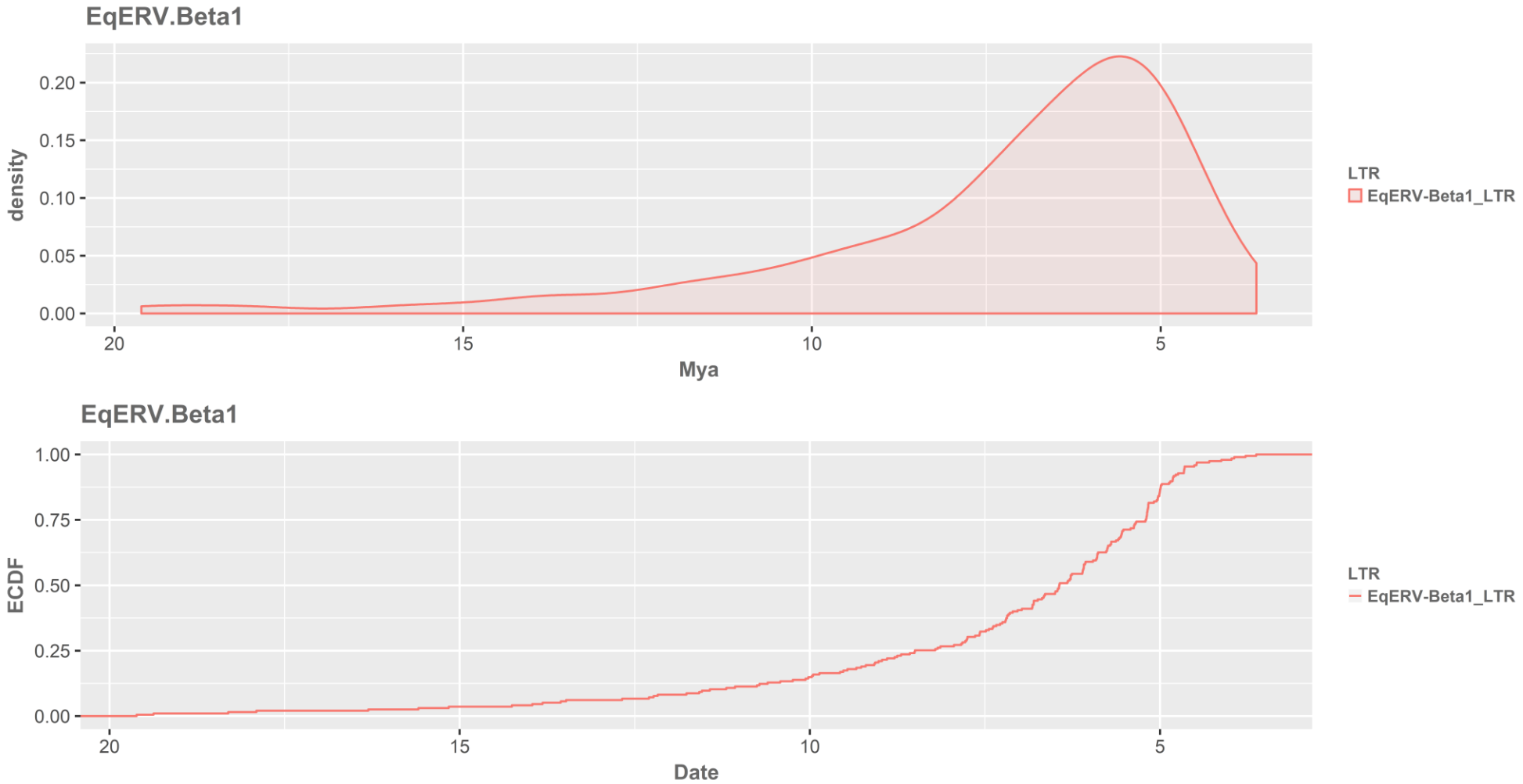


Figure 5-8 Density and ECDF plots of Beta1 solo LTRs. (Left) Density plot for the distribution along the time scale; (Right) ECDF plots for the cumulative proportion of observed LTRs versus time scale. The x-axis shows time in millions of years before present, and the y-axis shows the cumulative proportion. LTRs from the same ERV lineage are shown in the same plot with different colours. All X axes are adjusted to the same scale.

5.4.3 Clade II: Kappa1 and Kappa2

Both Kappa1 and Kappa2 lineages have nearly complete proviruses, but the copy number of proviruses was very low (4 for Kappa1, 3 for Kappa2) (Table 4-4). Three pairs of flanking LTR were recovered from the Kappa1 provirus loci. All three paired LTRs were highly similar to the LTR record 'ERV2-2-EC_LTR' of Repbase. LTRs identified from the Kappa2 provirus loci were not included in the Repbase.

Two LTR pairs of Kappa1 were used to estimate the integration time; one pair was discarded due to the long indels. The distances of Kappa1 paired LTRs were 0.04 and 0.042 base substitutions per site, respectively. Divided by the neutral mutation rate, these Kappa1 proviruses were estimated to integrate into the horse genome at 9.09 and 9.54 Mya. Paired LTR dating was only possible for one Kappa2 provirus. The uncorrected genetic distance of Kappa2 LTRs was 0.035 base substitutions per site which were 7.95 Mya. Comparing the age of proviruses indicated that the Kappa2 lineage was slightly younger than Kappa1 lineage. However, both of them integrated into the equid genome before the divergence of donkey and horse.

A total of 72 solo LTRs from the Kappa1 lineage and 55 solo LTRs from the Kappa2 lineage were aligned. The average genetic distance between Kappa1 LTRs and consensus sequences was 0.06 base substitutions per site, which equates to 17.04 Myr when assuming a neutral rate. For Kappa2 LTRs, the average distance was only 0.03 base substitutions per site, equating to 6.81 Myr of neutral evolution. Notably, dates obtained from solo LTRs of the Kappa1 and Kappa2 lineages were consistent with those obtained from orthologs. The maximum integration age of the Kappa lineages was between 25.95 Mya and 17.38 Mya, for Kappa1 and Kappa2 respectively.

The Kappa1 lineage was older than Kappa2 in general (Figure 5-9). Most of the Kappa1 integrated into the host genome before 10 Mya, but the majority of Kappa2 appear after 10 Mya. Furthermore, the copy number of Kappa1 increased steadily in the horse genome over time. In contrast, the copy number of Kappa2 ERVs only remained at a low level, and then abruptly expanded to the current number after 10 Mya.

Phylogenetic trees (Figure 5-10A) were inferred separately using alignments of Kappa1 and Kappa2 solo LTR sequences. Notably, Kappa1 solo LTRs formed four major clades, two of them with bootstrap values > 75. By mapping integration time on the phylogeny, all four clades showed the similar trends. Every clade contained a certain number of old and recent integration time points. Therefore, I inferred that the copy number of Kappa1 ERVs raised by at least three major expansions. First expansion began at around 26 Mya and expanded until 11 Mya. Second expansion happened at approximately 25 Mya, and it kept increasing copy number until 11 Mya. The third expansion occurred later than the other two expansions as roughly 21 Mya but continued increasing copy number to 8 Mya. Based on this assumption, Kappa1 proviruses could be the result of the third expansion which is the most recent one.

However, all solo LTRs of Kappa2 lineages were more likely to be generated by the same expansion (figure 5-10b). The phylogeny of Kappa2 LTRs did not show any clades with high bootstrap values (Figure 5-11). It was also interesting that the description of integration age of Kappa2 LTRs completely differed from the description of Kappa1 LTR integration age. On the Kappa2 phylogeny, integration time points at the close period tended to cluster together. The mapped time points showed a gradient that gradually changes from early to recent along the tree topology.

In sum, these results suggest that the Kappa1 lineage originated at least 25 Mya. At least three Kappa1 expansions happened according to the phylogenetic reconstruction and annotation of integration time. By comparison, all Kappa2 originated from the same germline invasion around 21 Mya. The copy number of Kappa1 increased quickly for a long time-period, but the copy number of Kappa2 ERVs only grew fast after 10 Mya.

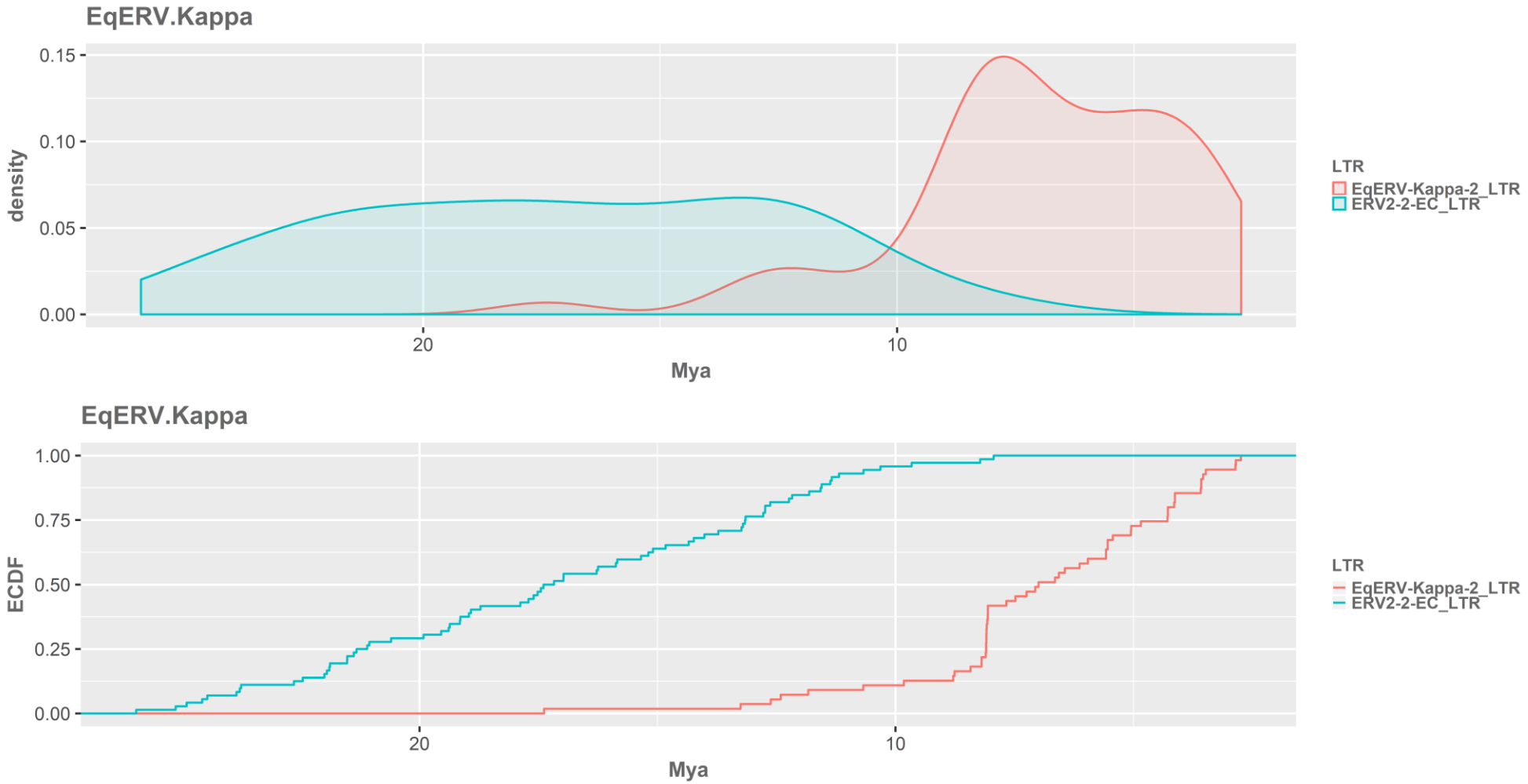


Figure 5-9 Density and ECDF plots of Kappa solo LTRs. (Left) Density plot for the distribution along the time scale; (Right) ECDF plots for the cumulative proportion of observed LTRs versus time scale. The x-axis shows time in millions of years before present, and the y-axis shows the cumulative proportion. LTRs from the same ERV lineage are shown in the same plot with different colours. All X axes are adjusted to the same scale.

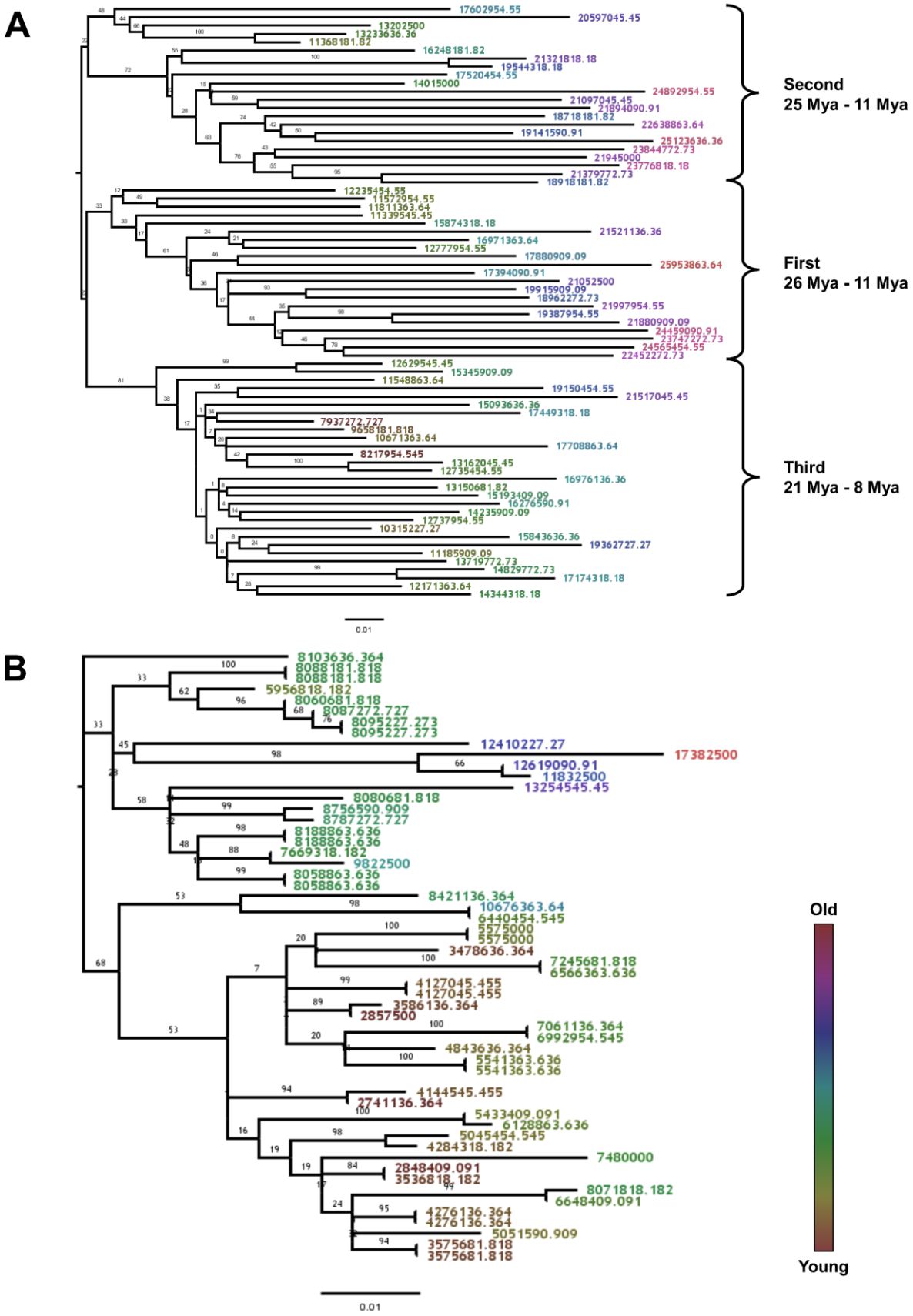


Figure 5-10 Maximum likelihood phylogenetic tree of Kappa solo LTRs. (A) Phylogeny of Kappa.1 solo LTRs; (B) Phylogeny of Kappa.2 solo LTRs; Phylogenetic reconstruction was inferred by RAxML using multiple sequence alignment of Kappa solo LTRs. Tips represent the integration time of each solo LTR. Phylogeneies are mid-rooted. Tips are coloured according to the associated integration age using a colour scale from red (old) to brown (young). Values shown on branches are bootstrap values.

5.4.4 Clade II: U1

Two genome organisations of U1 ERVs were found

The genome organisations found among proviruses of the U1 lineage are shown in Figure 5-11. The type I organisation was typical of a betaretrovirus and featured a dUTPase domain encoded at the junction of the *gag* and *pro*. In total, 11 U1 proviruses were identified that had this type I genome organisation, of which nine were identified on unmapped chromosomal regions. Based on the consensus sequence, the N-terminal segment of dUTPase was approximately 111 aa (~ 333 bp) long. Moreover, the N-terminal segment overlapped the whole NC domain. The C-terminal segment was roughly 120 aa (~ 360 bp) long. The length of the dUTPase was typical of those found in betaretroviruses.

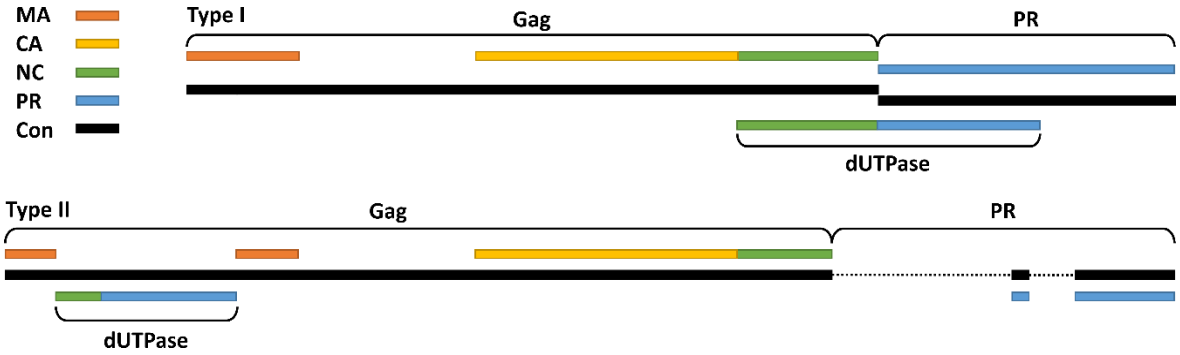


Figure 5-11 The genomic organisations of U1. Basket shows the range of *gag*, *pro* and dUTPase ORFs. Coloured frames show protein products: MA (orange), CA (yellow), NC (green), PR (blue). The consensus sequences of type I and II proviruses are shown as black frame. Deletions of PR of type II genomic organisation are shown as dash line. The figure is shown on the scale.

The second type of genomic organisation (referred to type II) was found in 18 proviruses that had been mapped to specific chromosomes and nine proviruses that had not. The dUTPase in these proviruses was encoded 120 bp downstream of the *gag* start codon. The total length of the dUTPase in these proviruses was 478 bp (i.e. truncated relative to that found in type I). The alignment of dUTPase sequences of type I and type II proviruses indicated that the whole N-terminal segment of dUTPase in the type II provirus could be aligned to 108 bp C-terminal of the NC protein within *gag*. The C-terminal segment of dUTPase in the type II

provirus was 40 bp shorter than the C-terminal segment of which of in the type I provirus. Instead, dUTPase in the type II provirus has a 51 bp MA domain tail.

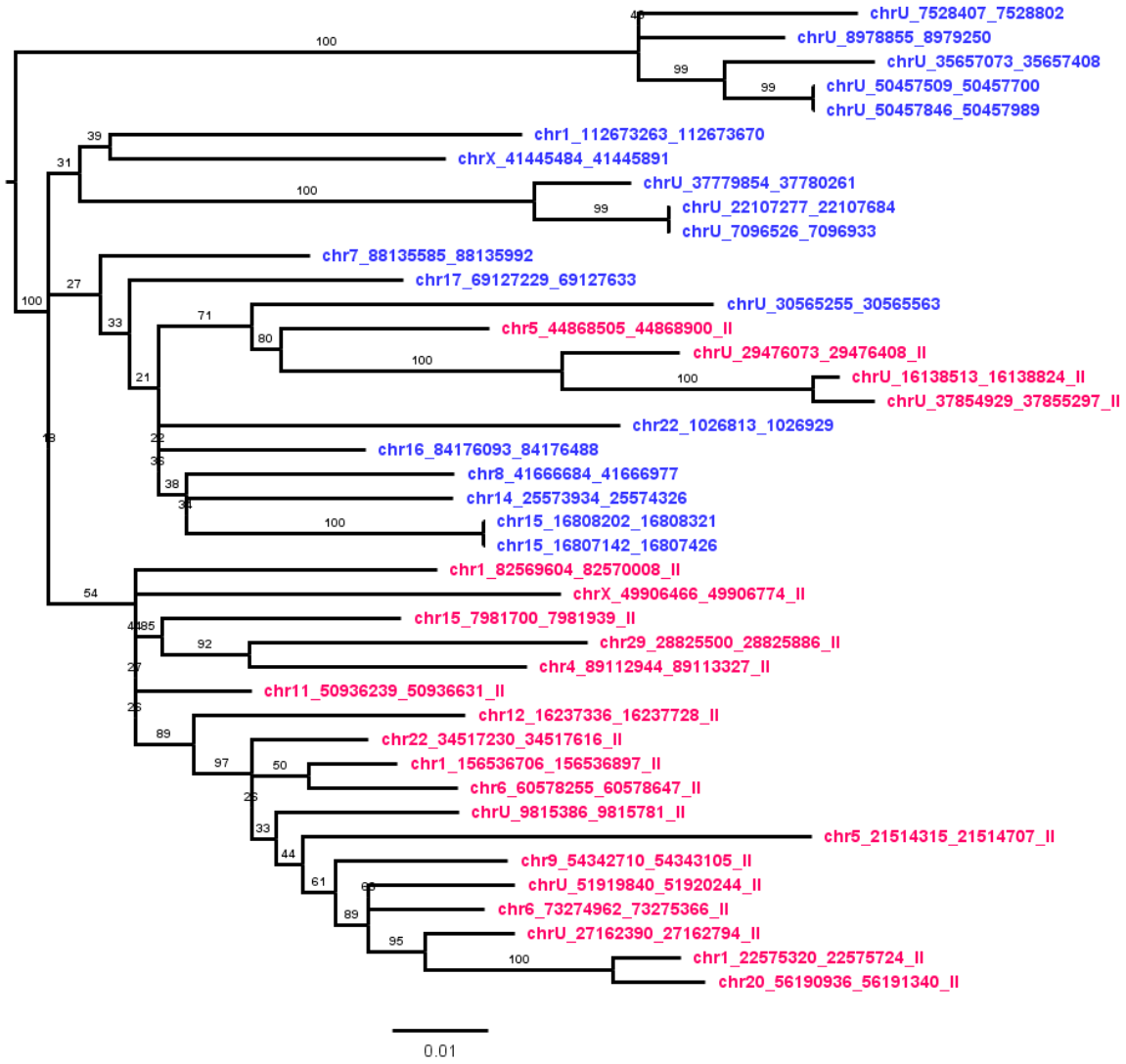


Figure 5-12 Maximum likelihood phylogenetic tree of U1 dUTPase. Phylogenetic reconstruction was inferred by RAxML using multiple sequence alignment of type I and type II dUTPase. Type I and type II dUTPase are shown as red and blue, respectively. Values shown on branches are bootstrap values. Tips represent the location of U1 proviruses.

In type II proviruses, the presumably relocated dUTPase interrupts the *gag* reading frame. The 51 bp MA domain tail was a duplicate of 51 bp 5'flanking region of the dUTPase domain in the type II proviruses. This duplication could not be observed from the type I proviruses. Also, the translation frame shift between *gag-pro* junction was found in the dUTPase domain in the type II proviruses. The NC domain of *gag* of type II proviruses was 21 bp shorter at the 3'end. This truncation also caused the loss of the stop codon in *gag*.

Furthermore, the type II provirus has an interrupted *pro* coding domain. Compared to type I *pro*, type II *pro* consisted of one 48 bp and one 236 bp fragment. These two fragments were concatenated in the type II *pro*, but a 100 bp sequence was observed to separate them in type I.

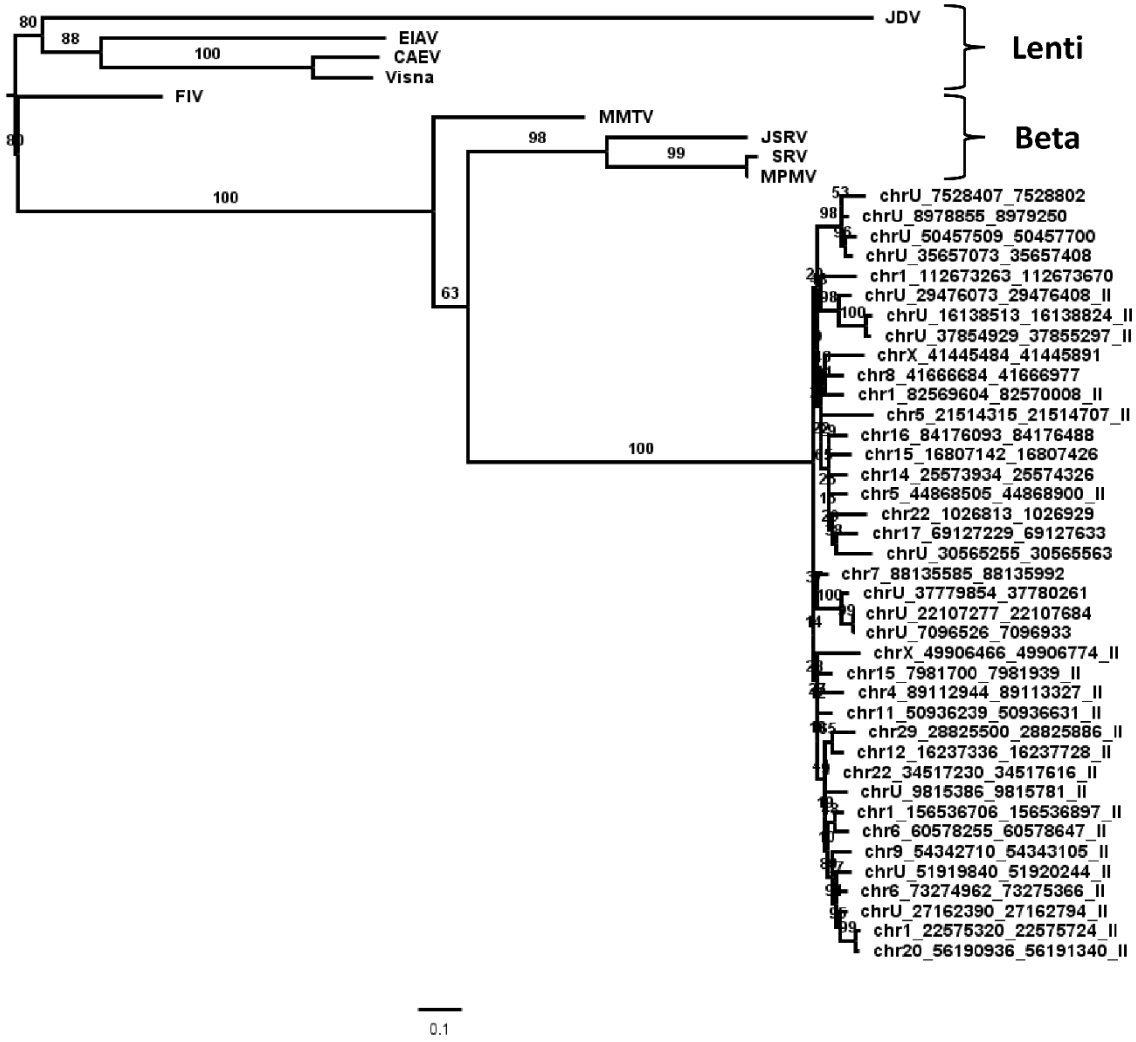


Figure 5-13 Maximum likelihood phylogenetic reconstruction of U1 dUTPase. Phylogenetic reconstruction was inferred by RAxML using multiple sequence alignment of dUTPases of U1, known betaretroviruses and lentiviruses. Values shown on branches are bootstrap values.

Sequence comparisons showed that dUTPase of the type I and type II proviruses are closely related. Phylogenetic reconstruction of dUTPase roughly split into two clades using mid root (Figure 5-12). However, non-parametric bootstrap replication did not provide strong support for this split. The phylogenetic tree of exogenous retroviral dUTPase with U1 dUTPase demonstrated that all dUTPase from U1 formed a monophyletic clade (Figure 5-13). Thus, all dUTPase in U1 proviruses clearly have a common origin.

Complete coding regions found in the U1 proviruses

Annotations of genomic structure suggested that most of U1 proviruses had relatively complete genomes with two flanking LTRs. In total, 15 potential coding regions of 12 proviral loci were found to be over 300 aa in length (eight regions on the chromosome unknown). The translations of long coding regions were further checked by BLASTp against the NCBI protein database. BLAST results indicated that five potential regions were gag-relative - including partial gag and complete NC domains. Moreover, seven regions were relative to *pol*. Also, two regions were found to be LINE1-relative. One *env* were found at the chromosome unknown.

Unfortunately, none of these coding regions was completed. The longest coding regions were found at chromosome X: 41,445,484-41,445,891, it was a 744-aa long partial *pol*. This region was still 30% shorter than the normal class II pol (around 1000 aa). One small region (104 aa) was found at the immediately downstream of the long region. The separation of long and short *pol* coding regions was due to a frame-shift caused by indels. All the others were much shorter than any known proviral genes. Another interesting finding was a relatively intact *gag* (486 aa in length). It was identified in a type II provirus encoding both dUTPase and gag. This was the only example of a provirus that encoded dUTPase and *gag* in the same frame.

The most complete provirus among the U1 lineages was a type I provirus identified on chromosome X (41,445,484-41,445,891) (Figure 5-14). It encodes 468 aa *gag*, 173 aa *pro*, 744 aa *pol* and a 225 aa *env*. However, all proviral genes had at least one in-frame stop codon and/or frame-shift.

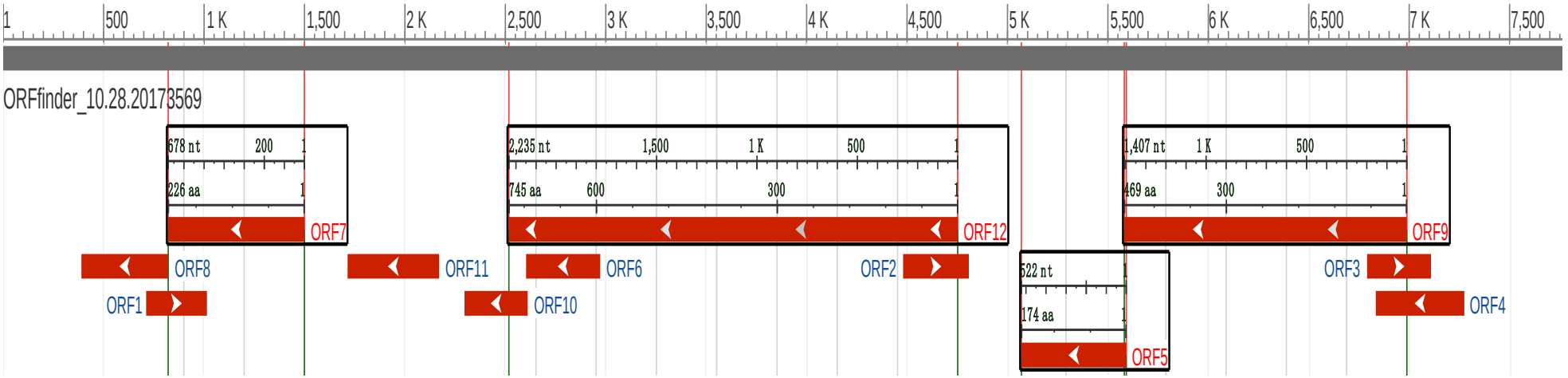


Figure 5-14 Detection of ORFs on chromosome X: 41,445,484-41,445,891. Detection of ORFs was performed by ORFfinder on NCBI website. Gag, pro and pol ORFs are shown as red, blue and purple. Potential ORFs are shown as red frame and strand is shown by the white arrow.

Recent activity of the U1 lineages

The existence of orthologous loci indicated that the date of U1 integration was not earlier than 54 Mya but also not later than 4.5 Mya. 18 pairs of flanking LTRs had been checked, and all LTRs found on the chromosome "unknown" were not included (Table 5-5). The most divergent paired LTRs were dated to 15.23 Mya. One pair of identified LTRs was observed at chromosome 1 and dated to recent (0 Mya). Only two loci were estimated to be less than 5 Mya (0 and 2.5 Mya). These two loci were, therefore, more likely to integration into the horse genome after the divergence of horse and donkey. The 1kb flanking region of these two loci was extracted and BLASTed against the donkey genome. The empty integration sites were identified at the orthologous loci in the donkey genome. All other integrations happened between 5 Mya and 15 Mya. Thus, most of integration events of U1 occurred before the divergence of horse and donkey. Estimations of integration times for U1 based on LTRs also suggested recent activity (0 to 2.5 Mya) (Table 5.5).

It is interesting that proviruses with the rearranged type II proviruses were younger than type I proviruses. I annotated the integration age and genome structure onto a phylogenetic tree which was inferred based on the alignment of whole proviral sequences (Figure 5-15). The dUTPase-encoding regions were removed. Notably, the midpoint-rooted phylogeny showed that both type I and type II proviruses had the same origin. However, insertions with the more type I proviruses were found almost exclusively toward the mid-pointed root whereas type II proviruses clustered together in a single derived clade with robust bootstrap support. The density plot showed that most of type I proviruses appeared between 5 to 15 Mya. By contrast, type II proviruses were relatively young, with the majority arising within the last 10 Myr.

Table 5-5 Integration time of U1 proviruses using paired LTR dating

Label	Distance	Mya (Neutral)	Type
chr1_22575320_22575724	0	0.00	II
chr6_60578255_60578647	0.011	2.50	II
chr1_29482117_29482521	0.025	5.68	Neither
chr20_56190936_56191340	0.029	6.59	II
chr15_7981700_7981939	0.033	7.50	II
chr9_54342710_54343105	0.034	7.73	II
chr5_44868505_44868900	0.035	7.95	II
chrX_41445484_41445891	0.035	7.95	I
chr5_21514315_21514707	0.036	8.18	II
chr1_82569604_82570008	0.046	10.45	II
chr22_1026813_1026929	0.051	11.59	I
chr29_28825500_28825886	0.051	11.59	II
chr17_69127229_69127633	0.052	11.82	I
chr12_16237336_16237728	0.054	12.27	II
chr8_41666684_41666977	0.062	14.09	I
chrX_49906466_49906774	0.065	14.77	II
chr16_84176093_84176488	0.066	15.00	I
chr7_88135585_88135992	0.067	15.23	I

Label: chromosome_start_end; Distance: pair-wise maximum likelihood distance between 5' and 3' LTRs; MYA(Netural): million years ago estimated using neutral mutation rate

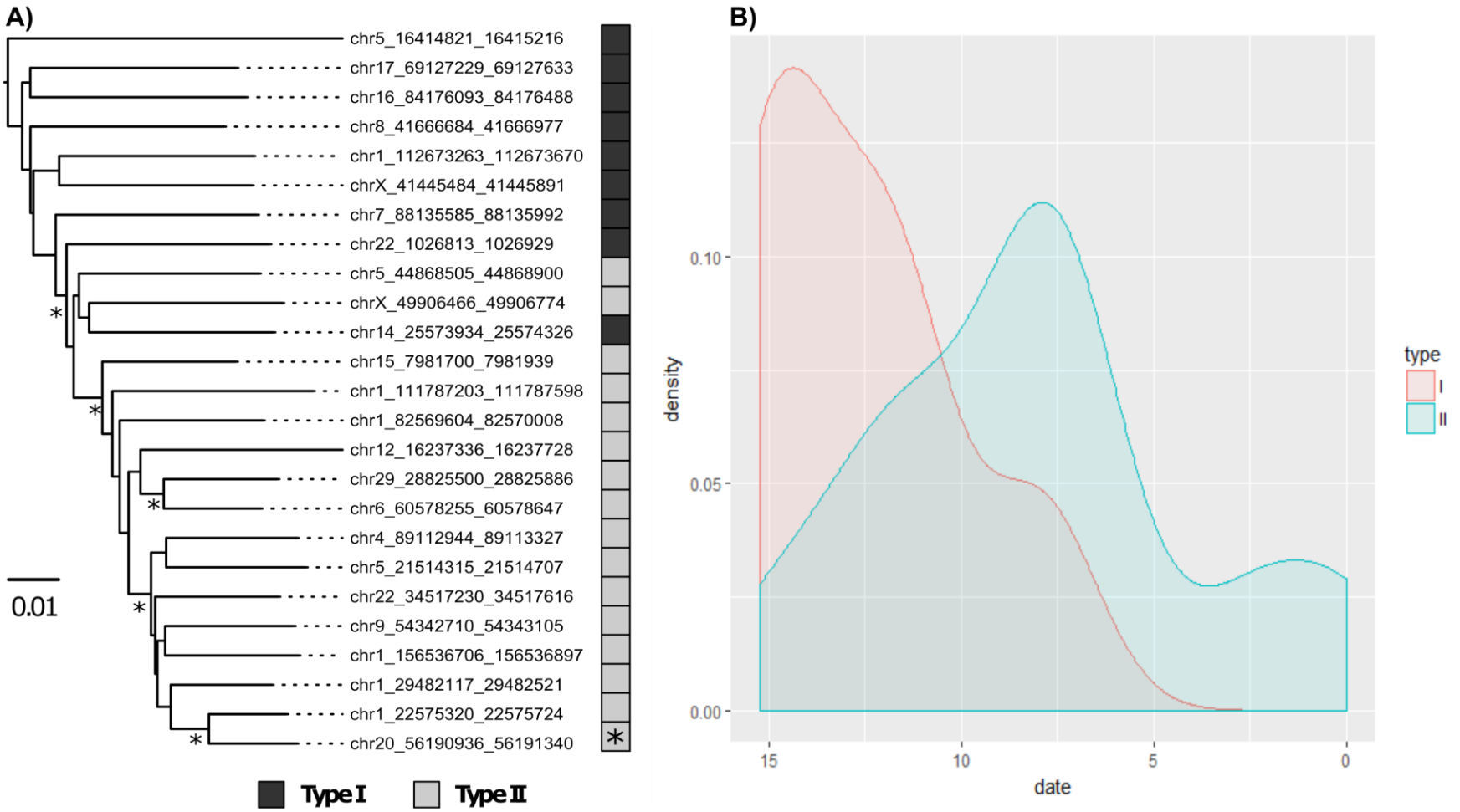


Figure 5-15 Phylogeny and density plot of full-length U1 proviruses. (A) Phylogenetic reconstruction of full-length U1 proviruses. Phylogenetic reconstruction was inferred by RAxML using multiple sequence alignment of full-length U1 proviruses. Asterisks marked branches that have bootstrap value over 90. Asterisk on sidebar shows the youngest provirus based on the paired LTR dating. Type I and Type II proviruses are marked by sidebar as black and grey, respectively; (B) Density plot for the distribution along the time scale. The x-axis shows time in millions of years before present, and the y-axis shows the density distribution. LTRs from the same ERV lineage are shown in the same plot with different colours. All X axes are adjusted to the same scale.

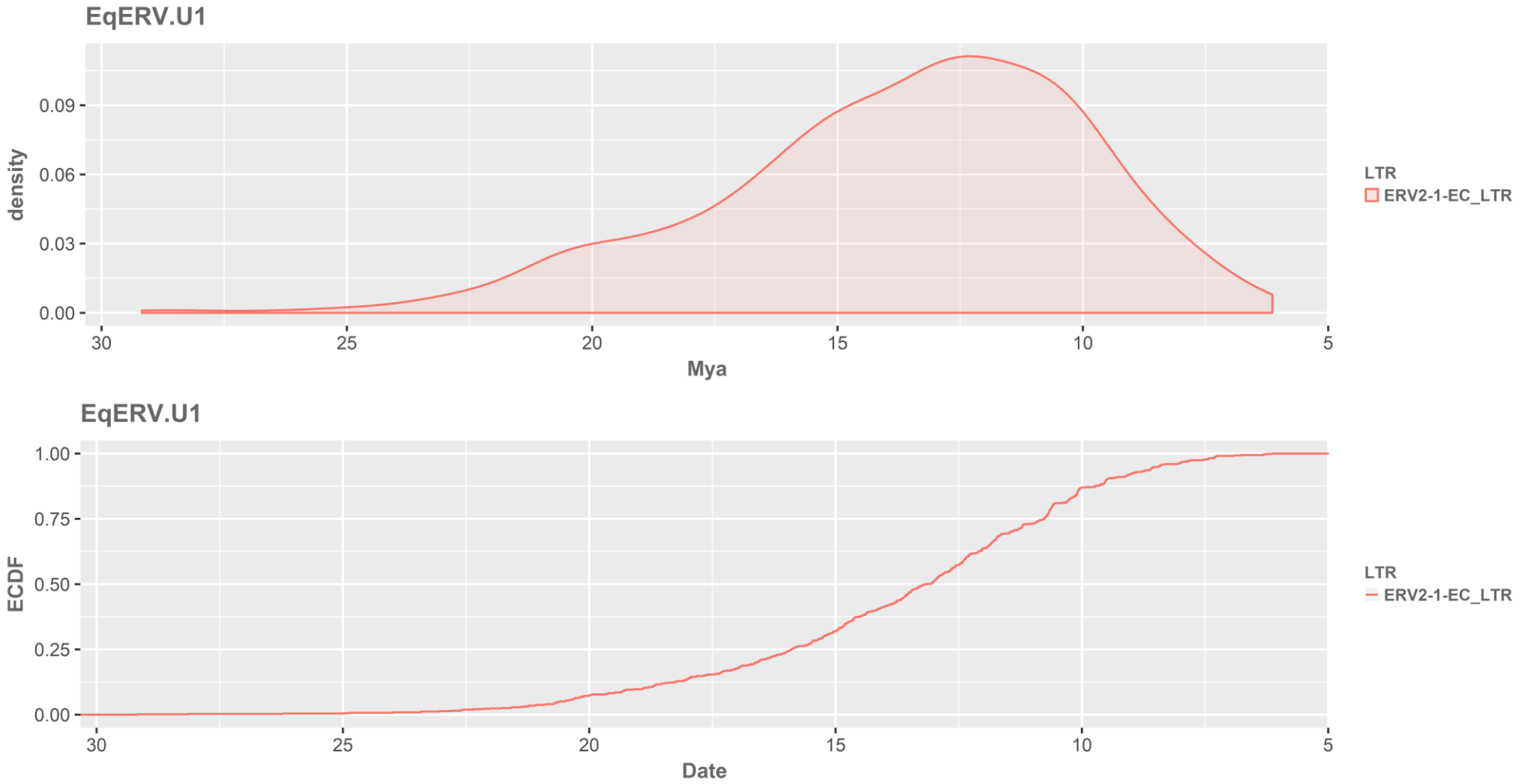


Figure 5-16 The ECDF plot of U1 solo LTRs. ECDF plots for the cumulative proportion of observed LTRs versus time scale. The x-axis shows time in millions of years before present, and the y-axis shows the cumulative proportion. LTRs from the same ERV lineage are shown in the same plot with different colours. All X axes are adjusted to the same scale.

The U1 lineage is the most abundant lineage among all modern ERV lineages. Based on the sequence similarity, all flanking LTRs found in proviruses were highly similar to 'ERV2-1-EC_LTR' in the Repbase. RepeatMasker further identified 669 solo LTR loci in the horse genome, giving the largest copy number for any modern perissodactyl ERV lineage.

Estimations of integration time based on the solo LTRs indicated that the majority of insertions happened no earlier than 29 Mya (Figure 5-16). The Density plot of solo LTR insertions showed a peak around 12 Mya, which suggested integration happened more frequently during this period. This was also the integration time of the majority of proviruses. The ECDF plot suggested that U1 began to accumulate in the horse genome with high speed since 25 Mya. Around 12 Mya, it had a sharp growth until 6 Mya. Compared to the early stage (12-25 Mya), the cumulative rate of insertions was much faster during the later stage (6-12 Mya).

Together, these data indicated that the germline invasion event that originally generated the U1 lineage happened somewhere between 25-30 Mya. The initial expansion of this lineage involved ERVs with type I genome structure. The copy number increased rapidly. Moreover, around 15 Mya, it reached the peak of growth. All identified proviruses were dated back to this period. Also, one copy underwent the genome rearrangements that generated a novel (type II) genome structure, and this element gave rise to a lineage that has been expanding up until relatively recently.

Transcriptome of U1 loci

Molecular dating results suggested that U1 was active until relatively recently. Also, nearly intact proviruses were observed in the horse genome. Thus, it is feasible to think that the U1 is transcriptionally active. To check this, I first examined the transcriptome of E.derms, an equine-derived cell line. Only provirus and solo LTR loci that had an expression level about fragments per kilobase of transcript per millions mapped reads (FPKM) were taken into account. The provirus on chromosome 29:28,825,500~28,825,886(-) was found to have low expression values but reads were able to cover the whole proviral locus, suggesting that U1 is actively transcribed in the E.derm cell line (Figure 5-17).

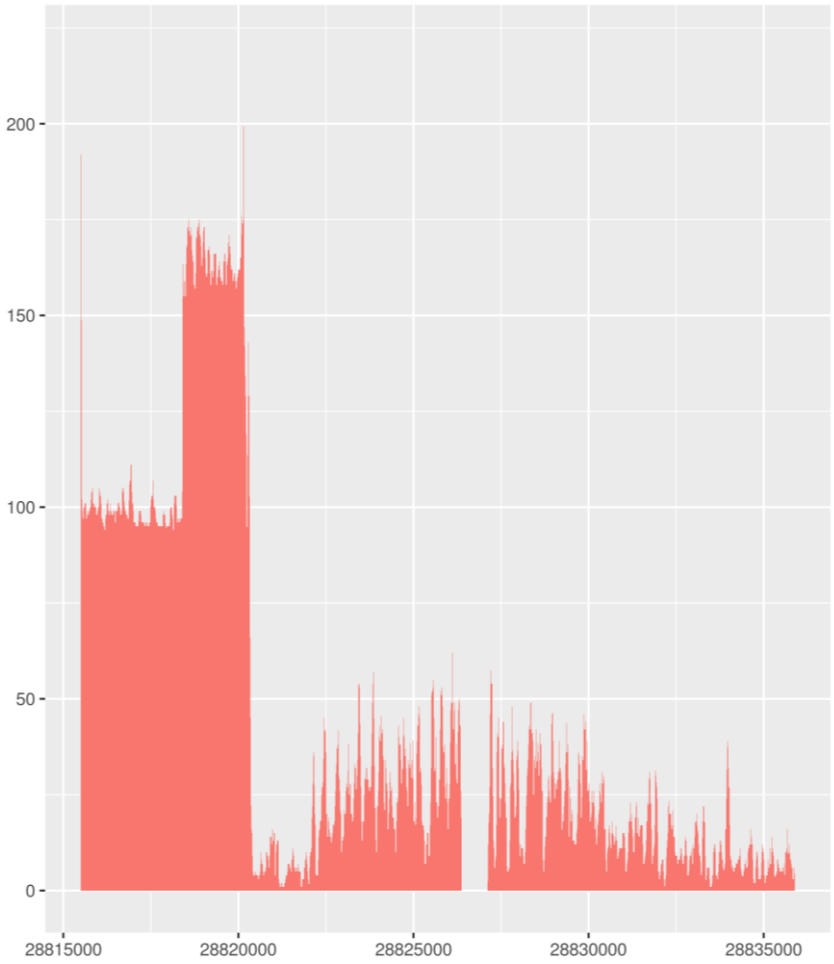


Figure 5-17 Read coverage plot of ERV locus in the E.Derm cell line. The x-axis shows coordinates of chromosome 29 (28,825,500~28,825,886(-)) of the horse genome, and the y-axis shows the read coverage of E.Derm cell line transcriptome dataset.

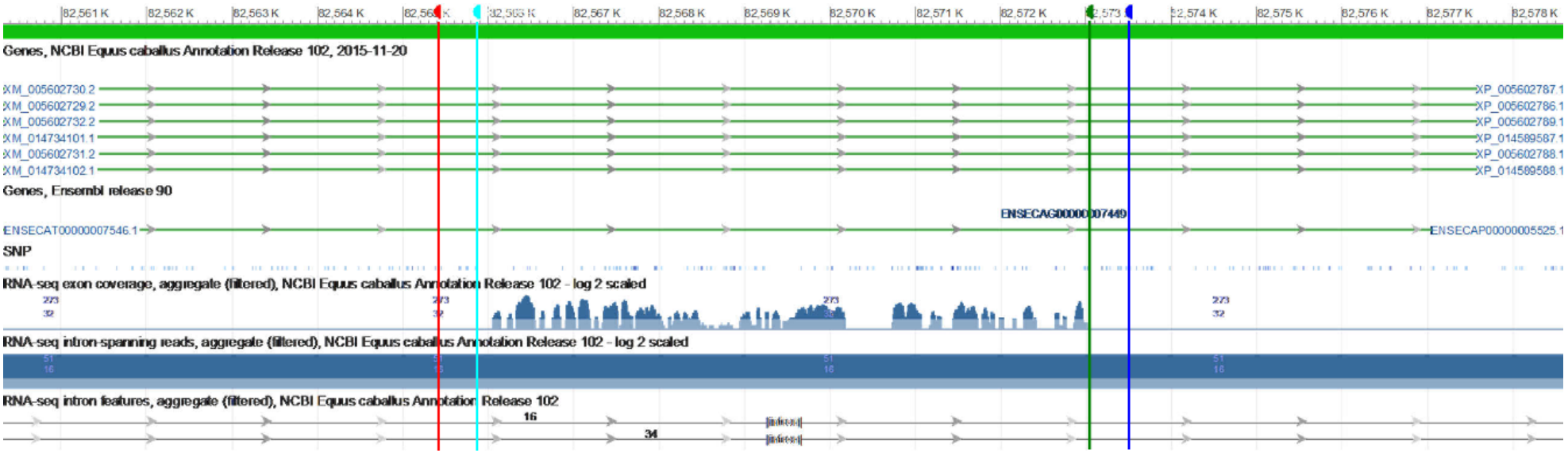


Figure 5-18 Genomic regions, transcripts of PTPN20 and U1 provirus. The figure is automatically generated by NCBI Graphics View. Colour lines are used to show borders of flanking LTR regions. Red and cyan-blue lines flank the 5'LTR, green and blue lines flank 3'LTR. Cyan-blue and green lines flank the internal coding region of the provirus. Blue peaks shown in the RNA-seq exon coverage section represent the exon coverage of RNA-seq alignments, the coverage values are scaled with a log2 scaled transform.

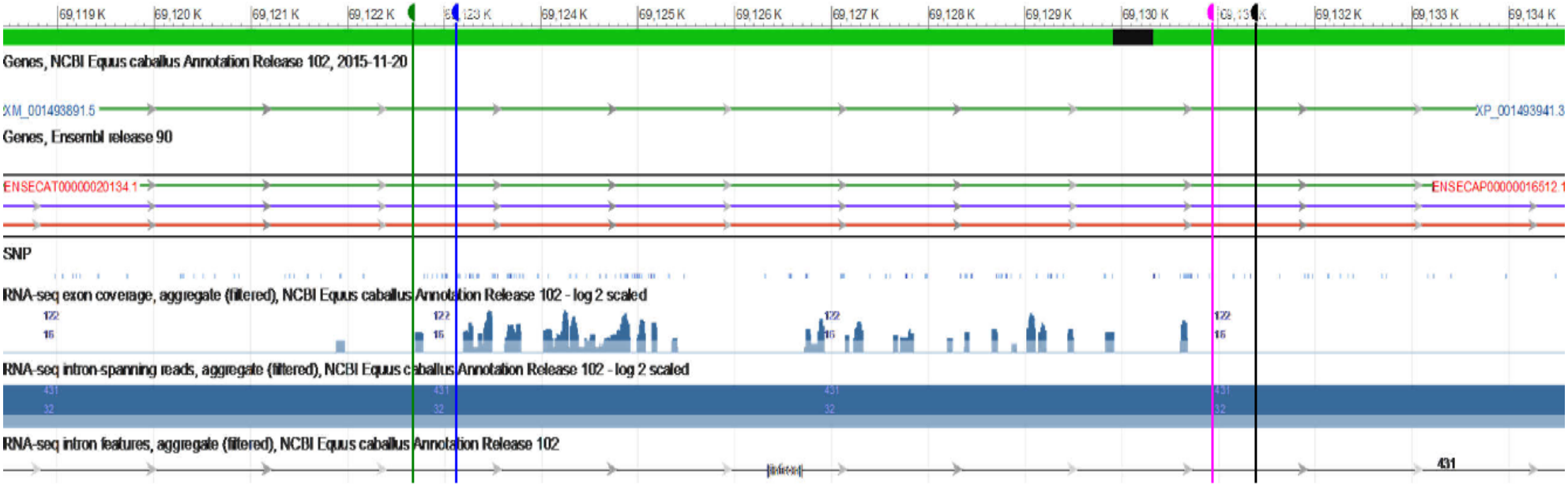


Figure 5-19 Genomic regions, transcripts of PCCA and U1 provirus. The figure is automatically generated by NCBI Graphics View. Colour lines are used to show borders of flanking LTR regions. Green and blue lines flank the 5'LTR, Pink and black lines flank 3'LTR. Blue and pink lines flank the internal coding region of the provirus. Blue peaks shown in the RNA-seq exon coverage section represent the exon coverage of RNA-seq alignments, the coverage values are scaled with a log2 scaled transform.

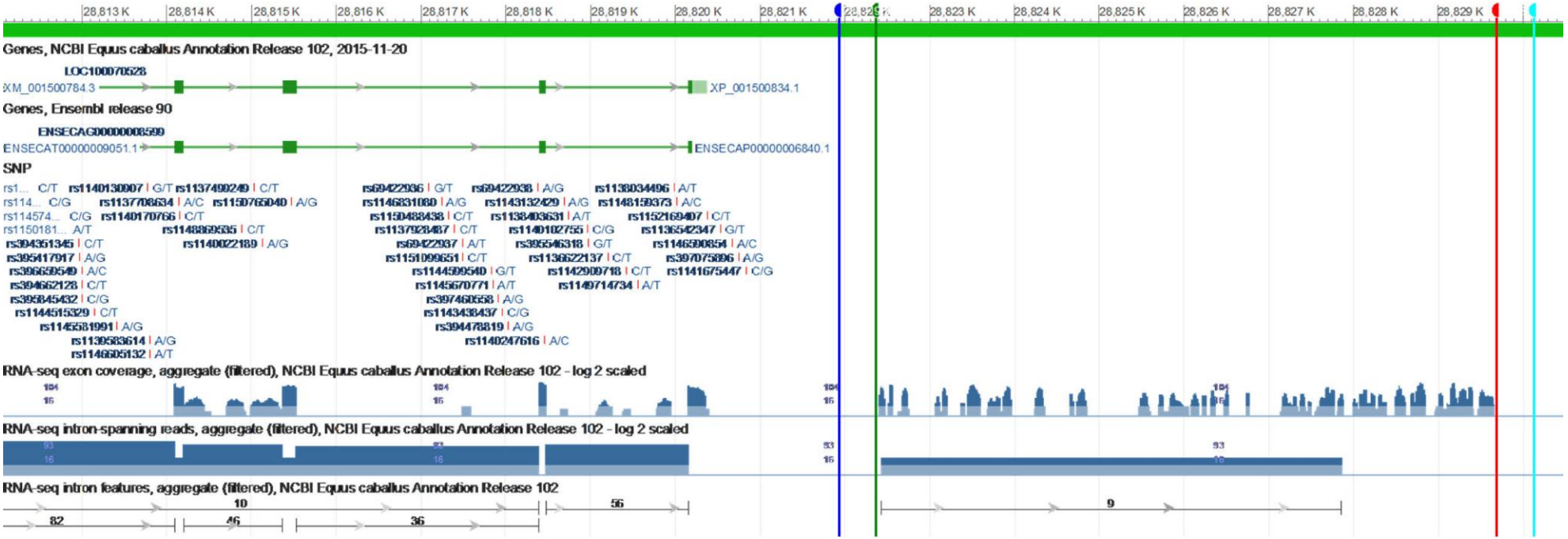


Figure 5-20 Genomic regions, transcripts of AK1CO and U1 provirus. The figure is automatically generated by NCBI Graphics View. Colour lines are used to show borders of flanking LTR regions. Blue and green lines flank the 5'LTR, red and cyan-blue lines flank 3'LTR. Green and red lines flank the internal coding region of the provirus. Blue peaks shown in the RNA-seq exon coverage section represent the exon coverage of RNA-seq alignments, the coverage values are scaled with a log2 scaled transform.

This result urged me to examine the transcriptome of horse tissues. A public transcriptomic dataset of 17 equine tissues was investigated. Approximately 4551 million reads were obtained from the ENA database, which was then mapped to the equine reference. Mapping to ensemble and ERV annotation results in 80.91% (~ 3683 million reads) of reads assigned to genes and ERV loci. Of all 885 of solo LTRs and provirus loci, 182 ERV have expression level over 1 FPKM. 21 of 182 loci were identified as U1. Six and 14 loci are type I and type II, and one locus Is undetermined type. Of these 21 U1 loci, nine proviruses are almost fully covered, which suggested all U1 genes were transcribed.

Table 5-6 Expressions of U1 in horse tissues

Tissues	Type I	Type II
Bone Marrow	-	-
Brain	-	-
Brain Stem	+	-
Donkey placental	-	-
E.derm	-	+
Hinny placental	-	-
Horse placental	-	-
Inner Cell Mass	-	-
Kidney	-	-
Lamellar	-	-
Skin	-	+
Liver	-	-
Mute placental	-	-
Oviduct	+	-
Peripheral blood mononuclear cell	-	-
Spinal Cord	+	-
Trophectoderm	+	+
Uterus	-	-

Among nine provirus loci covering by reads, two loci located within gene intron. The first provirus located in the intron 3-4 of the propionyl-CoA carboxylase alpha subunit (PCCA) on the reverse strand of chromosome 17 (69,127,229-69,127,633) (Figure 5-18). Another provirus located in the intron 2-3 of protein tyrosine phosphatase, non-receptor type 20 (PTPN20) on the reverse strand of chromosome 1 (Figure 5-19). Both of PCCA and PTPN20 were on the forward strand. Reads also covered the provirus found on chromosome 29 (28,825,500~28,825,886). This

provirus was found at the 1575 bp downstream of aldo-keto reductase family one member C23-like protein (AK1CO) (Figure 5-20). Notably, the provirus on chromosome 29 located at the same strand of AK1CO. The coverage plot (Figure 5-18) indicated that reads completely cover the junction between AK1CO gene and downstream U1 locus, which suggests that downstream U1 locus may transcribe associated with AK1CO gene.

Of these tissues examined above, transcripts related to Type I proviruses were found in the brainstem, spinal cord and oviduct, whereas E.derms and skin only expressed type II proviruses (Table 5-6). Trophectoderm has both kinds of type I and type II provirus transcripts. In E.derms, only one completed U1 locus on chromosome 29 was transcribed.

5.5 Discussion

5.5.1 The evolutionary history of perissodactyl ERVs

In this chapter, I investigated the activity of distinct perissodactyl ERV lineages in the horse. For each of the nine major lineages, I determined the minimum age and inferred the overall retrotranspositional activity over time.

Perissodactyl ERV activity before 54 Mya

In the period before 54 Mya, only ancestral lineages were active. Solo LTR dating suggested that the initial invasions of Rho, Theta, Sigma and Lambda began before the divergence of major perissodactyl groups. However, it seems that species living in this period have just begun accumulating ERV insertions in their genomes, which suggest the virus expansion in the host genome still in the early stage.

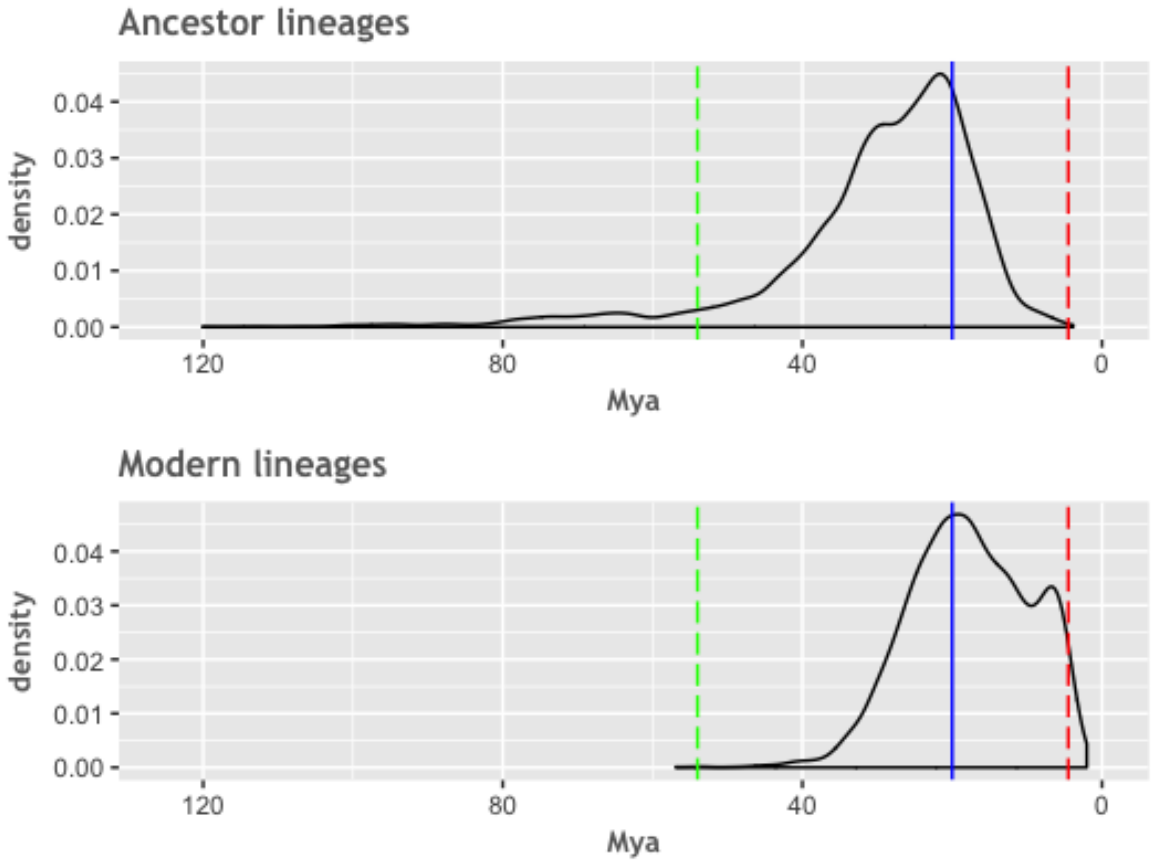


Figure 5-21 Density plot for the distribution along the time scale. (Upper) All ancestor lineages; (Lower) All modern lineages. The x-axis shows time in millions of years before present, and the y-axis shows the density distribution. X-axes are adjusted to the same scale. Speciation of perissodactyls (54 Mya), Equus genus (4.5 Mya are marked by green and red dash lines, respectively. 20 Mya is marked by a blue line.

Between 20 Mya and 54 Mya

From 54 Mya to 20 Mya (from early Eocene to the middle Miocene) early equids diverged from the other perissodactyl species. Early in this period (i.e. from 54 Mya to 40 Mya), several 'new' ERV lineages were established, leaving fixed insertions. At a later stage (from 40 Mya to 20 Mya), modern lineages 'Zeta' began invading the genomes of *equids*, but the majority of activity involved ancestral lineages. The copy number of both ancestral and modern ERVs was increasing rapidly during this period.

From 20 Mya to present

From 20 Mya until the present day (whole Miocene and Pliocene), early equids evolved into modern species with major adaptations to new habitat and climate. In this period, the activity of most ancestral lineages had ceased. Nevertheless, several new Rho and Theta sublineages were established in the host genome. Furthermore, all clade II lineages were established in this period. The invasion began around 25 Mya and reached the peak at approximately 10 Mya. At the time of speciation of *Equus* genus, activities of ancestral lineages had become very subdued. Almost no novel ancestral insertions were generated and/or fixed in the host germline, and some sublineages had stopped expansion over several million years. All ancestral insertions had accumulated multiple mutations, and none of their ORFs remained intact. However, many modern lineages still were activating at a high level. A large number of insertions of modern lineages were established in the host genome during this period, and most of them still kept the full-length genome structure or even long ORFs.

5.5.2 Only modern lineages were active until recent

Thus, in my study, I have analysed transcriptome of 17 horse tissue and E.derm cell line. Based on these data, several ERV loci with full-length proviruses were found to be transcribed in different biological condition. With relatively intact proviral genome structure and ORFs, these loci were more likely to have function and possible to co-opted with host genome. Similar results were also reported by multiple previous studies (Brown *et al.*, 2012; Moreton *et al.*, 2014; Stefanetti *et al.*, 2016; Gim and Kim, 2017). Furthermore, transcriptomic analysis suggests the

current activities of the ERVs in the genome. Together with classification of ERVs and generation of evolutionary timescale, a comprehensive description of ERV current activities and of that in the past were described.

Only modern lineages can be dated to recent times

Compared to the ancestor lineages, modern ERV lineages had more recently integrated elements. Some proviruses of modern ERVs in the horse genome were estimated to be no more than 2-3 Myr old. Consistent with this, the donkey genome lacked the corresponding insertions. More importantly, these recently integrated elements retained a relatively complete proviral structure. Some of them still had *env* genes (a characteristic of younger ERV lineages). Although most of *env* genes are presumed to be non-functional due to mutations and in-frame stop codons, the existence of relatively intact envelopes in many modern proviruses suggested that their recent expansion has been driven by reinfection.

Only modern lineages are transcribed

I identified transcripts of U1 proviruses in multiple horse tissues (Table 5-6). Read coverage spanned the complete proviral genome of U1. I also found that some loci had higher coverage and depth than other loci. Provirus loci with higher coverage and depth were more likely to be the genuine source of transcripts. In this case, three U1 loci were fully covered by reads.

These loci seemed to have tissue-specific expression. For example, a provirus on chromosome 29 only has an expression in the E.Derm cell line. Another interesting feature of U1 expression is that expressed proviruses are located within or near genes. Those that are within genes are located on introns in the antisense orientation. As transcript annotations are usually predictions *in silico* which may not reflect the real-life situation accurately, the locations of proviruses and genes cannot be used as crucial evidence to draw the conclusion that the nearby gene triggers transcripts of proviruses.

However, there is insufficient evidence to show how these proviruses transcribed. Since all identified proviral genes were interrupted by mutations and stop codons, none can express intact proteins. Thus, it appears unlikely that any of the lineages

described which also suggests that these proviruses are not able to generate virus particle and reinfect other cells. Furthermore, none of the U1 insertions described here possesses an intact *pol* or *env*, so it also seems unlikely that *trans*-complementation between distinct loci could lead to the generation of infectious particles. However, recent studies in humans have shown that polymorphic and intact ERVs may be present at a low level in the population – thus, it remains possible that the U1 lineage is active in a horse population somewhere.

5.5.3 Mode of copy number expansion

The lack of equine ERVs encoding intact *env* genes suggests that most have undergone recent expansion through mechanisms other than reinfection. One potential mechanism of copy number increase for proviruses that lack envelope is intracellular retrotransposition, wherein ERVs replicate without leaving the cell. I identified both modern and ancestral proviruses that had to flank LTRs and relatively complete *gag* and *pol* genes, but lack *env* genes or only had a truncated remnant of the *env* gene. This finding demonstrates that intracellular retrotransposition has been important in the evolution of equine ERVs.

However, for the majority of ancestral lineages, especially Lambda lineage, I also observed a large number of provirus loci that were flanked or adjacent to the LINE1 elements. One possible exploitation is that these ERV loci were acquiesced by the LINE1-mediated formation of processed pseudogenes. Such mechanism was observed from many HERV-W loci in the human genome (Pavlicek *et al.*, 2002; Pavlícek *et al.*, 2002). Interestingly, most of these loci lacked paired LTRs but still contained one or two genes, gag and pol. Moreover, all of them did not have env genes. These proviruses were more likely to be replicated together with LINE1 elements as non LTR-retrotransposons. However, as most of such loci were highly degraded, it was a challenge to obtain the exact range of these loci. Thus it was hard to align these loci and identify their polyadenylation signal (AATAAA) and poly-A tail.

5.5.4 Limits of the different dating method

I identified several ERV loci that were orthologous across several species, providing robust minimum age for particular perissodactyl ERV lineages. While this

approach provides a very robust minimum age, it has some limitations. Firstly, orthologous loci can only provide a minimum age - the real age may be much greater. Secondly, dating on this basis requires that the divergence times of host species are well-established, and in some cases, they are not (uncertainty can be in the range of several million years).

Dating methods based on sequence divergence also have limitations. Firstly, sequences may not evolve in a clock-like manner. Furthermore, poorly understood processes such as gene conversion may produce artefactual results. Furthermore, even assuming that sequences evolve in a clock-like way, date estimates rely on an accurate rate estimate. This is difficult as mutation rates may vary across genomic loci, - for example, some functional LTRs and proviruses may be evolving under negative (purifying) selection. When the molecular clock is used to date solo LTRs, estimation of the ancestral sequence will exert an influence on dating. Thus, accurate reconstruction is vital, yet this is hard to verify.

Another issue is the detection of solo LTRs. The *de novo* detection can identify all repeats from the genome. However, the confirmation that particular solo LTR sequences were associated with ERVs required the knowledge of references or internal coding regions. Thus, many genomic repeats are suspected to be LTRs, but they cannot be confirmed as being retroviral LTRs.

5.6 Conclusion

In this chapter, I investigated the retrotranspositional activity of equine ERV lineages during the evolutionary history of ERVs in the horse genome. Ancestral ERV lineages (i.e. those that invaded the perissodactyl germline prior to the divergence of the *Hippomorpha* and *Ceratomorpha*) were actively expanding in the period from 54-20 Mya. The activity of 'modern' ERV lineages overlapped that of ancestral ERV lineages to a large extent. However, these lineages were active up until more recently, including after the divergence of donkeys and horses. By contrast, no ancestral ERV lineage appears to have generated novel insertions after this point. Transcriptomic analysis indicated that some loci within one modern lineage are transcribed, potentially in a tissue-specific manner.

6 Discussion

In this PhD project, I developed a novel pipeline for ERV annotation that integrates a 'phylogenetic screening' approach to ERV characterisation with other software tools for ERV annotation. I then used this pipeline to characterise ERVs in the *E.caballus* genome and those of two other perissodactyls: the donkey (*Equus asinus*) and white rhinoceros (*Ceratotherium simum*). Through comparative analysis of these three genomes, I derived a calibrated timeline describing the process through which ERV diversity has been generated in the equine germline. I provide an overview of retrotranspositional activity among distinct perissodactyl ERV lineages and identify individual ERV loci that show evidence of involvement in physiological processes and/or pathological conditions.

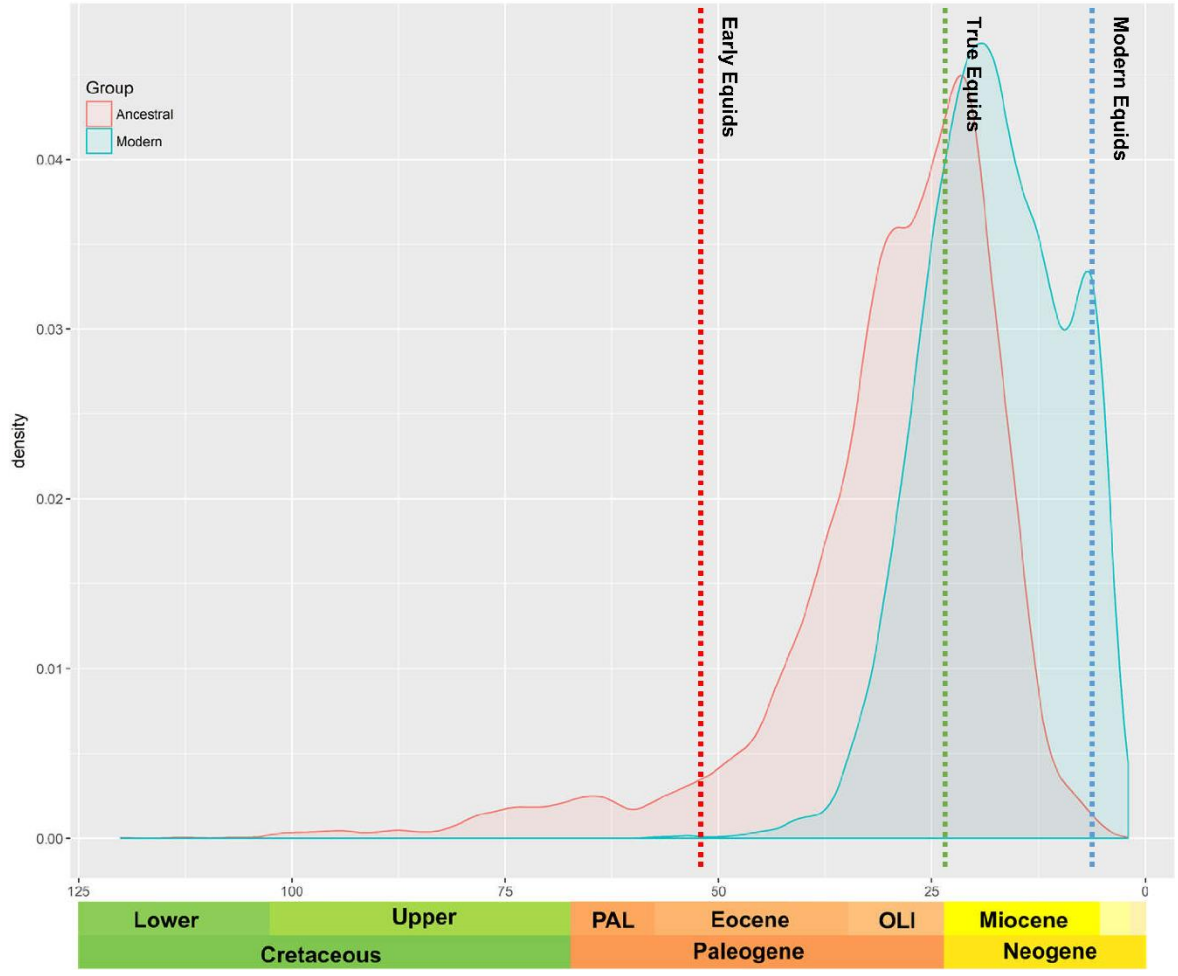


Figure 6-1 Co-evolution of perissodactyl ERVs and equids. Density plots are showing copy number of ancestral and modern ERV lineages around the time axis (X-axis**).** The evolution of equids diverged into three periods: early equids (23~54 Mya, red line), true equids (5~23 Mya, green line) and modern equids (present~5 Mya, blue line). The geologic timescale was shown under the time axis.

6.1 ERVAP – a novel pipeline for characterising ERVs

The ‘phylogenetic screening’ approach to describing ERV diversity was first applied to human ERVs (Tristem, 2000). The power of this approach is the importance that it places on establishing the evolutionary relationships between different ERV lineages. Once these have been resolved to some degree, it becomes easier to interpret the genomic diversity of ERVs, as this can be placed in context concerning the process that generated it.

In early studies, phylogenetic screening was performed manually (Tristem, 2000). In this project, the DIGS tool was used to provide a mechanism for performing phylogenetic screening in a semi-automated, relatively high-throughput way. This approach is also relatively efficient, as it directs attention toward loci that are highly likely to be retroviral. It is, therefore, less computationally intensive than scanning entire genome assemblies in a more inclusive, but naïve way.

Moreover, I created the ERVAP pipeline, which integrates a DIGS-based phylogenetic screening approach with other tools for ERV annotation. ERVAP provides automatic annotation functions that allow RT loci and lineages identified by RT-based phylogenetic screening to be characterised in greater depth. These include tools that use hidden Markov models (HMMS) to detect retroviral protein domains, regardless of whether these occur in full-length proviruses with LTRs, or in fragmented retrovirus genomes. This approach has fulfilled a gap of other ERV detection programs. Current detection programs initiate the screening for detection of paired LTRs. When LTRs are not within the expected size-range or are highly degenerated, some ERV loci will be missed. ERVAP avoids this limitation. The information recovered by ERVAP not only benefits the study of ERV classification based on reference retroviral elements and understanding of ERV characterisation but also find the missing elements which can aid these analyses.

In sum, ERVAP is a pipeline that is designed specifically for evolutionary analysis. The annotation and extracted sequences generated by ERVAP are highly valuable for ERV classification or investigation of ERV evolution.

6.2 Characterisation of nine distinct perissodactyl ERVs using ERVAP

I used the ERVAP pipeline to investigate ERV diversity in 17 perissodactyl genomes. A total of 18,290 RT loci were identified. At least nine major ERV lineages were detected, and their relationships to other known ERVs and retroviruses was reconstructed. Interestingly, comparison of the diversity of ERVs in the perissodactyl species suggested that gammaretroviruses and epsilonretroviruses are absent in all perissodactyls, and clade II ERVs are only present in equids, being completely absent from rhinoceroses.

Next, I characterised the genome structure for each identified perissodactyl ERV lineage. The ERVAP pipeline was used to investigate the genomic regions flanking each RT locus identified by DIGS. Representative genome structures and consensus sequences were generated based on the recovered proviral sequences of each major ERV lineage.

Some retroviruses encode auxiliary or "accessory" genes in addition to the standard gag, pol, and env coding domains. ERVAP did not detect the presence of accessory genes in most of the consensus genome structures recovered here. This included the Beta1 lineage - which is closely related to MMTV. Whereas MMTV encodes a sag gene in the LTR, there was no evidence for a related gene being present in the Beta1 lineage. However, the Kappa1 lineage apparently encodes a homolog of rec, a trans-activating regulator of transcription.

In general, the ERV landscape of the horse genome resembles that of other large-bodied *Boreoeutherian* mammals (e.g. hominids, cetaceans and bovids). In all of these groups, studies have reported a relatively low number of intact ERVs, and furthermore, most ERVs are derived from groups that have no closely-related exogenous counterparts. By contrast, the genomes of many small-bodied mammal species (e.g. rodents, bats) harbour large numbers of relatively intact ERVs that group closely with exogenous Gamma- and Betaretroviruses in phylogenetic trees.

Strikingly, perissodactyl genomes exhibit a total absence of ERVs grouping within the *Gammaretrovirus* genus (as this genus is defined by exogenous isolates). In addition, the rhinoceros genome exhibits a total absence of clade II

(Betaretrovirus-related) ERVs, despite these being present in the genome of almost every other mammal species. At present, I can only speculate as to the underlying causes of these observations. However, it is clear from work in other systems that mammals harbour numerous genes that function specifically in antiviral defence against retroviruses. For example, some proteins encoded by APOBEC3 family genes are potent inhibitors of retroviruses. Interestingly, these genes are expanded in the horse genome (Bogerd et al., 2008; Zielonka et al., 2009).

6.3 Inferences about ancient retroviruses

The comparative analysis also reveals much about the history of exogenous retroviruses. To begin with, the ancient retroviruses that gave rise to clade II ERVs in equid genomes were circulating in ancestral mammals at the very beginning of the Miocene epoch (~23 Mya). These include the Kappa.1 and Kappa.2 lineages, which are closely related to the HERV.K supergroup found in primates (Figure 4-4). Consistent with the idea that these ERV lineages derive from infectious retroviruses that circulated in the some of these primate ERV lineages seem to have been established by distinct germline colonisation events that occurred in approximately the same geological time period (Hohn, Hanke and Bannert, 2013). Therefore, it seems that these viruses circulated during the Aquitanian stage of the early Miocene (20-23 Mya). In addition, the "B-type" lineage of betaretroviruses, which includes mouse mammary tumour virus (MMTV), as well as related ERVs in bats and cattle, apparently entered the equid germline around this time. This is the oldest age estimate yet obtained for a betaretrovirus in the B-type lineage, and also establishes that long LTR sequences associated with these viruses (Hayward, Grabherr and Jern, 2013), have been a defining characteristic for at least this long.

A recent study showed that the ancient clade I retrovirus that generated ERV.Fc lineages in diverse mammals also circulated during the early Miocene epoch (Diehl *et al.*, 2016). While there is no ERV.Fc lineage in the perissodactyl germline, there is a closely related lineage – Zeta. This lineage, which is closely related to the HERV.W and ERV.9 lineages in primates, entered the perissodactyl germline prior to the *Ceratomorpha-Hipporpha* divergence but carried on expanding long after (Figure 5-7). Interestingly, the expansion of this lineage in horses seems to mirror that of the HERV.W lineage in primates (Grandi *et al.*, 2018). The oldest perissodactyl ERV lineages – including Rho, Theta, and Sigma – presumably derive from ancient viruses that circulated over 54 Mya (and potentially much earlier than this).

6.4 Timeline of ERV activity in the horse

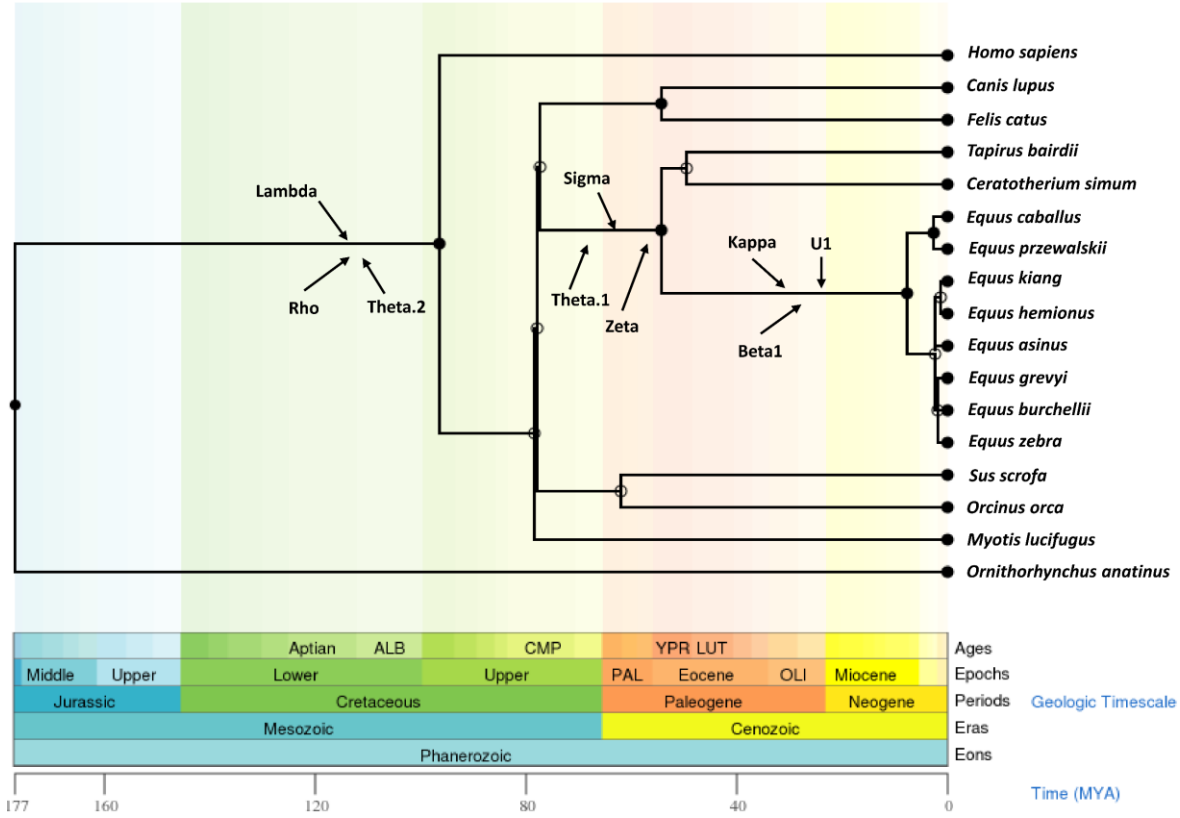


Figure 6-2 Summary of nine major germ-line invasion on taxonomy tree. The topology of timetree was obtained from the TimeTree resource (Kumar *et al.*, 2017). It is summarised based on the published studies. Arrows with labels represent the estimated initial germ-line invasion of each major lineage.

Data recovered using ERVAP was used to infer a calibrated timeline of activity for perissodactyl ERVs in the horse germline. I estimated the integration time of ERV loci based on orthology and on molecular clock-based analysis of paired LTRs and solo LTRs. This revealed that ancestral ERV lineages (i.e. those that invaded the perissodactyl germline prior to the divergence of the *Hippomorpha* and *Ceratomorpha*) were actively expanding in the period from 54-20 Mya. The activity of 'modern' ERV lineages overlapped that of ancestral ERV lineages to a large extent. However, these lineages were active up until more recently, including after the divergence of donkeys and horses. By contrast, no ancestral ERV lineage appears to have generated novel insertions after this point.

I investigated the transcriptional activity of equine ERV lineages, revealing that one modern lineage (U1) is actively transcribed, potentially in a tissue-specific

manner. I did not identify any proviral loci within this lineage that were replication competent regarding encoding intact genes. Furthermore, although some loci have nearly intact ORFs, the U1 provirus population examined here did contain within it the capacity to express the full set of retroviral proteins required to produce an infectious viral particle. It remains possible, however, that more intact proviruses are present within the horse population, but as polymorphic alleles present only at a low frequency (Subramanian *et al.*, 2011).

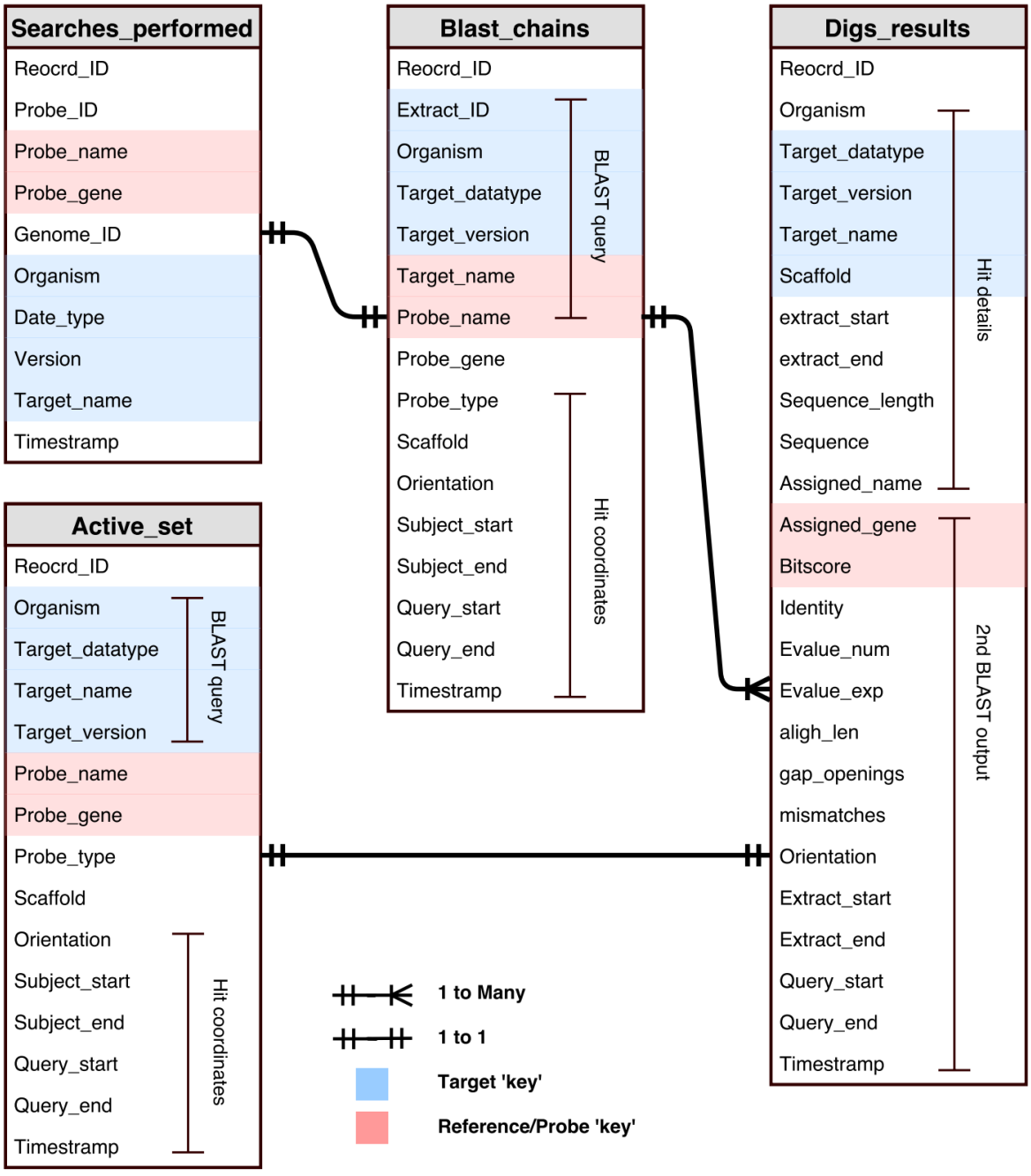
Alternatively, the detection of actively described loci within the U1 lineage might reflect the co-option or exaptation of these loci to perform physiological functions. For example, studies in humans and mice have shown that ERVs have important roles regulating gene expression, particularly during early development (Mi *et al.*, 2000; Dupressoir *et al.*, 2009). In theory, the dramatic expansion during the Miocene (15-20 Mya) of certain modern ERV lineages in equid genomes, could be associated with the evolution of physiological adaptations that occurred as these species shifted from being small forest-dwelling animals feeding on leafy vegetation into larger-bodied herbivores adapted for life in open grassland (MacFadden, 2005). The dataset generated in this project will be of great utility to future studies aiming to investigate the potential functional roles of equine ERVs and their impact on equine evolution.

Appendix I

Taxa Labels	Virus Name	Tree name
Mammalia_Mus.Gamma.MDEV_RT	Mus_dunni-ERV	MDEV
Actinopterygii_GaERV.1.Ocean_RT	GaERV-1	GaERV.1
Actinopterygii_Sander.Epsilon.WDSV_RT	WDSV	WDSV
Actinopterygii_Sander.Epsilon.WEHV_RT	WEHV-1	WEHV.1
Actinopterygii_SnRV.Omega_RT	SnRV	SnRV
Actinopterygii_SSSV.Xi_RT	SSSV	SSSV
Actinopterygii_ZFERV.Xi_RT	ZFERV	ZFERV
Aves_Anas.Rho_RT	Anas-R	Anas.R
Aves_Charadrius.Rho_RT	Melopsittacus-H	Melopsittacus.H
Aves_Falc.ch.Rho_RT	Falco-R	Falco.R
Aves_Gallus.ALV.Rous.Alpha_RT	RSV	RSV
Aves_Gallus.Beta.B1(29)_RT	GGERV-29	GGERV.29
Aves_Gallus.Chi2_RT	GGERV-LA	GGERV.LA
Aves_Gallus.ChiRV1.Iota_RT	ChiRV1	ChiRV1
Aves_Gallus.Gamma.REV_RT	REV	REV
Aves_LPDV.Beta.b3_RT	LPDV	LPDV
Aves_Meleagris.ALV.Alpha_RT	MgALV	MgALV
Aves_Melopsittacus.Iota.H_RT	Melopsittacus-H	Melopsittacus.ERV
Aves_Pseudopodoces.Iota.H_RT	Pseudopodoces-H	Pseudopodoces.ERV
Aves_Taenopygia.F.Nu_RT	Taeniopygia-F	TgERV.F
Aves_Taenopygia.Rho_RT	Taeniopygia-R	Taeniopygia.R
Latimeria_CoeFV.Spuma_RT	CoeEFV	CoeEFV
Mammalia_Beta.d1.SMRVH_RT	SMRV-H	SMRV.H
Mammalia_Beta.d1.SRV1_RT	SRV-1	SRV.1
Mammalia_Beta.k1.BERV_RT	BERV-beta3	BERV.beta3
Mammalia_BFV.Spuma_RT	BFV	BFV
Mammalia_BIV.Lenti_RT	BIV	BIV
Mammalia_EFV.Spuma_RT	EFV	EFV
Mammalia_EIAV.Lenti_RT	EIAV-Am	EIAV
Mammalia_EqERV.Beta.b1_RT	EqERV-b1	EqERV.b1
Mammalia_Felis.Gamma.RD114_RT	RD114	RD114
Mammalia_FFV.Spuma_RT	FFV	FFV
Mammalia_Gamma.CERV1_RT	CERV-1	CERV.1
Mammalia_Gamma.CERV2_RT	CERV-2	CERV.2
Mammalia_Gamma.KoRV_RT	KoRV	KoRV
Mammalia_Gamma.KwERV_RT	KwERV	KwERV
Mammalia_Gamma.PERV.A_RT	PERV-A	PERV.A
Mammalia_Gamma.PERV.C_RT	PERV-C	PERV.C

Mammalia_HERV.E.Eta_RT	HERV-E	HERV.E
Mammalia_HERV.Fc.Zeta.Fc_RT	ERV-Fc	ERV.Fc
Mammalia_HERV.H.Zeta.I_RT	HERV-H	HERV.H
Mammalia_HERV.I.Iota_RT	HERV-I	HERV.I
Mammalia_HERV.L.Lambda_RT	HERV-L	HERV.L
Mammalia_HERV.R.Eta_RT	HERV-R	HERV.R
Mammalia_HERV.W.Zeta.II_RT	HERV-W	HERV.W
Mammalia_HERVFb.Zeta.I_RT	HERV-Fb	HERV.Fb
Mammalia_HERVXA.Zeta.I_RT	HERV-XA	HERV.XA
Mammalia_HIV1B.Lenti_RT	HIV-1	HIV.1
Mammalia_Homo.ERV9.Zeta.II_RT	ERV-9	ERV.9
Mammalia_Homo.Gamma.HERV.T_RT	HERV-T	HERV.T
Mammalia_Homo.Sigma.HERV.S.con_RT	HERV-S	HERV.S
Mammalia_JDV.Lenti_RT	JDV	JDV
Mammalia_JSRV.Beta.d2_RT	JSRV	JSRV
Mammalia_Kappa.HERVK.HML2_RT	HERV-K-HML2	HERV.K.HML2
Mammalia_KERV.Beta.Australasia_RT	KERV	KERV
Mammalia_MMTV.Beta.b1_RT	MMTV	MMTV
Mammalia_Mus.Beta.a.IAPE_RT	IAPE	IAPE
Mammalia_Mus.Gamma.MoMLV_RT	MLV	MLV
Mammalia_Myotis.Gamma.RfRV_RT	RfRV	RfRV
Mammalia_Oryctolagus.Beta.Outlier.RERVH_RT	RERV-H	RERV.H
Mammalia_PERV.E.Eta_RT	PERV-E	PERV.E
Mammalia_PSIVgml.Lenti_RT	pSIVgml	pSIVgml
Mammalia_RELIK.con.Lenti_RT	RELIK	RELIK
Mammalia_SFVspid.Spuma_RT	SFVspi	SFVspi
Mammalia_SRLV.A.Lenti_RT	SRLV-A	SRLV.A
Reptilia_Beta.pyERVmol_RT	PyERV	PyERV

Appendix II



Entity-Relationship diagram of MySQL database generated by the DIGS tool.
For each DIGS screening project, the DIGS tool creates a new schema in the MySQL database. Each schema has four tables: BLAST_chains, Digs_results, Seaches_performed, and Active_set. Crossbars show the range of information section in the table; the relationship between each table are linked by relational arrows.

Bibliography

Aken, B. L. *et al.* (2016) 'The Ensembl gene annotation system', *Database*, 2016, p. baw093. doi: 10.1093/database/baw093.

Andrake, M. D. and Skalka, A. M. (1996) 'Retroviral Integrase, Putting the Pieces Together', *Journal of Biological Chemistry*, 271(33), pp. 19633–19636. doi: 10.1074/jbc.271.33.19633.

Andrews, S. (2010) *FastQC A Quality Control tool for High Throughput Sequence Data*. Available at: http://www.bioinformatics.babraham.ac.uk/projects/fastqc/.

Arnaud, F. *et al.* (2008) 'Endogenous retroviruses', *Cellular and Molecular Life Sciences*, 65(21), pp. 3422–3432. doi: 10.1007/s00018-008-8500-9.

Babaian, A. and Mager, D. L. (2016) 'Endogenous retroviral promoter exaptation in human cancer.', *Mobile DNA*, 7, p. 24. doi: 10.1186/s13100-016-0080-x.

Baillie, G. J. *et al.* (2004) 'Multiple Groups of Endogenous Betaretroviruses in Mice, Rats, and Other Mammals', *Journal of Virology*, 78(11), pp. 5784–5798. doi: 10.1128/JVI.78.11.5784-5798.2004.

Bao, Z. and Eddy, S. R. (2002) 'Automated de novo identification of repeat sequence families in sequenced genomes.', *Genome research*, 12(8), pp. 1269–76. doi: 10.1101/gr.88502.

Barboni, P. *et al.* (2001) 'Evidence for the presence of two bovine lentiviruses in the cattle population of Bali.', *Veterinary microbiology*, 80(4), pp. 313–27. Available at: http://www.ncbi.nlm.nih.gov/pubmed/11348768.

Barre-Sinoussi, F. *et al.* (1983) 'Isolation of a T-lymphotropic retrovirus from a patient at risk for acquired immune deficiency syndrome (AIDS)', *Science*, 220(4599), pp. 868–871. doi: 10.1126/science.6189183.

Basta, H. A. *et al.* (2009) 'Evolution of Teleost Fish Retroviruses: Characterization of New Retroviruses with Cellular Genes', *Journal of Virology*,

83(19), pp. 10152–10162. doi: 10.1128/JVI.02546-08.

Bendrey, R. (2012) 'From wild horses to domestic horses: a European perspective', *World Archaeology*, pp. 135–157.

Bénit, L. *et al.* (1999) 'ERV-L elements: a family of endogenous retrovirus-like elements active throughout the evolution of mammals.', *Journal of virology*, 73(4), pp. 3301–3308. Available at: http://www.ncbi.nlm.nih.gov/pubmed/10074184.

Bénit, L., Dessen, P. and Heidmann, T. (2001) 'Identification, phylogeny, and evolution of retroviral elements based on their envelope genes.', *Journal of virology*, 75(23), pp. 11709–19. doi: 10.1128/JVI.75.23.11709-11719.2001.

Best, S. *et al.* (1996) 'Positional cloning of the mouse retrovirus restriction gene Fvl', *Nature*, 382(6594), pp. 826–829. doi: 10.1038/382826a0.

Bhatia, S., Patil, S. S. and Sood, R. (2013) 'Bovine immunodeficiency virus: a lentiviral infection', *Indian Journal of Virology*, 24(3), pp. 332–341. doi: 10.1007/s13337-013-0165-9.

Blanco-Melo, D., Gifford, R. J. and Bieniasz, P. D. (2017) 'Co-option of an endogenous retrovirus envelope for host defense in hominid ancestors', *eLife*, 6. doi: 10.7554/eLife.22519.

Blomberg, J. *et al.* (2009) 'Classification and nomenclature of endogenous retroviral sequences (ERVs)', *Gene*, 448(2), pp. 115–123. doi: 10.1016/j.gene.2009.06.007.

Bogerd, H. P. *et al.* (2008) 'Equine Infectious Anemia Virus Resists the Antiretroviral Activity of Equine APOBEC3 Proteins through a Packaging-Independent Mechanism', *Journal of Virology*, 82(23), pp. 11889–11901. doi: 10.1128/JVI.01537-08.

Brown, K. *et al.* (2012) 'Characterisation of retroviruses in the horse genome and their transcriptional activity via transcriptome sequencing', *Virology*. Elsevier,

433(1), pp. 55–63. doi: 10.1016/j.virol.2012.07.010.

Brown, K., Emes, R. D. and Tarlinton, R. E. (2014) 'Multiple Groups of Endogenous Epsilon-Like Retroviruses Conserved across Primates', *Journal of Virology*, 88(21), pp. 12464–12471. doi: 10.1128/JVI.00966-14.

Brown, P. (1997) *Integration, Retroviruses*. Available at: http://www.ncbi.nlm.nih.gov/pubmed/21433344.

Cai, D. *et al.* (2009) 'Ancient DNA provides new insights into the origin of the Chinese domestic horse', *Journal of Archaeological Science*, 36(3), pp. 835–842. doi: 10.1016/j.jas.2008.11.006.

Camacho, C. *et al.* (2009) 'BLAST+: architecture and applications', *BMC Bioinformatics*, 10(1), p. 421. doi: 10.1186/1471-2105-10-421.

Cartellieri, M. *et al.* (2005) 'Determination of the relative amounts of Gag and Pol proteins in foamy virus particles.', *Retrovirology*, 2, p. 44. doi: 10.1186/1742-4690-2-44.

Caspi, A. (2005) 'Identification of transposable elements using multiple alignments of related genomes', *Genome Research*, 16(2), pp. 260–270. doi: 10.1101/gr.4361206.

Chan, P. P. and Lowe, T. M. (2009) 'GtRNAdb: a database of transfer RNA genes detected in genomic sequence.', *Nucleic acids research*, 37(Database issue), pp. D93-7. doi: 10.1093/nar/gkn787.

Cherkasova, E., Weisman, Q. and Childs, R. W. (2013) 'Endogenous retroviruses as targets for antitumor immunity in renal cell cancer and other tumors.', *Frontiers in oncology*, 3, p. 243. doi: 10.3389/fonc.2013.00243.

Clamp, M. *et al.* (2004) 'The Jalview Java alignment editor', *Bioinformatics*, 20(3), pp. 426–427. doi: 10.1093/bioinformatics/btg430.

Coffee, L. L., Casey, J. W. and Bowser, P. R. (2013) 'Pathology of Tumors in Fish Associated With Retroviruses', *Veterinary Pathology*, 50(3), pp. 390–403. doi:

10.1177/0300985813480529.

Coffin, J. M. (1990) *Retroviridae and their replication*. Fields Vir. New York: Raven Press.

Coffin, J. M., Hughes, S. H. and Varmus, H. E. (1997) *Retroviruses*. Cold Spring Harbor Laboratory Press. Available at: https://www.ncbi.nlm.nih.gov/books/NBK19376/.

Conley, A. and Hinshelwood, M. (2001) 'Mammalian aromatases', *Reproduction*, 121(5), pp. 685–695. doi: 10.1530/rep.0.1210685.

Cremona, M. A. *et al.* (2017) 'IWTomics: Interval-Wise Testing for Omics Data.' Available at: https://bioconductor.org/packages/release/bioc/html/IWTomics.html.

Cui, J. *et al.* (2014) 'Low frequency of paleoviral infiltration across the avian phylogeny', *Genome Biology*, 15(12), p. 539. doi: 10.1186/s13059-014-0539-3.

Damgaard, C. K. *et al.* (2004) 'RNA interactions in the 5' region of the HIV-1 genome.', *Journal of molecular biology*, 336(2), pp. 369–79. Available at: http://www.ncbi.nlm.nih.gov/pubmed/14757051.

Diehl, W. E. *et al.* (2016) 'Tracking interspecies transmission and long-term evolution of an ancient retrovirus using the genomes of modern mammals', *eLife*, 5, p. e12704. doi: 10.7554/eLife.12704.

Doolittle, R. F. *et al.* (1989) 'Origins and evolutionary relationships of retroviruses.', *The Quarterly review of biology*, 64(1), pp. 1–30. Available at: http://www.ncbi.nlm.nih.gov/pubmed/2469098.

Douek, D. C., Roederer, M. and Koup, R. A. (2009) 'Emerging concepts in the immunopathogenesis of AIDS.', *Annual review of medicine*, 60, pp. 471–84. doi: 10.1146/annurev.med.60.041807.123549.

Dunn, B. M. *et al.* (2002) 'Retroviral proteases.', *Genome biology*, 3(4), p. REVIEWS3006. Available at: http://www.ncbi.nlm.nih.gov/pubmed/11983066.

Dupressoir, A. *et al.* (2009) 'Syncytin-A knockout mice demonstrate the critical role in placentation of a fusogenic, endogenous retrovirus-derived, envelope gene', *Proceedings of the National Academy of Sciences*, 106(29), pp. 12127–12132. doi: 10.1073/pnas.0902925106.

Dupressoir, A., Lavialle, C. and Heidmann, T. (2012) 'From ancestral infectious retroviruses to bona fide cellular genes: Role of the captured syncytins in placentation', *Placenta*, 33(9), pp. 663–671. doi: 10.1016/j.placenta.2012.05.005.

Eddy, S. R. (1998) 'Profile hidden Markov models.', *Bioinformatics*, 14(9), pp. 755–63. Available at: http://www.ncbi.nlm.nih.gov/pubmed/9918945.

Eddy, S. R. (2001) *HMMER: biosequence analysis using profile hidden Markov models*. Available at: http://hmmer.org/.

Edgar, R. C. (2004) 'MUSCLE: multiple sequence alignment with high accuracy and high throughput', *Nucleic Acids Research*, 32(5), pp. 1792–1797. doi: 10.1093/nar/gkh340.

Edgar, R. C. and Myers, E. W. (2005) 'PILER: identification and classification of genomic repeats.', *Bioinformatics*, 21 Suppl 1, pp. i152-8. doi: 10.1093/bioinformatics/bti1003.

Ellinghaus, D., Kurtz, S. and Willhoeft, U. (2008) 'LTRharvest, an efficient and flexible software for de novo detection of LTR retrotransposons', *BMC Bioinformatics*, 9(1), p. 18. doi: 10.1186/1471-2105-9-18.

Enssle, J. *et al.* (1996) 'Foamy virus reverse transcriptase is expressed independently from the Gag protein.', *Proceedings of the National Academy of Sciences*, 93(9), pp. 4137–4141. doi: 10.1073/pnas.93.9.4137.

Fauquet, C. M. and Fargette, D. (2005) 'International Committee on Taxonomy of Viruses and the 3,142 unassigned species.', *Virology journal*, 2(1), p. 64. doi: 10.1186/1743-422X-2-64.

Finn, R. D. *et al.* (2016) 'The Pfam protein families database: towards a more sustainable future', *Nucleic Acids Research*, 44(D1), pp. D279–D285. doi: 10.1093/nar/gkv1344.

Finn, R. D., Clements, J. and Eddy, S. R. (2011) 'HMMER web server: interactive sequence similarity searching', *Nucleic Acids Research*, 39(suppl), pp. W29–W37. doi: 10.1093/nar/gkr367.

Flockerzi, A. *et al.* (2005) 'Human endogenous retrovirus HERV-K14 families: status, variants, evolution, and mobilization of other cellular sequences.', *Journal of virology*, 79(5), pp. 2941–9. doi: 10.1128/JVI.79.5.2941-2949.2005.

Flügel, R. M. and Pfrepper, K. I. (2003) 'Proteolytic processing of foamy virus Gag and Pol proteins.', *Current topics in microbiology and immunology*, 277, pp. 63–88. Available at: http://www.ncbi.nlm.nih.gov/pubmed/12908768.

Franchini, L. F., Ganko, E. W. and McDonald, J. F. (2004) 'Retrotransposon-Gene Associations Are Widespread Among D. melanogaster Populations', *Molecular Biology and Evolution*, 21(7), pp. 1323–1331. doi: 10.1093/molbev/msh116.

Franzen, J. (2011) 'Book reviews', *mammalia*. Johns Hopkins University Press, 75(2). doi: 10.1515/mamm.2011.002.

Frerichs, G. N. *et al.* (1991) 'Spontaneously productive C-type retrovirus infection of fish cell lines.', *The Journal of general virology*, 72 (Pt 10, pp. 2537–9. doi: 10.1099/0022-1317-72-10-2537.

Fushan, A. A. *et al.* (2015) 'Gene expression defines natural changes in mammalian lifespan', *Aging Cell*, 14(3), pp. 352–365. doi: 10.1111/acel.12283.

Gallo, R. *et al.* (1983) 'Isolation of human T-cell leukemia virus in acquired immune deficiency syndrome (AIDS)', *Science*, 220(4599), pp. 865–867. doi: 10.1126/science.6601823.

Garcia-Etxebarria, K. and Jugo, B. M. (2012) 'Detection and characterization of endogenous retroviruses in the horse genome by in silico analysis.', *Virology*,

434(1), pp. 59–67. doi: 10.1016/j.virol.2012.08.047.

Garcia-Etxebarria, K., Sistiaga-Poveda, M. and Jugo, B. M. (2014) 'Endogenous retroviruses in domestic animals.', *Current genomics*, 15(4), pp. 256–65. doi: 10.2174/1389202915666140520003503.

Gel, B. and Serra, E. (2017) 'karyoploteR: an R/Bioconductor package to plot customizable genomes displaying arbitrary data', *Bioinformatics*, 33(19), pp. 3088–3090. doi: 10.1093/bioinformatics/btx346.

Gifford, R. and Tristem, M. (2003) 'The evolution, distribution and diversity of endogenous retroviruses.', *Virus genes*, 26(3), pp. 291–315. doi: 10.1023/A:1024455415443.

Gim, J.-A. and Kim, H.-S. (2017) 'Identification and Expression Analyses of Equine Endogenous Retroviruses in Horses.', *Molecules and cells*, 40(10), pp. 796–804. doi: 10.14348/molcells.2017.0141.

Goff, S. P. (2007) 'Host factors exploited by retroviruses', *Nature Reviews Microbiology*, 5(4), pp. 253–263. doi: 10.1038/nrmicro1541.

Goff, S. P. (2013) *Retroviridae*. 6th ed, *Fields Virology*. 6th ed. Philadelphia: Lippincott Williams & Wilkins.

Goldschmidt, V. *et al.* (2002) 'Direct and indirect contributions of RNA secondary structure elements to the initiation of HIV-1 reverse transcription.', *The Journal of biological chemistry*, 277(45), pp. 43233–42. doi: 10.1074/jbc.M205295200.

Goodenow, M. M. *et al.* (2002) 'Naturally Occurring Amino Acid Polymorphisms in Human Immunodeficiency Virus Type 1 (HIV-1) Gag p7NC and the C-Cleavage Site Impact Gag-Pol Processing by HIV-1 Protease', *Virology*, 292(1), pp. 137–149. doi: 10.1006/viro.2001.1184.

Goto, H. *et al.* (2011) 'A massively parallel sequencing approach uncovers ancient origins and high genetic variability of endangered Przewalski's horses.', *Genome biology and evolution*, 3, pp. 1096–106. doi: 10.1093/gbe/evr067.

Gouy, M., Guindon, S. and Gascuel, O. (2010) 'SeaView Version 4: A Multiplatform Graphical User Interface for Sequence Alignment and Phylogenetic Tree Building', *Molecular Biology and Evolution*, 27(2), pp. 221–224. doi: 10.1093/molbev/msp259.

Grandi, N. *et al.* (2018) 'HERV-W group evolutionary history in non-human primates: characterization of ERV-W orthologs in Catarrhini and related ERV groups in Platyrrhini', *BMC Evolutionary Biology*, 18(1), p. 6. doi: 10.1186/s12862-018-1125-1.

Gremme, G., Steinbiss, S. and Kurtz, S. (2013) 'GenomeTools: A Comprehensive Software Library for Efficient Processing of Structured Genome Annotations', *IEEE/ACM Transactions on Computational Biology and Bioinformatics*, 10(3), pp. 645–656. doi: 10.1109/TCBB.2013.68.

Gross, L. (1951) '"Spontaneous" Leukemia Developing in G3H Mice Following Inoculation, In Infancy, with AK-Emkemic.', *Experimental Biology and Medicine*, 76(1), pp. 27–32. doi: 10.3181/00379727-76-18379.

Han, G.-Z. (2015) 'Extensive retroviral diversity in shark', *Retrovirology*, 12(1), p. 34. doi: 10.1186/s12977-015-0158-4.

Hashimoto, K. *et al.* (2015) 'CAGE profiling of ncRNAs in hepatocellular carcinoma reveals widespread activation of retroviral LTR promoters in virus-induced tumors.', *Genome research*, 25(12), pp. 1812–24. doi: 10.1101/gr.191031.115.

Hayward, A., Grabherr, M. and Jern, P. (2013) 'Broad-scale phylogenomics provides insights into retrovirus-host evolution', *Proceedings of the National Academy of Sciences*, 110(50), pp. 20146–20151. doi: 10.1073/pnas.1315419110.

Hayward, J. A. *et al.* (2013) 'Identification of diverse full-length endogenous betaretroviruses in megabats and microbats', *Retrovirology*, 10(1), p. 35. doi: 10.1186/1742-4690-10-35.

Hayward, W. S., Neel, B. G. and Astrin, S. M. (1981) 'Activation of a cellular onc

gene by promoter insertion in ALV-induced lymphoid leukosis', *Nature*, 290(5806), pp. 475–480. doi: 10.1038/290475a0.

Hazewinkel, M. (1994) 'Density of a probability distribution', in *Encyclopaedia of Mathematics*. Springer Netherlands.

Hecht, S. J. *et al.* (1996) 'Distribution of endogenous type B and type D sheep retrovirus sequences in ungulates and other mammals.', *Proceedings of the National Academy of Sciences*, 93(8), pp. 3297–3302. doi: 10.1073/pnas.93.8.3297.

Heidmann, O. *et al.* (2009) 'Identification of an endogenous retroviral envelope gene with fusogenic activity and placenta-specific expression in the rabbit: a new "syncytin" in a third order of mammals.', *Retrovirology*, 6, p. 107. doi: 10.1186/1742-4690-6-107.

Henderson, L. E., Krutzsch, H. C. and Oroszlan, S. (1983) 'Myristyl amino-terminal acylation of murine retrovirus proteins: An unusual post-translational protein modification', *Proceedings of the National Academy of Sciences*, 80(2), pp. 339–343. doi: 10.1073/pnas.80.2.339.

Heneine, W. *et al.* (2003) 'Human infection with foamy viruses.', *Current topics in microbiology and immunology*, 277, pp. 181–96. Available at: http://www.ncbi.nlm.nih.gov/pubmed/12908773.

Hillier, L. W. *et al.* (2004) 'Sequence and comparative analysis of the chicken genome provide unique perspectives on vertebrate evolution', *Nature*, 432(7018), pp. 695–716. doi: 10.1038/nature03154.

Hindmarsh, P. and Leis, J. (1999) 'Retroviral DNA integration.', *Microbiology and Molecular Biology Reviews*, 63(4), p. 836–43, table of contents. Available at: http://www.ncbi.nlm.nih.gov/pubmed/10585967.

Hizi, A. and Herzig, E. (2015) 'dUTPase: the frequently overlooked enzyme encoded by many retroviruses', *Retrovirology*, 12(1), p. 70. doi: 10.1186/s12977-015-0198-9.

Hohn, O., Hanke, K. and Bannert, N. (2013) 'HERV-K(HML-2), the Best Preserved Family of HERVs: Endogenization, Expression, and Implications in Health and Disease', *Frontiers in Oncology*, 3. doi: 10.3389/fonc.2013.00246.

Holl, H. M. *et al.* (2015) 'Generation of a de novo transcriptome from equine lamellar tissue', *BMC Genomics*, 16(1), p. 739. doi: 10.1186/s12864-015-1948-8.

Holl, H. M. *et al.* (2016) 'Variant in the RFWD3 gene associated with PATN1 , a modifier of leopard complex spotting', *Animal Genetics*, 47(1), pp. 91–101. doi: 10.1111/age.12375.

Holmes, E. C. (2011) 'The evolution of endogenous viral elements.', *Cell host & microbe*, 10(4). doi: 10.1016/j.chom.2011.09.002.

Holzschu, D. L. *et al.* (1998) 'The nucleotide sequence and spliced pol mRNA levels of the nonprimate spumavirus bovine foamy virus.', *Journal of virology*, 72(3), pp. 2177–82. Available at: http://www.ncbi.nlm.nih.gov/pubmed/9499074.

Huang, J. *et al.* (2014) 'Analysis of horse genomes provides insight into the diversification and adaptive evolution of karyotype.', *Scientific reports*, 4(1), p. 4958. doi: 10.1038/srep04958.

Huang, J. *et al.* (2015) 'Donkey genome and insight into the imprinting of fast karyotype evolution', *Scientific Reports*. Nature Publishing Group, 5, p. 14106. doi: 10.1038/srep14106.

Hunter, E. (1997) *Viral Entry and Receptors*, *Retroviruses*. Available at: http://www.ncbi.nlm.nih.gov/pubmed/21433347.

Hunter, E. and Swanstrom, R. (1990) 'Retrovirus Envelope Glycoproteins', in, pp. 187–253. doi: 10.1007/978-3-642-75218-6_7.

Ikeda, H. and Sugimura, H. (1989) 'Fv-4 resistance gene: a truncated endogenous murine leukemia virus with ecotropic interference properties.', *Journal of virology*, 63(12), pp. 5405–12. Available at:

http://www.ncbi.nlm.nih.gov/pubmed/2555565.

Imbeault, M., Helleboid, P.-Y. and Trono, D. (2017) 'KRAB zinc-finger proteins contribute to the evolution of gene regulatory networks.', *Nature*, 543(7646), pp. 550–554. doi: 10.1038/nature21683.

Iqbal, K. *et al.* (2014) 'RNA-Seq Transcriptome Profiling of Equine Inner Cell Mass and Trophectoderm', *Biology of Reproduction*, 90(3), pp. 61–61. doi: 10.1095/biolreprod.113.113928.

Jacks, T. and Varmus, H. (1985) 'Expression of the Rous sarcoma virus pol gene by ribosomal frameshifting', *Science*, 230(4731), pp. 1237–1242. doi: 10.1126/science.2416054.

Jern, P. *et al.* (2005) 'Sequence Variability , Gene Structure , and Expression of Full-Length Human Endogenous Retrovirus H', *Journal of virology*, 79(10), pp. 6325–6337. doi: 10.1128/JVI.79.10.6325.

Jern, P., Sperber, G. O. and Blomberg, J. (2004) 'Definition and variation of human endogenous retrovirus H.', *Virology*, 327(1), pp. 93–110. doi: 10.1016/j.virol.2004.06.023.

Jo, H. *et al.* (2012) 'Identification and classification of endogenous retroviruses in the canine genome using degenerative PCR and in-silico data analysis.', *Virology*, 422(2), pp. 195–204. doi: 10.1016/j.virol.2011.10.010.

Jolicoeur, P. (1991) 'Neuronal loss in a lower motor neuron disease induced by a murine retrovirus.', *The Canadian journal of neurological sciences. Le journal canadien des sciences neurologiques*, 18(3 Suppl), pp. 411–3. Available at: http://www.ncbi.nlm.nih.gov/pubmed/1657349.

Jónsson, H. *et al.* (2014) 'Speciation with gene flow in equids despite extensive chromosomal plasticity', *Proceedings of the National Academy of Sciences*, pp. 18655–18660. doi: 10.1073/pnas.1412627111.

Jurka, J. *et al.* (1996) 'Censor—a program for identification and elimination of

repetitive elements from DNA sequences', *Computers & Chemistry*, 20(1), pp. 119–121. doi: 10.1016/S0097-8485(96)80013-1.

Jurka, J. *et al.* (2005) 'Repbase Update, a database of eukaryotic repetitive elements', *Cytogenetic and Genome Research*, 110(1–4), pp. 462–467. doi: 10.1159/000084979.

Kalyaanamoorthy, S. *et al.* (2017) 'ModelFinder: fast model selection for accurate phylogenetic estimates', *Nature Methods*, 14(6), pp. 587–589. doi: 10.1038/nmeth.4285.

Kalyanaraman, A. and Aluru, S. (2006) 'Efficient algorithms and software for detection of full-length LTR retrotransposons.', *Journal of bioinformatics and computational biology*, 4(2), pp. 197–216. Available at: http://www.ncbi.nlm.nih.gov/pubmed/16819780.

Kamat, A. *et al.* (1998) 'Characterization of the Regulatory Regions of the Human Aromatase (P450arom) Gene Involved in Placenta-Specific Expression', *Molecular Endocrinology*, 12(11), pp. 1764–1777. doi: 10.1210/mend.12.11.0190.

Katzourakis, A. and Gifford, R. J. (2010) 'Endogenous Viral Elements in Animal Genomes', *PLoS Genetics*, 6(11), p. e1001191. doi: 10.1371/journal.pgen.1001191.

Katzourakis, A. and Tristem, M. (2005) 'Phylogeny of human endogenous and exogenous retroviruses', in *Retro-viruses and Primate Genome Evolution*. Georgetown: Landes Bioscience.

Kaye, J. F., Richardson, J. H. and Lever, A. M. (1995) 'cis-acting sequences involved in human immunodeficiency virus type 1 RNA packaging.', *Journal of virology*, 69(10), pp. 6588–92. Available at: http://www.ncbi.nlm.nih.gov/pubmed/7666564.

Kent, W. J. (2002) 'BLAT---The BLAST-Like Alignment Tool', *Genome Research*, 12(4), pp. 656–664. doi: 10.1101/gr.229202.

Kim, S. *et al.* (2008) 'Integration Site Preference of Xenotropic Murine Leukemia Virus-Related Virus, a New Human Retrovirus Associated with Prostate Cancer', *Journal of Virology*, 82(20), pp. 9964–9977. doi: 10.1128/JVI.01299-08.

King, A. M. *et al.* (2011) *Virus taxonomy: ninth report of the International Committee on Taxonomy of Viruses*. Elsevier.

Kitamura, Y., Lee, Y. M. and Coffin, J. M. (1992) 'Nonrandom integration of retroviral DNA in vitro: effect of CpG methylation.', *Proceedings of the National Academy of Sciences*, 89(12), pp. 5532–5536. doi: 10.1073/pnas.89.12.5532.

Klymiuk, N. *et al.* (2003) 'Characterization of Endogenous Retroviruses in Sheep', *Journal of Virology*, 77(20), pp. 11268–11273. doi: 10.1128/JVI.77.20.11268-11273.2003.

Kozak, C. A. *et al.* (1984) 'A unique sequence related to the ecotropic murine leukemia virus is associated with the Fv-4 resistance gene.', *Proceedings of the National Academy of Sciences*, 81(3), pp. 834–837. doi: 10.1073/pnas.81.3.834.

Krueger, F. (2015) *Trim Galore!: A wrapper tool around Cutadapt and FastQC to consistently apply quality and adapter trimming to FastQ files*. Available at: https://www.bioinformatics.babraham.ac.uk/projects/trim_galore/.

Kumar, S. *et al.* (2017) 'TimeTree: A Resource for Timelines, Timetrees, and Divergence Times', *Molecular Biology and Evolution*, 34(7), pp. 1812–1819. doi: 10.1093/molbev/msx116.

Kumar, S. and Subramanian, S. (2002) 'Mutation rates in mammalian genomes', *Proceedings of the National Academy of Sciences*, 99(2), pp. 803–808. doi: 10.1073/pnas.022629899.

Kurtz, S. *et al.* (2001) 'REPuter: the manifold applications of repeat analysis on a genomic scale.', *Nucleic Acids Research*, 29(22), pp. 4633–4642. doi: 10.1093/nar/29.22.4633.

van der Kuyl, A. C. (2011) 'Characterization of a full-length endogenous beta-

retrovirus, EqERV-beta1, in the genome of the horse (Equus caballus).', *Viruses*, 3(6), pp. 620–8. doi: 10.3390/v3060620.

van de Lagemaat, L. N. *et al.* (2003) 'Transposable elements in mammals promote regulatory variation and diversification of genes with specialized functions', *Trends in Genetics*, 19(10), pp. 530–536. doi: 10.1016/j.tig.2003.08.004.

Lander, E. S. *et al.* (2001) 'Initial sequencing and analysis of the human genome.', *Nature*, 409(6822), pp. 860–921. doi: 10.1038/35057062.

Langmead, B. and Salzberg, S. L. (2012) 'Fast gapped-read alignment with Bowtie 2', *Nature Methods*, 9(4), pp. 357–359. doi: 10.1038/nmeth.1923.

LaPierre, L. A. *et al.* (1998) 'Two closely related but distinct retroviruses are associated with walleye discrete epidermal hyperplasia.', *Journal of virology*, 72(4), pp. 3484–90. Available at: http://www.ncbi.nlm.nih.gov/pubmed/9525688.

Larruskain, A. and Jugo, B. (2013) 'Retroviral Infections in Sheep and Goats: Small Ruminant Lentiviruses and Host Interaction', *Viruses*, 5(12), pp. 2043–2061. doi: 10.3390/v5082043.

Larsson, A. (2014) 'AliView: a fast and lightweight alignment viewer and editor for large datasets', *Bioinformatics*, 30(22), pp. 3276–3278. doi: 10.1093/bioinformatics/btu531.

Lavie, L. *et al.* (2004) 'Human endogenous retrovirus family HERV-K(HML-5): status, evolution, and reconstruction of an ancient betaretrovirus in the human genome.', *Journal of virology*, 78(16), pp. 8788–98. doi: 10.1128/JVI.78.16.8788-8798.2004.

Lee, A. *et al.* (2013) 'Identification of an ancient endogenous retrovirus, predating the divergence of the placental mammals', *Philosophical Transactions of the Royal Society B: Biological Sciences*, 368(1626), pp. 20120503–20120503. doi: 10.1098/rstb.2012.0503.

Lee, Y. N. and Bieniasz, P. D. (2007) ‘Reconstitution of an infectious human endogenous retrovirus.’, *PLoS pathogens*, 3(1), p. e10. doi: 10.1371/journal.ppat.0030010.

Lepa, A. and Siwicki, A. K. (2011) ‘Retroviruses of wild and cultured fish.’, *Polish journal of veterinary sciences*, 14(4), pp. 703–9. Available at: http://www.ncbi.nlm.nih.gov/pubmed/22439348.

Leroux, C., Cador, J.-L. and Montelaro, R. C. (2004) ‘Equine Infectious Anemia Virus (EIAV): what has HIV?s country cousin got to tell us?’, *Veterinary Research*, 35(4), pp. 485–512. doi: 10.1051/vetres:2004020.

Levitsky, V. G. (2004) ‘RECON: a program for prediction of nucleosome formation potential’, *Nucleic Acids Research*, 32(Web Server), pp. W346–W349. doi: 10.1093/nar/gkh482.

Levy, J. A. (1973) ‘Xenotropic Viruses: Murine Leukemia Viruses Associated with NIH Swiss, NZB, and Other Mouse Strains’, *Science*, 182(4117), pp. 1151–1153. doi: 10.1126/science.182.4117.1151.

Li, H. *et al.* (2009) ‘The Sequence Alignment/Map format and SAMtools’, *Bioinformatics*, 25(16), pp. 2078–2079. doi: 10.1093/bioinformatics/btp352.

Li, R. *et al.* (2005) ‘ReAS: Recovery of ancestral sequences for transposable elements from the unassembled reads of a whole genome shotgun.’, *PLoS computational biology*, 1(4), p. e43. doi: 10.1371/journal.pcbi.0010043.

Lilly, F. and Pincus, T. (1973) ‘Genetic Control Of Murine Viral Leukemogenesis’, in, pp. 231–277. doi: 10.1016/S0065-230X(08)60532-1.

Löchelt, M. and Flügel, R. M. (1996) ‘The human foamy virus pol gene is expressed as a Pro-Pol polyprotein and not as a Gag-Pol fusion protein.’, *Journal of virology*, 70(2), pp. 1033–40. Available at: http://www.ncbi.nlm.nih.gov/pubmed/8551561.

Lowe, T. M. and Eddy, S. R. (1997) ‘tRNAscan-SE: A Program for Improved

Detection of Transfer RNA Genes in Genomic Sequence', *Nucleic Acids Research*, 25(5), pp. 955–964. doi: 10.1093/nar/25.5.0955.

Macfadden, B. J. (1997) 'Pleistocene horses from Tarija, Bolivia, and validity of the genus † Onohippidium (Mammalia: Equidae)', *Journal of Vertebrate Paleontology*, 17(1), pp. 199–218. doi: 10.1080/02724634.1997.10010964.

MacFadden, B. J. (1986) 'LATE HEMPHILLIAN MONODACTYL HORSES (MAMMALIA, EQUIDAE) FROM THE BONE VALLEY FORMATION OF CENTRAL FLORIDA', *Journal of Paleontology*, 60(2), pp. 466–475. Available at: http://www.jstor.org/stable/1305172.

MacFadden, B. J. (2000) 'Cenozoic Mammalian Herbivores From the Americas: Reconstructing Ancient Diets and Terrestrial Communities', *Annual Review of Ecology and Systematics*, 31(1), pp. 33–59. doi: 10.1146/annurev.ecolsys.31.1.33.

MacFadden, B. J. (2005) 'EVOLUTION: Fossil Horses--Evidence for Evolution', *Science*, 307(5716), pp. 1728–1730. doi: 10.1126/science.1105458.

Maksakova, I. A. *et al.* (2006) 'Retroviral elements and their hosts: insertional mutagenesis in the mouse germ line.', *PLoS genetics*, 2(1), p. e2. doi: 10.1371/journal.pgen.0020002.

Mann, R., Mulligan, R. C. and Baltimore, D. (1983) 'Construction of a retrovirus packaging mutant and its use to produce helper-free defective retrovirus.', *Cell*, 33(1), pp. 153–9. Available at: http://www.ncbi.nlm.nih.gov/pubmed/6678608.

Marth, C. D. *et al.* (2015) 'Deep sequencing of the uterine immune response to bacteria during the equine oestrous cycle', *BMC Genomics*, 16(1), p. 934. doi: 10.1186/s12864-015-2139-3.

Martin, J. *et al.* (1997) 'Human endogenous retrovirus type I-related viruses have an apparently widespread distribution within vertebrates.', *Journal of virology*, 71(1), pp. 437–43. Available at: http://www.ncbi.nlm.nih.gov/pubmed/8985368.

Martin, J. *et al.* (1999) 'Interclass transmission and phyletic host tracking in murine leukemia virus-related retroviruses.', *Journal of virology*, 73(3), pp. 2442–9. Available at: http://www.ncbi.nlm.nih.gov/pubmed/9971829.

Martin, M. A. *et al.* (1981) 'Identification and cloning of endogenous retroviral sequences present in human DNA.', *Proceedings of the National Academy of Sciences*, 78(8), pp. 4892–4896. doi: 10.1073/pnas.78.8.4892.

Mayer, B. J., Hamaguchi, M. and Hanafusa, H. (1988) 'A novel viral oncogene with structural similarity to phospholipase C.', *Nature*, 332(6161), pp. 272–5. doi: 10.1038/332272a0.

Mayer, J. and Meese, E. U. (2002) 'The human endogenous retrovirus family HERV-K(HML-3).', *Genomics*, 80(3), pp. 331–43. Available at: http://www.ncbi.nlm.nih.gov/pubmed/12213204.

McCallin, A. J., Maertens, G. N. and Bangham, C. R. (2015) 'Host determinants of HTLV-1 integration site preference', *Retrovirology*, 12(Suppl 1), p. O27. doi: 10.1186/1742-4690-12-S1-O27.

McCann, E. M. and Lever, A. M. (1997) 'Location of cis-acting signals important for RNA encapsidation in the leader sequence of human immunodeficiency virus type 2.', *Journal of virology*, 71(5), pp. 4133–7. Available at: http://www.ncbi.nlm.nih.gov/pubmed/9094696.

McCarthy, E. M. *et al.* (2002) 'Long terminal repeat retrotransposons of Oryza sativa', *Genome Biology*, 3(10), p. research0053.1. doi: 10.1186/gb-2002-3-10-research0053.

McCarthy, E. M. and McDonald, J. F. (2003) 'LTR_STRUC: a novel search and identification program for LTR retrotransposons', *Bioinformatics*, 19(3), pp. 362–367. doi: 10.1093/bioinformatics/btf878.

McCarthy, E. M. and McDonald, J. F. (2004) 'Long terminal repeat retrotransposons of Mus musculus.', *Genome biology*, 5(3), p. R14. doi: 10.1186/gb-2004-5-3-r14.

McKenna, M. C. and Bell, S. K. (1997) *Classification of mammals: above the species level*. Columbia University Press.

Medstrand, P. and Blomberg, J. (1993) 'Characterization of novel reverse transcriptase encoding human endogenous retroviral sequences similar to type A and type B retroviruses: differential transcription in normal human tissues.', *Journal of virology*, 67(11), pp. 6778–87. Available at: http://www.ncbi.nlm.nih.gov/pubmed/7692084.

Meredith, R. W. *et al.* (2011) 'Impacts of the Cretaceous Terrestrial Revolution and KPg Extinction on Mammal Diversification', *Science*, 334(6055), pp. 521–524. doi: 10.1126/science.1211028.

Mi, S. *et al.* (2000) 'Syncytin is a captive retroviral envelope protein involved in human placental morphogenesis.', *Nature*, 403(6771), pp. 785–9. doi: 10.1038/35001608.

Miller, A. D. (1996) 'Cell-surface receptors for retroviruses and implications for gene transfer.', *Proceedings of the National Academy of Sciences*, 93(21), pp. 11407–11413. doi: 10.1073/pnas.93.21.11407.

Miller, J. M. and Van Der Maaten, M. J. (1977) 'Use of glycoprotein antigen in the immunodiffusion test for bovine leukemia virus antibodies.', *European journal of cancer*, 13(12), pp. 1369–75. Available at: http://www.ncbi.nlm.nih.gov/pubmed/202468.

Moreton, J. *et al.* (2014) 'Characterisation of the horse transcriptome from immunologically active tissues', *PeerJ*, 2, p. e382. doi: 10.7717/peerj.382.

Mouse Genome Sequencing Consortium *et al.* (2002) 'Initial sequencing and comparative analysis of the mouse genome.', *Nature*, 420(6915), pp. 520–62. doi: 10.1038/nature01262.

Murphy, W. J. *et al.* (2007) 'Using genomic data to unravel the root of the placental mammal phylogeny', *Genome Research*, 17(4), pp. 413–421. doi: 10.1101/gr.5918807.

Naville, M. and Volff, J.-N. (2016) ‘Endogenous Retroviruses in Fish Genomes: From Relics of Past Infections to Evolutionary Innovations?’, *Frontiers in Microbiology*, 7, p. 1197. doi: 10.3389/fmicb.2016.01197.

NCBI Resource Coordinators (2018) ‘Database resources of the National Center for Biotechnology Information.’, *Nucleic acids research*, 46(D1), pp. D8–D13. doi: 10.1093/nar/gkx1095.

Nexø, B. A. *et al.* (2016) ‘Are human endogenous retroviruses triggers of autoimmune diseases? Unveiling associations of three diseases and viral loci’, *Immunologic Research*, 64(1), pp. 55–63. doi: 10.1007/s12026-015-8671-z.

Nguyen, L.-T. *et al.* (2015) ‘IQ-TREE: A Fast and Effective Stochastic Algorithm for Estimating Maximum-Likelihood Phylogenies’, *Molecular Biology and Evolution*, 32(1), pp. 268–274. doi: 10.1093/molbev/msu300.

Orlando, L. *et al.* (2013) ‘Recalibrating Equus evolution using the genome sequence of an early Middle Pleistocene horse’, *Nature*, 499(7456), pp. 74–78. doi: 10.1038/nature12323.

Paces, J., Pavlícek, A. and Paces, V. (2002a) ‘HERVd: database of human endogenous retroviruses.’, *Nucleic acids research*, 30(1), pp. 205–6. Available at: http://www.ncbi.nlm.nih.gov/pubmed/11752294.

Paces, J., Pavlícek, A. and Paces, V. (2002b) ‘HERVd: database of human endogenous retroviruses.’, *Nucleic acids research*, 30(1), pp. 205–6. doi: 10.1093/nar/gkv1157.

Pacholewska, A. *et al.* (2015) ‘The Transcriptome of Equine Peripheral Blood Mononuclear Cells’, *PLOS ONE*. Public Library of Science, 10(3), p. e0122011. doi: 10.1371/journal.pone.0122011.

Palmarini, M. *et al.* (1999) ‘Jaagsiekte sheep retrovirus is necessary and sufficient to induce a contagious lung cancer in sheep.’, *Journal of virology*, 73(8), pp. 6964–72. Available at: http://www.ncbi.nlm.nih.gov/pubmed/10400795.

Paul, T. A. *et al.* (2006) 'Identification and characterization of an exogenous retrovirus from atlantic salmon swim bladder sarcomas.', *Journal of virology*, 80(6), pp. 2941–8. doi: 10.1128/JVI.80.6.2941-2948.2006.

Pavlicek, A. *et al.* (2002) 'Processed Pseudogenes of Human Endogenous Retroviruses Generated by LINEs: Their Integration, Stability, and Distribution', *Genome Research*, 12(3), pp. 391–399. doi: 10.1101/gr.216902.

Pavlícek, A. *et al.* (2002) 'Length distribution of long interspersed nucleotide elements (LINEs) and processed pseudogenes of human endogenous retroviruses: implications for retrotransposition and pseudogene detection.', *Gene*, 300(1–2), pp. 189–94. doi: 10.1093/oxfordjournals.molbev.a004108.

Payne, G. S. *et al.* (1981) 'Analysis of avian leukosis virus DNA and RNA in bursal tumours: viral gene expression is not required for maintenance of the tumor state.', *Cell*, 23(2), pp. 311–22. doi: 10.1016/0168-9525(86)90247-7.

Payne, L. N. *et al.* (1991) 'A novel subgroup of exogenous avian leukosis virus in chickens.', *The Journal of general virology*, 72 (Pt 4), pp. 801–7. doi: 10.1099/0022-1317-72-4-801.

Payne, L. N. (1992) 'Biology of Avian Retroviruses', in *The Retroviridae*. Boston, MA: Springer US, pp. 299–404. doi: 10.1007/978-1-4615-3372-6_6.

Petropoulos, C. (1997) 'Retroviral Taxonomy, Protein Structures, Sequences, and Genetic Maps', in *Retroviruses*. Cold Spring Harbor: Cold Spring Harbor Laboratory Press. Available at: https://www.ncbi.nlm.nih.gov/books/NBK19417/.

Pincus, T., Rowe, W. P. and Lilly, F. (1971) 'A major genetic locus affecting resistance to infection with murine leukemia viruses. II. Apparent identity to a major locus described for resistance to friend murine leukemia virus.', *The Journal of experimental medicine*, 133(6), pp. 1234–41. Available at: http://www.ncbi.nlm.nih.gov/pubmed/4325133.

Prothero, D. R. and Schoch, R. M. (1989) *The evolution of perissodactyls*. Oxford

University Press, USA.

Pruitt, K. D. *et al.* (2014) 'RefSeq: an update on mammalian reference sequences', *Nucleic Acids Research*, 42(D1), pp. D756–D763. doi: 10.1093/nar/gkt1114.

Purchase, H. G. *et al.* (1973) 'A new group of oncogenic viruses: reticuloendotheliosis, chick syncytial, duck infectious anemia, and spleen necrosis viruses.', *Journal of the National Cancer Institute*, 51(2), pp. 489–99. Available at: http://www.ncbi.nlm.nih.gov/pubmed/4358134.

Quinlan, A. R. and Hall, I. M. (2010) 'BEDTools: a flexible suite of utilities for comparing genomic features', *Bioinformatics*, 26(6), pp. 841–842. doi: 10.1093/bioinformatics/btq033.

Quinn, J. H. (1955) *Miocene Equidae of the Texas Gulf Coastal Plain*.

Rabson, A. and Graves, B. (1997) *Synthesis and Processing of Viral RNA*, *Retroviruses*. Available at: http://www.ncbi.nlm.nih.gov/pubmed/21433339.

Radinsky, L. B. (1966) 'The Adaptive Radiation of the Phenacodontid Condylarths and the Origin of the Perissodactyla', *Evolution*, 20(3), p. 408. doi: 10.2307/2406639.

Raudsepp, T. *et al.* (2004) 'A detailed physical map of the horse Y chromosome', *Proceedings of the National Academy of Sciences*, 101(25), pp. 9321–9326. doi: 10.1073/pnas.0403011101.

Redelsperger, F. *et al.* (2016) 'Genetic Evidence That Captured Retroviral Envelope syncytins Contribute to Myoblast Fusion and Muscle Sexual Dimorphism in Mice', *PLOS Genetics*, 12(9), p. e1006289. doi: 10.1371/journal.pgen.1006289.

dos Reis, M. *et al.* (2012) 'Phylogenomic datasets provide both precision and accuracy in estimating the timescale of placental mammal phylogeny', *Proceedings of the Royal Society B: Biological Sciences*, 279(1742), pp. 3491–3500. doi: 10.1098/rspb.2012.0683.

Rhee, S. S., Hui, H. X. and Hunter, E. (1990) ‘Preassembled capsids of type D retroviruses contain a signal sufficient for targeting specifically to the plasma membrane.’, *Journal of virology*, 64(8), pp. 3844–52. Available at: http://www.ncbi.nlm.nih.gov/pubmed/2370682.

Rice, P., Longden, I. and Bleasby, A. (2000) ‘EMBOSS: The European Molecular Biology Open Software Suite’, *Trends in Genetics*, 16(6), pp. 276–277. doi: 10.1016/S0168-9525(00)02024-2.

Rombel, I. T. *et al.* (2002) ‘ORF-FINDER: a vector for high-throughput gene identification’, *Gene*, 282(1–2), pp. 33–41. doi: 10.1016/S0378-1119(01)00819-8.

Rowe, H. M. *et al.* (2010) ‘KAP1 controls endogenous retroviruses in embryonic stem cells’, *Nature*, 463(7278), pp. 237–240. doi: 10.1038/nature08674.

Ryder, O. A. (1993) ‘Przewalski’s Horse: Prospects for Reintroduction into the Wild’, *Conservation Biology*, 7(1), pp. 13–19. doi: 10.1046/j.1523-1739.1993.07010013.x.

Sala, M. and Wain-Hobson, S. (2000) ‘Are RNA viruses adapting or merely changing?’, *Journal of molecular evolution*, 51(1), pp. 12–20. Available at: http://www.ncbi.nlm.nih.gov/pubmed/10903368.

Salmons, B. and Günzburg, W. H. (1987) ‘Current perspectives in the biology of mouse mammary tumour virus.’, *Virus research*, 8(2), pp. 81–102. Available at: http://www.ncbi.nlm.nih.gov/pubmed/2823501.

Santillana-Hayat, M. *et al.* (1996) ‘Inhibition of the in vitro infectivity and cytopathic effect of human foamy virus by dideoxynucleosides.’, *AIDS research and human retroviruses*, 12(15), pp. 1485–90. doi: 10.1089/aid.1996.12.1485.

Schröder, A. R. W. W. *et al.* (2002) ‘HIV-1 Integration in the Human Genome Favors Active Genes and Local Hotspots’, *Cell*, 110(4), pp. 521–529. doi: 10.1016/S0092-8674(02)00864-4.

Sepkowitz, K. A. (2001) ‘AIDS — The First 20 Years’, *New England Journal of*

Medicine, 344(23), pp. 1764–1772. doi: 10.1056/NEJM200106073442306.

Shinnick, T. M., Lerner, R. A. and Sutcliffe, J. G. (1981) 'Nucleotide sequence of Moloney murine leukaemia virus', *Nature*, 293(5833), pp. 543–8. doi: 10.1038/293543a0.

Singer, J. B. *et al.* (2018) 'GLUE: A flexible software system for virus sequence data', *bioRxiv*. doi: 10.1101/269274.

Sinzelle, L. *et al.* (2011) 'Characterization of a Xenopus tropicalis endogenous retrovirus with developmental and stress-dependent expression.', *Journal of virology*, 85(5), pp. 2167–79. doi: 10.1128/JVI.01979-10.

Slater, G. S. C. and Birney, E. (2005) 'Automated generation of heuristics for biological sequence comparison.', *BMC bioinformatics*, 6(1), p. 31. doi: 10.1186/1471-2105-6-31.

Smit, AFA, Hubley, R & Green, P. (2013) *RepeatMasker Open-4.0*. Available at: http://www.repeatmasker.org.

Smit, A. F. (1999) 'Interspersed repeats and other mementos of transposable elements in mammalian genomes.', *Current opinion in genetics & development*, 9(6), pp. 657–63. Available at: http://www.ncbi.nlm.nih.gov/pubmed/10607616.

Smits, K. *et al.* (2016) 'The Equine Embryo Influences Immune-Related Gene Expression in the Oviduct', *Biology of Reproduction*, 94(2), pp. 36–36. doi: 10.1095/biolreprod.115.136432.

Sommer, R. S. *et al.* (2011) 'Holocene survival of the wild horse in Europe: a matter of open landscape?', *Journal of Quaternary Science*, 26(8), pp. 805–812. doi: 10.1002/jqs.1509.

Song, N. *et al.* (2013) 'Identification and classification of feline endogenous retroviruses in the cat genome using degenerate PCR and in silico data analysis', *Journal of General Virology*, 94(Pt_7), pp. 1587–1596. doi: 10.1099/vir.0.051862-0.

Sperber, G. *et al.* (2009) 'RetroTector online, a rational tool for analysis of retroviral elements in small and medium size vertebrate genomic sequences.', *BMC bioinformatics*, 10 Suppl 6, p. S4. doi: 10.1186/1471-2105-10-S6-S4.

Sperber, G. O. *et al.* (2007) 'Automated recognition of retroviral sequences in genomic data—RetroTector©', *Nucleic Acids Research*, 35(15), pp. 4964–4976. doi: 10.1093/nar/gkm515.

Stamatakis, A. (2014) 'RAxML version 8: a tool for phylogenetic analysis and post-analysis of large phylogenies', *Bioinformatics*, 30(9), pp. 1312–1313. doi: 10.1093/bioinformatics/btu033.

Stefanetti, V. *et al.* (2016) 'High Expression of Endogenous Retroviral Envelope Gene in the Equine Fetal Part of the Placenta', *PLOS ONE*, 11(5), p. e0155603. doi: 10.1371/journal.pone.0155603.

Steinbiss, S., Gremme, G., *et al.* (2009) 'AnnotationSketch: a genome annotation drawing library', *Bioinformatics*, 25(4), pp. 533–534. doi: 10.1093/bioinformatics/btn657.

Steinbiss, S., Willhoeft, U., *et al.* (2009) 'Fine-grained annotation and classification of de novo predicted LTR retrotransposons', *Nucleic Acids Research*, 37(21), pp. 7002–7013. doi: 10.1093/nar/gkp759.

Steiner, C. C. and Ryder, O. A. (2011) 'Molecular phylogeny and evolution of the Perissodactyla', *Zoological Journal of the Linnean Society*, 163(4), pp. 1289–1303. doi: 10.1111/j.1096-3642.2011.00752.x.

Storch, T. G. *et al.* (1985) 'Proliferation of infected lymphoid precursors before Moloney murine leukemia virus-induced T-cell lymphoma.', *Journal of the National Cancer Institute*, 74(1), pp. 137–43. Available at: http://www.ncbi.nlm.nih.gov/pubmed/3871491.

Subramanian, R. P. *et al.* (2011) 'Identification, characterization, and comparative genomic distribution of the HERV-K (HML-2) group of human endogenous retroviruses.', *Retrovirology*, 8, p. 90. doi: 10.1186/1742-4690-8-90.

Sugimoto, J. and Schust, D. J. (2009) 'Review: Human Endogenous Retroviruses and the Placenta', *Reproductive Sciences*, 16(11), pp. 1023–1033. doi: 10.1177/1933719109336620.

Sverdlov, E. D. (1998) 'Perpetually mobile footprints of ancient infections in human genome.', *FEBS letters*, 428(1–2), pp. 1–6. Available at: http://www.ncbi.nlm.nih.gov/pubmed/9645463.

Sverdlov, E. D. (2000) 'Retroviruses and primate evolution', *BioEssays*, 22(2), pp. 161–171. doi: 10.1002/(SICI)1521-1878(200002)22:2<161::AID-BIES7>3.0.CO;2-X.

Swanstrom, R. and Wills, J. (1997) 'Synthesis, Assembly, and Processing of Viral Proteins', in *Retroviruses*. Available at: http://www.ncbi.nlm.nih.gov/pubmed/21433349.

Tallmadge, R. L. *et al.* (2015) 'Bone marrow transcriptome and epigenome profiles of equine common variable immunodeficiency patients unveil block of B lymphocyte differentiation', *Clinical Immunology*, 160(2), pp. 261–276. doi: 10.1016/j.clim.2015.05.005.

Telesnitsky, A. and Goff, S. (1997) *Reverse Transcriptase and the Generation of Retroviral DNA*, *Retroviruses*. Available at: http://www.ncbi.nlm.nih.gov/pubmed/21433342.

Tempel, S. (2012) 'Using and Understanding RepeatMasker', in, pp. 29–51. doi: 10.1007/978-1-61779-603-6_2.

Trapnell, C. *et al.* (2012) 'Differential gene and transcript expression analysis of RNA-seq experiments with TopHat and Cufflinks', *Nature Protocols*, 7(3), pp. 562–578. doi: 10.1038/nprot.2012.016.

Trapnell, C., Pachter, L. and Salzberg, S. L. (2009) 'TopHat: discovering splice junctions with RNA-Seq', *Bioinformatics*, 25(9), pp. 1105–1111. doi: 10.1093/bioinformatics/btp120.

Tristem, M. *et al.* (1996) 'Characterization of a novel murine leukemia virus-

related subgroup within mammals.', *Journal of virology*, 70(11), pp. 8241–6. Available at: http://www.ncbi.nlm.nih.gov/pubmed/8892961.

Tristem, M. (2000) 'Identification and characterization of novel human endogenous retrovirus families by phylogenetic screening of the human genome mapping project database.', *Journal of virology*, 74(8), pp. 3715–30. Available at: http://www.ncbi.nlm.nih.gov/pubmed/10729147.

'UniProt: the universal protein knowledgebase' (2017) *Nucleic Acids Research*, 45(D1), pp. D158–D169. doi: 10.1093/nar/gkw1099.

Varmus, H. E., Quintrell, N. and Ortiz, S. (1981) 'Retroviruses as mutagens: insertion and excision of a nontransforming provirus alter expression of a resident transforming provirus.', *Cell*, 25(1), pp. 23–36. Available at: http://www.ncbi.nlm.nih.gov/pubmed/6268304.

Verdonck, K. *et al*. (2007) 'Human T-lymphotropic virus 1: recent knowledge about an ancient infection', *The Lancet Infectious Diseases*, 7(4), pp. 266–281. doi: 10.1016/S1473-3099(07)70081-6.

Villesen, P. *et al*. (2004) 'Identification of endogenous retroviral reading frames in the human genome.', *Retrovirology*, 1, p. 32. doi: 10.1186/1742-4690-1-32.

Vilstrup, J. T. *et al*. (2013) 'Mitochondrial Phylogenomics of Modern and Ancient Equids', *PLoS ONE*. Edited by C. Lalueza-Fox, 8(2), p. e55950. doi: 10.1371/journal.pone.0055950.

Vogt, V. (1997) *Retroviral Virions and Genomes*, *Retroviruses*. Cold Spring Harbor Laboratory Press. Available at: http://www.ncbi.nlm.nih.gov/pubmed/21433348.

Wade, C. M. *et al*. (2009) 'Genome Sequence, Comparative Analysis, and Population Genetics of the Domestic Horse', *Science*, 326(5954), pp. 865–867. doi: 10.1126/science.1178158.

Waku, D. *et al*. (2016) 'Evaluating the Phylogenetic Status of the Extinct

Japanese Otter on the Basis of Mitochondrial Genome Analysis', *PLOS ONE*, 11(3), p. e0149341. doi: 10.1371/journal.pone.0149341.

Walker, R. (1969) 'Virus associated with epidermal hyperplasia in fish.', *National Cancer Institute monograph*, 31, pp. 195–207. Available at: http://www.ncbi.nlm.nih.gov/pubmed/5393702.

Wallner, B. *et al.* (2013) 'Identification of Genetic Variation on the Horse Y Chromosome and the Tracing of Male Founder Lineages in Modern Breeds', *PLoS ONE*, 8(4), p. e60015. doi: 10.1371/journal.pone.0060015.

Wang, X. *et al.* (2012) 'Random X inactivation in the mule and horse placenta', *Genome Research*, 22(10), pp. 1855–1863. doi: 10.1101/gr.138487.112.

Warmuth, V. *et al.* (2011) 'European domestic horses originated in two holocene refugia.', *PloS one*, 6(3), p. e18194. doi: 10.1371/journal.pone.0018194.

Weiss, R. (1993) 'How does HIV cause AIDS?', *Science*, 260(5112), pp. 1273–1279. doi: 10.1126/science.8493571.

Weiss, R. A. (1996) 'Retrovirus classification and cell interactions.', *The Journal of antimicrobial chemotherapy*, 37 Suppl B, pp. 1–11. Available at: http://www.ncbi.nlm.nih.gov/pubmed/8818825.

Wheeler, T. J. and Eddy, S. R. (2013) 'nhmmer: DNA homology search with profile HMMs', *Bioinformatics*, 29(19), pp. 2487–2489. doi: 10.1093/bioinformatics/btt403.

Wickham, H. (2016) *ggplot2: Elegant Graphics for Data Analysis*. Springer-Verlag New York.

Wilkinson, D. A., Mager, D. L. and Leong, J.-A. C. (1994) 'Endogenous Human Retroviruses', in *The Retroviridae*. Boston, MA: Springer US, pp. 465–535. doi: 10.1007/978-1-4899-1730-0_9.

Williams, K. J. and Loeb, L. A. (1992) 'Retroviral Reverse Transcriptases: Error Frequencies and Mutagenesis', in, pp. 165–180. doi: 10.1007/978-3-642-77011-

1_11.

Wilson, D. E. and Reeder, D. M. (2005) *Mammal species of the world: a taxonomic and geographic reference*. JHU Press.

Withers-Ward, E. S. *et al.* (1994) 'Distribution of targets for avian retrovirus DNA integration in vivo.', *Genes & Development*, 8(12), pp. 1473–1487. doi: 10.1101/gad.8.12.1473.

Wu, X. (2003) 'Transcription Start Regions in the Human Genome Are Favored Targets for MLV Integration', *Science*, 300(5626), pp. 1749–1751. doi: 10.1126/science.1083413.

Xiong, Y. and Eickbush, T. H. (1990) 'Origin and evolution of retroelements based upon their reverse transcriptase sequences.', *The EMBO journal*, 9(10), pp. 3353–62. Available at: http://www.ncbi.nlm.nih.gov/pubmed/1698615.

Xu, Z. and Wang, H. (2007) 'LTR_FINDER: an efficient tool for the prediction of full-length LTR retrotransposons', *Nucleic Acids Research*, 35(Web Server), pp. W265–W268. doi: 10.1093/nar/gkm286.

Yoshinaka, Y. *et al.* (1985) 'Murine leukemia virus protease is encoded by the gag-pol gene and is synthesized through suppression of an amber termination codon.', *Proceedings of the National Academy of Sciences*, 82(6), pp. 1618–1622. doi: 10.1073/pnas.82.6.1618.

Zaitseva, L., Myers, R. and Fassati, A. (2006) 'tRNAs promote nuclear import of HIV-1 intracellular reverse transcription complexes.', *PLoS biology*, 4(10), p. e332. doi: 10.1371/journal.pbio.0040332.

Zhu, H. *et al.* (2018) 'Database-integrated genome screening (DIGS): exploring genomes heuristically using sequence similarity search tools and a relational database', *bioRxiv*. doi: 10.1101/246835.

Zielonka, J. *et al.* (2009) 'Restriction of Equine Infectious Anemia Virus by Equine APOBEC3 Cytidine Deaminases', *Journal of Virology*, 83(15), pp. 7547–

7559. doi: 10.1128/JVI.00015-09.