The *in silico* Identification and Analysis of Ancient and Recent Endogenous Retroviruses in Mammalian Genomes

ADAM LEE

IMPERIAL COLLEGE LONDON

DEPARTMENT OF BIOLOGY

PhD Thesis

ABSTRACT

Recent advances in DNA sequencing technologies have led to a vast plethora of vertebrate genomes being made available for bioinformatic analysis and investigation. This has presented retrovirologists with many new opportunities to study endogenous retroviruses (ERVs) – selfish genetic element (SGEs) endogenised within the genomic DNA of their hosts. Many of these ERVs exist as molecular fossils of past germline infections by their exogenous counterparts, representing approximately 8-10% of mammalian genomes. While the majority are thought to be inactive today, one particular retroviral group – HERV-K(HML-2) – has been implicated in recent activity. In this thesis, efficient, synergistic *in silico* techniques have been implemented, with which intensive, genome-wide retroviral screens were performed. This has culminated in the identification of 11 novel, insertionally polymorphic human ERVs (HERVs), belonging to the HERV-K(HML-2) lineage, in two high-coverage archaic hominid genomes. This thesis also identifies the oldest ERV described to date – orthologous across all placental mammals – estimated to have endogenised in the germline of an ancestral mammal, 128-140 million years ago. Three SGEs, found to be endogenised within this ancient ERV, have also been described and assigned a minimum age of 104 million years, making these the oldest, definitively dated SGEs. This thesis also presents a computer program for renaming all identified ERVs in vertebrate genomes, according to a newly designed nomenclature standard to be implemented globally, that aims to unambiguously catalogue all the ERVs identified, to date.

Declaration of Originality

I hereby declare that the work presented in this PhD thesis is my own, and that all other scientific literature drawn upon is cited in the text, and referenced in the reference list. Exceptional to this, are the perl computer programming scripts in appendices G and L, which were modified by Dr. Derek Huntley and Dr. Steve Cook, where stated in the text. Furthermore, chapter 4 reports my input on a collaborative project to reinvent existing nomenclature conventions, imposing a new standard by which all endogenous retrovirus insertions will be named hereforth. The work and results presented in this chapter are my own.

Copyright Declaration

The copyright of this thesis rests with the author and is made available under the Creative Commons Attribution Non-Commercial No Derivatives licence. Researchers are free to copy, distribute or transmit the thesis on the condition that they attribute it, that they do not use it for commercial purposes and that they do not alter, transform or build upon it. For any reuse or distribution, researchers must make clear to others the licence terms of this work.

Acknowledgements

I would like to firstly thank my supervisor, Dr. Michael Tristem, for his help, care, unwavering support and guidance throughout my PhD. I could not have asked for a better supervisor. I would also like to thank Dr. Derek Huntley for his invaluable help with computing, sharing his programming knowledge, and for granting me access to Imperial's high-performance computing facilities; also for his unwavering support and friendship over the last year. I would also like to thank Dr. Steve Cook for always offering me invaluable insight into approaching programming, and for helping me get started, and continuing to support me throughout my PhD. It is much appreciated. I also thank Dr. Magda Charalambous for her valuable advice and encouragement. I would also like to thank my family and, in particular my mother Mona Lee, for her unwavering support and meaningful discussion, throughout. To everyone else who has been there for me and offered me encouragement – I thank you, sincerely.

Table of Contents

Title Page	1
Abstract	2
Declaration of Originality	3
Copyright Declaration	3
Acknowledgements	4

CONTENTS:

Chapter 1 - Introduction	9
1.1 General Introduction and Overview	10
1.2 Retroviral Genome and Life Cycle	14
1.2.1 *gag*	18
1.2.2 *pro* and *pol*	18
1.2.3 *env*	22
1.2.4 Accessory Genes	23
1.2.5 Packaging Signal	24
1.2.6 Long Terminal Repeats (LTRs)	24
1.3 ERV Classification and Nomenclature	26
1.4 Examples of Retroviral Groups	30
1.4.1 ERV-L	30
1.4.2 HERV-K	31
1.5 Retroelements	31
1.5.1 LINEs	33
1.5.2 SINEs	35
1.5.3 LTR Retrotransposons	35
1.6 Age Estimation and Dating of ERVs	37
1.7 ERV Activity and Implications in Disease	40
1.7.1 Activity	40
1.7.2 Disease	41
1.8 Co-Option and Exaption of Retroviral Genes	43
1.9 Next Generation Sequencing	45
1.10 Bioinformatic Tools	49
Chapter 2 - Identification of an Ancient Endogenous Retrovirus, Predating the Divergence of the Placental Mammals	53

2.1 Introduction 54

2.2 Materials and Methods 58
2.2.1 Element Retrieval 58
2.2.2 Sequence Alignment 61
2.2.3 SGE Identification and Annotation 62
2.2.4 Species-Specific Insertions 62
2.2.5 Sequence Homology 63
2.2.6 Phylogeny Construction 64
2.2.7 Dating and Initial Infection 64
2.2.8 Measuring Selection 65
2.2.9 Synteny 66

2.3 Results 67
2.3.1 Screening 67
2.3.2 Orthology 67
2.3.3 Orthologous SGE Insertions 68
2.3.4 Afrotherian Region of Low Sequence Homology 75
2.3.5 Unidentified Sequence 76
2.3.6 Sequence Homology 76
2.3.7 Co-Phylogenies 77
2.3.8 Dating the Initial Infection Event 79
2.3.9 Selection 81
2.3.10 Synteny 81

2.4 Discussion 82

Chapter 3 - Archaic Hominid Retroviral Screens 90
3.1 Introduction 91
3.2 Materials and Methods 95
3.2.1 Data Preparation and Formatting 95
3.2.2 Retroviral Screen 96
3.2.3 Initial Verification and Read Counts 96
3.2.4 Back-Filtering 99
3.2.5 Second-Stage HERV-K(HML-2) Screening 101
3.2.6 Matching TSDs 101
3.2.7 Fuzzy-Matching 101
3.2.8 Fuzzy-Match Filtering 102
3.2.9 Screen Comparisons 102
3.2.10 Locus Mapping 103

3.3 Results 103
3.3.1 Stringent-Matching Retroviral Screen
of High-Coverage Denisovan and Neanderthal Genomes. 103
3.3.2 Stringent-Matching Loci: Comparisons with
Other Insertionally Polymorphic HERVs 105

3.3.3 Fuzzy-Matching Retroviral Screen of High-Coverage Denisovan Genome 108
3.3.4 Fuzzy-Matching Loci: Comparisons with Other Insertionally Polymorphic HERVs 109
3.3.5 Template Switching 110

3.4 Discussion 111

Chapter 4 - ERV Nomenclature: Implementing a Revised System for Naming Endogenous Retroviruses 120

4.1 Introduction 121
4.2 Methods and Results 124
4.2.1 Data Used 124
4.2.2 Data Capture 125
4.2.3 Leading Zeros 127
4.2.4 Genus Assignment 128
4.2.5 Name Printing 129
4.2.6 Output Files 130
4.2.7 Feedback 130
4.2.8 Sorting 131
4.2.9 ERV Names 138
4.2.10 Unassigned Genera 139
4.2.11 Final Tabulated Output 139

4.3 Discussion 140
4.3.1 General Discussion 140
4.3.2 Future Developments and Considerations 141

Chapter 5 - General Discussion and Future Developments 144

5.1 General Discussion and Future Developments 145

References 152

Appendix A 159
Appendix B 172
Appendix C 173
Appendix D 176
Appendix E 180
Appendix F 181
Appendix G 182
Appendix H 183
Appendix I 184
Appendix J 189
Appendix K 190
Appendix L 195

Figures:

Figure 1.1 11
Figure 1.2 13
Figure 1.3 16
Figure 1.4 17
Figure 1.5 20
Figure 1.6 21
Figure 1.7 27
Figure 1.8 34
Figure 1.9 39
Figure 1.10 43
Figure 1.11 51
Figure 2.1 55
Figure 2.2 60
Figure 2.3 69
Figure 2.4 70
Figure 2.5 71
Figure 2.6 72
Figure 2.7 74
Figure 2.8 78
Figure 2.9 79
Figure 2.10 82
Figure 3.1 98
Figure 3.2 100
Figure 3.3 104
Figure 3.4 108
Figure 3.5 110
Figure 4.1 123
Figure 4.2 127
Figure 4.3 128
Figure 4.4 129
Figure 4.5 132

Tables:

Table 1.1 32
Table 2.1 59
Table 2.2 76
Table 3.1 106
Table 3.2 109
Table 4.1 139

Chapter 1: Introduction

1.1 General Introduction and Overview:

Retroviruses are a family of non-icosahedral enveloped RNA viruses, which share a unique method of replication, centred on a reversal of the usual flow of biological information. Exogenous retroviral particles comprise two copies of a single-stranded RNA genome which, upon infection, is reverse transcribed into a DNA copy by a virally encoded reverse transcriptase (RT) enzyme. This DNA copy is subsequently inserted into the chromosomal DNA of the host cell, facilitated by a viral integrase (IN) enzyme. Endogenous retroviruses (ERVs) occur when an exogenous retrovirus (XRV) infects the chromosomal DNA of the host organism in this way. This host-integrated retroviral element is termed a DNA provirus. ERV proviruses make up a significant fraction of mammalian genomes, comprising approximately 8% of euchromatin in humans, numbering a total of some 450,000 loci (Lander *et al*., 2001). If infection occurs in a germ cell, the resulting proviral element can be vertically passed down from parent to offspring, remaining in the genome over generations (Boeke & Stoye, 1997). Unlike the replication cycles of other viruses, retroviruses do not typically lyse the host cell, enabling a more long-term association between parasite and host (Coffin, 1997). This is particularly advantageous to the persistence of viral DNA in the germline. Thus, ERV insertions can provide an informative molecular fossil record captured within the genomes of their hosts, and are thought to vastly outnumber their exogenous counterparts (Jern *et al*., 2005a).

Amongst the first retroviral diseases to be described, were two belonging to the genus *Alpharetrovirus*. These were Avian Leukemia Virus (ALV), discovered by Ellen and Bang in 1908 and Rous Sarcoma Virus (RSV) discovered in chickens by Peyton Rous in 1911. It wasn't until 1936 that the first mammalian retrovirus was discovered in mice, by John Bittner. This virus – mouse mammary tumour virus (MMTV) – was first noticed when adoptive nursing appeared to transmit murine breast cancer to offspring (Cohen & Varmus, 1979). Further research led to Temin's DNA provirus hypothesis in 1976, where he proposed that RNA tumour viruses would create and integrate

a cDNA copy of their genomes into the chromosomal DNA of the host – a notion supported by the earlier characterisation of the reverse transcriptase enzyme in 1970 (Temin, 1964). The first documented discovery of a germline-integrated ERV occurred in the late 1960s, where inherited cellular genes with the ability to encode retroviral gene products were observed (Weiss & Vogt, 2011). Hybridisation experiments confirmed the heritable nature of retroviral genomes, with possible involvements in various cancers spurring researchers to find out more about this newly discovered genetic element (Temin & Rubin, 1958). It was these initial, pivotal discoveries that fuelled academic interest in retroviruses, at large.

Mature, infective exogenous retrovirus particles are termed virions. They are spherical in shape, and are enveloped in a host cell-derived lipid bilayer, with diameters ranging from 80-100 nm. The viral envelope contains ~8 nm glycoprotein surface projections throughout, giving these virions a characteristic knob-studded appearance (Figure 1.1). Exogenous retroviruses contain an internal core encapsidating the viral RNA genome and its associated replicative enzymes (Johnson, 2008).

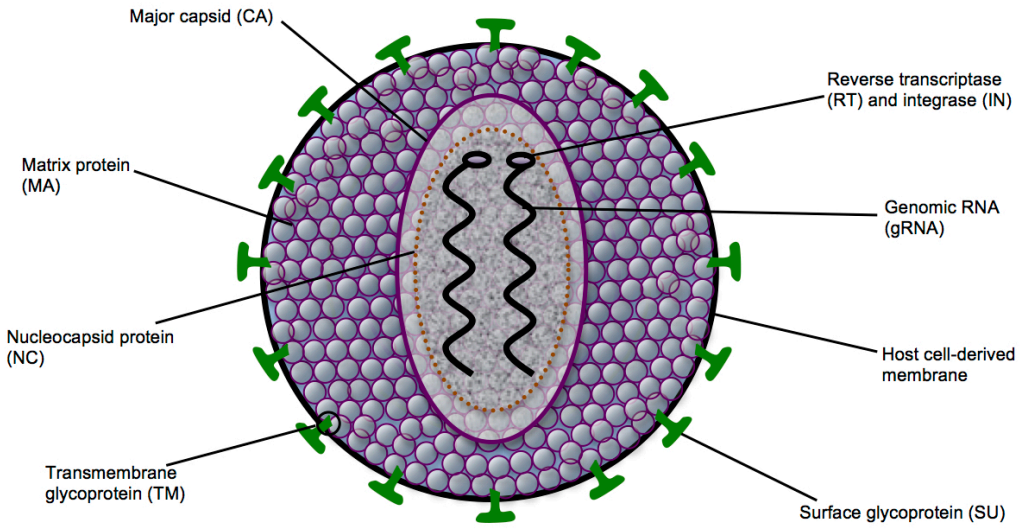


Figure 1.1: **Retrovirus virion morphology**. An mature, exogenous retroviral particle. Two copies of the viral RNA genome are bound to Pol products – RT and IN. This is encased in nucleocapsid (NC) protein, which is further encased within the major capsid (CA). This viral core exists within a matrix (MA), which is enveloped in a host cell-derived surface membrane, studded with Env-encoded surface (SU) and transmembrane (TM) glycoproteins, giving the virus a knob-studded appearance. Modified from *Virus Taxonomy: Seventh Report of the International Committee on Taxonomy of Viruses* (van Regenmortel *et al*., 2000).

In order for an ERV to become heritable within its host, a replication-competent XRV must infect a germ cell, resulting in the integration of its DNA provirus into the chromosomal DNA of oocytes or spermatozoa. This initial endogenisation is termed colonisation (Figure 1.2). Once germline integration is complete, the DNA provirus can be transmitted vertically to host progeny. Each independent colonisation event defines a new ERV lineage (Tristem, 2000). These lineages comprise paralogous copies of the colonising element, and can arise either by retrotransposition, or as a result of active ERV integrations producing replication-competent XRVs, which can reinfect the host genome. Within a relatively short period of time, it may be possible for ERVs to achieve high copy numbers in this way. However, the rate of this amplification typically declines with time. Purifying selection pressures can lead to ERVs eventually accruing an increasing number of deactivating mutations incurred by a variety of different mechanisms. Conversely, declining ERV amplification rates can also arise from an absence of purifying selection pressures that might otherwise have maintained the integrity of viable ERV sequences (Gifford & Tristem, 2003). It has been shown that ERVs can replicate to form a pool of infectious viral particles, and it is this pool upon which purifying selection can act (Belshaw *et al*., 2004). Many ERV lineages can also reset their fitness, either by replication, or by recombination occurring between the diploid RNA genomes. In fact, expression and replication of some ERV lineages may still be possible – even after millions of years – and has been implicated in certain human endogenous retrovirus (HERV) lineages (Belshaw *et al*., 2004).

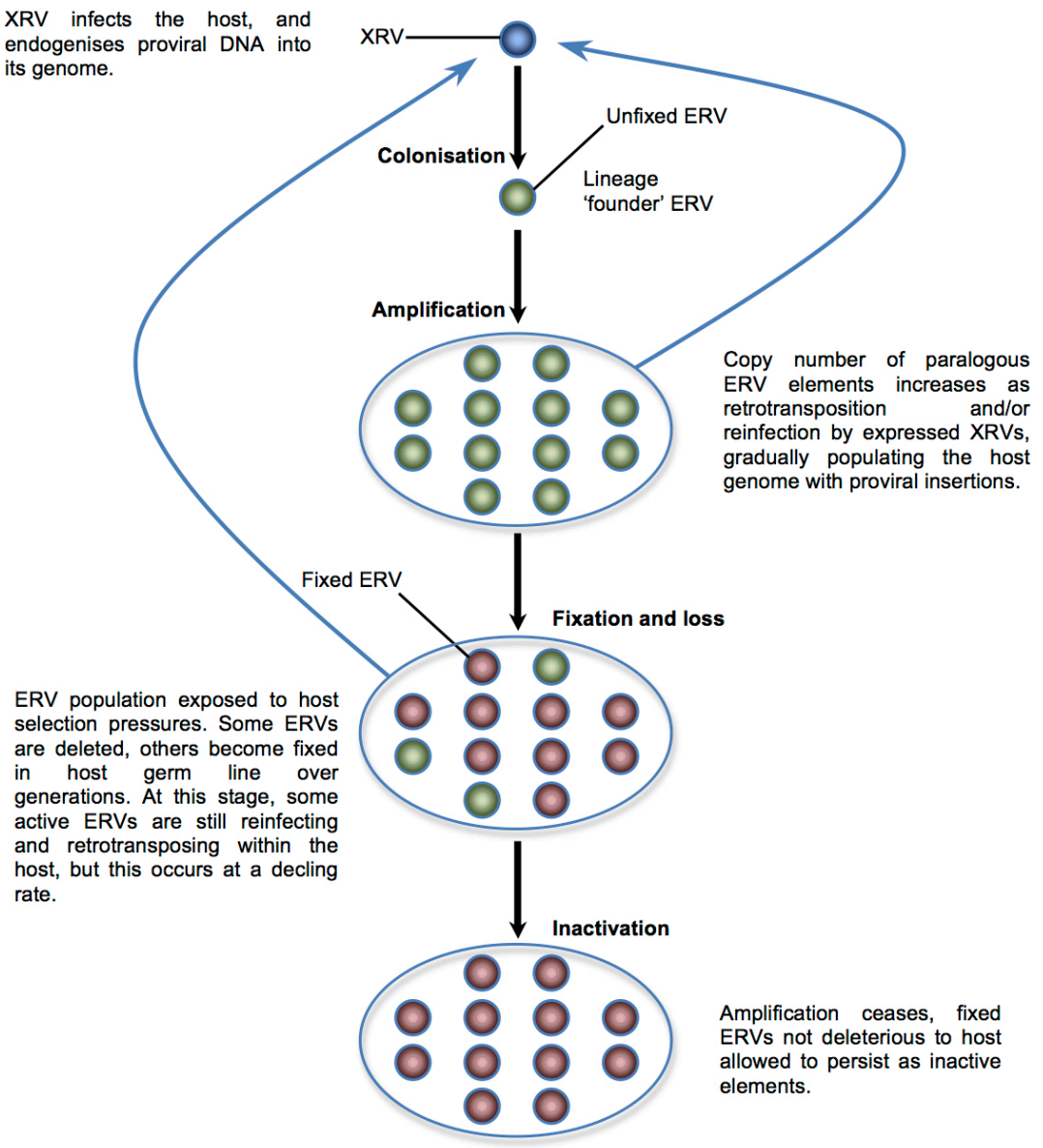


Figure 1.2: **Formation and fixation of an ERV lineage**. The colonization, amplification, fixation, loss and inactivation stages of an ERV lineage (modified from Gifford & Tristem, 2003).

It was previously thought that mutations resulting from normal genetic processes within the host would accumulate irreversibly within ERVs to the point of inactivation (Barbulescu *et al*., 2001). However, complete deactivation can also result from the interruption of coding regions of ERV insertions by large indels (insertions and deletions) often caused by recombination events or disruption of ERVs by other selfish genetic elements (SGEs). Once an ERV insertion is made defective, it is usually neutral to the host, and may persist in the germline across generations, with many ERVs eventually drifting to

fixation within the host species (Gifford & Tristem, 2003). Given that the majority of existing ERV insertions are defective, their evolution is largely constricted to the neutral evolution rate of the host organism itself. In host species with long generation times, these heritable ERVs act to slow the rapid rates of evolution that retroviruses would undergo as exogenous particles, potentially providing valuable insight into retroviral evolution.

Retroviral insertion is essentially a non-targeted, random process, such that the chromosomal locus of each retroviral element is unique (Stoye, 2012). While it has long been assumed that all ERV insertions in mammalian genomes are defective, the distinction between ERVs and XRVs has recently become less clear, since some ERVs have been demonstrated to retain activity, and regain their capability to express exogenous viral particles and reinfect host cells (Belshaw *et al*., 2004; Giminez *et al*., 2010).

1.2 Retroviral Genome and Life Cycle

Relative to many other viruses, retroviruses have small genomes, typically around 9-10 kilobases (Kb) in length. Despite this, they efficiently encode several protein products, and do so by utilising frame shifts and multiple overlapping open reading frames (ORFs), maximising the potential of their compact genomes (Temin, 1964; Nelson *et al*., 2003). Since retroviruses are RNA viruses, and are dependent on integration into host genomic DNA for gene expression, their promoter sequences are typically located at the 3′ end of the viral genomic RNA (gRNA) sequence. Thus, the DNA copy of the genome that results from reverse transcription, contains these sequences in the appropriate position, just ahead of the retroviral genes.

Most ERV genomes (Figure 1.3) characteristically comprise at least four genes – *gag, pro*, *pol* and *env* – flanked on either side by two identical, long terminal repeats (LTRs). The LTRs are typically between 300-1200 base pairs (bp) in length, while the enclosed viral genes total to anywhere from 5-10 Kb (Temin, 1964). In addition to these, retroviruses also have internal promoters; the first is referred to as the primer binding site (PBS), and the second as the

polypurine tract (PPT), with the latter being required for second strand DNA synthesis (Temin, 1985). The *gag* gene encodes structural proteins *via* a gag polyprotein precursor. Its derived name – '*g*roup specific *a*nti*g*en' – reflects its antigenic properties. The gag polyprotein is cleaved primarily into three mature proteins: these are the capsid (CA), nucleocapsid (NC) and matrix (MA). *pro* encodes viral aspartyl protease which acts last, to proteolytically process polyproteins encoded by the other viral genes, namely *gag* and *pol*, although it also affects the protein product of the *env* gene, in some cases. The *pol* gene encodes the viral reverse transcriptase (RT) enzyme, responsible for producing a DNA copy of the RNA genome, a process characteristic of retroviruses and other retroelements. Lastly, the *env* gene encodes viral glycoproteins that are expressed on the host cell-derived membranous envelope of infectious viral particles; products of *env* translation comprise surface glycoprotein (SU) and trans-membrane protein (TM) units. The latter governs important virus-host cell interactions, forming a complex that interacts with host cell surface receptors, mediating fusion of the virus and host cell membranes. This is the first crucial step of viral infection (Coffin *et al*., 1997).

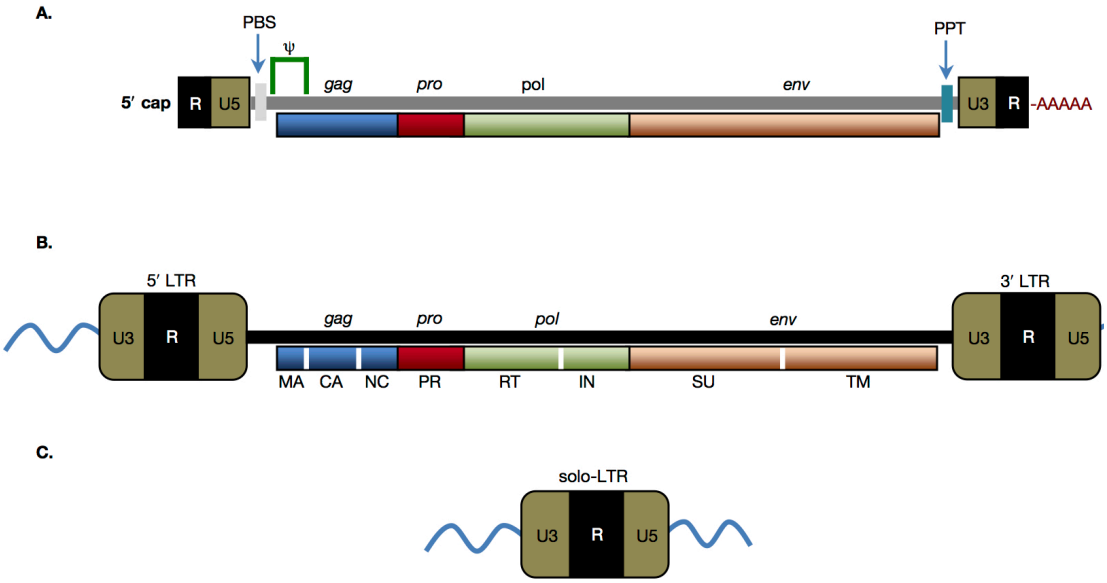


Figure 1.3: **ERV genomic structures**. (A) The viral RNA genome, with its 5′ cap, R and U5 repetitive and regulatory regions, PBS promoter, ψ packaging signal and coding genes (*gag, pro, pol* and *env*). The PPT second-strand DNA synthesis promoter is shown, together with U3, R and 3′ poly-adenylation signal or poly-A tail. (B) An integrated DNA provirus, with *gag, pro, pol* and *env* genes labeled on the top, and their encoded protein products labeled along the bottom. These include matrix protein (MA), major capsid (CA), nucleocapsid core (NC), protease (PR), reverse transcriptase (RT), integrase (IN), surface (SU) and transmembrane (TM) glycoproteins. (C) A solitary LTR (solo-LTR) structure that results from the deletion of ERV internal region, as a result of homologous recombination between an ERV's two LTRs (modified from Gifford & Tristem, 2003).

In order to gain entry into its host cell, the retrovirus' cell surface glycoproteins interact with receptors on the surface of the target cell. If the target cell exhibits the correct receptor for the infecting virus, fusion of the cell and virus membranes occur, so that the contents of the viral particle are released into the cytosol of the host cell. The viral gRNA is then reverse transcribed by RT into cDNA copies, which are trafficked to the nucleus, together with viral IN which integrates the DNA provirus into the chromosomal DNA of the host cell. Host cell machinery then transcribes viral RNA transcripts, which are translated on host cell ribosomes into the various viral protein products. Meanwhile, Env glycoproteins are embedded into the host cell membrane, in preparation for the emerging viral particles. Finally, the protein products together with newly translated viral gRNA, assemble at the cell surface, where infectious viral particles can now bud off from the host cell, ready to infect again (Figure 1.4) (Stoye, 2012).

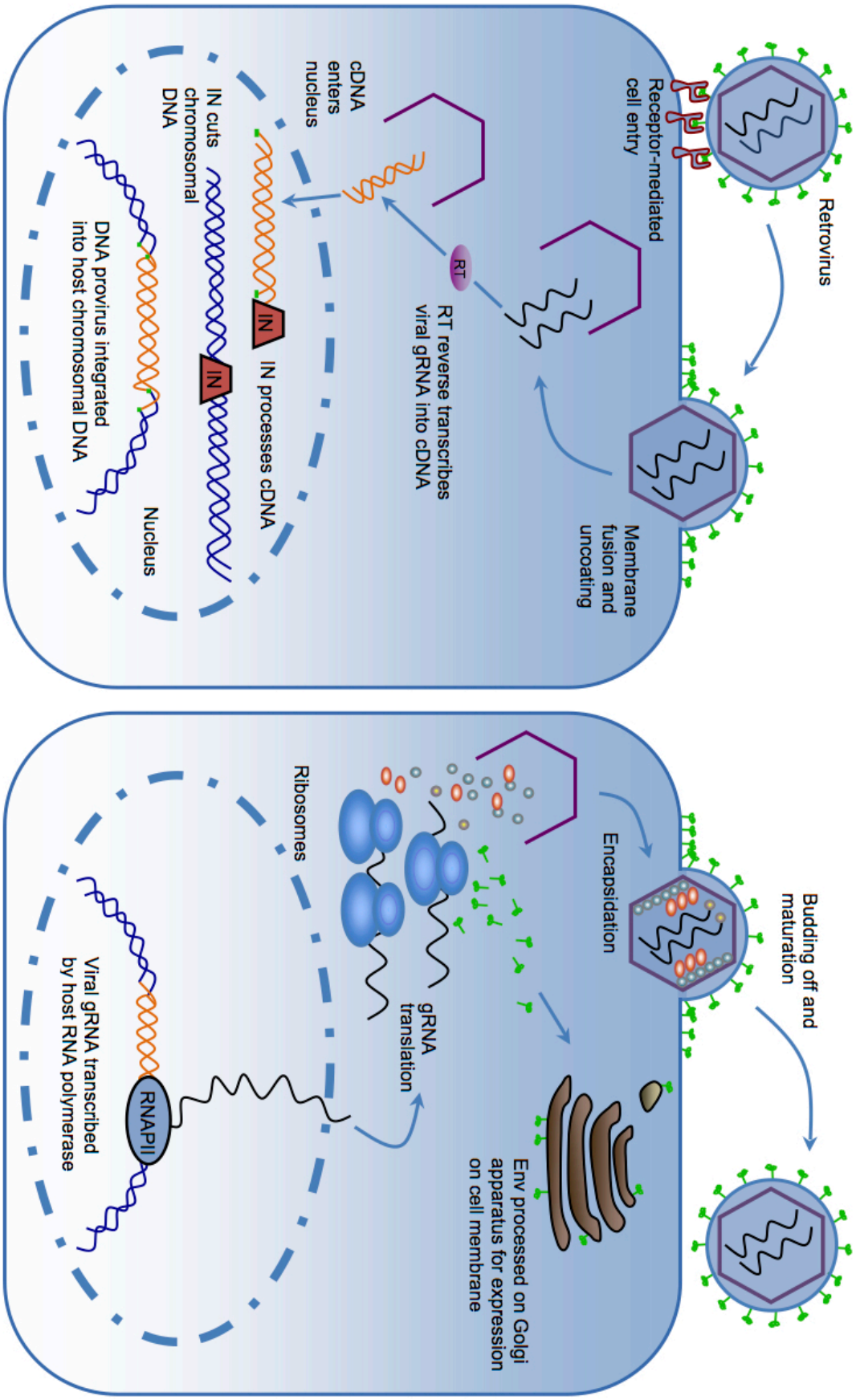


Figure 1.4: **Retroviral infection of a host cell** and replication cycle. Viral entry is shown on the left, and exit on the right. An infectious retrovirus approaches and infects a cell, reverse transcribing its RNA genome, and integrating a DNA provirus into the nuclear DNA of the host cell. This is transcribed and viral products are translated using host cell machinery, prior to assembly at the cell surface. The new infectious virus buds off from the host cell, and is ready to infect again (modified from Stoye, 2012).

1.2.1 *gag*: Proteins derived from the *gag* gene are proteolytically cleaved late in virion assembly, and it is thought that some 2,000 *gag* polyprotein precursors are contained in a single virion, comprising over 75% of the total retroviral protein content. Gag polyproteins are typically universally organised in a configuration of MA-CA-NC. The Gag polyprotein is functionally cleaved in this way in order to incorporate it into the virion. However, it has been found in its native form within the host cell in some instances, which is thought to help diffuse attempts by the host cell to localise it (Coffin *et al*., 1997).

Essential to retrovirus assembly, Gag proteins are typically myristoylated at the amino terminus, which allows them to be targeted to the host cell plasma membrane. Without myristoylation, viral budding does not occur, and Gag protein gradually accumulates within the host cell. In certain retroviruses, however, myristoylation does not occur, and other means of localising Gag proteins to the cell membrane are employed. Such examples include Equine Infectious Anemia Virus (EIAV) and Avian Sarcoma Leukosis Virus (ASLV), where modification of the amino terminal methionine residue by acetylation or phosphorylation has been observed (Coffin *et al*., 1997).

Furthermore, additional functions of Gag proteins have been proposed more recently. Prior to plasma membrane localisation, it has been observed that Gag proteins undergo transient nuclear trafficking in some retroviruses, such as RSV. It is thought that this can aid efficient packaging of viral gRNA into nascent virions. The mechanism is thought to involve viral gRNA inducing a conformational change in Gag, such that its Nuclear Export Signal (NES) is exposed, allowing for targeted nuclear export of Gag to the plasma membrane (Parent, 2011). Moreover, it has also been reported that viral gRNA appears to have a regulatory effect on Gag proteins, as well, whereby it is thought to regulate Gag distribution in the nucleus and cytosol of the host cell (Gudleski *et al*., 2010).

1.2.2 *pro* and *pol*: The *pro* and *pol* genes both follow the *gag* gene positionally in the retroviral genome, and encode important enzymology

required for replication. *pro* encodes a viral aspartyl protease, which has important roles in cleaving polyproteins into mature end products, whereas *pol* encodes a polypeptide from which RT and IN are derived (Boeke & Stoye, 1997; Jern *et al*., 2005b). These mature RT and IN molecules are cleaved from the Pol polypeptide by Pro once the virion has budded off from the host cell.

Pro and Pol initially coexist as part of a polyprotein, following translation. This polyprotein is linked to the Gag polyprotein, and both are trafficked to the cell surface membrane together. It is here, just prior to budding, that the viral protease cleaves viral polyproteins into their mature forms (Oroszlan & Luftig, 1990). In the majority of retroviruses, translation of the *gag-pol* gRNA results in the production of a ~55 kilodalton (kDa) Gag polyprotein, which is separated off prior to the translation of *pro* and *pol*. However, in approximately 5% of cases, a frame-shift occurs during translation, which can result in two possible fusion peptides. Frameshifting signals can occur in one of two locations: either before *pro* or just before *pol*. In the former, a Gag-Pro fusion peptide arises, since the Gag stop codon is no longer in frame. In the latter scenario, a Gag-Pro-Pol fusion peptide arises (Figure 1.5). In both cases, Pro exhibits autocatalytic activity, separating itself away from Gag and/or Pol by self-cleaving the peptide bonds holding them together. Following this separation event, Pro catalyses cleaving of Gag and Pol polyproteins in the normal manner (Dunn *et al*., 2002).

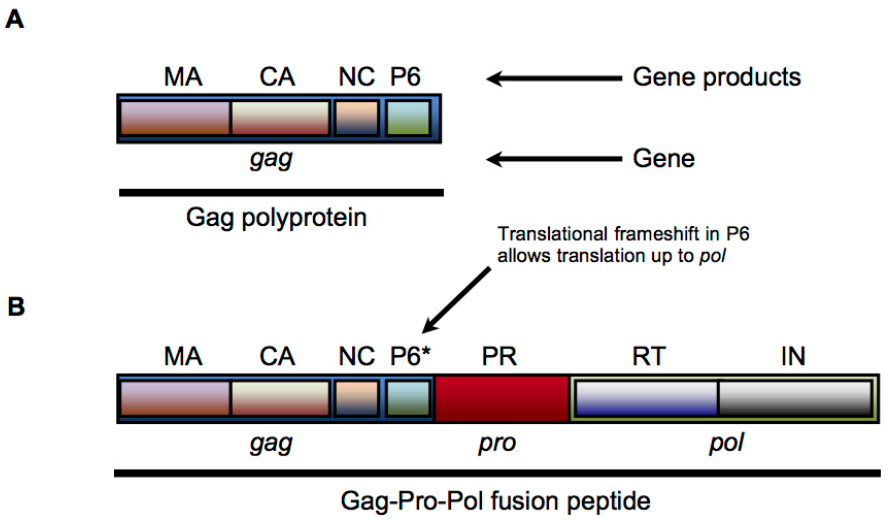


Figure 1.5: **Products of *gag* translation**. (A) A Gag polyprotein as would be expected without frameshifts. The stop codon at the end of the *gag* gene is intact, and so translation ends before the *pro* ORF. (B) One of two possible examples of fusion peptides generated when a frameshift occurs in the P6 region. The example shown is the Gag-Pro-Pol fusion peptide (modified from Dunn *et al*., 2002).

As aforementioned, the *pol* gene encodes a Pol polyprotein, from which RT and IN are cleaved by Pro. RT has two catalytic domains: one reverse transcribes RNA from DNA, and the other exhibits RNase H activity. During reverse transcription, viral gRNA is reverse transcribed by RT, using a intracellular supply of host cell nucleotides. This process is primed by the virus' first internal promoter: the PBS. As it does this, a temporary DNA-RNA intermediate is formed. This is separated by RT's RNase H activity, allowing for the gRNA's second internal promoter (the PPT) to prime second strand DNA synthesis. This results in a double-stranded DNA copy of the genome (Figure 1.6). RT typically has poor proof-reading activity giving high error rates of 10^{-4}-10^{-6} substitutions per base, per round of replication (Katz & Skalka, 1990). Contrastingly, DNA polymerase typically has an error rate of 10^{-7}-10^{-11}. This high error rate of viral RT is one of the factors influencing the rapid evolution of infectious retroviruses – a marked contrast from their fixed endogenised counterparts.

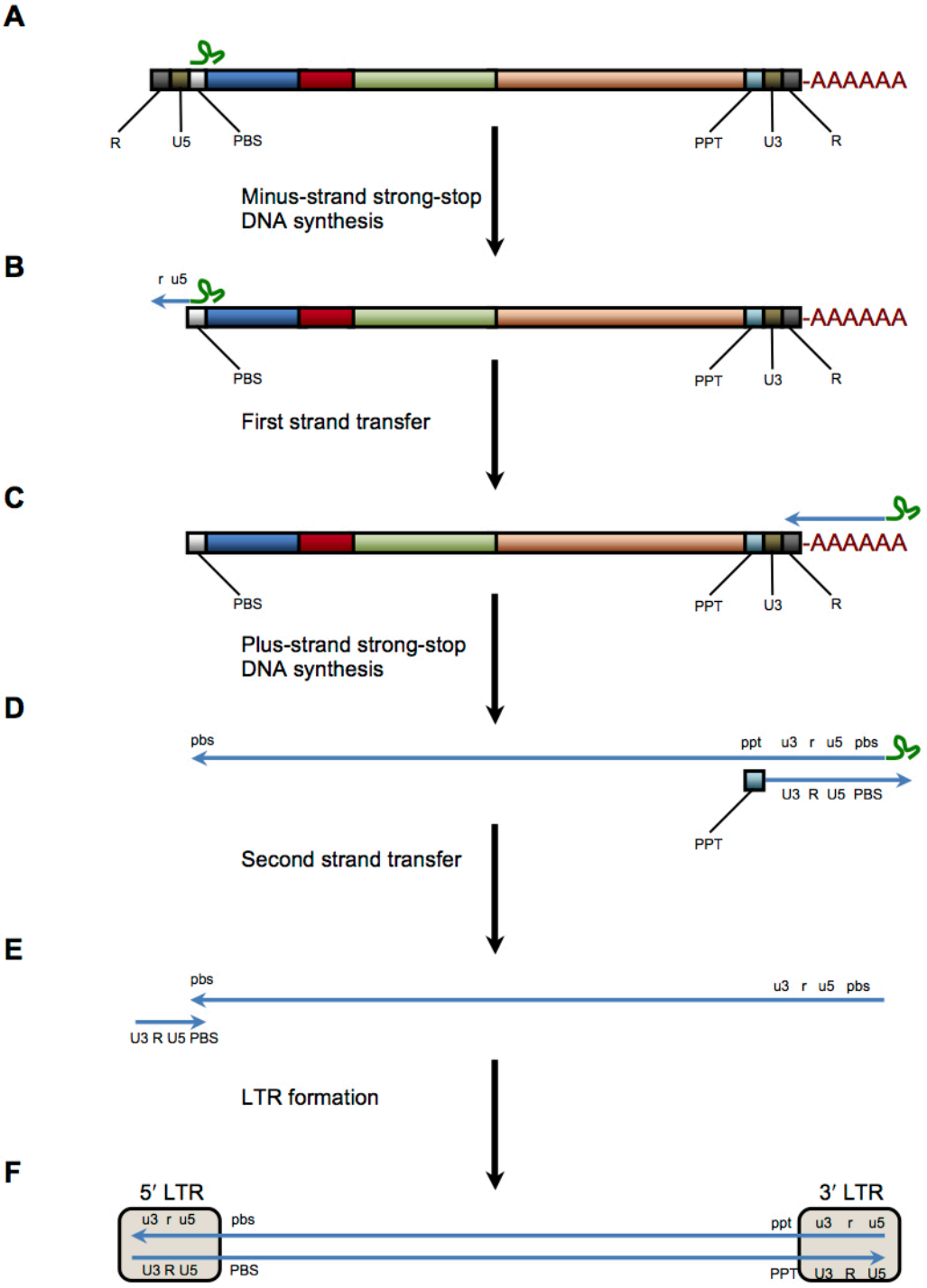


Figure 1.6: **Reverse transcription** – a critical step in retroviral replication. Viral RNA genome is coloured (A-C), with emerging DNA provirus depicted in blue (D-F). The tRNA primer is shown in green. cDNA synthesis is segmental, occurring in a stop-start fashion, and is closely followed by RNase-H removing residual nucleic acid templates (modified from Hunter, 2008).

The function of IN is to facilitate the integration of viral DNA into the chromosomal DNA of the host, where it can be transcribed by host cell's machinery (Temin, 1985). Following on from reverse transcription, the two ends of the viral genomic dsDNA bind to integrase, arranging the molecules in a close circle conformation as part of a preintegration complex, which is subsequently trafficked to the host cell nucleus. IN monomers or multimers are thought to bind to each end of the viral dsDNA, and subsequently form peptide bonds with one another, keeping the closed ring conformation intact (Murphy & Goff, 1992). Once inside the nucleus, the conformation is opened in close proximity to the host chromosomal DNA. This is accompanied by deleting the four central bases of the closed circle dsDNA duplex – a '-TG' from the 5′ end and '-CA' from the 3′ end of the linear dsDNA, creating an uneven, two base overhang at each end of the element. Both of these two-base sequences are highly conserved among many retrotransposable elements (Katz & Skalka, 1994).

This site-specific deletion of two bases from either end of the viral dsDNA – mediated by IN's endonuclease activity – exposes 3′ hydroxyl groups, recessed by two nucleotides. This is termed 'processing' the viral DNA. IN then creates a 4-6 bp staggered cleavage in the host chromosomal DNA, exposing 5′ phosphate groups. The processed viral dsDNA is then ligated into the cleaved chromosomal DNA. Since the overhang on the viral DNA is only 2 bp, and the staggered cut made by integrase in the chromosomal DNA is larger, host DNA repair mechanisms fill in the gaps, resulting in small target site duplications (TSDs) or direct repeats, either side of the viral insertion. The final integrated viral cDNA genome is termed a provirus, and is the final form in which ERVs exist within their hosts (Katz & Skalka, 1994).

1.2.3 *env*: *env* is one gene that delineates retroviruses from other retroelements. Responsible for encoding viral coat proteins expressed on the surface of the host cell-derived viral membrane, Env has also been shown to rely on Gag for trafficking to the cell surface during virion budding. Its most important function is enabling infectivity, and spreading of viruses between cells and individuals. Furthermore, a retrovirus' specific *env* gene can

determine the type of host cell to which it binds, since Env products mediate host cell receptor interactions (Löwer *et al*., 1996).

The Env polyprotein is cleaved into two functional, mature proteins – SU and TM. Prior to this, nascent Env polyproteins bind to signal recognition particles on the surface of the host cell endoplasmic reticulum (ER). They remain anchored to the ER membrane by a trans-membrane hydrophobic segment at the carboxyl terminus, where they are then trafficked through the host Golgi apparatus. It is here that a host cell protease cleaves the Env polyprotein into its mature SU and TM conformations. If uncleaved Env polyproteins are expressed on the viral surface, they can bind cell receptors, but cannot induce membrane fusion. Thus, this cleavage step is critical to infectivity of the resulting virion (Coffin *et al*., 1997).

1.2.4 Accessory Genes: While the four aforementioned genes are found across most of the *Retroviridae*, certain genera and retroviral groups contain additional accessory genes. One such example is the *dut* gene, which is common to all except for one lineage of HERV-K elements, and some *lentiviruses*. This gene encodes a dUTPase enzyme that catalyses the hydrolysis of dUTP into dUMP and pyrophosphate. It has been shown that encoding viral dUTPase has benefits during replication of the viral genome. dUMP is a precursor for the synthesis of TTP, reducing the cellular concentration of dUTP and maintaining a low dUTP:TTP ratio. This prevents RT from incorporating potentially mutagenic uracil into the emerging viral cDNA. In some lentiviruses, dUTPase-deficient mutants – while still able to replicate – were proven to replicate slower in non-dividing primary macrophages (Turelli *et al*., 1996).

Unlike many other genes, the *dut* accessory gene does not commonly occur at the same genomic locus, but is often found at different positions in the genome. In non-primate *lentiviruses*, for example, it is not found in the *pro* ORF, but in the *pol* ORF in B- and D-type retroviruses (Elder *et al*., 1992). In other ‘complex’ retroviruses, such as the *lentiviruses* and *spumaretroviruses*, accessory genes with regulatory effects on viral gene expression have also

been observed. Conversely, these accessory genes are often found in the distal segments of the genome, either between *pol* and *env* or downstream of *env* ahead of the U3 promoter region of the 3′ LTR. Oncogenic retroviral genes – termed *onc* genes – have also been observed, being named 'transforming viruses' due to their ability to induce oncogenic tumour formation in cultured cells (Coffin *et al*., 1997). There are many different potential accessory genes among different members of the *Retroviridae*, each adding further functional dimensions and complexity to such a compact genome.

1.2.5 Packaging Signal: Just before budding off, viral particles encapsidate newly translated viral gene products, together with two copies of their gRNA. This gRNA is therefore dimerised, and the determinants of this dimerisation reaction are encoded by a retroviral packaging signal (ψ) – a *cis*-acting RNA element approximately 80-150 bp long. This is located at the 5′ untranslated region (UTR), just downstream of the PBS, ahead of the *gag* gene. ψ contains a series of intermittently palindromic sequences, enabling it to form a contiguous series of four stem-loops, each separated by a short, double-stranded sequence linker (Skripkin *et al*., 1994).

These stem-loops are named SL1, SL2, SL3 and SL4. SL1 is a highly conserved hairpin, containing a primary dimer initiation site (DIS), which forms a 'kissing loop' intermediate that facilitates RNA dimerisation (Skripkin *et al*., 1994). According to the retrovirus group, each stem-loop may have slightly different properties, and even interactions with virus-specific accessory genes, as with HIV (Abbink *et al*., 2008).

1.2.6 Long Terminal Repeats (LTRs): Flanking the genetic and regulatory components of endogenised proviruses, are two identical LTRs. Upon integration, LTR sequences are identical, and serve as promoters that enable the expression of viral genes, using host cell machinery. The sizes of each LTR component can vary drastically between different retrovirus lineages, but rarely total to more than ~1.2 Kb per LTR (Coffin *et al*., 1997).

The 5′ end of retroviral gRNA is structured R-U5 and the 3′ end is structured R-U3, where R is a repetitive sequence common to both ends, U5 is the 5′-most region of the genome, and U3 is the 3′-most region. However, since each LTR in the provirus is identical – with the structure U3-R-U5 – their synthesis is somewhat unique relative to the rest of the viral genome. During reverse transcription, RT must undergo a series of two template switches, jumping from the 5′-most to the 3′-most end of the RNA genome. These jumps are termed strand-transfer reactions, and it is these that result in identical LTRs being generated (Figure 1.6). It is because these LTRs are synthesised afresh with each replication cycle, that they are always identical upon insertion, and remain unaffected by point mutations that may otherwise be picked up by retroviruses after several rounds of replication (Coffin *et al*., 1997).

Functionally, LTRs are intended to prime, promote, enhance and regulate the expression of proviral genes. Inadvertently, however, it has been observed and suggested that LTRs may also contribute to the regulation of genes within the host's chromosomal DNA. For example, ERV-derived LTRs can be exapted by the host, which can make use of their promoting properties to generate new patterns of gene expression. Alternatively, if gene expression is enhanced next to an oncogene within a dividing cell, it is quite possible that ERV-derived LTRs could be a causative agent of some cancers (Cohen *et al*., 2009).

Despite a great number of full-length ERV insertions remaining in the genomes of many species, these have always been vastly outnumbered by solitary LTR (solo-LTR) structures resulting from internal homologous recombination between the two LTRs of an ERV element (see Figure 1.3, C). This recombination event results in the enclosed viral genome being excised, leaving behind a single, solo-LTR in place of a complete provirus. It is thought that these solo-LTRs outnumber full-length elements at a ratio of at least 10:1 in mammalian genomes (Benachenhou *et al*., 2009).

Towards the final stages of the retroviral life cycle, newly synthesised viral protein products and two strands of gRNA assemble at the cell surface membrane, in preparation for budding off of the viral particle. Shortly after this, maturation of the nascent viral particle into a mature virion begins. Prior to this stage, the new XRV still contains its structural proteins as part of uncleaved polyproteins. Once the virus is independent of the host cell, polyproteins are cleaved by PR, and self-assembled into the internal matrix, major capsid and nucleocapsid comprising the viral core. Once this is complete, the virus is said to be mature, and is ready to infect a new host cell (Coffin *et al.,* 1997).

1.3 ERV Classification and Nomenclature:

The family *Retroviridae* contains two subfamilies: *Orthoretrovirinae* and *Spumaretrovirinae*. Retroviruses are further loosely incorporated into three phyletic classes: I, II and III (Figure 1.7). These are groups based on sequence similarity to infectious XRVs, and are further subcategorised into genera. The retroviral genera include: gammaretroviruses (mammalian type C), betaretroviruses (mammalian types B and D), spumaretroviruses (also known as foamy viruses), alpharetroviruses (avian type C), epsilonretroviruses (found in fish hosts), lentiviruses and deltaretroviruses; (Sverdlov, 2000). Class I ERVs are most similar to gammaretroviruses; class II ERVs are similar to beta- and alpharetroviruses, and class III ERVs are similar to spumaretroviruses (Griffiths, 2001; Doolitle *et al*., 1989). Whilst there are exceptions, most ERVs are no longer closely related to circulating XRVs, and are incapable of expressing viral particles. However, many ancient ERVs are interleaved with extant XRVs in phylogenies, suggesting that they are extinct sibling species of these XRVs, rather than their evolutionary ancestors (Johnson, 2008).

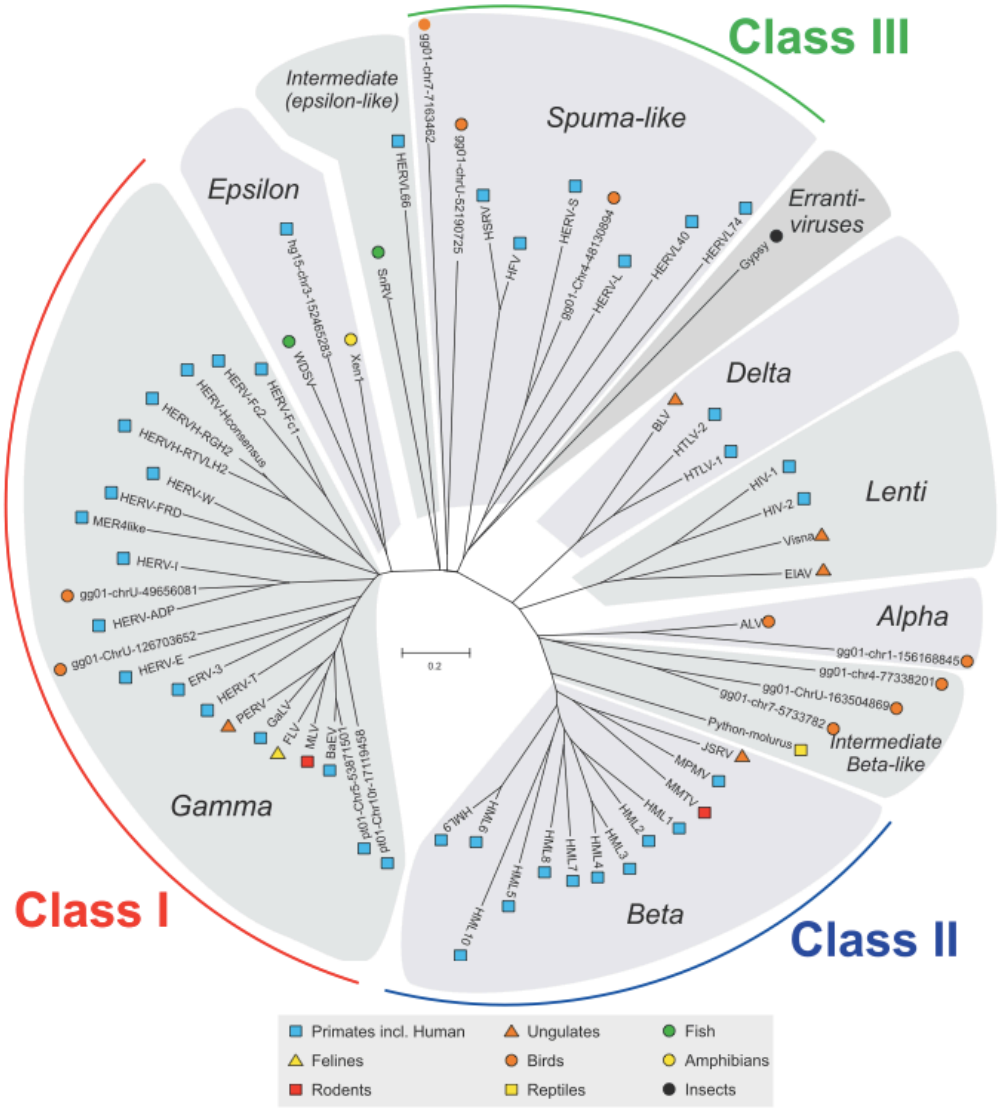


Figure 1.7: **Unrooted neighbour-joining (NJ) retroviral tree** / dendogram of the three classes, and seven genera of ERVs, based on the *pol* gene. Each branch on the tree represents a distinct retroviral lineage. The tree backbone structure is based on retrieved exogenous retroviral sequences from GenBank, to provide useful taxonomic reference (Jern *et al*., 2005).

Retroviral phylogenies are typically constructed from the *pol* gene, which is widely accepted to be the most conserved gene across the *Retroviridae*. Such phylogenies show clear clustering of ERVs into the three described classes, however the relationships between ERVs within a class are not always accurately represented by sequence similarities in the *pol* gene. Certain retroviruses may exhibit highly homologous *pol* genes, yet their LTRs may show virtually no homology at all; they are thus defined as distinct sub-groups based on this. Amongst other factors, recombination can effect such

differences, as it can happen between retroviral LTRs, or even full-length insertions (Johnson & Coffin, 1999). Classification attempts can be further hampered by the fact that many full-length ERV insertions may retain both LTRs, but lose much of the internal regions flanked by them (Mager, 2005).

Of the retroviral genera, *alpharetroviruses* include many avian retroviruses, and the very first discovered ERVs belonged to this genus. *Alpharetroviruses* are C-type, class II elements, with a simple genomic organisation. Some examples contain one or more accessory genes, but this is often not the case. The ERV internal region is approximately 6.8-9 Kb in length, and their LTRs are typically quite short (0.3 Kb). Phylogenetic analyses suggest a close relatedness between *alpharetroviruses* and *betaretroviruses*. *Betaretroviruses* are class II B- and D- type retroviruses, with LTRs varying between 0.3-1 Kb, and internal regions totalling approximately 7.5-9 Kb in size. They also encode the *dut* accessory gene, encoding a dUTPase enzyme, and can encode other accessory genes as well. *Deltaretroviruses* have, to date, only been described based on their exogenous form, and have not yet been identified as endogenised proviruses. They are characterised as class II, complex C-type retroviruses, with internal regions of 8-9 Kb and LTRs of 0.7 Kb in length. *Gammaretroviruses* are a class I, C-type simple retroviruses, with a 8-9 Kb internal region, and LTRs approximately 0.5 Kb long. They don't encode accessory genes. *Lentiviruses* are named based on the first member being discovered to induce slowly progressive paralytic symptoms in sheep, when discovered by Sigurdsson in 1954. They are also class II, complex retroviruses with internal regions of 8-9 Kb and a varied LTR size range from 0.2-1 Kb. They have all the typical retroviral genes, but can also express a variety of different accessory genes, several of which are common to all *lentiviruses* and others being unique to certain *lentiviral* groups. *Spumaretroviruses* are a class III, C-type element with an unusually large genome, consisting of an internal region measuring approximately 13 Kb in length, and LTRs of approximately 1.7 Kb. They, too, are capable of expressing various accessory genes. *Epsilonretroviruses* are common to fish hosts, and are class I, C-type retroviruses which exhibit a degree of relatedness to *gammaretroviruses*. However, unlike the latter,

epsilonretroviruses also express a specific set of accessory genes (Llorens *et al*., 2010). An outlier to these genera is the HERV-L retroviral group – which is considered to be an evolutionary intermediate between infectious retroviruses and LTR-retrotransposons, since it lacks the *env* gene (Bénit *et al*., 1999).

Within the human lineage, class I and III HERVs appear to be oldest, being present throughout the primates, while class II includes more recently active HERVs; such as the human-specific subgroup HERV-K(HML-2), active within the last 6 million years (Medstrand & Mager, 1998). HERV groups are mostly named according to the tRNA primer used at their PBS: HERV-L elements, for instance, are primed by tRNALys, whereas HERV-H elements are primed by tRNAHis; thus, they are designated a hyphenated, appended single-letter amino acid code (Renoux-Elbé *et al*., 2002).

ERVs are further subdivided into separate lineages – often erroneously termed 'families' – containing structurally heterogeneous, yet phylogenetically related elements (Lander *et al*., 2001). Whilst *Retroviridae* is also termed a family, many retrovirologists view this as a historical term, and accept the term family in reference to a single ERV lineage of paralogous copies, stemming from a single germline infection event. The uncertainty surrounding this term is one among many other uncertainties prevalent within the retroviral nomenclature and classification systems, and has sparked a long-emerging interest among retrovirologists to develop a new classification system. A system in which no ambiguities lie, and in which every bioinformatically identified ERV insertion is named, with an entirely unique identifier. This will allow for a comprehensive catalogue of every existing endogenised retrovirus identified in presently available genomic databases. This is further discussed in chapter 4 of this thesis, together with a program developed to run in conjunction with an existing ERV locus-mining pipeline. This program automatically generates names for retroviral insertions, in accordance with the new proposed ERV nomenclature standard (Blomberg *et al*., unpublished). Given these taxonomic developments, paralogous groups of related ERVs, such as HERV-K(HML-2), will not be referred to as families in this thesis, but instead as retroviral groups or lineages.

While there exists a sound taxonomic system for the classification of exogenous retroviruses (devised by the International Committee on Taxonomy of Viruses (ICTV)), the same cannot be said for ERVs, whose classification system is loosely derived from similarities of ERVs to their exogenous counterparts. Further complications arise where XRVs are pathogenic, since they are often named after their disease-causing phenotype. Moreover, since related ERVs often lack any phenotype, mirroring nomenclature off of their pathogenic exogenous counterparts is misleading (Blomberg *et al*., 2009).

1.4 Examples of Retroviral Groups

While there are many examples of retroviral lineages integrated in mammalian genomes, some have been researched in greater detail than others. Two examples of biologically important retrovirus groups are summarised in the following section.

1.4.1 ERV-L: Of all the described retroviral groups, one contains the most ancient virus dated at present. This group is ERV-L; the –L designation is appended to reflect its leucine tRNA primer. First discovered in 1995 (Cordonnier *et al*., 1995), HERV-L was characterised in humans, and found to be present across many mammalian orders, with copy numbers varying from 1-200. Unlike the majority of retroviruses, ERV-L proviruses are missing their *env* gene, and are thus incapable of producing viral particles. In chapter 2 of this thesis, and published in a paper in appendix A, the ERV-L group is shown to also contain the oldest discovered virus, with an estimated age of approximately 128 million years old. This element has a definitive minimum age of at least 104 million years, as demonstrated by orthology of the element in question across the four main mammalian superorders, suggesting it infected a common ancestor prior to the placental radiation (Lee *et al*., 2013).

1.4.2 HERV-K: Another particular retroviral group that is of consistent academic interest, is HERV-K, so named due to its lysine tRNA primer. The HERV-K group is thought to have first infected the germline of Old World primates, following their divergence from New World primates approximately 35 million years ago (Mayer & Meese, 2003; Cordonnier *et al*., 1995; Gifford & Tristem, 2003). Since this infection event, HERV-K elements are thought to have undergone several periods of expansion within primate genomes, evidenced by the large numbers of HERV-K elements in humans and African great apes; far smaller copy numbers are observed in lesser apes and orangutans. These lines of evidence point towards an integration time of approximately 8-15 million years ago (Johnson & Coffin, 1999; Barbulescu *et al*., 1999), with a number of human-specific elements indicating even more recent activity of this retroviral group (Barbulescu *et al*., 1999; Buzdin *et al*., 2003).

1.5 Retroelements

ERVs, together with another group of parasitic genetic elements known as retrotransposons, are collectively termed retroelements. Like ERVs, retrotransposons are a class of non-infectious selfish genetic elements (SGEs) that also frequently accumulate in the germlines of their hosts (Boeke & Stoye, 1997). They were first discovered in 1956, in maize (McClintock, 1956), and are thought to amount to some 42.2% of the human genome (Lander *et al*., 2001). Common to all retroelements is their incorporation of an RNA intermediate at some stage of their replicative life cycle, which is largely quite similar (Wilkinson *et al*., 1994). Endogenised retroviruses do differ from retrotransposons in a number of ways, however. The former encode a functional *env* gene, which encodes viral envelope proteins expressed on the surface of host cell-derived viral membrane, allowing for the formation of infectious viral particles. In most retrotransposons, this gene is either absent or non-functional (Lerat & Capy, 1999) and these retroelements self-propagate throughout the genome solely *via* retrotransposition. Conversely, ERVs are capable of propagating both by retrotransposition and, in the case of active elements, reinfection of the host by emerging infectious virions.

Retrolelements flanked by identical LTRs are termed LTR-retrotransposons, whilst those lacking LTRs are termed non-LTR retrotransposons; further subdivisions are based on characteristics such as self-encoding RT, or dependency on the RT generated by other elements during replication (Coffin *et al.*, 1997). Table 1.1 shows a list of the most commonly occurring types of retroelements within mammalian genomes.

Table 1.1: **Types of retroelements** found within mammalian genomes. Adapted from Boeke & Stoye, 1997; Strauss & Strauss, 2007.

Retroelement type	Description
ERV	Endogenous Retroviruses, found within vertebrates. Integrated DNA provirus stage. Often vertically transmitted *via* the germline. Elements are largely defective, and rarely infectious. Comprise large fraction of vertebrate genomes.
LTR retrotransposon	Retrotransposons flanked by identical (at the point of integration) LTRs. Not known to be infectious. Comprise large fraction of mammalian genomes.
Poly(A) retrotransposon	Non-LTR retrotransposons. Poly-A region at 3′ end.
Pararetrovirus	Infectious DNA viruses replicating *via* an RNA intermediate. Well-known examples include human hepadnavirus and caulimovirus.
Retrotranscript	Elements that do not encode RT, but transpose *via* RNA intermediate. May rely on poly(A)-retrotransposons for mobility.

While it's largely thought that retroelements are endogenised DNA SGEs, there are some that exist in extrachromosomal DNA, such as retrointrons and retroplasmids. These have not yet been identified in mammals. There are even infectious DNA elements that do not integrate into the host genome, but use a cellular RNA intermediate to replicate. These are pararetroviruses, and most famously include caulimovirus and hepadnavirus; the latter being a distant relative of retroviruses and an important human pathogen (Stoye, 2012).

Examples of different types of retrotransposon include abundant long-interspersed nuclear elements (LINEs) and short-interspersed nuclear elements (SINEs). They lack the extracellular life cycle stages common to all retroviruses, and are present at far higher frequencies within vertebrate genomes (Goodlier & Kazazian, 2008). Formerly viewed as largely parasitic

'junk' DNA, retroelements have been demonstrated to influence the function of host genes, and may have more involved roles in shaping genomes over time (Figure1.8) (Gogvadze & Buzdin, 2009).

1.5.1 LINEs: LINEs comprise the greatest fraction of the vertebrate genome relative to other retroelements, and achieve copy numbers of approximately 850,000 in the human genome. They are autonomous non-LTR retrotransposons, capable of self-replicating throughout the genome by retrotransposition, and they often have polyadenylated stretches of DNA at the 3′ terminus. LINEs total approximately 6 Kb in length, and contain an internal polymerase promoter, two primary ORFs, flanked on either side by untranslated regions (UTRs) and target site duplications (TSDs). The second ORF – known as ORF2 – encodes the RT needed for replication. Within the human genome, there are three families of LINE: LINE1, LINE2 and LINE3 (Lander *et al*., 2001). The most abundant of these is LINE1 (L1), which is the only one that remains active to the present day (Sassaman *et al*., 1997; Naas *et al*., 1998).

Once translated, LINE RNA is trafficked to the nucleus as part of a ribonucleoprotein complex containing LINE-encoded proteins. The RT in LINE elements is *cis*- acting, so its RNA template is the same RNA molecule from which it was itself translated. Here it is subjected to endonuclease activity, creating a single-stranded nick. This modified RNA is then used to prime RT to begin reverse transcription, which begins at the 3′ end, but often fails to continue all the way to the 5′ end, resulting in large numbers of truncated LINE repeats. These are often rendered non-functional as a direct result of this truncation. In L1 LINEs, active elements are found to average a larger 1,070 bp median length, while defective elements are often found at a mean of only 900 bp (Lander *et al*., 2001).

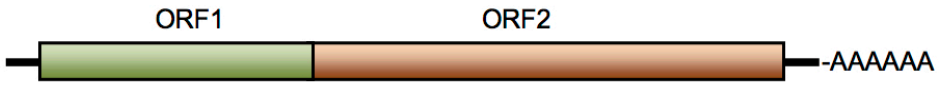
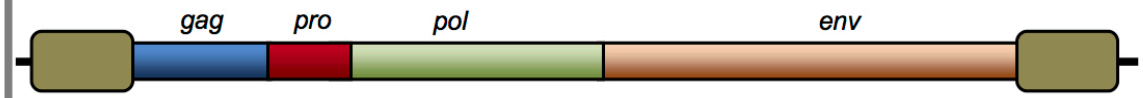
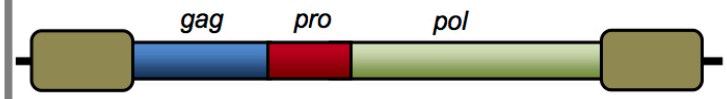
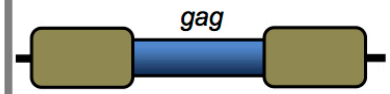
Retroelement	Schematic	Size	Copy Number	Human Genome Fraction
LINE		6 Kb	850,000	21%
SINE		1-400 bp	1.5M	13%
ERV		5-10 Kb	450,000	8-10%
Autonomous LTR Retrotrans-poson		6-11 Kb		
Non-autonomous LTR Retrotrans-poson		1.5-3 Kb	250,000	

Figure 1.8: **Examples of retroelements**. These five types of retroelement are abundant in mammalian genomes. Details of the range of sizes, copy number and occupied fraction of the human genome (as an example) are given, together with schematic diagrams showing the endogenised structures found in host DNA. LINEs, SINEs, ERVs and LTR retrotransposons are all depicted (modified from Lander *et al*., 2001).

1.5.2 SINEs: LINE replicative machinery can also be an integral replication cycle stage for other non-autonomous SGEs, such as non-autonomous SINEs. SINE elements comprise the other main type of non-LTR retotrotransposon, also abundant in the genomes of mammals. While they are far more numerous than LINEs – with copy numbers of approximately 1,500,000 elements in humans – they are typically far shorter at approximately 100-400 bp in length. SINEs contain an internal promoter region, yet do not encode any functional proteins of their own. They instead hijack the LINE machinery in order to retrotranspose throughout the genome. The promoter regions of SINEs are largely tRNA-derived. Most SINE families are defective, however the Alu family has been found to remain active in humans. It is this Alu family of elements that does not have a tRNA-derived internal promoter; it is instead derived from a signal recognition particle component, 7SL. Alu repeats also comprise a single monophyletic family (Lander *et al*., 2001).

SINE integrations, as with any other SGE, can occasionally be deleterious to the host. Over time, they are either deleted or made defective by the action of purifying selection pressures acting against them. When defective SINEs, or other SINE insertions that are neutral to the host arise, they can reach fixation like other SGEs. Aside from Alu's, some of the most prevalent families of SINEs – both of which are defective – are mammalian-wide interspersed repeats (MIRs) and Ther2/MIR3 elements (Lander *et al*., 2001). MIRs, in particular, are extremely abundant in mammalian DNA. They are estimated to have begun amplifying in mammalian genomes since the Mesozoic era, with this particular family occurring at copy numbers as high as 10^7 (Jurka *et al*., 1995). Using orthology reconstructions, MIR elements have been estimated to have minimum ages of approximately 65 million years, although further light will be shed on this, in chapter 2 and appendix A of this thesis.

1.5.3 LTR Retrotransposons: LTR retrotransposons – like non-LTR retrotransposons – have autonomous and non-autonomous members. Both types are flanked by two identical LTRs, with the difference between them being that autonomous elements contain *gag* and *pol* genes, whereas non-

autonomous elements only contain *gag* (Lander *et al*., 2001). As with retroviruses, *gag* encodes structural proteins, and *pol* encodes the necessary enzymology for replication. Accordingly, the autonomous elements are compositionally similar to ERVs, except that they lack the *env* gene. Thus, LTR retrotransposons do not produce infective exogenous particles (Havecker *et al*., 2004). That said, ERV-L is an ancient retroviral lineage entirely comprising defective elements, which also lacks *env*, but is phylogenetically distinct from autonomous LTR retrotransposons (see chapter 2).

With a genome superseded only by ERVs in terms of complexity, LTR retrotransposons are able to produce their own protease, RT, IN and RNase H enzymes, and have internal promoter and enhancer sequences. Much like retroviruses, reverse transcription is primed by tRNA within a cytoplasmic particle in the cell. LTR retrotransposons contain families of elements that are thought to be amongst the most ancient of all selfish genetic elements. One particularly common family is the mammalian-apparent LTR-retrotransposon (MaLR) family. These have been found widely distributed across the Mammalia, and it has been suggested that they have been amplifying in their genomes for the past 80-100 million years (Smit, 1993). Older still, are the *Ty3/Gypsy* family. This family differs from MaLRs, in that it has been found in both vertebrate and invertebrate genomes. Constructed phylogenies of *Ty3/Gypsy* retrotransposons has implied that this family first began propagating approximately 600 million years ago (Volff *et al*., 2001).

It remains unclear whether retroviruses or retrotransposons came first, however it is quite probable that retroelements in general, hark back to an ancient world, when biosynthesis was primarily RNA-mediated. RT, despite being a DNA polymerase, is thought to have evolved from an ancestral RNA polymerase, since it exhibits high levels of similarity with RNA polymerases found in many extant RNA viruses. This notion was further corroborated by phylogenetic analysis of these two groups of enzymes, suggesting RT and viral RNA polymerases evolved from a common ancestor (Xiong & Eickbush, 1990). In fact, RT-like telomerase enzymes are essential to cell division in extant eukaryotes. Telomerase is a ribonucleoprotein, which contains an RNA

template component. The enzyme functions to synthesise repetitive DNA telomeres that are appended to the ends of chromosomes; a process involving the reverse transcription of telomerase's own RNA template. This reverse transcriptase activity is similar to that of RT in retroelements (Strauss & Strauss, 2007).

1.6 Age Estimation and Dating of ERVs

Studying ERVs has helped to gain a deeper understanding of both host and virus evolution. Unlike dating organisms from fossil data, viruses are comparatively difficult to date, as the only informative record available, is endogenised provirus sequence. Ordinarily, the age of ERV elements can be estimated by calculating sequence divergence between the two LTRs flanking the virus sequence, and dividing this by the substitution rate of the host genome (Dangel *et al*., 1995). This takes into account the fact that the two LTRs are identical in sequence upon integration of the virus, and therefore any sequence difference observed between them is expected to have resulted from genomic mutational processes occurring within the host, at the host's neutral rate of evolution (Kijima & Innan, 2010). Once a provirus inserts into a host genome, the two LTRs are subject to independent evolution, as though they were paralogous to one another. However, the neutral rate of host evolution is not the only factor to take into account, since integrated proviruses may be subjected to selective pressures. Substitution rates among ERV lineages may also vary, affecting the reliability of such LTR-based age estimations (Martins & Villesen, 2011).

Once integrated, ERVs are no longer subjected to the fast substitution rates of exogenous particles, and are typically confined to the neutral rate of the host organism. This is typically a far slower evolutionary rate than that of an exogenous retroviral particle, many of which are characterized by substitution rates three to six times quicker than those of their host (Figure 1.9) (Duffy *et al*., 2008). Using phylogenies, previous estimations of substitution points based on more distant calibration points can be updated. Congruence between these and additional calibration points can also be used to more

accurately affirm the age estimates of retroviral elements. Long-term viral evolution rates can also be calculated, by dividing the substitutions of an ancestral ERV sequence by those of its circulating XRV counterpart (Gilbert & Feschotte, 2010). Other methods of estimating the age of retroviruses include measuring the frequency of stop codons within the retroviral element's various ORFs (Belyi *et al*, 2010). Such dating methods can also aid the calibration of existing retroviral phylogenies, assisting the determination of viral evolution rates.

One of the main limitations for dating ERVs is the degeneracy of the elements themselves, over time. ERVs older than 250 million years cannot be detected in genomes using current technologies, although selective pressures on certain insertions can allow them to persist as a genetically detectable sequence, for longer than might ordinarily be expected (Jern *et al*., 2005). For such cases, screening the genomes of several organisms evolutionarily related to the host can be used to search for orthologous elements. This can help to infer a minimum age for the element, since the presence of an orthologus ERV insertion in two related species indicates that the infection event predates the divergence of those species, since ERV insertion is random. It is also possible to infer upper age boundaries for the element, by including additional species in which only solo-LTR structures are observed at the same genomic locus (Gilbert & Feschotte, 2010).

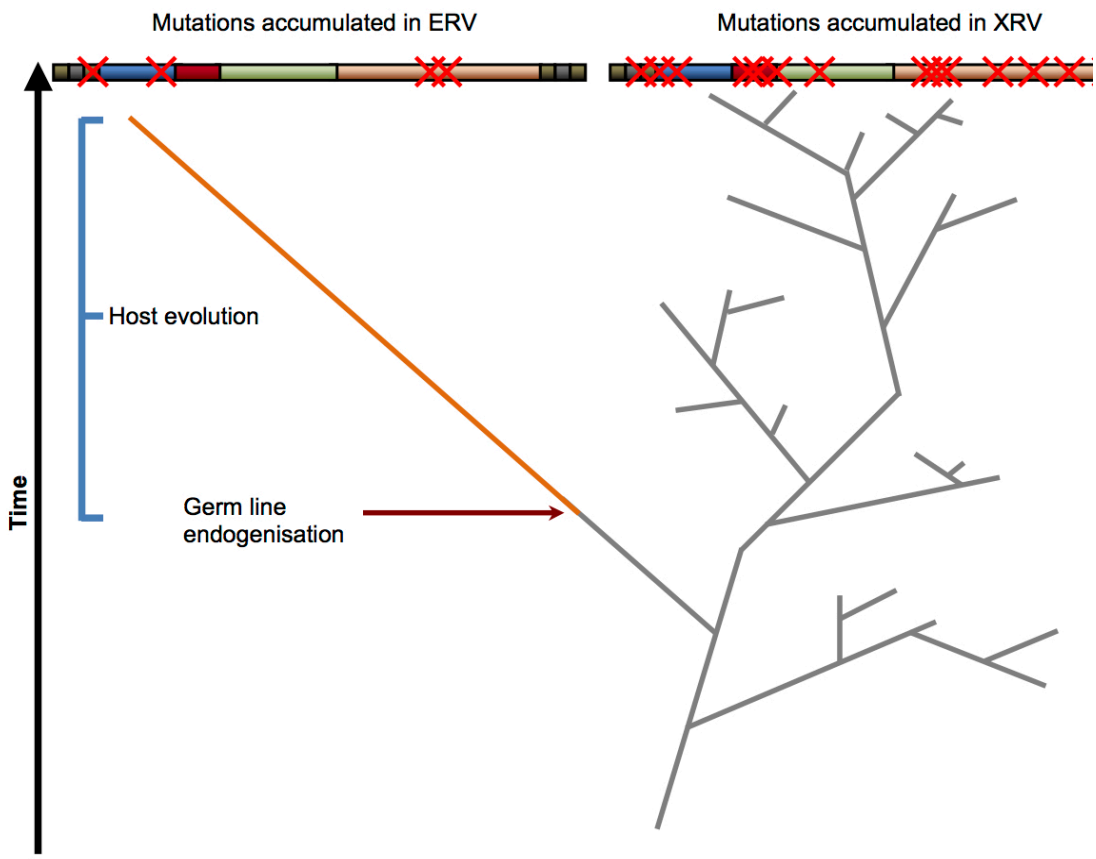


Figure 1.9**: A schematic diagram illustrating substitutions** accumulated in an endogenous retrovirus (left) and its circulating, exogenous counterparts (right). Mutations are represented by red crosses. Typically, XRVs accumulate more substitutions, since ERVs are constrained to the neutral evolution rate of the host species. The substitutions in the endogenous element can be divided by those in its corresponding exogenous element to give an estimate of long-term viral evolution rate (modified from Feschotte & Gilbert, 2012).

Another viable technique for dating retroviral sequences, is through examining the insertion of SGEs within a retroviral element. Using SGEs in order to date sequences was first proposed by Murata *et al*., in 1996. The insertion events of SINEs, for instance, can be used as clade markers, since all SINE hosts should, in principle, contain SINE insertions from a single monophyletic clade (Cook & Tristem, 1997). SINEs are class I elements; they remain at the genomic locus of insertion and, unlike other elements, transfer horizontally to other species less frequently. These features alone favour their use as clade markers over other transposable elements. The presence of a single SINE family, for instance, can be used as a clade marker for a single node (Cook & Tristem, 1997).

Whilst SINE elements were traditionally observed using PCR with degenerate primers and Southern hybridizations, advances in sequencing technologies have now made it possible to use sequence data to identify their whereabouts *in silico*, and estimate their age, by the rate of accumulation of various mutations and insertions (Cook & Tristem, 1997).

Furthermore, phylogenetic trees can be constructed from orthologous SINE elements, as these should form part of a single monophyletic group. SINE elements also offer advantages when constructing trees (Cook & Tristem, 1997). For instance, insertion site analyses can look at several independent loci, which reduces tree/species tree incongruence, or incompatibility between individual genes, since different loci on the same gene are being compared (Murata *et al*., 1996).

1.7 ERV Activity and Implications in Disease

Retroviral studies are both important, and highly relevant today, since ERVs – originally thought to be relatively harmless in almost all extant mammals – have been implicated in several diseases over the past two decades. They have been considered as potential mediators of cancers, neurological and autoimmune disorders (Lower, 1999). Investigations have, however, proven cumbersome due to difficulty in identifying coding ERV lineages, since there is often abundant background noise from defective elements. Despite this, advances in next generation sequencing technologies continually provide increasingly useful information for identifying ERV loci potentially causative of disease (Griffiths, 2001).

1.7.1 Activity: Despite an overwhelming majority of non-functional elements observed in host genomes, active, replication-competent ERVs have been identified in some species (Stocking & Kozak, 2008). Among these, are Australian koalas. Analysis of koala genomes has revealed a spreading endogenisation by a retrovirus known as KoRV, which exhibits great sequence similarity to Gibbon Ape Leukemia Virus (GALV). KoRV was previously shown to be a pathogenic retrovirus causing neoplasia (Tarlinton *et*

al., 2005), and was subsequently found in its endogenised form in the genomes of some koalas, but not others. This insertional polymorphism strongly suggests that the virus is active, and that it may be existing in both its endogenous and exogenous forms. Further supporting this possibility, it was found that KoRV's number of paralogous copies (copy number) varied between individual animals. Observation of *env* genes between different insertions (both within the same animals and between individuals) also revealed high levels of mutation. None of these were found to be nonsense mutations, and deletions of up to 9 bp in length were all found to be in frame. All of these evidences are suggestive of a virus that is actively infecting and propagating within its host (Tarlinton *et al*., 2006).

1.7.2 Disease: In 2010, research by Gimenez *et al*. attempted to confirm hypotheses implying that active HERVs were thought to contribute to various human cancers, using microarrays to specifically detect the presence of transcripts containing viral LTRs. They identified 6 HERV-W loci that were overexpressed in testicular cancer and this included a ERVWE1 transcript – previously thought to be placenta-restricted. Comparative analyses of retroviral elements from tumoural tissue versus normal tissue, demonstrated that hypomethylation of U3-region promoters (at the 3′ end) was required for their activation. However, so far, these studies are associative, and have not confirmed that it is definitely these HERVs that are causative of cancers. Further research in this area may hopefully confirm or refute this. Additional investigations linking HERVs to disease have arisen form degenerate primer PCR studies, which implicated HERV-W in multiple sclerosis patients (Perron *et al*., 1997). HERV-W was also found to be expressed in the brain and cerebrospinal fluid samples of schizophrenics (Karlsson *et al*., 2001).

One particular group of ERVs is the subject of continual research and interest: the human HERV-K(HML-2) retroviruses. This lineage makes up under 1% of HERV elements, and is the primary HERV group implicated to be currently active in humans (Paces *et al*., 2002). Whilst a pivotal pathogenic role for these HERVs is yet to be proven, it remains likely, given that multiple HERV-K(HML-2) elements have been identified as being insertionally polymorphic;

present in some individuals, but not others (Turner *et al*., 2001; Belshaw *et al*., 2005; Agoni *et al*., 2012; Marchi *et al*., 2014; Lee *et al*., 2014). This insertional polymorphism indicates that integration has occurred recently, since these HERVs are not manifest in the majority of individuals, and thus have not had sufficient time to drift to fixation in the global human population (Turner *et al*., 2001; Belshaw *et al*., 2005). Recent leaps in sequencing technologies have led to an increasingly rich wealth of human genome sequence data, all of which can be thoroughly screened for novel HERV-K(HML-2) insertions that may implicate emergent disease in humans.

In 2008, the products of 23 transcriptionally active HERV-K(HML-2) loci were identified in human tissues, demonstrating the capability to produce infectious particles. Out of a group of putative HERV-K(HML-2) loci, this study reported that the expression of an element on chromosome 22 (22q11.21) was strongly upregulated in germ cell tumours (GCTs), which was also correlated with high levels of HERV-K(HML-2) Gag protein in these tissues (Flockerzi *et al*., 2008). Further studies have linked GCTs with the impairment of spermatogenesis in affected males. This was correlated with a regulatory protein, known as c-terminal open reading frame (cORF) protein, produced by the *env* gene of certain HERV-K(HML-2) elements. cORF proteins can then bind to a promyelocytic leukemia zinc finger protein integral to spermatogenesis, impairing its function (Boese *et al*., 2000). Patients with germ cell tumours also appear to express anti-HERV-K-Gag and -Env antibodies further implying retroviral involvement in disease (Boller *et al*., 1997).

HERV-K(HML-2) elements have also been shown to express virus particles within teratocarcinoma-derived cell lines (Figure 1.10) (Boller *et al*., 2008). While many of these particles bud at the cell surface, it is unlikely that they are infectious (Bieda *et al*., 2001). Despite this, molecular biology and serological analyses have collectively presented several lines of evidence suggesting that at least one of the 31 HERV-K subsets known today is responsible for production of complete, intact viral particles found within tumour and melanoma cells (Bieda *et al*., 2001; Muster *et al*., 2003). As with many such studies, all these observations are associative, and there has not

yet been any definitive proof that HERV-K(HML-2) elements play a specifically instrumental role in human disease.

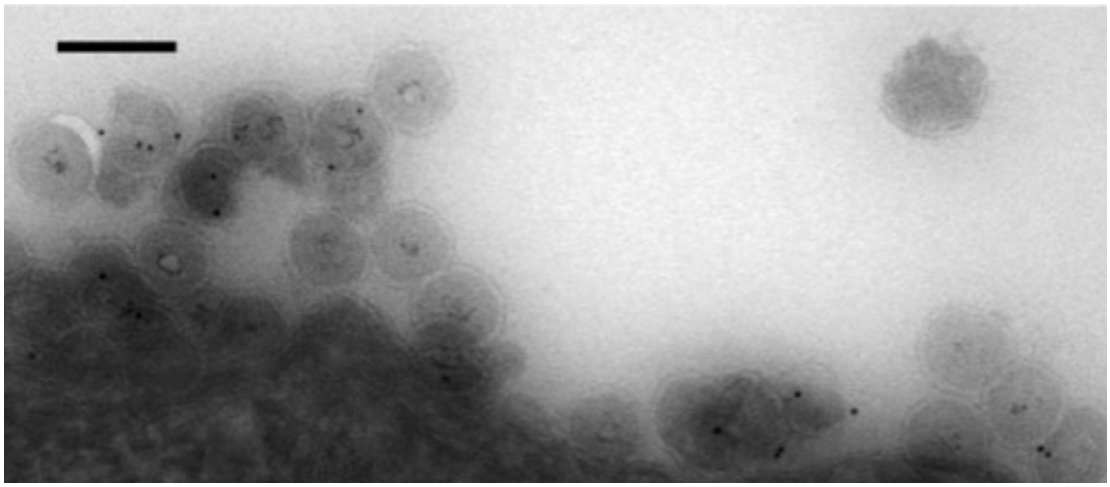


Figure 1.10: **Immunoelectron microscopically labeled HERV-K113 virions** expressed artificially from the HERV-K113 sequence in the human genome. These have also been shown to bud from human teratocarcinoma cells. The bar indicates 150nm scale length (Boller *et al*., 2008).

1.8 Co-Option and Exaption of Retroviral Genes

ERV expression is not solely implicated in disease and producing deleterious phenotypes. There are also instances where retroviral genes have been exapted and selected for in their hosts, with time. Additionally, as previously mentioned, there are many examples of the exaption of ERV LTRs for their gene regulation properties. The vast majority of these retroviral elements are, however, defective, due to an accumulation of nonsense mutations, frameshifts, and recombinational deletion to form solo-LTRs (Boeke & Stoye, 1997). These solo-LTRs often confer benefits to the host *via* their internal promoter sequences.

When it comes to the exaption of retroviral genes, one of the most remarkable examples are the *syncytin* genes. These are fully functional, coding genes with protein products derived from the *env* gene of an inserted retroviral provirus. Within the genomes of extant placental mammals, it is possible to observe that these ancestral retroviral insertions first endogenised before the

placental radiation. In fact, it is their conferred biological function that gives this mammalian infraclass its name (Lavialle *et al*., 2013).

As part of the life cycle of an infectious retrovirus, the *env* gene encodes surface-projected glycoproteins. An important function of these is to mediate the fusion of cell and virus membranes required for administering viral proteins and gRNA into the host cell. Similarly, the ERVWE1 locus – a member of the HERV-W family – is an ERV whose Syncytin-1 envelope glycoprotein is known to mediate trophoblast membrane fusion during placental formation in humans (Mallet *et al*., 2004). The trophoblast is formed by the outermost cells of the blastocyst, known as cytotrophoblasts, and it is these that attach to the uterine wall. Cytotrophoblasts are part of a small subset of mammalian cells that can undergo cell-to-cell fusion, fusing together into the syncytiotrophoblast – a specialised syncytial layer which allows for implantation of the developing embryo into the maternal endometrium (Harris, 2010). The syncytiotrophoblast is pivotal in function, allowing for exchange of nutrients and hormones between the embryo and mother, whilst simultaneously providing an immune response to foreign pathogens (Gude *et al*., 2004). *syncytin-1* was found to be ubiquitously expressed across trophoblasts, whereas *syncytin-2* – another virally-derived, exapted *env* gene – was found to be expressed in only a fraction of cytotrophoblasts. This *syncytin-2* gene was shown to confer two functions: it drives the fusion of cytrotrophoblasts into the adjacent syncytiotrophoblast tissue, while also inhibiting fusion between cytotrophoblasts themselves. It is therefore thought to regulate and allow for the regeneration and growth of the syncytiotrophoblast tissue through the term of pregnancy, as the foetus develops (Esnault *et al*., 2008). Furthermore, the ERVWE1 locus is a remnant of a retroviral infection whose 5′ LTR is still intact, and has been proven to promote and confer tissue specificity of the *syncytin* genes (Mallet *et al*., 2004). Studies using *syncytin* knockout mice revealed that these genes are critical to placenta formation and embryo survival. Thus, these particular *syncytin* genes – initially a chance infection event by an exogenous retrovirus – may have been critical in the evolution of placenta formation from ancestral egg-laying mammals (Lavialle *et al*., 2013). Interestingly, the same retroviral

element is also thought to contribute to epithelial cancers (Bjerregaard *et al*., 2006) and autoimmunity disorders.

Certain ERVs in other mammalian species, such as mice, have been demonstrated to restrict retroviral infection, whereby Gag-like proteins interact with incoming retroviral cores to inhibit infection. The gene in question is the ERV-derived *Fv1* gene in mice, which acts to restrict murine leukemia virus (MLV) infection. Other mammals – including humans – have also been shown to restrict retroviral infection in a similar way (Towers *et al*., 2000), and it may be possible that ERV Gag proteins are selected for as a mechanism for retroviral infection control (Griffiths, 2001).

1.9 Next Generation Sequencing

Since the advent of the human genome project, demand for accurate, abundant genome sequence information has steadily increased. Technologies for sequencing DNA from biological samples are constantly being refined, improved and made more economical. In particular, the advances of Next Generation Sequencing (NGS) technologies, has allowed for the compiling of rich libraries of sequence information, which is particularly well suited to the studies of ERV integrations, and genome evolution. Prior to the availability of these technologies, genomes were sequenced using the automated Sanger method (termed a 'first generation' technology), which was responsible for the completion of the first human genome sequence (among others). However, first generation methods presented many limitations, which heightened the need to develop more efficient technologies (Metzker, 2010).

One of the most obvious benefits of NGS over Sanger methods, is cost. Some NGS technologies are capable of producing billions of short DNA sequence fragments – termed 'reads' – in a single run. This is facilitated by immobilising billions of DNA molecules on a solid surface, where they are easily read from a localised area. Since multiple areas can be read simultaneously on a large surface, the process is faster and more efficient than first generation methods (Metzker, 2010).

Broadly, NGS methods are divided into four stages: template preparation, DNA sequencing, imaging and the analysis of resulting data. There are two main methods for generating templates from biological samples. The first uses a single DNA molecule, and is termed single-nucleotide addition (SNA), and the second involves clonally amplifying single DNA molecules, termed cyclic reversible termination (CRT) (Metzker, 2005; Fan & Gunderson, 2006).

Firstly, sample DNA molecules are shortened by fragmenting, with resulting cleaved ends being repaired and adenylated. These molecules are referred to as mate-pair templates or fragment templates. In most technologies, these templates are then amplified, and immobilised, so that huge numbers of sequencing reactions can occur simultaneously, while being easily traced to a restricted physical space on the reaction surface (Metzker, 2010).

Most systems rely on clonal amplification of sample DNA molecules in order to generate the required signal strength for sequencing. Emulsion PCR (emPCR) and solid-phase amplification are frequently used for this, with the former involving a cell-free system where each DNA molecule binds to a single substrate bead, and the latter localises the sample to a glass slide (Dressman *et al*., 2003; Fedurco *et al*., 2006). During emPCR, the objective is to create a library of mate-pair targets, which allow for the ligation of oligonucleotide adapters onto them. These adapters contain universal priming sites, which facilitate amplification using commonly available PCR primers (Metzker, 2010). Once DNA has been amplified using emPCR methods, there are different options for immobilising the DNA. Technologies such as Life/APG (Kim *et al*., 2007) involve cross-linking the beads to an amino-coated surface, whereas the popular Roche/454 technology involves positioning beads inside PicoTiterPlate (PTP) wells. Polonator technologies, alternatively, can sometimes use a system whereby the emPCR beads are bound into a polyacrylamide gel matrix, which can then be fixed onto a glass slide surface (Shendure *et al*., 2005).

Conversely, solid-phase amplification systems center around clonally amplified clusters of DNA, which are generated on a glass slide, to which forward and reverse primers are annealed. Here the amount of primer relative to the amount of template, determines the extent of amplification. In systems such as those developed by Illumina, hundreds of millions of template clusters are generated, and spatially separated on the glass slide. Since one end of the molecules binds to surface-bound primer molecules, the other end is left free, allowing for subsequent sequencing reactions to occur in a restricted location on the slide, per molecule of DNA (Metzker, 2010).

Most NGS reactions are luminscent, generating light of different wavelengths according to the presence of a different base. Therefore, it would seem that clonal amplification of sample DNA would be the best option for generating the necessary signal strength for sequencing. However, there are benefits to using single-molecule templates, and bypassing clonal amplification steps. The primary advantage is the sample size can be small - < 1 µg, whereas clonal amplication typically requires 3-20 µg samples. A second advantage is that – by eliminating the need for PCR – the accuracy of the resulting sequence data is potentially improved, since polymerase error rates (and other factors associated with PCR) can often introduce mutations and thus, a degree of unwanted error (Metzker, 2010).

With the single-molecule approach, there are three main methods used. One approach uses covalently bonded primers, distributed evenly across a solid surface. These are used to prime and extend single molecule templates, thus binding them to the same substrate. Another approach also makes use of surface-bound primer molecules. With this approach, however, the single-molecule templates generated from fragmenting the sample DNA have universal adaptors ligated to their ends, which allows them to be hybridised with the surface-bound primers, immobilising them this way (Harris *et al*., 2008). With these two approaches, sequencing is initiated by reacting the immobilised primers with DNA polymerase. A third approach – employed by Pacific Biosciences, Life/VisiGen and LI-COR Biosciences – utilises evenly distributed, surface-immobilised molecules of polymerase, which have primers

bound to them (Eid *et al*., 2009). Pre-primed template molecules can then be bound to these for immobilisation. One main advantage of this third approach is that it is known to generate longer sequence reads. This can be particularly helpful in sequencing genomes for which there are poor reference sequences against which to construct assemblies, as there is more sequence information within each read, and potentially greater read overlap for easier assembly.

The main component of NGS sequencing technologies revolves around a process known as Cyclic Reversible Termination (CRT). The first step in CRT makes use of a DNA polymerase, which is specifically intended for complementing a single nucleotide base only. This base is fluorescently tagged and bound to a terminating group, which ensures that no further DNA synthesis can take place on the template strand. A washing step then discards free excess nucleotides from the sequencing environment and, based on the fluorescent signal detected off of the newly incorporated, tagged nucleotide, imaging is performed. Each of the four nucleotides will be tagged with fluorescence of different wavelengths, which are detected and computed by the sequencing analyser. In order to progress to the next base on the template strand, both the terminating group and fluorescent dye are cleaved away from the complimented nucleotide. A further washing step is performed, before the cycle is repeated for the next base in the sequence (Metzker, 2005).

The actual sequencing reaction differs between technologies, but is principally similar. The process has two main differences when using clonally amplified templates versus single-molecule templates. With the former, a group of identical template molecules is imaged, with the resulting sequence being a consensus of each sequenced template. This lends to potential problems, and shorter resultant reads. As all the template molecules are identical, the addition process must be efficient and synchronous, so that each template is generating the same base signal. However, this is rarely the case, and leads to lagging-strand dephasing, where some of the templates are not fully extended. A similar issue – leading-strand dephasing – can also occur when too many bases are added to certain molecules. Such dephasing issues,

create a less uniform fluorescent signal, with a lower signal to noise ratio, which can result in erroneous base calls, and truncated read lengths. These problems do not exist with single-molecule templates; this instead places higher emphasis on the requirement of more accurate sequencing reactions, since only a single template molecule is used, and error is not spread over a consensus, as with clonally amplified templates. However, if multiple additions occur erroneously in a single-molecule template system, this can result in quenching of adjacent dye molecules, where 'dark nucleotides' are added (Metzker, 2010).

One of the most commercially successful sequencing technologies is the Illumina/Solexa Genome Analyser, which relies on clonal amplification together with a four-colour CRT system. Imaging is performed using a two-laser system, using total internal reflection fluorescence (TIRF) (Metzker, 2010). Recent anthropological findings and studies have been conducted that make extensive use of the Illumina system, and have resulted in genome sequences being made available at very high sequence coverage levels, providing a rich library of accurate sequence information for the associated genomes (Meyer *et al*., 2012; Prüfer *et al*., 2014). This rich sequence data is particularly useful for the study of retroviral insertions, as it provides complete sequence records at crucial host-virus junctions that may have previously gone unnoticed in inferior genome data.

1.10 Bioinformatic Tools

The availability of rich genome sequence data gives researchers invaluable opportunity to analyse the genetic components of living organisms, *in silico*. Advancing sequencing technology has run parallel with comparable advances in computing technologies, processing power and data storage; although the former often supercedes the development of the latter. Such technological advances are a basic requirement for processing the voluminous data produced by NGS protocols, especially when dealing with high coverage genomic sequence data. As an example, the first draft of the Denisova man genome – an archaic hominid from the Altai mountain range in Siberia –

totalled ~30 gigabytes (GB) of plain text FASTA format data when sequenced at relatively low, 1.9-fold coverage (Reich *et al*., 2010). The same genome sequenced at higher coverage (approximately 30-fold) totalled ~600 GB (Meyer *et al*., 2012). The computational requirements for processing and analysing data of this volume are substantially increased. In the case of the more recent Denisova sequence, analysis is almost impossible on a standard specification personal computer, as processing power and random access memory (RAM) demands are too high to make any large-scale analysis of this data feasible (see chapter 3).

In much the same way as a search engine allows users to search databases of websites using a choice phrase or keyword, sequence alignment algorithms have been developed, where users can search genomic libraries using nucleic or amino acid probe sequences. The most popular sequence alignment technology used in this way is the Basic Local Alignment Search Tool (BLAST) (Altschul *et al*., 1990), which is equipped for handling both amino and nucleic acid sequences. When aligning DNA sequences, algorithms generally fit into two categories: global and local alignment. A global aligner, such as EMBOSS' Needle (Rice *et al*., 2000), takes into account large stretches of sequence, and may output long sequences of low homology. Conversely, BLAST and other local aligners are designed to search for smaller regions of high sequence conservation, and results are subsequently built from groups of subsequence alignments that contribute to the overall sequence match (Altschul *et al*., 1990). With BLAST, these small nucleotide strings are termed 'words'; the word size is a customisable parameter – often defaulted at 11 bp – which can be altered to skew results. While this word-based approach is often fast and efficient, it is better suited to higher levels of sequence homology, where matches are unlikely to be afflicted by large indels. BLAST is often inefficient at correctly identifying sequence homologues that are poorly conserved, as might be observed in evolutionarily ancient DNA between distantly related species. For such applications, the use of global alignment tools may be preferable.

The BLAST suite contains five programs: BLASTP, BLASTN, BLASTX, TBLASTN and TBLASTX. BLASTN – which searches a nucleotide database with a nucleotide query - is typically the most suited to the study of ERV insertions. This is because most ERVs are defective, and there is rarely interest in the protein products of defective elements.

In order to generate a sequence alignment BLAST awards scores to matches, mismatches and introduced gaps. This allows many different possible alignments to be evaluated, and the best one to be chosen for the final result (Pertsemlidis & Fondon III, 2001). Matches are typically always given the same score. They are scored against a scoring matrix, which gives a set of values describing the probability of a relevant nucleotide base pairing presenting in the alignment. This means that BLAST is capable of allowing for the matching of words that are not 100% stringent, and can include some mismatching residues. Scoring matrices for nucleotide residues are typically more simplified than those designed for amino acid alignments. While positive scores are awarded to matches, negative penalty scores can be applied to mismatches and gap insertions. While these penalties are usually equal for each mismatch, they differ for gaps. The first, single residue gap that is inserted is termed a 'gap open' penalty. Gaps inserted subsequent to this initial gap are awarded 'gap extension' penalties (Figure 1.11). By default, gap extension penalties are usually much smaller than gap open penalties (Pertsemlidis & Fondon III, 2001).

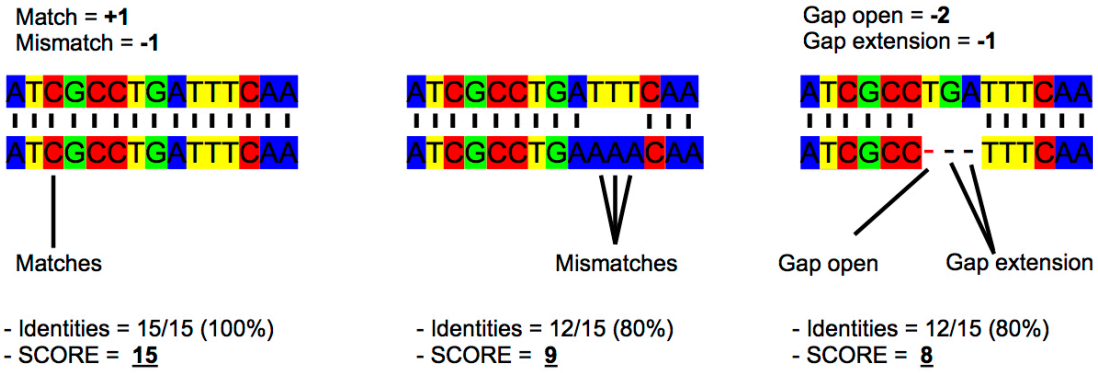


Figure 1.11: **Diagram illustrating BLAST scoring**. Example scores and penalties are given above the sequences. Matching bases between query (top) and subject sequences (bottom) are represented by small black lines. Sequence identities and scores are given, showing that the relationship between these two values is not always the same.

When using default parameters, the first step of a BLAST search, is the filtering of low-complexity regions of sequence (such as certain simple repeats) which would typically yield unimportant results that would otherwise make it into the final output. BLAST then divides each sequence (both query and database) into shorter subsequences called segments. When there is strong (gap-free) homology between a query and subject segment, these pairs of short sequences are termed segment pairs. BLAST can then recover many similar segment pairs, and check them against a given threshold score termed the neighbourhood word-score threshold. If they exceed this threshold, they are referred to as high-scoring segment pairs (HSPs). The highest scoring HSP is termed the maximal-scoring segment pair (MSP). Thus, the program generates a list of words, which it then compares against the target BLAST database, which is already pre-prepared into all possible word occurrences (Altschul *et al*., 1990). Alignments are then extended from adjacent occurrences of these short sequence matches. BLAST also provides statistical estimates of expect value (e-value) – which is a measure of the number of times each complete sequence match is expected to occur by chance, given the database size. Other values such as query coverage (the amount of query represented by a match), percentage identities (the percentage of total individual, identical matches) are all given as part of the final output of a BLAST search (Pertsemlidis & Fondon III, 2001).

Aside from the myriad of available bioinformatic programs available, it is often desirable to be able to custom design simple programs for dealing with biological data. To do this, bioinformticians use programming languages such as Python and Perl to enable custom analysis of biological data. Perl has been readily adopted by biologists, since it provides a very verbose and flexible platform to batch-process large volumes of data, and customise workflows as required for specific applications. Anything from simple string identification, through to more complex applications, can be automated using programming languages like perl, saving valuable time and resources, while also potentially allowing functionality outside the scope of many existing bioinformatic programs.

CHAPTER 2: Identification of an Ancient Endogenous Retrovirus, Predating the Divergence of the Placental Mammals

2.1 Introduction:

Studying ERV orthologues across several species can be an accurate method of assigning a minimum age to endogenised retroviral insertions (Gilbert & Feschotte, 2010). Unlike ERVs, many of their vertebrate hosts have a detailed fossil record, allowing for the construction of robust, accurate phylogenies. This chapter investigates the orthologous insertions of an ancestral ERV element present within several extant placental mammals, for which phylogenies are available and documented in the literature (a placental mammal phylogeny is shown in Figure 2.1). Using the sequence of an intact ancient HERV-L retrovirus that was found in the human genome, it was possible to identify orthologous HERV-L-like (ERV-L) elements within other placental mammals, in order to establish how far back in evolutionary time the initial infection event was (Lee *et al*., 2013).

There has been much speculation regarding retroviral evolution across the eras, and for the most part, only rough estimations have been available when dating ancient elements. As discussed in chapter 1, the retroviral life cycle involves an exogenous stage whose evolution is difficult to study and analyse. The rates of evolution observed in XRVs are often so fast, that they give little insight into the origins of any particular lineage of retroviruses (Feschotte & Gilbert, 2012). It is in this respect that endogenised retroviral elements offer great advantage, being fundamentally confined to mutating at the host rate of evolution (Duffy *et al*., 2008). If the germline into which an ERV integrates belongs to a host with a slow generation time, for example, it might be expected that the ERV would accumulate mutations at a lesser frequency than if it were endogenised in a host with a shorter generation time (Laird *et al*., 1969). This is due to the fact that most species with a shorter life-span have to copy their genomes more frequently, and thus the rate of nucleotide substiutions is often higher, allowing for more rapid molecular evolution (Thomas *et al*., 2010).

If a SGE of any kind had inserted and persisted in the germline of an organism for a long period of time, such an element would likely pose no

deleterious risk to the host organism. Having reached fixation, these SGEs can often remain endogenised long enough for an organism to evolve, to the point of speciation. In this scenario, it is possible to reconstruct an orthology, whereby orthologous SGEs between two or more species are identified by the unique flanking sequence present at either end of the element. In the case of ERVs, the initial infection of the host genome is entirely random (Gilbert & Feschotte, 2010). Thus, it is extremely unlikely that two independent infection events, in two different host species, would occur at the same genomic locus. It is therefore possible to conclude that two ERVs in two separate host species are orthologous, and that the initial infection was ancestral to their divergence from one another (Katzourakis & Gifford, 2010).

Retroviruses have been observed in a wide range of vertebrate hosts. The genomes of many mammals comprise approximately 8% ERV proviral DNA, derived from molecular fossils of past retroviral infections (Lander *et al*., 2001). Approximately 147 million years ago, ancient mammals began speciating into two major clades, one of which is the Eutheria, containing the highly diverse Placentalia infraclass, commonly referred to as the placental mammals. The majority of extant mammals belong to this infraclass, with the remainder belonging to Marsupialia and Monotremata (Murphy *et al.*, 2001a).

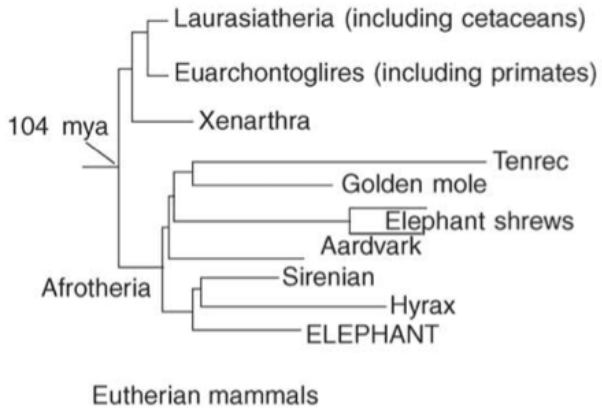


Figure 2.1: **Phylogeny of the four placental mammal superorders**: Afrotheria, estimated to have diverged from Xenarthra and Boreoeutheria approximately 104-110 million years, are indicated (Roca & O'Brien, 2005). The elephant, manatee (a sirenian), aardvark and golden mole are among the most basal placental mammals, and are two of the most basal Afrotherian mammals in which orthologus ERV-L elements have been identified, in this study.

Placental mammals are characterised by their reliance on the placenta during gestation – an organ which connects the developing foetus to the uterine wall during development, allowing for the free flow of nutrients and waste. This organ incorporates a co-opted ERV, whose gene function is critical for membrane fusion events that occur during foetal development (Mallet *et al*., 2004). The Eutheria are grouped into four main superorders. These are:

- Afrotheria (superorder I)
- Xenarthra (superorder II)
- Euarchontoglires (superorder III)
- Laurasiatheria (superorder IV)

Superorders III-IV are together referred to as the Boreoeutheria which, together with Xenarthrans, diverged from Afrotherians approximately 104-110 million years ago (Roca & O'Brien, 2005). Following this, a burst of evolution and speciation occurred, referred to as the placental radiation, that ultimately led to the high diversity seen in extant placentals (Springer *et al*., 2003). This diversity has resulted in an infraclass of mammals of many different sizes, with a great range of generation times. Those with slower generation times, in particular, are valuable to studying the evolution of retroviruses, endogenised as a result of ancient infection events. Thus, several of the species selected in this study generally have slow generation times and long gestation periods that tend to favour slower rates of genome evolution. Elephants and manatees are two such mammals, and are classified within the mammalian superorder Afrotheria (Figure 2.1) – the most basal placental mammal clade. The divergence of Afrotheria from Boreoeutherian and Xenarthran mammals was thought to have occurred approximately 104-110 million years ago, corresponding with the vicariance resulting from separation of the African continent from other land masses (Roca & O'Brien, 2005; Murphy *et al*., 2001b; Springer *et al*., 2003).

The ERV-L retroviral lineage was first identified in humans, with HERV-L initially being discovered in the placenta. Here, HERV-L elements were present at a copy number of approximately 200 (Cordonnier *et al.*, 1995).

Together with autonomous MaLR (mammalian apparent LTR retrotransposons) elements, ERV-Ls comprise class III ERVs. Both classes of element share identifiable ORFs with similarity to *pol* and contain LTRs, and both also lack the retroviral *env* gene. Based on 51% sequence homology between their LTRs, it is thought both ERV-L and MaLRs have a common ancestor and diverged from one another at some point (McCarthy & McDonald, 2004).

Further investigations have since identified members of the ERV-L group across the Placentalia, although lacking orthology studies have not allowed for a definitive age to be assigned to this lineage, as of yet. ERV-L retroelements have previously been found to occur as far back as 45-65 million years ago within primates, at the prosimian-simian branchpoint, and the occurrence of ERV-L elements in many placental mammals has also been used to suggest that ERV-L elements predate the placental radiation, which occurred approximately 70 million years ago (Bénit *et al*., 1999). Other studies where orthologies were used as a means of determining a minimum age boundary for a virus, include the identification of Rabbit Endogenous Lentivirus type-K (RELIK) across the Lagomorpha, which assumes a minimum age of 12 million years for this virus (Keckesova *et al*., 2009).

Orthologies have also been constructed for considerably older, non-retroviral endogenised viral elements (EVEs). The viral family *Bornaviridae*, for instance, was shown to be at least 93 million years old, while orthologies reconstructed for EVEs belonging to the *Parvoviridae* and *Circoviridae* revealed minimum ages of 30 and 60 million years, respectively (Katzourakis & Gifford, 2010). Similar techniques have also been used on other SGEs. Like ERVs, SGEs that insert into the germline, remaining neutral to the host, can also be inherited vertically, persisting across generations for millions of years. Among these, mammalian-wide interspersed repeats (MIRs) are estimated to have been amplifying within mammalian genomes for approximately 130 million years. This was assumed from constructing an orthology assigning MIRs a minimum age of 65 million years, with the 130 million year estimate being achieved through the further application of a molecular clock (Jurka *et*

al., 1995). This study, however, uncovers evidence of a far older MIR element, orthologous across all four placental superorders, assigning MIRs a definitive minimum age of 104 million years (Lee *et al*., 2013).

Other transposable elements abundant in the genomes of extant mammals are the medium-reiteration frequency interspersed repeats (MERs). Of these, the MER3 family is notably diverse, with poor sequence conservation between elements (Jurka, 1990). While MER3 orthologies have not yet previously been used to infer a minimum age for this family, it has been assumed that they are at least 55 million years old (Labuda & Striker, 1989). Orthologous MERs have, however, been identified across all four Eutherian superorders in this investigation, assigning them a minimum age of 104 million years.

This study presents the oldest definitively dated virus published to date, and contains within it a number of selfish genetic elements, also orthologous across all investigated species (Lee *et al*., 2013). Through presenting a conserved proviral orthologue, it is hoped that the techniques demonstrated here can set a new benchmark for dating ancient orthologous SGEs, allowing for a deeper understanding of SGE and retroviral evolution and origins.

2.2 Materials and Methods:

A series of different analyses were performed in order to confirm orthology, estimate age and evaluate potential selection pressures acting on the ERV-L element identified in this study. Each set of analyses performed on the element are organised into individual subsections, under the relevant headings.

2.2.1 Element Retrieval: The first step involved retrieving the HERV-L sequence to be used as a starting probe. This sequence was available online in NCBI's Nucleotide database, under the accession number X89211. This was then used as a probe sequence with which subsequent BLAST searches were conducted on various mammalian genomes available in NCBI's online repository. The BLASTN algorithm was selected, with all other algorithm

parameters set to default, and repeat filters turned off. A number of specific mammals representative of superorders I-IV were selected for the master alignment and further study, based on factors such as generation time and sequence conservation of the identified ERV-L orthologue (see Table 2.1).

Table 2.1: **Species** in which the orthologous ERV-L was identified, representing all four placental superorders.

SUPERORDER I	SUPERORDER II	SUPERORDER III	SUPERORDER IV
Elephant *Loxodonta africana*	**Armadillo** *Dasypus novemcinctus*	**Human** *Homo sapiens*	**Horse** *Equus caballus*
Manatee *Trichechus manatus*		**Chimpanzee** *Pan troglodytes*	**Pig**
Golden mole *Chrysochloris asiatica*			**Seba's short-tailed bat** *Carollia perspicillata*
Aardvark *Orycteropus afer*			**Flying fox** *Pteropus vampyrus*

A phylogenetic tree showing the placental mammals and marsupials is shown in Figure 2.2. Host species selected for study here, that are present on the tree, are boxed to indicate the evolutionary range across which the ERV-L was studied.

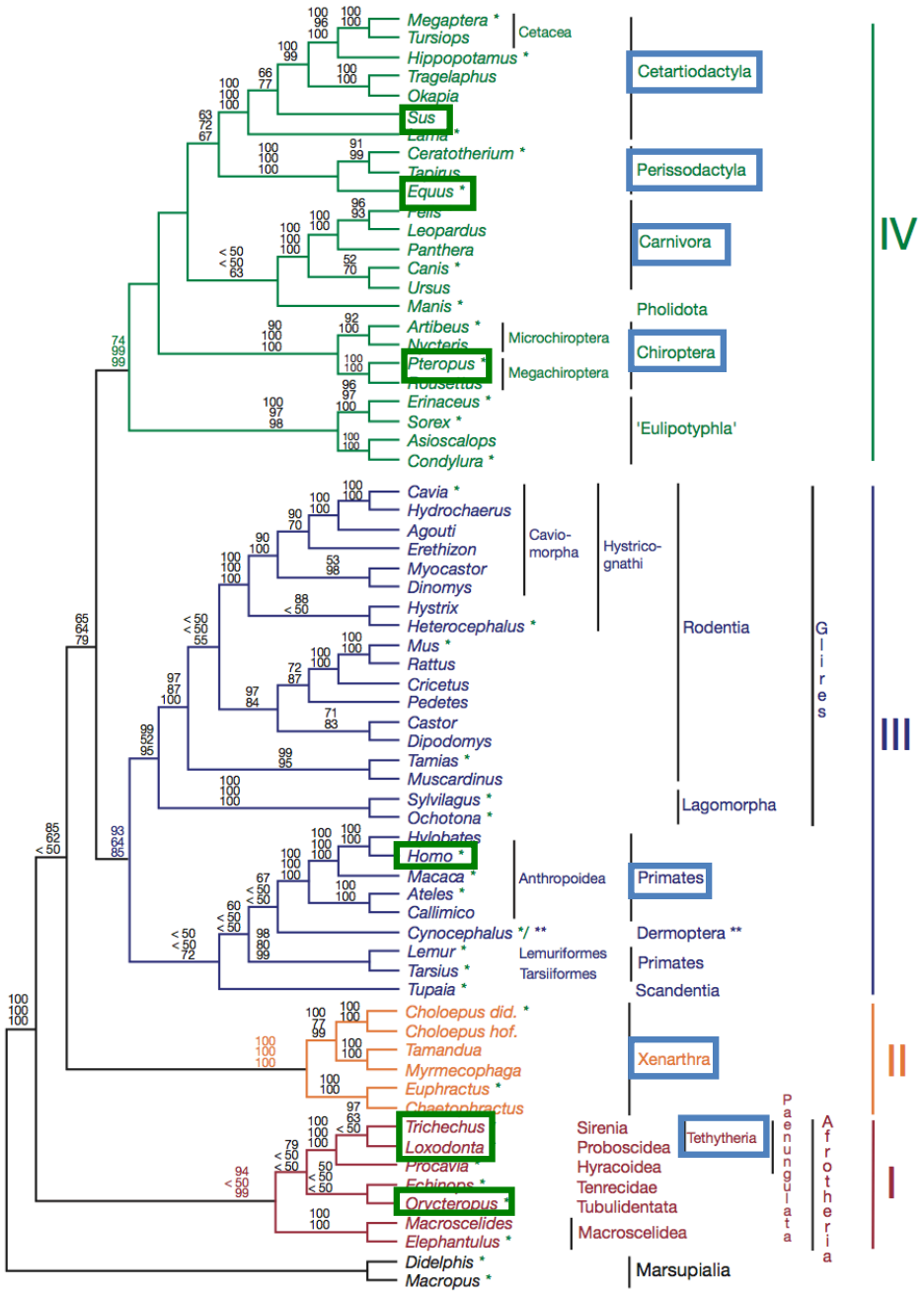


Figure 2.2: **Mammalian phylogeny** depicting the placental mammals and marsupials. An unrooted, neighbour-joining phylogenetic tree built by Murphy *et al.* (2001a) resolving ancestry of 64 placental mammals from the four Eutherian superorders. Numbers represent bootstrap values. Similar trees were constructed using maximum parsimony and maximum-likelihood methods, with their bootstrap values being represented by the middle and lower values. Species in which orthologous ERV-L insertions were identified are highlighted in green boxes, while their encompassing clades are highlighted in blue. The golden mole and Seba's short-tailed bat were not incorporated into this phylogeny, and are therefore not included in this figure (adapted from Murphy *et al*., 2001a).

2.2.2 Sequence Alignment: Initial sequence alignments were constructed from approximately 1.2 Kb of homologous 5′ flanking sequence, slightly upstream (70 bp) of the ERV-L insertion. In order to download flanking sequence, the first method used was NCBI's URL-based system known as eFetch, where parameters such as chromosome number and locus start/stop values, among others, were entered into a specific eFetch URL. However, a more recent update to NCBI's BLAST interface has made it possible to download whole contigs on which matches are located, using the interface available on the online BLAST results page. Further analyses were re-done using flank obtained in this way. A BLAST search using the full-length ERV-L element as a probe was conducted against each of the selected genomes, and flanking sequence was obtained using the outlined method, for the top match in each species. Provided that at least 500 bp of sequence flanking the ERV-L virus was available on the same contig, the sequence was used. In the case of the species selected in this study, at least 1.2 Kb was available, with the exception of the golden mole, which was instead verified as an ortholgoue using just the 3′ flank due to insufficient sequence data at the 5′ end of the element. The 3′ flank was obtained in the same way, and loaded into an alignment of its own, comprising ~2 Kb of sequence. Insufficient sequence data for the Seba's short-tailed bat and flying fox meant that they were not included in this alignment.

Accordingly, three alignments were made in total – one comprising ~1.2 Kb of the 5′ flank, one comprising ~2 Kb of the 3′ flank, and one of the full-length ERV-L insertion across all eleven mammalian species. Sequences were aligned using EBI's multiple sequence comparison by log-expectation (MUSCLE) alignment algorithm (Edgar, 2004). This provided a useful guide, since the ERV-L elements in some of the species were interrupted by long insertions. Alignments were then adjusted manually in CLC's Main Workbench, where necessary, to give the optimal alignment. Using the same software, sequence conservation per base was recorded across all eleven species, and plotted onto a histogram. Once alignment had been established across all species, the trailing ends were trimmed for each of the three final

alignments, so that all sequences started and ended as synchronously as possible, provided that homology was intact across all species towards the proximal and distal ends of each alignment.

2.2.3 SGE Identification and Annotation: Once the main ERV-L alignment was complete, each species' orthologue was run through RepeatMasker (Smit *et al*., 1996-2010) to identify repetitive genetic elements present within the sequence. Species-specific repeat filters were applied, where applicable, and the locus of each identified repetitive element was annotated onto the main alignment in CLC's Main Workbench. In some species, shared SGEs were erroneously identified by RepeatMasker as being shorter than their orthologues in other mammals in which they were present. For such species, these annotations were extended manually, providing the underlying nucleotide sequence exhibited clear homology with the identified orthologous mammalian SGEs.

Certain species contained very large insertions within the ERV-L element, that were not common to all the other mammals. Therefore, following annotation of identified SGEs, the alignments were manually edited once more to make sure that sequence homology was optimised – particularly in regions separated by such large insertions. These manual adjustments to the alignment were made by eye, where patterns were visible. In more cumbersome areas, a combination of BLAST's Bl2Seq BLASTN aligner, ClustalW, EMBOSS' Needle and ClustalO were used. In order to attain the best results, more cumbersome regions were broken down into smaller units in one species, and aligned against longer units in another. Global alignment algorithms such as Needle provided the required sensitivity to identify regions of low homology.

2.2.4 Species-Specific Insertions: Once aligned and annotated, a second, shortened copy of the master ERV-L alignment was made. To do this, any insertions that were present in either one or two species, that amounted to longer than 10 bp, were deleted, in order to bring blocks of stronger, more universal sequence homology together in a way that would allow easier

visualisation of the orthologous ERV-L, as a whole. At every point where an insertion was deleted, a small flag was annotated onto the corresponding position on the alignment, labelled with the size of the deletion made. Insertions less than 10 bp were retained, and can still be seen in this shortened alignment. The flagged shortened alignment, and unshortened alignments can be found in the electronic supplementary information for the published study (details in appendix B).

One particular structure identified, was common only to the four Afrotherian mammals in the alignment. This was identified by RepeatMasker as an ERV LTR, and is located approximately 2,300 bp into the shortened master alignment. Using BLASTN, the 5′ and 3′ LTR of the HERV-L were selected, and were searched against the human genome. The highest-scoring paralogous LTRs were then obtained from these results. These paralogues were aligned, together with the inserted LTR taken from the elephant, along with its 5′ and 3′ LTRs. These sequences were aligned using MUSCLE, and used to construct a maximum-likelihood phylogeny, with branch support in the form of 1,000 bootstrap replicates.

2.2.5 Sequence Homology: Further analyses were conducted in the form of several pairwise alignments made using CLC Main Workbench. Two regions were selected for comparison, based on these being best represented across all eleven mammals, without significant interruption from insertions. The first was a 997 bp region from the start of the element: designated region A. The second region totalled 830 bp downstream of position 4,417 in the alignment: designated region B. Data was complete for all species in region B, but there was missing sequence data for the flying fox in region A.

Once the regions were selected, all gaps were manually removed, and these sequences were compared in all possible pairwise combinations, using CLC Main Workbench. The comparisons were outputted in the form of percentage identities between each sequence pair, to illustrate sequence conservation.

2.2.6 Phylogeny Construction: Subsequent analyses were in the form of two phylogenetic trees. Both phylogenies were maximum-likelihood trees, built using the MEGA bioinformatic suite (Tamura *et al*., 2011) and the TN93 substitution model (Tamura & Nei, 1993).

The first phylogeny was constructed from a region of the alignment just downstream of the 5′ LTR, comprising 1 Kb of ERV internal region (designated region 1). This region was selected based on showing good sequence conservation, while also representing nine out of the eleven mammals. Seba's short-tailed bat and flying fox were omitted due to insufficient sequence data in this region.

The second phylogeny was constructed from a shorter, 600 bp region starting at position 3,220 in the alignment (designated region 2). The purpose of this was to confirm the phylogenetic relationships represented by the first tree, and to include Seba's short-tailed bat and the flying fox, both of which were absent from the first phylogeny. Furthermore, this second region of the alignment insufficiently represented the Afrotheria, so these were omitted from this second tree.

Two co-phylogeny figures were then constructed from the first tree, using established mammalian phylogenies obtained from the literature as a reference (Eizrik *et al*., 2001; Delsuc *et al*., 2002; Jones *et al*., 2005; Perelman *et al*., 2011).

2.2.7 Dating the Initial Infection: LTR divergence calculations were performed in order to estimate the age of the initial infection event for this ERV-L element, past the minimum age boundary assigned by orthology and co-phylogenetic inference. This assumed the divergence of Eutherian mammals to have occurred 104 million years ago (Roca & O'Brien, 2005), and involves calculating the percentage of non-identical residues between LTRs of three different mammal combinations.

These are:

- Human (superorder III) vs. elephant (superorder I)
- Human (superorder III) vs. armadillo (superorder II)
- Human (superorder III) vs. horse (superorder IV)

The following calculations were done for each of these three combinations. The methodology outlined below uses the first species combination as an example:

1. 5′ LTRs of the human (superorder III) and elephant (superorder I) ERV-L orthologue underwent pairwise comparison using a global alignment algorithm: EMBOSS' Needle (Rice *et al*., 2000).
2. The same was then performed on the 5′ LTR of the elephant, and the human 3′ LTR, and the percentage of non-identical bases for both pairs were obtained.
3. Subtracting the percentage of non-identical bases of the human-elephant 5′-5′ and elephant-human 5′-3′ comparisons allow estimation of LTR divergence prior to the placental radiation.
4. The mouse was then used as a model for mammalian substitution rate, and the estimated amount of divergence between 5′ and 3′ LTRs could then be divided by this (see results, section 2.3.8).

2.2.8 Measuring Selection: In order to ascertain whether any selection was operating on the ERV-L element, d_N/d_S ratios were calculated. This is a measure of non-synonymous to synonymous substitution ratios. Non-synonomous substiutions are nucleic acid substiutions that alter the amino acid translated, whereas synonomous substiutions do not.

Measuring selection in this way requires a region of the ERV-L element to be aligned with another element from the same retroviral group, in which coding ORFs could still be identified. For these purposes, MuERV-L was chosen, and obtained from NCBI's Nucleotide database under accession number Y12713.

MuERV-L was then BLAST searched against the ERV-L element, and blocks of sequence homology were identified. For this, BLASTN was selected, and NCBI's Bl2Seq – a pairwise BLAST tool – was used. Leading and trailing sequence in between these blocks was deleted, with these deletions being made in multiples of three in order to keep the alignment in the same coding frame. Aligning blocks of the ERV-L sequence to MuERV-L, aimed to restore the original coding ORFs as best as possible – a requirement of d_N/d_S calculation. The species successfully aligned to MuERV-L and included in these calculations were:

- Armadillo (superorder II)
- Human (superorder III)
- Chimpanzee (superorder III)
- Pig (superorder IV)
- Horse (superorder IV)
- Seba's short-tailed bat (superorder IV)

None of the Afrotheria aligned sufficiently in the regions selected, which fell within a ~2.5 Kb stretch of sequence in this clade that did not align with the other mammals. Codonwise d_N/d_S ratios were calculated using HyPhy (Kosakovsky *et al*., 2005) as part of the MEGA bioinformatic tool suite (Tamura *et al*., 2011). *p*-values were also calculated in HyPhy, and values less than 0.05 were deemed significant, and used to reject a null hypothesis which states that sequences are evolving at a neutral rate (Kovakovsky *et al*., 2005; Suzuki & Gojobori, 1999). 108 codons were identified in the alignment, and the values for *p* and d_N/d_S were averaged.

2.2.9 Synteny: The HERV-L element under investigation is located on human chromosome 17. Chromosome-wide synteny of genetic blocks on chromosome 17 was analysed, and compared between the four species available in ENSEMBL's database, using their SyntenyView tool. These species were human, chimpanzee, horse and pig, representing superorders III and IV respectively. At the time of analysis, none of the other investigated mammals were available in the SyntenyView database.

2.3 Results

2.3.1 Screening: Several of the mammalian genomes available in NCBI's repository were screened, and orthologous ERV-L elements were identified in several of these. However, representative members of each of the four placental mammal superorders were selected – based on a few factors. These included good alignment to the ERV-L element, good alignment at the two flanking sequences to other orthologues identified and, in the case of unassembled genomes, presence of the entire element, and at least 500 bp of sequence of flanking the element on the same contig or sequence read.

The ERV-L probe was also BLAST searched against the genomes of available marsupials, but returned no significant, positive matches. Two monotreme genomes were also examined, and while several very small regions (>300 bp) of well-conserved SINE and LINE sequence were observed in the platypus (*Ornithorhyncus anatinus*), further traces of the ERV-L element itself could not be found.

2.3.2 Orthology: Two alignments of the flanking sequence were made, which also included the start of the ERV-L LTR. For the 5′ flank, all species except for the Golden Mole had sufficient sequence intact for alignment. Figure 2.3 shows a region of strongly aligning sequence flanking the 5′ LTR in these species. Golden Mole does, however, align with good homology at the flank of the 3′ end of the ERV-L element. For easy visualisation, a 457 bp insertion in the three included Afrotheria – just upstream of the 5' LTR – was excised from Figure 2.3.

For the 3′ flank alignment, all species except for Seba's short-tailed bat and flying fox exhibit good homology. These two species had insufficient sequence spanning the 3' flanking region, and were omitted for this reason. They are, however, confirmed orthologous by the presence of strongly aligning 5′ flanking sequence. Figure 2.3 also shows a region of the 3′ flank

alignment, where the nine included mammals share strong sequence homology.

2.3.3 Orthologous SGE Insertions: Analyses of each ERV-L orthologue using RepeatMasker revealed that a number of other selfish genetic elements had inserted into the ERV-L itself, many of these SGEs also being orthologous across the investigated mammals. Two elements in particular, showed very high sequence homology across all eleven mammals; these were a MIR3 retrotransposon (Figure 2.4) and a MER3 DNA transposon (Figure 2.5).

The main, shortened alignment, from which large insertions over 10 bp in two or less species were removed, is shown in Figure 2.6. The conservation histogram, shows good homology across the ERV-L element, where there is sequence data across all eleven mammals; each bar represents average sequence homology per 50 bp of sequence. However, at approximately 450 bp into the alignment, the four Afrotherian mammals do not align with the remaining seven mammals for a region spanning just over 3 Kb. This region is shaded green in Figure 2.6. In the elephant and manatee, both of which have the slowest generation times of the selected Afrotherian mammals, there appears to be a region near the start of the green shaded area in which two SINE elements have inserted, together with the remnants of a LINE. Neither are present in the aardvark or golden mole. Furthermore, a solo LTR structure – confirmed by phylogenetic analysis – is interrupted by a MER, SINE and LINE element in elephant and manatee, and by a SINE in the aardvark. This sub-region is lacking sequence data in the golden mole.

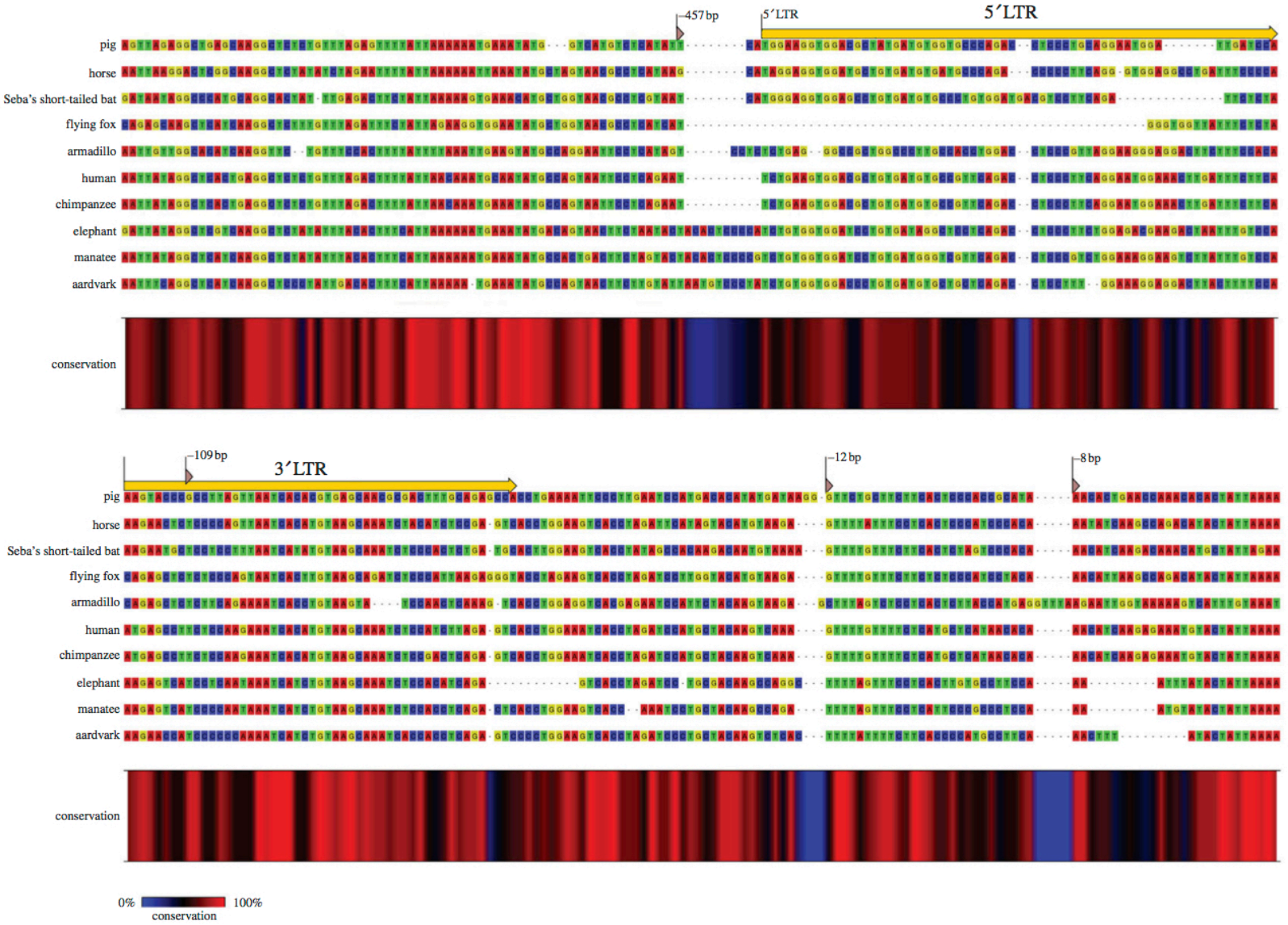


Figure 2.3: **ERV-L orthologue**, showing the extreme 5′ (top) and 3′ (bottom) ends of the element, together with a region of homologous flanking sequence across the ten of the eleven investigated mammals. The golden mole is not included as its genome did not include sufficient sequence data at these regions. For illustration purposes, insertions from certain species were deleted; this is indicated by flags in the sequence showing the amount of removed sequence. A blue-red colour gradient beneath the alignment represents per-base sequence conservation.

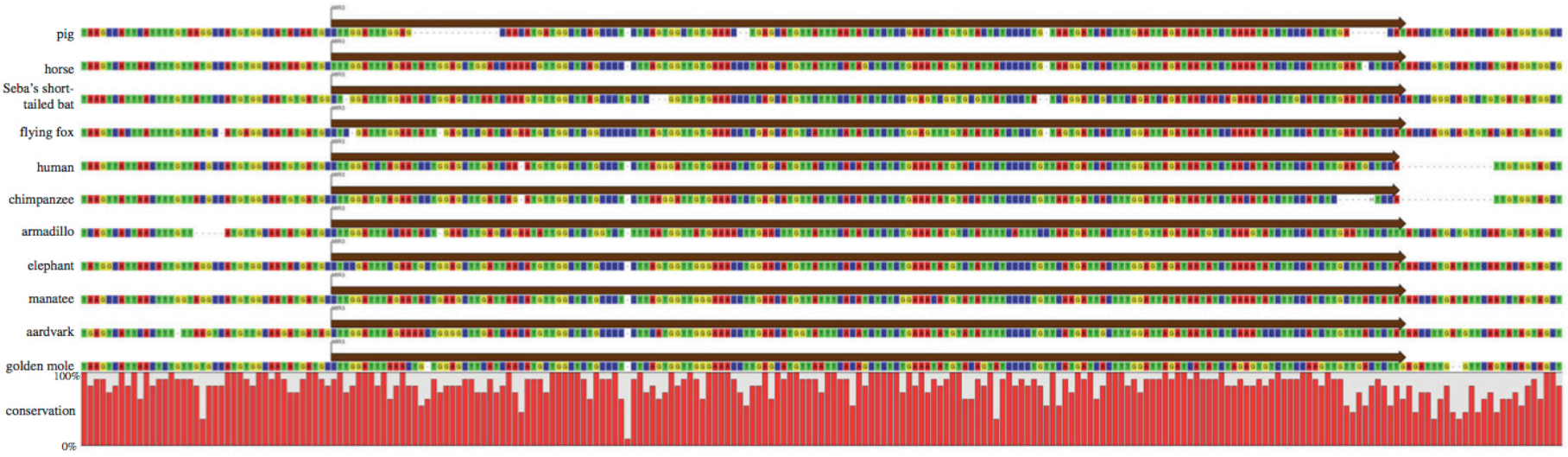


Figure 2.4: **MiR3 retrotransposon**, orthologous to all eleven investigated mammals. Two small regions of DNA flanking these are also shown. Sequence conservation is illustrated using a histogram beneath the alignment.

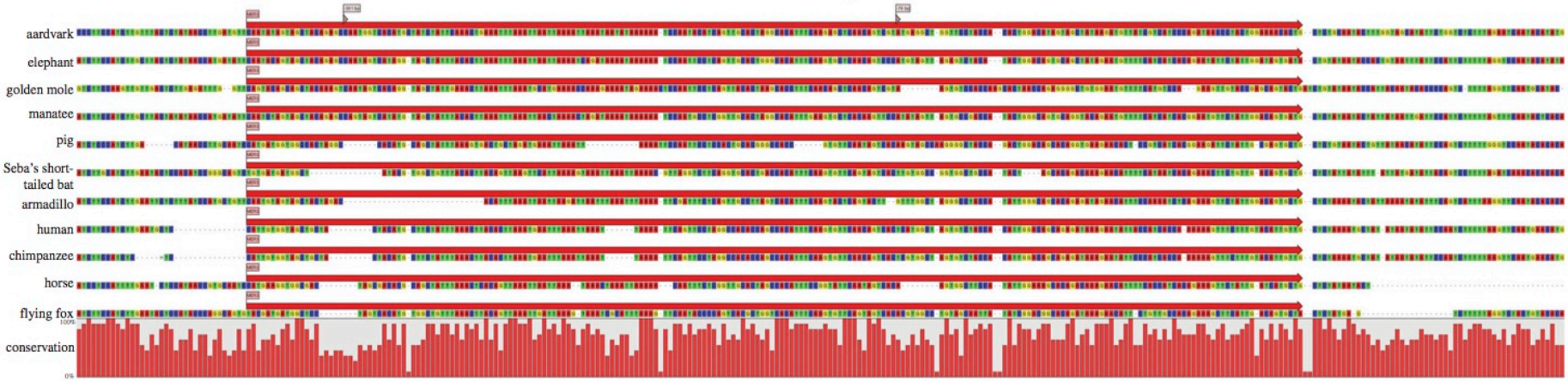


Figure 2.5: **MER3 DNA transposon**, orthologous across all eleven mammalian species, with the DNA alignment shown. A small region of sequence flanking these elements is shown on either side. A sequence conservation histogram runs along the underside of the alignment.



Figure 2.6: **ERV-L schematic**, showing the aligned element, together with identified SGEs and genomic repeats. A conservation histogram at the bottom indicates sequence homology per 50 bases. A scale bar along the top illustrates alignment position, in bases, for reference. A solo-LTR structure is identified in beige in the four Afrotheria; the region of low homology (referred to in the text) is shaded light green. Two regions, 1 and 2, are demarcated above the figure and correspond to the sequence selected for building phylogenies. The presence of sequence information is indicated by a black line, to clarify contiguous regions where gaps have been made to maintain alignment.

Examination of the ERV-L across the eleven mammals indicates that they have most likely been inactive for a considerable length of time, with no clearly identifiable ORFs still intact. All eleven species exhibit intact, identifiable 5′ LTRs, and all except for the golden mole have conserved 3′ LTRs, as well.

Despite no obvious ORFs, there were several points in the sequence identified by RepeatMasker as ERV-L internal region. This was particularly well conserved close to the 5′ LTR, approximately 500-1,000 bp into the alignment; conservation of this region was clear in all species except for the armadillo, in which there were deletions here. Approximately 3,200-3,800 bp into the alignment, ERV-L internal region was also conserved among the Boreoeutherian and Xenarthran mammals.

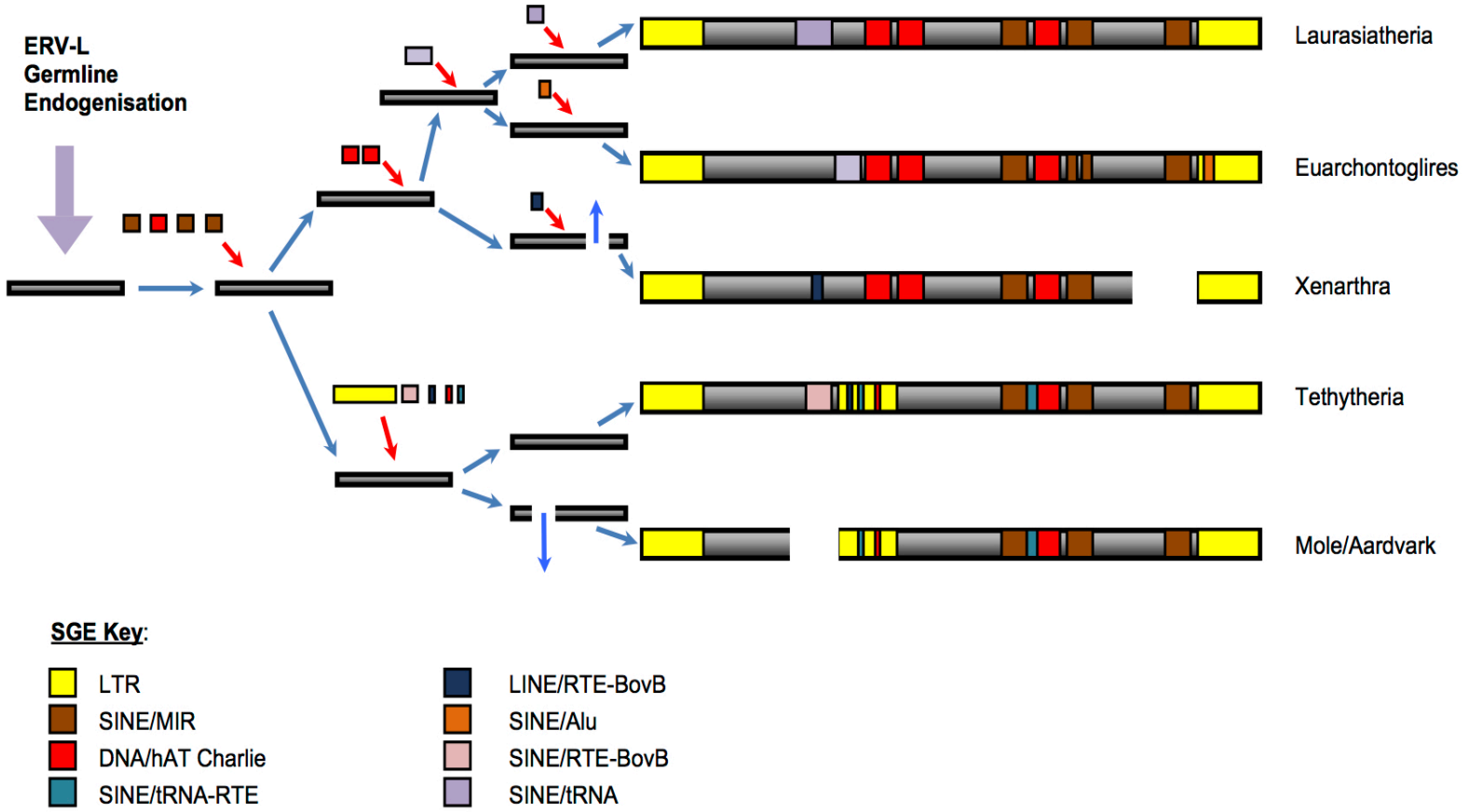


Figure 2.7: **Schematic of SGE insertion**, showing the likely pattern of SGE insertion into the ERV-L, following the initial integration event. The main SGEs are highlighted here, and explained in the key. Red arrows denote an SGE insertion, and blue upwards and downwards-pointing arrows (emerging from an element) denote a deletion event.

Other highly conserved regions included shared SGEs: a MIR at approximately 4,000 bp (Figure 2.4) and a MER3 approximately 4,750 into the alignment (Figure 2.5). Both of these SGEs were strongly conserved across all eleven mammalian species. Other SGEs occurred with strong conservation, but not across all species. A schematic representing a likely order of SGE insertion is shown in Figure 2.7. Notably, at an approximate position of 5,800 bp into the alignment, there is another MIR element, found in all but the armadillo and golden mole; these two species exhibited deletions in this region.

Full-length – shortened and un-shortened – flank and element alignments are available online (details in appendix B).

2.3.4 Afrotherian Region of Low Sequence Homology: The light green shaded area on Figure 2.6 highlights a region in which the four Afrotherian mammals chosen exhibit low sequence homology when compared to the other Eutherian species. Within this region is the solo-LTR structure, confirmed by the aforementioned maximum-likelihood phylogenies using paralogous HERV-L LTRs. It is likely that this was due to a second retroviral endogenisation, and the subsequent recombination of its homologous 5′ and 3′ LTRs, resulting in excision of the retroviral genome, while leaving behind a solo-LTR.

Interestingly, this solo-LTR was itself the subject of further retrotransposition by SINE and LINE insertions. In the Tethytherian mammals, between approximately 1,100-1,850 bp, an ART2A SINE is present, followed by the remnants of a RTE1_LA LINE, followed again by a SINE2-1_Pca. Also present in both Tethytherians, is a MER5A DNA/hAT-Charlie transposon at approximately 2,450 bp, and a L1 LINE at approximately 2,800 bp into the alignment. Common to both Tethytherians and the aardvark, is a AFROSINE1B SINE element at approximately 2,650 bp into the alignment, although there is a deletion encompassing this region in the golden mole. The

same class of SINE can be seen inserted approximately 4,200 bp into the alignment, and is orthologous to all four Afrotherian mammals in this study.

2.3.5 Unidentified Sequence: A region of just under 800 bp in length (~5,000-5,750 bp in the alignment) was largely unidentified by RepeatMasker, excepting the remnants of a SINE element in the aardvark. Despite this, the sequence is aligned, albeit with poorer, but not negligible, homology (averaging approximately 57% sequence similarity). It is unclear as to whether this region was subjected to heavy mutational processes, or whether this is the remnants of an insertion; both are likely to have occurred prior to the divergence of the Afrotheria from other Eutherians. This unidentified region is flanked on either side by highly conserved MIR repeats; the one flanking its 5′ end being present in all eleven species, and its 3′ MIR present in ten.

2.3.6 Sequence Homology: The sequence identities calculated in all pairwise combinations for the eleven species analysed, are presented in Table 2.2. Per-base identities were averaged across two regions – A and B – from which gaps were removed.

Table 2.2: **Table showing pairwise, percentage sequence identities** for all combinations of the eleven mammals, across two regions of the ERV-L orthologue. Values calculated from region A (positions 1-997) are shaded in blue, while those calculated from region B (positions 4,417-5,247) are in red. The flying fox was omitted from comparisons for region A, since it presented insufficient sequence data.

observed identities (%)		1	2	3	4	5	6	7	8	9	10	11
pig	1		68	64	62	65	65	57	62	65	59	51
horse	2	75		72	69	72	72	66	69	73	65	59
Seba's bat	3	62	67		64	65	65	60	63	67	59	55
flying fox	4					63	63	58	60	62	58	53
human	5	68	72	59			98	62	67	69	64	55
chimpanzee	6	67	72	59		98		62	67	69	64	55
armadillo	7	34	39	35		36	35		63	66	60	55
elephant	8	58	63	63		59	59	37		83	73	65
manatee	9	60	64	65		62	61	38	83		71	65
aardvark	10	56	60	61		59	58	39	70	73		59
golden mole	11	45	48	50		46	46	25	58	61	55	

For region A, homology of up to 65% sequence identities was observed between superorders I and IV, and the same is reflected in region B. As expected, human-chimpanzee exhibited high 98% sequence identities for both regions. Notably, the golden mole exhibited 25-50% sequence identities relative to the Boreoeutheria and Xenarthra for region A, while this spanned 51-59% for region B. A uniquely low homology was observed for this species relative to the armadillo in region A, but this could be explained by a lack of internal region in the armadillo, and a slightly shortened 5′ LTR in the golden mole. In region A, the armadillo showed sequence identities between 34-39% when compared to superorders III and IV, while this ranged from a notably higher 57-66% for region B.

Elephant-human comparisons revealed high sequence conservation in region B in particular, reaching identities of 67%. Considering these two species diverged more than 104 million years ago, this level of sequence conservation may be telling of the evolutionary rates of these two mammalian species. Notably, these two species also have particularly long gestation periods and generation times. This idea is supported by a similarly highly conserved region B between horse and elephant: 69% sequence identities. Like humans and elephants, horses too share a longer gestation period and generation time, which may act to slow the evolution of the ERV-L insertion, over time.

2.3.7 Co-Phylogenies: A simplified mammalian phylogeny depicting published phylogenetic relationships between the eleven investigated species was drawn from the literature, and arranged beside the retroviral phylogeny constructed from the ERV-L orthologue. This was done twice, each time using a different region of the alignment for tree building. Figure 2.8 depicts all mammals excepting the flying fox, which was ommitted due to insufficient sequence data; the retroviral phylogeny was drawn from region 1 of the alignment (indicated at the top of Figure 2.6). Figure 2.9 depicts just the Boreoeutherian and Xenarthran mammals, and is drawn from region 2 (indicated at the top of Figure 2.6).

The first co-phylogeny (Figure 2.8), shows that the retroviral tree correctly represents the phylogenetic relationships of the ten species investigated, when compared to the mammalian tree drawn from the literature. The topology of the Afrotherian and Xenarthran clades differs, however. In the second co-phylogeny (Figure 2.9), the relationships on the retroviral tree are less representative of those on the existing mammalian tree.

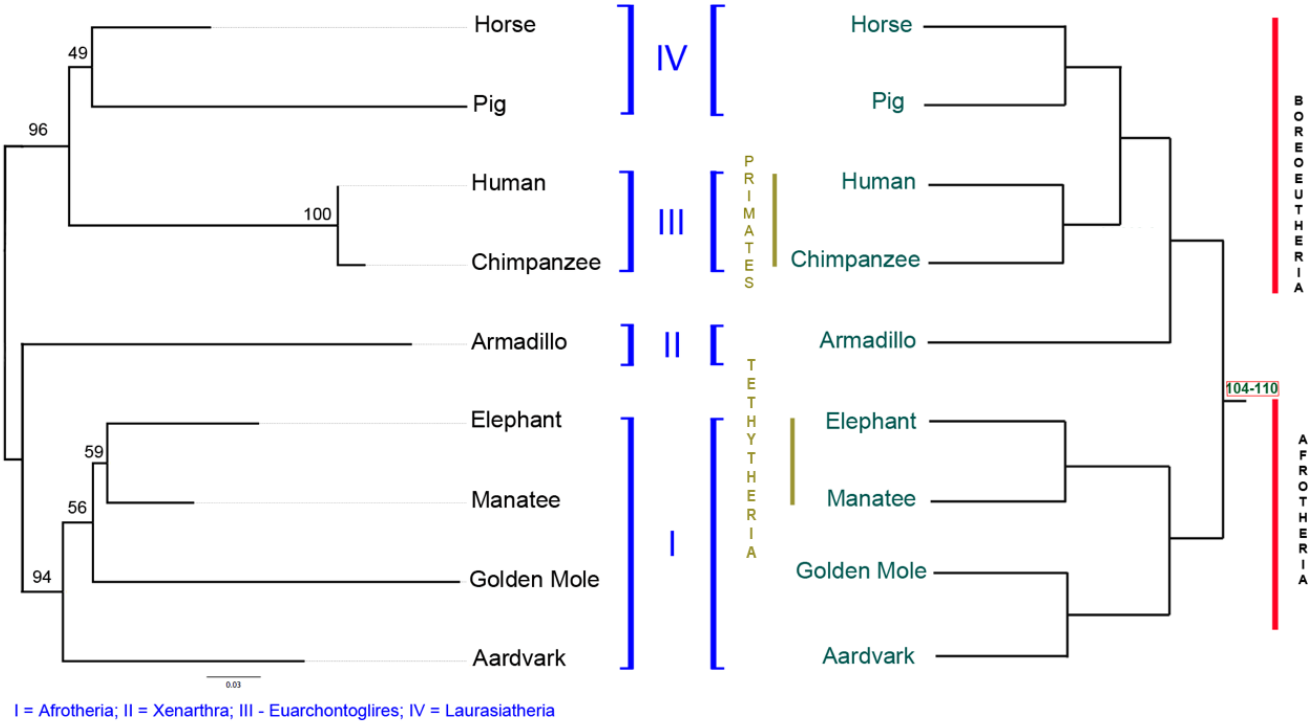


Figure 2.8: **First co-phylogeny** built from region 1 of the alignment, showing the retroviral tree (left) and mammalian phylogeny (right), derived from the literature (Roca & O'Brien, 2005; Eizirik *et al*., 2001; Delsuc *et al.* 2002; Jones *et al*., 2005; Perelman *et al*., 2011). Eight of the eleven investigated species are shown, excepting Seba's short-tailed bat and flying fox. The branch lengths on the retroviral tree are scaled, indicating the number of substitutions per site, with numbers corresponding to bootstrap values. The four superorders are bracketed in blue, while Afrotheria and Boreoeutheria are bracked in red., Primate and tethytherian clades are also indicated in beige.

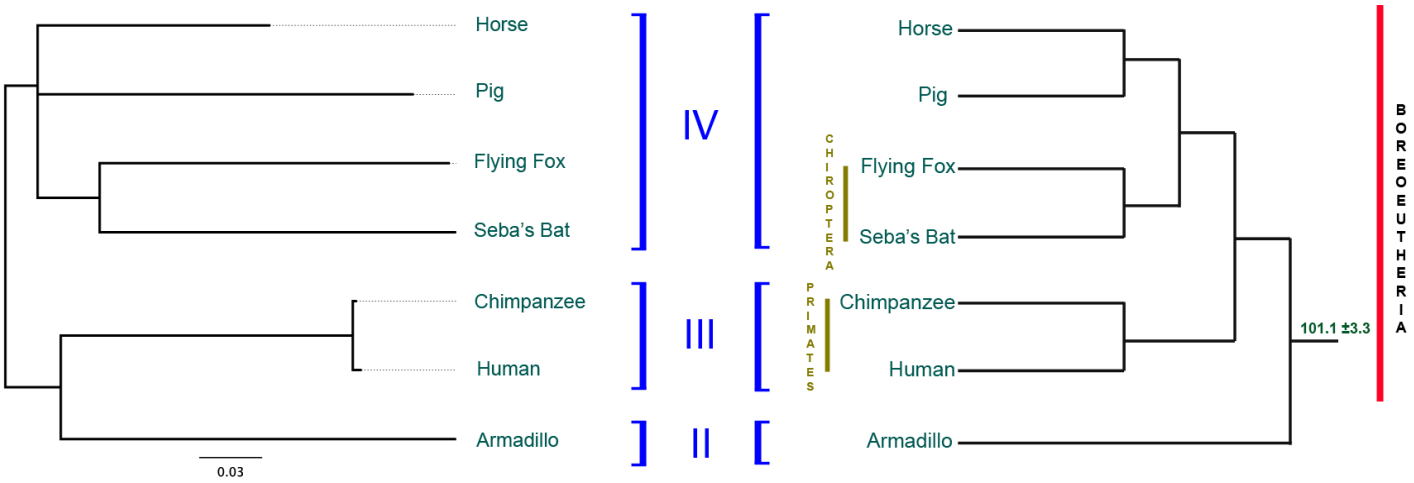


Figure 2.9: **Second co-phylogeny** built from region 2 of the alignment. The retroviral tree is shown (left) together with a mammalian phylogeny (right), derived from the literature (Roca & O'Brien, 2005; Eizirik *et al*., 2001; Delsuc *et al.* 2002; Jones *et al*., 2005; Perelman *et al*., 2011). Species included Xenarthran and Boreoeutherian mammals only. The branch lengths on the retroviral tree are scaled, indicating the number of substitutions per site, with numbers corresponding to bootstrap values. The four superorders are bracketed in blue, while Boreoeutheria is bracked in red. Chiropteran and primate clades are shown in beige.

2.3.8 Dating the Initial Infection Event: While reconstructing an orthology for the ERV-L retrovirus is an accurate way of assigning the element a minimum age, it is quite likely that the date of the initial infection is prior to this. Orthology in all four superorders suggests that this ERV-L infected an ancestral mammal prior to the divergence of the Eutheria. Given that it is present in all the Eutheria indicates that the ERV-L underwent deactivation, in order to become neutral to the host, and reach fixation in the population of this ancestral, pre-Eutherian mammal. Thus, it is unlikely that the date of the initial infection event was too close to the minimum age deduced from orthology and comparison against existing mammalian phylogenies. Further corroborating this is the fact that there are shared SGEs – two of which are present across all eleven mammals. This would suggest that the virus had been integrated and defective for long enough for these SGEs to insert, reach fixation and be vertically transmitted in the ancestral Eutheria.

In order to estimate how long prior to the placental radiation the ERV-L infection occurred, divergence between the two retroviral LTRs was

calculated. This makes the valid assumption that, upon integration, the two LTRs are identical in sequence, and accumulate mutations at random, at a steady rate, as a result of normal genetic mutational processes (Gilbert & Feschotte, 2010). Calculating the percentage divergence between pairs of species' 5′ LTRs – or their 3′ LTRs – can give an estimate of the initial date of infection. Using EMBOSS' Needle, pairwise sequence identities were obtained for human and elephant 5′ LTRs, and it was observed that they were divergent from one another by 31%. This 31% of change is assumed to be steady nucleotide substitution occurring since the divergence of placental mammals. In the same fashion, the 5′ LTR and 3′ LTR of the elephant ERV-L were compared using Needle, which revealed that they diverged by 53%. By subtracting 31% from 53%, an estimate for the amount of sequence change that occurred prior to the Afrotherian divergence from other Eutheria (104-110 million years ago) is obtained (Roca & O'Brien, 2005; Eizirik *et al*., 2001;Springer *et al*., 1997; Katzourakis *et al*., 2007). This value was 22%.

The next step of the calculation makes an inaccurate assumption that the neutral mutation rate of a species is the same across evolutionary time. However, owing to the lack of information available on neutral mutation rates of the mammals investigated, this assumption will be made in order to estimate the infection date. Another factor to take into account is the possibility of gene conversion and recombination events affecting the LTRs of retroviral insertions. Irrespective of this, here it will be assumed that the ancestral placental mammal was physically small, and had a relatively fast generation time, as might be comparable to extant mice. A neutral mutation rate of 4.5×10^{-9} substitutions per site per year – derived from the mouse genome – was applied, together with the lower boundary of the Eutherian divergence date. The calculation is therefore:

$$22/(0.45 \times 2) + 104$$

The calculated value for the human-elephant pair was 128 million years. This calculation was then repeated for pairs of human and horse, giving an

estimate of 139 million years; human and armadillo gave an estimate of 140 million years.

It is therefore suggested that the ERV-L retrovirus infected the genome of a pre-Eutherian ancestral mammal by up to 24 million years prior to the divergence of the Eutheria, potentially preceding this by up to 36 million years. It is, however, important to note that these age estimates assume a fast generation time for all mammals. This could mean that the initial integration dates given here may not be extrapolated back as far back as they could be. Estimating this with more certainty, would require a deeper understanding of ancestral mammal sizes, evolution and generation times.

2.3.9 Selection: Given that the ERV-L retains sequence homology of up to 69% between elephants and humans, it may seem possible that this ERV-L provirus was under selection. In order to ascertain this, d_N/d_S ratios were calculated across a small, 324 nucleotide region of the ERV-L alignment in the seven investigated Boreoeutherian and Xenarthran mammals; this was the longest region retaining homology to MuERV-L (the only ERV-L with recognisable ORFs). Using HyPhy, codonwise d_N and d_S values were calculated, and averaged to give a d_N/d_S ratio whose p-value was 0.71. Deciding that $p \leq 0.05$ would be the threshold for significance, a p-value of 0.71 is unable to reject the null hypothesis assuming sequences are evolving at neutrality. Therefore, despite only being able to perform this selection test on such a small region of the ERV-L sequence, it is unlikely that any form of selection is acting on this ERV, and that it is indeed evolving at neutrality.

2.3.10 Synteny: Analysis of human chromosome 17 – on which the ERV-L insertion is present – revealed that the syntenic regions containing the ERV-L were clearly intact, when compared against chromosome 17 of the pig, horse and chimpanzee (Figure 2.10). Since humans and chimpanzees are the least divergent of the three species pairs in Figure 2.10, the arrangement of the ERV-L syntenic block is in the same orientation, and the same position on the chromosome arm. Gene orders are scrambled between human, pig and horse, however. In both species, relative to human, the ERV-L-containing

block is on the opposite arm of the corresponding chromosome. The ERV-L-containing syntenic block is also in the same orientation as human in both pig and chimpanzee, but is inverted in the horse.

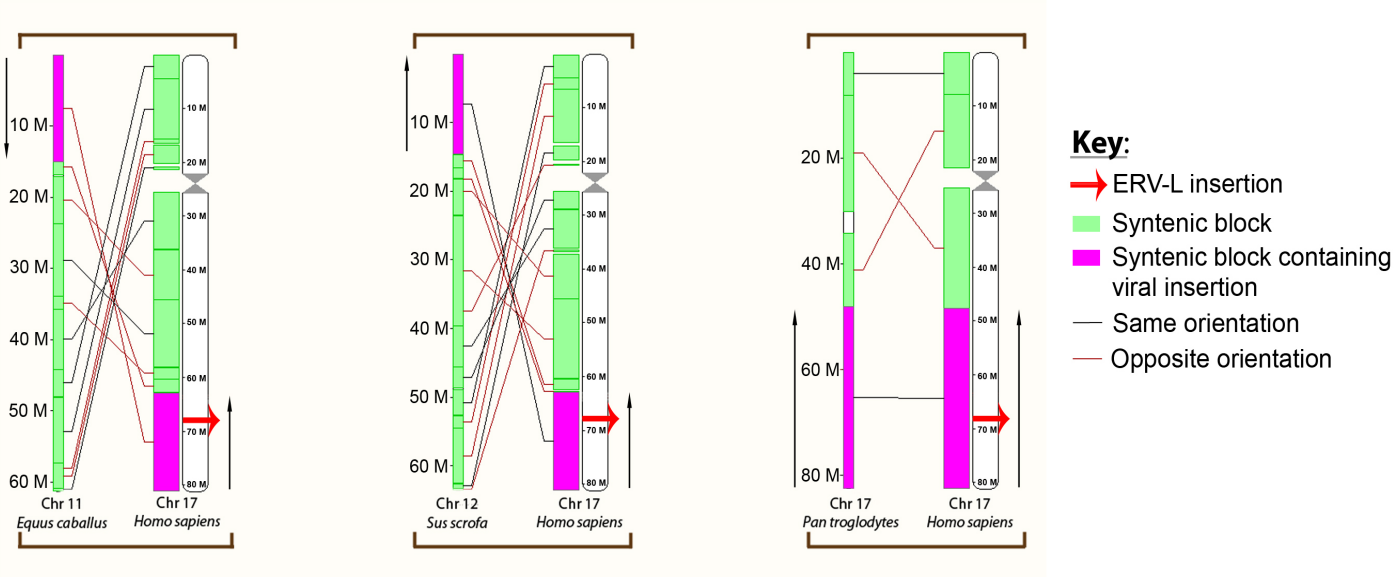


Figure 2.10: **SyntenyView analysis** of paired combinations of human against horse, pig and chimpanzee. Chromosome ideograms are shown and numbered along the bottom, with a scale in megabases shown on the left. Syntenic blocks containing the ERV-L insertion are shaded pink. Black arrows on the sides correspond to the orienation of the block, also indicated by thin, coloured lines connecting syntenic blocks.

2.4 Discussion

One of the most reliable methods for dating genetic elements, is to reconstruct orthologies across divergent host species, and compare these to well-resolved phylogenies, and existing divergence dates. This enables relatively accurate minimum age estimations. There are many investigations that have estimated integration dates for ERVs, but few have used orthology as a line of evidence. Thus, the ERV-L described here is the oldest virus dated with any degree of certainty, since it exhibits a clear shared flanking sequence across all four superorders of placental mammals. A minimum age of 104-110 million years is assigned based on comparison to existing mammalian phylogenies, and the date of divergence of Afrotheria from the Xenarthran and Boreoeutherian mammals (Roca & O'Brien, 2005).

Other studies using orthologies to date ERVs have only been conducted on primates, but have not included more basal mammalian groups (Johnson & Coffin, 1999). Previous investigations have also recorded abundant ERV-L elements across different placental mammals, leading to inference that this retroviral group is at least 70 million years old; although this was not corroborated by any proof of orthology between these elements (Bénit *et al*., 1999).

In this study, several evidences supporting the orthology of this element are provided. Host genome infection by retroviruses is a random, non-targeted process, meaning that orthology is highly unlikely to arise from two infection events in independent hosts, at the same genomic locus. Therefore, the observation of an ERV-L at the same locus across multiple mammalian hosts is almost certainly due to infection of a mammal ancestral to these species. Paralagous elements are more readily identified, but would not share the same flanking sequence, unless they inserted within a transposable element that later proliferated. By running the 5′ human flanking sequence through RepeatMasker (Smit *et al*., 1996-2010), no SGEs were identified, except for a single 111 bp SINE insertion. Similar analysis of the 3′ human flanking sequence revealed a LINE, two smaller LINE fragments and a SINE, all of which did not amass to the total flank of just over 2 KB. These began approximately 500 bp into the sequence. This, together with no identifiable LINE fragments in the 5′ flank suggests that this ERV-L did not insert within another repetitive element, and that identification of similar flanking sequence in other mammalian species – whose genomes are unassembled – is sufficient evidence of orthology. These inferences can also be applied to the remaining mammalian species, since there were several regions of strong sequence homology between all mammals in the 5′ flank and 3′ flank alignments (Figure 2.3). Synteny analyses of mammals with assembled genomes also revealed that syntenic blocks containing the ERV-L were intact; further evidencing orthology between these speices (Figure 2.10).

Estimations of the initial infection date based on LTR divergence concluded that the ERV-L element first infected an ancestral mammal at least 24 million years prior to the divergence of Afrotherians from Boreoeutherians and Xenarthrans (see results, section 2.3.8). Corroborating this, is the fact that the ERV-L insertion is not uninterrupted, but instead shares multiple orthologous SGEs, which would likely require time to become fixed in a pre-Eutherian ancestral mammal, prior to being vertically transmitted into Eutherians. For instance, the time taken for a gene to drift to fixation in humans is approximately 800,000 years (Kimura, 1984). The fact that several SGEs are present, many of which are themselves interrupted by subsequent insertions, suggests that it was unlikely they integrated at the same time, and that the infection of the ERV-L element was likely considerably earlier.

The most notable orthologous SGEs identified were a MIR element (Figure 2.4) and a MER element (Figure 2.5) in all eleven species. These were all found in the same orientation relative to the ERV-L element, and RepeatMasker identified these to be from corresponding SGE families. Interestingly, the conservation histogram shows that these three SGEs retain some of the highest levels of sequence conservation observed in the entire alignment. Furthermore, another MIR present in nine mammals (superorders I, III and IV) and two DNA transposons present in seven mammals (superorders II, III and IV) were also identified. All these SGEs are conserved within the ERV-L element, which supports the likelihood of the original germline retroviral infection having occurred approximately 128-140 million years ago.

In order to confirm that the element had not inserted prior to the divergence of non-Eutherians and Eutherian mammals, BLAST was used to screen the genomes of available monotremes. No significant matches were found in these mammals. Two marsupial genomes were also screened, also returning negative results. Since monotremes are estimated to have diverged from placental mammals 167 million years ago, and marsupials 147 million years ago (Bininda-Emonds *et al*., 2007), this lack of matching ERV-L sequence in these clades supports the estimated integration date of ~128 million years

derived from LTR divergence. Despite this, several of the inserted Alu SINEs and LINE elements were homologous in monotremes, but without the accompanying retroviral sequence. The limited size of the contig on which the element was present also prevented further analyses. Should these non-Eutherian mammals be sequenced at higher genome coverage in the future, it may be possible to more conclusively decide whether or not the ERV-L retroviral group infected these hosts.

As well as definitively identifying the oldest orthologous ERV-L, this investigation has also revealed the oldest MIR to be dated by orthology. Previously, it was thought that MIR elements were approximately 130 million years old, with orthology retracing elements as far back as 65 million years, and the remaining age calculated by application of a molecular clock (Jurka *et al*., 1995). However, two MIR elements – orthologous to all eleven mammals – are presented here which, like the ERV-L, suggests a pre-Eutherian origin. However, since these SINE elements do not have LTRs or a similarly homologous structure, it is only possible to assign them an upper age boundary of 104 million years. Previous SGE orthologies have shown that *Can* SINE elements have a minimum age of 40 million years (Zehr *et al.,* 2001), although – at present – none have been shown to be older than the SGEs presented here.

An orthologous MER DNA transposon was also identified (Figure 2.5). Previously thought to be 55 million years old without orthology, the MER3 orthologue identified here appears to be at least 104 million years old. It is documented that there is a high degree of variability within the MER3 family, leading to poor sequence conservation between elements (Jurka, 1990). This observation is confirmed by the sequence conservation histogram of the MER3 in Figure 2.5. When compared against that of the MIR in Figure 2.4, the sequence conservation of the MER3 orthologue is notably poorer than the highly conserved identified MIRs.

When a viral insertion is vertically transmitted from parent to offspring, it remains at the same genomic locus and is not subjected to selection

pressures. Any accumulated substitutions are most likely to be due to the error rate of the host's DNA replication mechanisms, as the virus evolves at its neutral rate. For an ERV to persist for over a hundred million years would imply that the virus was neutral to its host, and under no form of purifying selection. This is particularly characteristic of orthologous versus most paralogous ERVs. However, if a provirus exists as a paralogue in somatic cells, it is likely that these viruses exist as part of an infectious pool of replication-competent elements. These elements will likely be under host purifying selection, until successful integration into the germline occurs. Therefore, between germline integration events, ERVs infect somatic cells – a process which subjects them to purifying selection. However, once integrated into the germline – particularly when persistent over generations – they become exempt from selection pressures (unless co-opted), and are subjected only to the host's neutral rate of evolution (Belshaw *et al*., 2004).

Orthologous ERVs are therefore expected to have d_N/d_S ratios at unity. d_N/d_S ratios are a measure of non-synonomous (nucleotide changes altering the translated amino acid) and synonomous (nucleotide changes that don't alter the translated amino acid) substitutions (Mugal *et al.*, 2014). When d_S is greater than d_N, the resulting ratio is smaller than unity, and is indicative of purifying selection. Conversely, positive selection is indicated by d_N/d_S ratios above unity. Calculations of d_N/d_S across a small selection of the master alignment revealed that the null hypothesis assuming sequences are evolving at neutrality could not be rejected. This both supports orthology of the ERV-L elements, while also revealing it unlikely that recombination of the full ERV-L element could have occurred, while retaining the same flanks, since this would also be expected to reduce d_N/d_S. Since the values obtained in this investigation suggest the ERV-L element in question is evolving at the host's neutral rate, it would also be expected that phylogenetic relationships between ERV-L orthologues in the studied mammals would reflect those resolved by established mammalian trees. The retroviral sequence would, in effect, be no different, and should represent the same substitution rate as any other host-neutral sequences.

The first co-phylogeny (Figure 2.8) shows great similarity between retroviral and existing mammalian trees. The topology of Xenarthran and Afrotherian clades is, however, slightly different between the two trees, as would be expected when trying to resolve phylogenetic relationships from part of an endogenised retrovirus. This stands true for the second co-phylogeny (Figure 2.9), constructed from a smaller region of sequence, still. Further evidence supporting a similar relationship reflected by mammalian phylogenies and the ERV-L sequences reported in this chapter, is the pattern of SGE insertion. The SGEs appear to have inserted in a pattern that mirrors the speciation of the placental mammals (Figure 2.7), further augmenting the orthology of the ERV-L element under investigation.

A factor that is likely to have contributed to the ERV-L element being so highly conserved over such a long period of time, is the generation time of the mammals selected. With the exception of the bats and golden mole, larger mammals were chosen for these reasons. The higher sequence conservation observed between these confirms the importance of these choices when reconstructing orthologies. In the smaller mammals such as the golden mole, for example, the range of sequence identities shown in Table 2.2 was lower, when compared to larger mammals. Furthermore, the golden mole and flying fox both exhibited areas of apparent deletions where larger mammals were aligning with one another. The shorter generation times of smaller mammals tends, in general, to speed their rate of genome evolution; the net result being that ancient ERVs are not preserved as well as might be expected in larger mammals, with slower neutral evolution rates. This is confirmed by the generation time-effect hypothesis, which states that species with shorter generation times turnover their genomes at a faster rate, inducing a higher rate of DNA replication error than might be observed in a larger species (Laird *et al*., 1969). With the exceptions of the golden mole and Chiropteran species, many of the mammals selected here have long life spans, and thus slower nucleotide substiution rates; such as the elephant (50-70 years), and human (~75 years) (Goodman *et al*., 2009). Exceptional to this hypothesis, are ERV insertions under selection from the host organism, as would be expected in a co-opted element, or one from which certain genes or genomic regions had

been exapted (Lynch & Tristem, 2003; Blaise *et al*., 2005). The ERV-L reported here is, however, unlikely to have been under negative selection. Not only is this suggested by the d_N/d_S ratio calculations, but also by the lack of identifiable ORFs. If under some form of negative selection, it would be expected that retroviral genes persisting in the germline for this length of time would encode a functional protein product, thus exhibiting identifiable ORFs. Upon analysis, the reported ERV-L contained multiple in-frame stop codons and frame-shifting indels, which refute this possibility.

Aside from the importance of selecting the slowest-evolving hosts possible, there are other considerations to take into account when reconstructing ERV orthologies. While genome sequencing projects are continually advancing, many of the less scientifically important species are sequenced at low coverage, and are left unassembled. In the case of identifying ERVs – many of which are up to 10 Kb in length – this is far from ideal, since an individual contig or sequence read may not contain either the element itself, or the flank necessary to assume an ortholgue.

Another consideration is the alignment algorithms used. Ancient, evolutionarily disparate DNA proviruses are likely to have been affected by all manners of indels, SGEs, gene conversions and recombination events. Using local alignment algorithms such as BLAST – with its word-size-based approach of detecting adjacent, small regions of high conservation – is not always favourable. Breaking down probe sequences into smaller units and using these can often help, but more effective still is the use of more powerful algorithms such as MUSCLE for retrieving sequence homology. While this is the most effective method, it is not computationally as efficient to search an entire genomic database using MUSCLE, as can be done easily with BLAST. In fact, MUSCLE does not currently provide a 'search' facility, but is instead suited to aligning sequences that are within similar genomic regions. Conversely, BLAST incorporates technologies that allow fast searching of large amounts of data, but at the cost of sensitivity. As a compromise, identifying the broader genomic location of an element using smaller probes with BLAST, can allow for extraction of larger blocks of sequence that can be

more easily computed by alternative algorithms like MUSCLE. Once a whole element is identified, similar methods can be employed to compare flanks in order to ascertain orthology. Within an ERV alignment, cumbersome regions should further be adjusted manually, aided by global sequence aligners such as Needle, Bl2Seq and ClustalO.

Conclusions: With the use of several different bioinformatic tools and techniques, and after several different analyses, this investigation presents the oldest, definitively dated virus, and the oldest published MIR and MER3 elements. These studies may set a new benchmark for the field of paleovirology and paleogenomics in general, where the use of several approaches together with an ever-increasingly diverse global library of genomes, can facilitate the identification of evolutionarily ancient DNA. Using such techniques and orthology reconstruction could even take us closer to finding out more about the elusive origins of endogenous retroviruses, and retrotransposons, at large.

CHAPTER 3: Archaic Hominid Retroviral Screens

3.1 Introduction

In 2008, the remains of two archaic hominids were discovered in Denisova Cave in the Siberian Altai mountain range. The first was a ~50,000 year-old fossilised finger phalanx, from which an archaic hominid genome was sequenced. The first attempt to sequence this sample resulted in a 1.9-fold, low-coverage dataset (Reich *et al*., 2010), while a second, subsequent effort resulted in a 30-fold, high-coverage genome published in 2012 (Meyer *et al*., 2012). This genome belonged to an extinct hominid subspecies previously unknown to anthropologists, termed Denisova man. Denisovans are thought to be closely related to a second extinct hominid subspecies: Neanderthals. While several fossil samples from around the world have resulted in low-coverage genomes – and many partial, draft genomes – being made available for Neanderthals, it was only in late 2013 that a second specimen was recovered from Denisova cave, from which a high-coverage genome was derived. This genome was sequenced to approximately 52-fold coverage, and was derived from a proximal toe phalanx belonging to a female 'Altai Neanderthal' (also found in Denisova Cave). While both phalanxes were found at the same site, both the stratigraphy of the cave sublayer from which the bones were extracted, and phylogenetic analyses suggest that the Neanderthal toe phalanx is the older sample (Prüfer *et al*., 2014).

From these genomes, it was estimated that the divergence of anatomically modern humans from Denisovans and Neanderthals, occurred approximately 553-589,000 years ago (Prüfer *et al*., 2014). The subsequent divergence of Denisovans and Neanderthals from their common ancestor was thought to be approximately 381,000 years ago (Prüfer *et al*., 2014). Considering one retroviral group – HERV-K(HML-2) – has been implicated to be active within the last 100,000 years (Turner *et al*., 2001), it is feasible to consider that certain HERV loci may differ across these three hominid lineages. In 2012, Agoni *et al*. conducted a retroviral screen of HERV-K(HML-2) elements on two low-coverage genomes: one on Denisova man (Reich *et al*., 2010) and the other on Neanderthal genomic sequence (Green *et al*., 2010). They presented 14 HERV-K proviruses absent from the human reference sequence, and

concluded that these retroviruses were infecting their hosts at the point of, or following, their divergence from modern humans. Since the question of HERV-K(HML-2) retroviruses as possible human pathogens has yet to be definitively resolved, any evidence pointing towards this retroviral group being active today, is of recurrent academic interest.

In light of this, ongoing research continually presents a growing number of HERV-K proviruses that appear to have inserted after the divergence of chimpanzees and humans, approximately 6 million years ago (Reus *et al*., 2001). HERV-K(HML-2) elements are the most medically relevant of these, since their expression has been observed to be upregulated in certain tumours, and in patients suffering rheumatoid arthritis and schizophrenia (Subramanian *et al*., 2011). This group gets the parenthesised part of its name due to its phylogenetic similarity to mouse mammary tumour virus (MMTV); HML is an abbreviation of *h*uman *M*MTV-*l*ike (Andersson *et al*., 1999). The insertionally polymorphic HERV-K(HML-2) elements that have been reported in several different studies of human genomes (Belshaw *et al*., 2005; Turner *et al*., 2001; Hughes & Coffin, 2004), suggests a retroviral lineage which is still replicating, since these loci have not yet drifted to fixation within the human species. As well as implied recent activity, HERV-K(HML-2) elements have been linked to various cancers. It has been demonstrated that these elements encode accessory genes that are thought to play a pivotal role in inducing germ cell tumours. It was observed that certain HERV-K(HML-2) elements encode a cORF protein in the C-terminal open reading frame of the *env* gene. This was found to support the growth of tumours in mice, through impairing spermatogenesis by interacting with a promyelocytic leukemia zinc finger (PLZF) protein (Boese *et al*., 2000).

The HERV-K retroviral group is thought to have arisen approximately 35 million years ago, where it infected the germline of catarrhine primates subsequent to their divergence from platyrrhine primates (New World monkeys). The Catarrhini encompass cercopithecoids (Old World monkeys) and hominoids (humans and apes). While catarrhine primates contain all the known HERV-K elements, sequences exhibiting slight homology to these

have been observed in plattyrhine primates, as well (Turner *et al*., 2001). Despite this, it is generally accepted that the initial HERV-K germline endogenisation occurred after platyrrhine and catarrhine primates diverged from one another – approximately 35 million years ago. Furthermore, there is evidence to suggest that this retroviral group underwent several periods of expansion in hominoid lineages, suggesting a more recent integration time of 8-15 million years ago (Johnson & Coffin, 1999; Barbulescu *et al*., 1999). HERV-K elements exhibit similarity to exogenous betaretroviruses, and contain ten groups: HML-1 through 10. The most recently active is the -HML-2 group, which has been estimated to comprise some 60 proviruses, and approximately 2,500 solo-LTRs (Macfarlane & Simmonds, 2004; Mager & Medstrand, 2005). Of all the HERV-K viruses, only the -HML-2 group has exhibited insertional polymorphism, where an element may be present in one individual within a population, but not another. Moreover, the rate of insertion since the human-chimpanzee divergence – approximately 6 million years ago – appears to be constant, further implicating the present-day activity of this retroviral lineage. While an infectious HERV-K(HML-2) viral particle has not yet been identified, studies involving an engineered provirus, comprising a genome ancestral to this lineage, have resulted in exogenous viruses with weak infectivity (Dewannieux *et al*., 2006; Lee & Bieniasz, 2007).

It has been demonstrated that copy numbers of HERV-K(HML-2) elements have increased over time, predominantly *via* reinfection of the germline by exogenous particles (Belshaw *et al*., 2004). In support of this, Belshaw *et al*. (2004) calculated d_N/d_S ratios – a measure of non-synonomous (d_N) and synonomous (d_S) nucleotide substiutions – across a number of proviruses, strongly suggesting that this lineage was likely under a significant degree of purifying selection. By its nature, purifying selection implies that elements are replicating, and presenting themselves at varying levels of fitness. Unfit elements deleterious to the survival of any specific HERV population are supposedly placed under a purifying selection pressure, and eventually eliminated from the gene pool. The fact that d_N/d_S ratios significantly less than 1 have revealed that HERV-K(HML-2) elements are indeed under negative selection, confirms the existence of an infectious pool of unfixed retroviral

elements, that would supposedly actively reinfect the host genome (Belshaw *et al*., 2004).

It is possible that such unfixed elements are still reinfecting human germlines to the present day. Thus, opportunities to screen human genomes for retroviruses are often an important step towards understanding the activity and potential pathogenic role of HERV-K(HML-2) elements. This chapter presents the results of several HERV-K(HML-2) screens of archaic hominid genomes. The first available dataset screened was the low-coverage (1.9-fold) Denisovan genome published by Reich *et al*. in 2010. Mid-way through this screen, Agoni *et al*. published 14 novel HERV-K(HML-2) loci – later recovered in modern humans by Marchi *et al*. (2013) – retrieved from the same genome, in June of 2012. Shortly afterwards, a high-coverage (30-fold) Denisovan genome sequence was published (Meyer *et al*., 2012). Accordingly, it was decided that this high-coverage genome should be screened instead, since continuing the screen of the low-coverage genome would likely return the same loci Agoni *et al*. (2012) had already published. The methods and results of this low-coverage Denisovan screen are presented in appendix C of this thesis.

In addition to Meyer *et al*.'s (2012) high-coverage Denisovan genome, yet another high-coverage genome (52-fold) – this time belonging to an Altai Neanderthal – was made available in late 2013 (Prüfer *et al*., 2014). The authors decided only to post high-quality sequence reads that had been aligned to the human reference sequence. Since this would be an unlikely source for insertionally polymorphic HERV loci, a request was made to Prüfer *et al*., (2014) for access to the unpublished, unmapped data in late 2013. Access to this data was kindly granted prior to its subsequent publication in early 2014.

Subsequently, a series of novel HERV loci were discovered (Lee *et al*., 2014; appendix D). These loci represent the first published set of novel HERV-K(HML-2) loci recovered from the high-coverage Denisovan and Altai Neanderthal genomes published by Meyer *et al*. (2012) and Prüfer *et al*.

(2014), respectively. Moreover, this chapter also includes the results of an additional screen, which presents even more novel HERV-K(HML-2) loci (unpublished data). The results of these studies also provide insight into approaching genomic screens of this vast scale, and the associated methods used to compute them.

3.2 Materials and Methods

Extensive screens of three archaic hominid genomes were performed. The first of these used the low-coverage (1.9-fold) Denisova genome published by Reich *et al* in 2010. Since this screen was aborted upon the publication of 14 HERV loci by Agoni *et al*. (2012), together with the availability of a high-coverage genome for Denisova man, the methods and results of the low-coverage screen have been made available in appendix C of this thesis.

A second set of screens were conducted on high-coverage archaic hominid genomes. Of these, a stringent-matching screen (published in the paper in appendix D) was conducted on Denisova (30-fold coverage; Meyer *et al*., 2012) and Neanderthal genomes (52-fold coverage; Prüfer *et al*., 2014), while a later fuzzy-matching screen (unpublished data) was conducted on the Denisovan genome, with subsequent loci being screened for in the Neanderthal genome to ascertain which insertions are shared, and which are unique to Denisovans. Both of these high-coverage genome screens employed the same methodology to verify candidate HERV insertions, but differ in how the genomes were initially mined for HERV-K(HML-2) loci. These different mining procedures are described in sections 3.2.1 and 3.2.7-8 for the stringent- and fuzzy-matching screens, respectively.

3.2.1 Data Preparation and Formatting: Sequence files for the high-coverage Denisovan genome were obtained from the European Nucleotide Archive (ENA), under the study accession number ERP001519. For the Neanderthal genome, BAM files were obtained from the following URL:

http://cdna.eva.mpg.de/Neanderthal/altai/AltaiNeanderthal/bam/

The downloaded reads were available in SRA (sequence read archive) format, and required an initial conversion to the FASTQ format, *via* the 'fastq.dump' script provided by NCBI. FASTQ to FASTA conversions were subsequently performed using a script available as part of the FAST-X Toolkit.

Both mapped and unmapped Neanderthal sequence reads were formatted in the BAM alignment format – a binary format for storing sequence data – which is not directly suitable for analysis. They were therefore converted to FASTA files using a dedicated Sequence Archive/Map script as part of the SAMtools utility suite.

3.2.2 Retroviral Screen: Due to the computational power required to search datasets of such a vast size – totalling approximately 1.2 terabytes (TB) – a perl script was used to do the initial screening, rather than a more conventional method, such as BLAST. This script was set to stringently pattern-match the first and last 20 bp of the HERV-K113 LTR, and was run on both Denisovan and Neanderthal genomes (see 1. of Figure 3.1). It also analysed the amount of flanking sequence available within the read, and discarded reads with ≤7 bp available. For ease of data handling, the script also numbered the hits, and added their orientation into the sequence headers. This script is available in appendix E. Even on the highest specification domestic machines, these analyses took a considerable amount of time and processing power. Therefore, the scripts were run on a high-performance computing cluster, comprising 19 machines, each with 2 quad-core processors (between 2.4-3 GHz), a total of 368 GB of RAM memory and over 150 TB of storage (courtesy of Dr. Derek Huntley at the Imperial College London Bioinformatics Support Service).

3.2.3 Initial Verification and Read Counts: Once provirus-containing reads were extracted, putative novel HERV insertions were identified in the human genome. This involved downloading the hg38 assembly of the human genome (made available in late 2013), from UCSC's online archive. The proviral reads were then BLAST searched locally against the hg38 genome, and matches

were manually inspected for any mismatching sequence surrounding the viral breakpoint (see 2. of Figure 3.1). Since all reads contained the host-virus junction, any that matched human DNA along their whole length represented HERVs present in both genomes, and were subsequently discarded.

Reads where only flanking sequence or LTR matched were put aside as candidate novel proviruses for further inspection. However, sequencing artefacts such as template switches marred many of the reads. Due to the plentiful data available in such high-coverage genomes, many of the proviruses and flanking sequences were represented by multiple reads. A fragment of approximately 20 bp of flank and 20 bp of LTR was taken from candidate reads, and using a simple find tool in a text editor, corresponding unique reads were counted. Novel HERV candidates represented by multiple, unique reads were unlikely to be as a result of sequencing artefacts, and were more likely to be representing the actual archaic hominid sequence.

All reads representing an individual sequence were then sorted by the amount of flank represented within them (see 2. of Figure 3.1). The read containing the longest length of flanking sequence was subsequently used to search the hg38 assembly of the human genome using UCSC's BLAST-like Alignment Tool (BLAT), in order to confirm that the flank matched along its entire length. Reads that only matched once in archaic homonid genomes were marked as such, and only kept if subsequent verification revealed the other end of the provirus, together with matching TSDs.

Given the multiple reads representing many of the obtained 5′ host-virus junctions, even if the corresponding 3′ end could not be retrieved by back-filtering, a high unique read count would be considered sufficient validation of a novel virus, or potentially a solo-LTR structure.

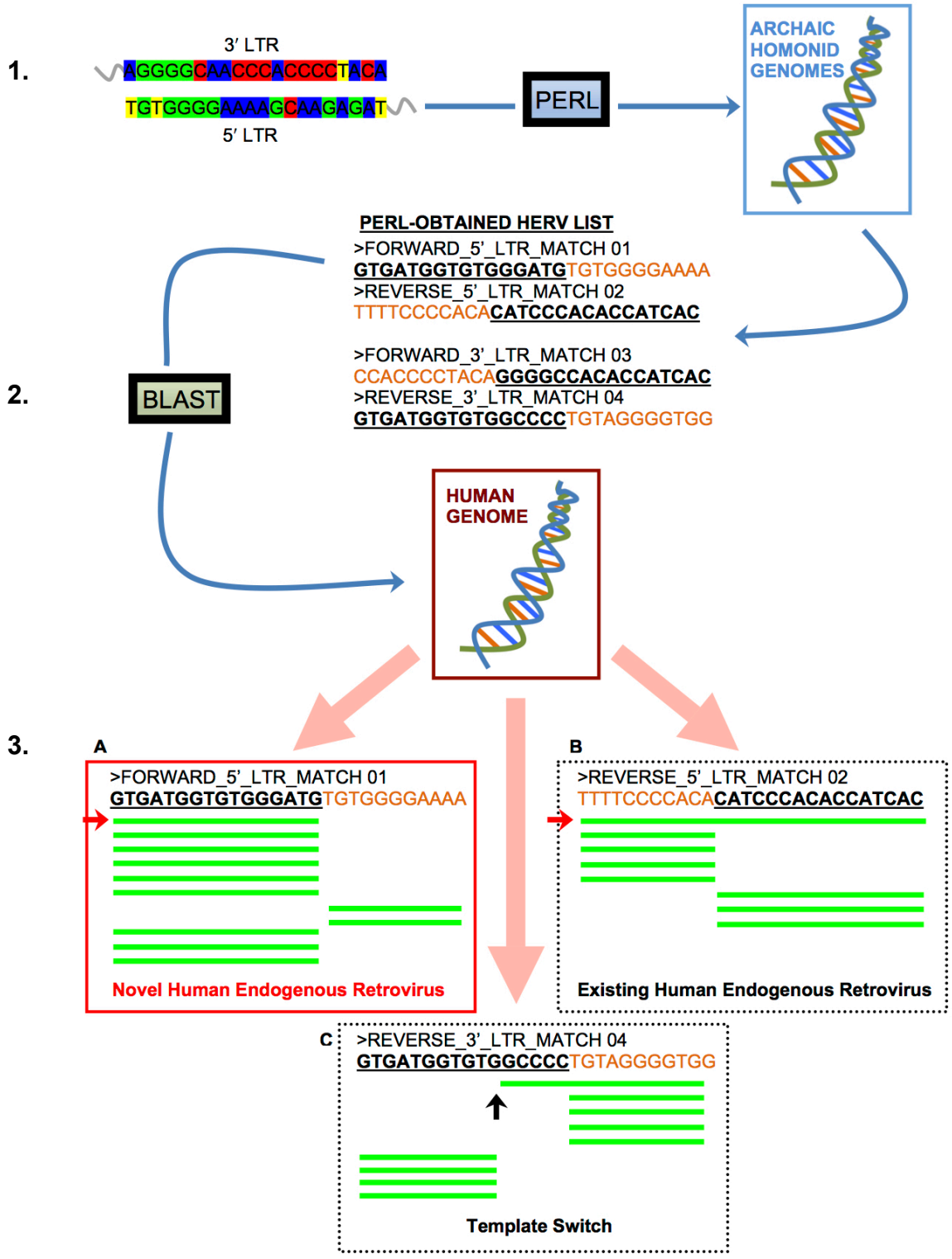


Figure 3.1: **Element retrieval and initial verification** of candidate proviruses. (1) Using either a 5′ or 3′ LTR as a starting probe, perl scripts process the archaic hominid genome data, stringently detecting reads containing HERVs and sufficient flanking sequence. (2) Resulting reads are BLAST searched against the human genome, with three general scenarios resulting. (3A) Novel matches are indicated by matching flank, and non-matching LTR (or *vice versa*). (3B) Full-length matches indicate viruses that are shared in both archaic and modern human genomes. (3C) A sequencing artefact known as a template switch results in an anomalous sequence read, comprising spliced sequence from two entirely separate sequences of chromosomal DNA.

3.2.4 Back-Filtering: Candidate 5′ ends of novel HERVs mined from the high-coverage archaic hominid genomes are verified by a process termed back-filtering (Aiewsakun & Brockman, 2011). This involves using the 5′ flanking sequence to extract a probable 3′ flank from the human reference genome by locating the empty preintegration site. This 3′ flank can then be used to recover the same sequence in the archaic hominid genomes. If, in the archaic hominid read, there is viral LTR adjacent to this 3′ flank, the virus has been successfully back-filtered, providing evidence for the other end of the element. In order to facilitate back-filtering, a library of 3′-LTR-containing reads was generated using perl scripts on the archaic hominid genomes.

Full-length candidate 5′ provirus reads were then run through a perl script, that trimmed back the LTR away from the reads, leaving behind only flanking sequence (see 1. of Figure 3.2; scripts available in the appendix F). These flanks were then BLAST searched against the hg38 assembly of the human genome, matching them to the corresponding empty preintegration sites (see 2. of Figure 3.2). 50 bp directly downstream of the matching sequence was then extracted, and was verified in the library of 3′-LTR-matching reads created during the first stages of back-filtering (see 3. of Figure 3.2). This was done, firstly, using a simple find tool in a text editor to match sequence slightly downstream of the breakpoint, allowing for potential mismatches in the TSD sequence directly adjacent to the LTRs. Matching 3′ ends were collated, and any 5′ ends for which corresponding 3′ ends could not be found, were further checked using BLAST as a final option.

These 5′ ends were then BLAST searched against the library of 3′ flank-containing reads, to confirm that they were not overlooked by the first, stringent back-filter using the text editor. In most cases, corresponding 3′ ends were not found in this library. They were therefore BLAST searched back against their entire respective source genomes, using a high-performance computer cluster (see 4. of Figure 3.2). 3′ ends could then be located, if present within the data. If they were not found in the source genomes, the data was noted as missing, and recorded as such in the results.

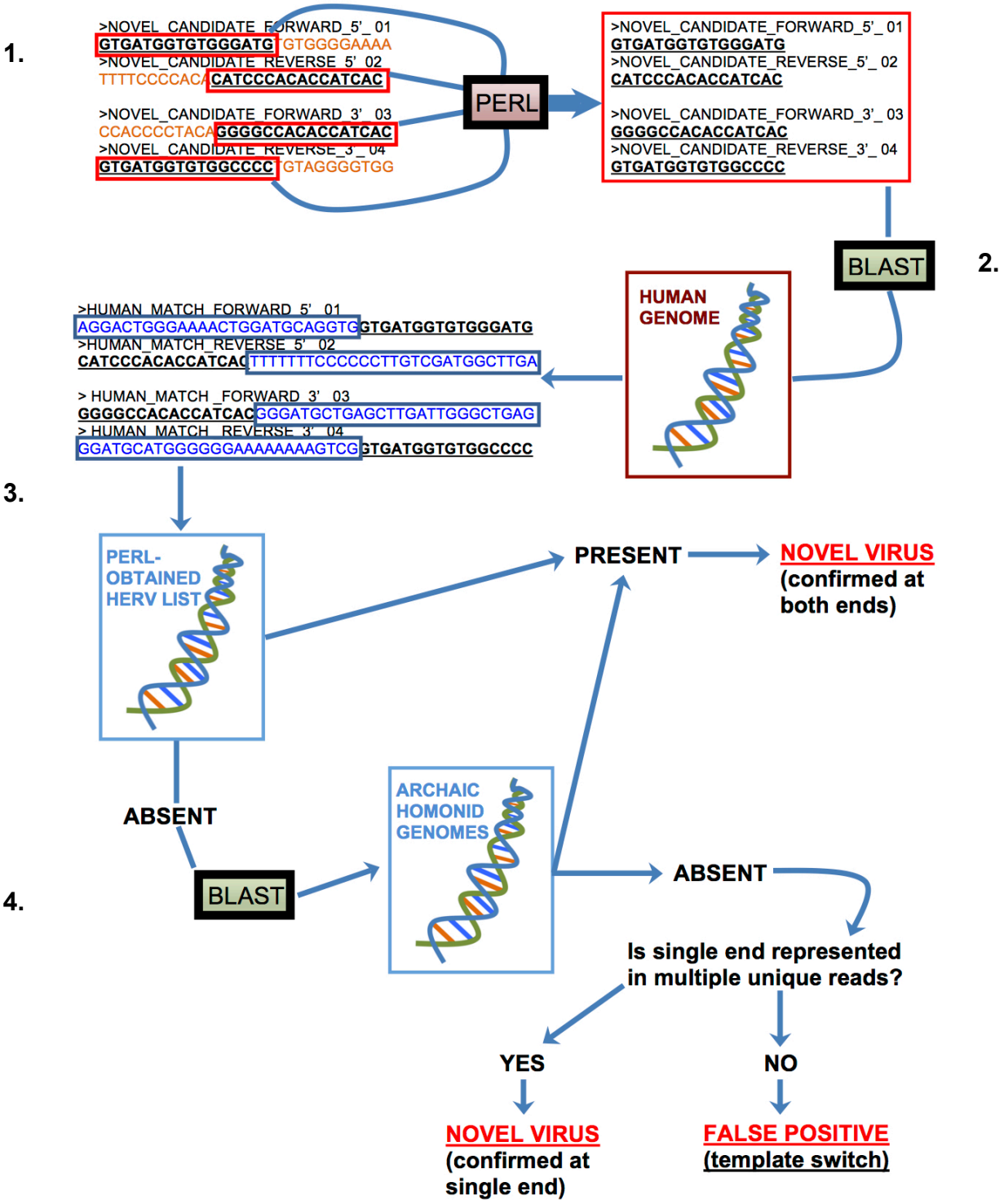


Figure 3.2: **Candidate novel provirus verification and back-filtering** schematic. (1) Candidate novel HERVs identified in the first stages of the screen, are processed by a perl script that removes LTRs, leaving behind a file of flanking sequence only. (2) These flanks are then BLAST searched against the human genome to reveal empty pre-integration sites. (3) 50 bp of human DNA that would flank the other side of the HERV insertion is then taken, and this is searched against a library of HERV-containing reads, derived from the opposite end of the virus. (4) Matches present in this library, are kept. Those absent from this library are subsequently BLAST searched against the archaic hominid genomes. If present, they are kept. If absent, they are subjected to read counts. If multiple, unique reads containing the same HERV locus are identified in the archaic hominid genomes, the provirus is considered novel, but represented only at a single end. If no additional reads are identified, the virus is discarded as a likely false positive resulting from template switching.

3.2.5 Second-Stage HERV-K(HML-2) Screening: Once the methods described in sections 3.2.2-4 were followed using the 5′ end of the HERV-K113 LTR as the starting probe, the same methods were repeated using the 3′ end of this LTR. Thus, any proviruses that were missed the first time round due to searches being too stringent, for instance, were more likely to be located from the opposite end of the virus. Parameters were kept the same throughout, and scripts were altered to handle opposing ends of the virus, where required.

3.2.6 Matching TSDs: Upon integration, endogenous retroviruses leave a signature TSD (also known as a direct repeat) at either end of the integration. These are usually 4-7 bp in length, and are usually visibly identical – although a slight loss of sequence homology can be observed in some cases. For novel proviruses that were successfully back-filtered and identified at both 5′ and 3′ ends, a final verification of these elements was the observation of near-identical – or completely identical – TSDs flanking both the 5′ and 3′ LTRs.

3.2.7 Fuzzy-Matching: The archaic hominid screens comprised a third component. This is a more recent development to the workflow, and was therefore not included in the published paper (appendix D). A perl script was written that made use of a fuzzy-matching perl module obtained from the CPAN archive (Hietaniemi, 2001-2005). This module – entitled String::Approx – allows for less stringent detection of the HERV-K113 LTR in the source data. The added advantage is that HERV elements within which substitutions, additions or deletions occur, would be picked up by using this perl script, where they may have otherwise been overlooked.

The script was set to detect a single base pair mismatch in the first (and last) 20 bp of the HERV-K113 LTR, and was run on the high-coverage Denisova genome only. To do this, the ‘amatch’ command of the String::Approx fuzzy-matching perl module was called within the perl script. The program then outputs the entire sequence read within which LTR sequence was detected. These sequence reads were then verified utilising the same methods outlined

in sections 3.2.3-6 of this chapter. This fuzzy-matching perl script can be found in appendix G.

3.2.8 Fuzzy-Match Filtering: While using the String::Approx module picked up HERVs with mismatches to the LTR probes used, it also picked up perfect matches, as well. Since these perfect matches had already been recovered in the first stage of the screening process by the stringent-matching perl scripts in appendix E, 100% matches were filtered away from this data, leaving only imperfect matches. This was done using a regex with a !~ binding operator, returning anything that was not 100% equal to the LTR probes used, and filtering them from the program output. This script is available in appendix H.

Since fuzzy-matching returns reads irrespective of the amount of flanking sequence, another script was written to filter the number of fuzzy matches down to reads containing 25 bp of flanking sequence, or greater. This script was designed for use together with the output of earlier scripts used in this workflow. Based on the sequence and file names assigned by the scripts used earlier in the workflow, the fuzzy match filtering script has two modes of operation. One is manual – where the orientation (forward or reverse) and end of the virus (5′ or 3′) is entered manually, into the scalars at the top of the script. The other mode is automatic, providing the inputted files were created using the previous fuzzy-matching and flank-filtering scripts. The filtering script detects both the orientation and the end of the virus represented by the sequence reads using the input file name, and processes them accordingly. The fuzzy match filter script is available in appendix I.

Due to the vast numbers of hits produced using the fuzzy-matching workflow, this was only run on the Denisovan genome. Subsequently, any novel proviruses were BLAST searched against the Neanderthal genome, to ascertain whether they are shared with this sister species, or whether they are unique to Denisova man. BLASTN was used for this, set to a word size of 11.

3.2.9 Screen Comparisons: Verified novel HERVs were finally compared against those recovered by published screens of other, high-coverage modern

human genomes. Loci were compared against those of HERV-K elements recovered by Lee *et al.*'s (2012) screen of 43 high-coverage genomes, and against those recovered in Marchi *et al*.'s (2014) screen of 358 high-coverage genomes. Comparisons were conducted by comparing mapped loci obtained from these studies, to those obtained from the archaic hominid screens in this study. Any matches that were found to be identical to these published reports, were discarded, leaving behind only novel HERV loci.

3.2.10 Locus Mapping: Once the final set of novel HERVs was obtained, their flanking sequences were mapped onto the hg38 assembly of the human genome using UCSC's BLAT. The loci of the empty preintegration sites was recorded from the top matches in each case. For proviruses providing strong matches across multiple loci, their flanks were subsequently run through RepeatMasker, and the class of repeat into which they integrated was recorded.

3.3 Results

As aforementioned, the first retroviral screen was conducted on the low-coverage Denisovan genome made available by Reich *et al*. (2010), but this was aborted upon both the subsequent release of Meyer *et al*.'s (2012) high-coverage Denisovan genome, and Agoni *et al*.'s (2012) publication, which presented a retroviral screen conducted on the former genome. While incomplete, the results and methods are available in appendix C.

3.3.1 Stringent-Matching Retroviral Screen of High-Coverage Denisovan and Neanderthal Genomes: A total of nine HERV-K(HML-2) elements were identified, that were found to be absent in the modern human reference genome (assembly hg38). Their loci and distribution is given in Figure 3.3A and full sequence reads are given in Table 3.1. Seven of these loci were successfully back-filtered, being represented at both ends of the insertion. Their TSDs were found to match, and are indicated on Figure 3.3B in an underlined, emboldened typeface. The remaining two loci were represented by multiple, unique sequence reads (Figure 3.3B).

A

HERV Insertion	Neanderthal		Denisova		hg38		Lee *et al.* / Marchi *et al.*		Breakpoint Locus (Mapped to hg38)
	5'	3'	5'	3'	5'	3'	5'	3'	
HERV-K-DeNe1	●	●	●	●	○	○	○	○	chr19:2890621 (Alu)
HERV-K-DeNe2	●	●	●	●	○	○	○	○	chr18:30138227
HERV-K-DeNe3	N.D.	●	N.D.	●	○	○	○	○	chrUn_K1270749v1:63347
HERV-K-De13	○	N.D.	●	N.D.	○	○	○	○	chr4:117182160
HERV-K-Ne4	●	●	○	○	○	○	○	○	chr8:74432832
HERV-K-Ne5	●	●	○	○	○	○	○	○	chr19:57485571
HERV-K-Ne6**	●	●	○	○	○	○	●	●	chr12:43919858 (L1)
HERV-K-Ne7**	●	●	○	○	○	○	●	●	chr5:65092618
HERV-K-De14*	○	○	●	●	○	○	●	●	chr20:12421693

B

HERV Insertion	+/-	5' Flanking Sequence	Endogenous Retrovirus 5' 3'	3' Flanking Sequence
HERV-K-DeNe1	+	~GATCCACCTGTCTCCGC**CTCCC**	TGTGGGGA - - → CCCCTACA	**CTCCC**AAAGTGCTGGGATGACA~
HERV-K-DeNe2	+	~GTATAATTCTACTAACC**CAAGTA**	TGTGGGGA - - → CCCCTACA	**CAAGTA**AAACCAAAAATTATTA~
HERV-K-DeNe3	+		- - → CCCCTACA	AAGAAGTGTGAAAAAAGAAATA~
HERV-K-De13	+	~TAACAGGCTCTGAAATTGTAGC	TGTGGGGA - - →	
HERV-K-Ne4	+	~ATCACAGTTTTTAAATTA**ACAAT**	TGTGGGGA - - → CCCCTACA	**ACAAT**AATATTGATACCAACAT~
HERV-K-Ne5	-	~CTTTAGGAATATTTACAT**GTCTA**	TGTGGGGA - - → CCCCTACA	**GTCTA**TAAGGAGTGTGAACATC~
HERV-K-Ne6*	+	~TGAGGTGATATCTCATT**GTGGT**	TGTGGGGA - - → CCCCTACA	**GTGGT**TTTGATTTTCATGTTCTT~
HERV-K-Ne7*	+	~AAAAAACGTAATAGTA**GACGTG**	TGTGGGGA - - → CCCCTACA	**GACGTG**ACATATGCATCACAGC~
HERV-K-De14*	-	~ATGGCTACTGGGAATTA**AGTGG**	TGTGGGGA - - → CCCCTACA	**AGTGG**GAAGGAGGATAAAGCT~

Figure 3.3: **Overview of novel proviral loci** obtained from the published, stringent-matching screen. A). Distribution of the archaic hominid insertions within Neanderthals and Denisovans, compared against hg38 and the HERV-K elements recovered in previous studies (Lee *et al*., 2012; Marchi *et al*., 2013; Marchi *et al*., 2014). Filled circles indicate that an element is present, while open circles denote absence. N.D. indicates that there was no sequence data available. The loci of corresponding empty preintegration sites, mapped to hg38, are also given. If this occurred within a repeat, the class of repeat is listed in parentheses. Single asterisks adjacent to the last three viruses represent those elements that were also recovered in some modern humans by Lee *et al*. (2012), while double asterisks represent those viruses also recovered by Marchi *et al.* (2013-14). The distribution of these three elements is most likely explained by ancestral polymorphism. B). HERV-K(HML2) flanking sequences for the nine endogenized retroviruses identified here. The 5′ and 3′ flanking regions are shown, together with the proximal and distal ends of the HERV LTR. Nomenclature follows the convention and numbering set by Agoni et al. (2012). Reads where flanking sequence was extensive were truncated and are provided, in full, in Table 3.1. + or − denotes the native orientation of the read against hg38. For HERVs for which both 5′ and 3′ flanks were obtained, the matching TSD sequence is underlined and in boldface.

Three of the identified proviruses appeared orthologous to both Denisovans and Neanderthals, whilst four were identified solely in Neanderthals, and two solely in Denisovans. Nomenclature of these elements followed the system laid out by Agoni *et al*., (2012); the first research group to publish the results

of genomewide retroviral screen conducted on low-coverage Denisovan (Reich *et al*., 2010) and Neanderthal genomes (Green *et al*., 2010). Thus, Denisovan-specific retroviruses were named to follow on from HERV-K-De13, and Neanderthal viruses followed on from HERV-K-Ne3. Agoni *et al*. (2012) did not, however, identify any of their recovered loci in both Denisovan and Neanderthal genomes. Since the screen presented here identified three novel HERV loci common to both Denisovans and Neanderthals, they have been designated HERV-K-DeNe1-3.

The loci of the corresponding human preintegration sites – mapped to hg38 – are shown in Figure 3.3A. Two proviruses – HERV-K-DeNe1 and HERV-K-Ne6 – were observed to have integrated within a SINE (Alu) and LINE (L1) element, respectively. For these, the top-matching locus from UCSC's BLAT was given, alongside the repeat class.

3.3.2 Stringent-Matching Loci: Comparisons with Other Insertionally Polymorphic HERVs: The flanking sequence and TSDs of the identified proviruses were then compared against those of HERV-K(HML-2) loci recovered in screens reported by two other research groups. Lee *et al*. (2012) screened a set of 43 high-coverage genomes, while Marchi *et al*. (2014) screened 358 high-coverage modern human genomes derived from the cancer genome atlas (TGCA). Lee *et al*. (2012) mapped their loci to the hg18 assembly of the human genome, while Marchi *et al*. (2014) mapped to hg19. Three archaic hominid HERV-(HML-2) loci were also recorded in these other studies. These were: HERV-K-Ne6, HERV-K-Ne7 and HERV-K-De14. All three were reported by Lee *et al*. (2012), whereas Marchi *et al*. (2014) only reported –Ne6 and –Ne7 (see Figure 3.3).

The remaining six loci appear to be absent from all 401 HERV-K-screened modern human genomes compared against in this study. As discussed by Marchi *et al*. (2013), it remains possible that these loci could be insertionally polymorphic, existing in some modern human individuals, albeit at very low allele frequencies.

Table 3.1: **Example HERV-K(HML-2)-containing reads** for the novel proviruses in the stringent-matching screen. The names of the reads, as given in the genome data for Denisovan and Neanderthal man, are given below, in correspondence with the nine published elements (Lee *et al*., 2014). Many of these elements were represented by multiple reads (often in excess of 30 for some of the retroviruses), however the list below comprises individual examples of reads containing flank and retroviral LTR. Proviruses found in Denisova and Neanderthal, are shaded pink; those unique to Neanderthal are shaded beige, and those unique to Denisova are shaded blue. Shaded regions include the name of the proviral insertion – following on from the nomenclature set by Agoni *et al*. (2012) – and two columns indicate the host genome ('D' for Denisova; 'N' for Neanderthal) and the end of the virus represented ('5' for 5′; '3' for 3′).

		DeNe1
D	5	>ERR141697_1_29695091 AGAGACGGGTTTCACCGTGTTGACCAGGCTGGTCTTGAACTCCTGACCTCATGTGATCCACCTGTCTCC GCCTCCCTGTGGGGAAAAGCAAGAGA
D	3	>ERR141700_5_76724767 CTTTTTGCTGAATTAAACAGCAAGTCGGCCAGGTATGGTGGCTCACGCCTGTCATCCCAGCACTTTGGG AGTGTAGGGGTGGGTTGCC
N	5	M_SN7001204_0130_AC0M6HACXX_PEdi_SS_L9302_L9303_1:6:2115:8893_39171 AATTTTTGCATTTTTAGTAGAGACGGGTTTCACCGTGTTGACCAGGCTGGTCTTGAACTCCTGACCTCAT GTGATCCACCTGTCTCCGCCTCCCTGTGGGGAAAAGCAAGAGAGAT
N	3	>M_NIOBE_0139_A_D0B5GACXX:8:1104:15932:187701 AATTAAACAGCAAGTCGGCCAGGTATGGTGGCTCACGCCTGTCATCCCAGCACTTTGGGAGTGTAGGG GTGGGTTGCCC

		DeNe2
D	5	>ERR141700_10_8936492 AGTAACAATCTGATCTCTCTTGCTTTTCCCCACATACTTGGGTTAGTAGAATTATACAAGGAAATTCGGTT GTTTCTGTGGTTTACAATAACTTAACATAATAACTATAATTGTGATTGGTAGCATATATTAGACATTAGAAT TTTAGAAAT
D	3	>ERR141700_3_62869174 TATTTTAAACAATTCTGGTATTAGCTGGTTTAACAGTATTTTCTTGATATTTAATAATTTTTGGTTTTACTTG TGAAGGGGTGGGTTGCC
N	5	>M_SN7001204_0130_AC0M6HACXX_PEdi_SS_L9302_L9303_1:5:1105:5693:78876 ACAATCTGATCTCTCTTGCTTTTCCCCACATACTTGGGTTAGTAGAATTATACAAGGAAATTCGGTTGTTT CTGTGGTTTACAATAACTTAACATAATAACTATAATTG
N	3	>M_SN928_0068_BB022WACXX:8:1302:13716:165743 No definition line GCAACCCACCCCTTCACAAGTAAAACCAAAAATTATTAAATATCAAGAAAATACTGTTAAACCAGCTAATA

		DeNe3
D	3	>ERR141692_39590621 TGGAGGGGCAACCCACCCCTACAAAGAAGTGTGAAAAAAGAAATATCTGTTGTTTTTTAGTCACCCGGT TTATGTTATTTTGTTATAAGAGTCCAAAATACACCAAGATATTCCACTTAATATG
N	3	>M_SN928_0073_BD0J78ACXX:6:2303:16328:73134 No definition line ACCCACCCCTACAAAGAAGTGTGAAAAAAGAAATATCTGTTGTTTTTTAGTCAACCGGTTTATGTTATTTG GTTATAAGAGTCCAAATACACCAAGATATTCCACTTAATATGTAGGGGAAGGCAAAAAATACT

		Ne4
N	5	>M_SN7001204_0130_AC0M6HACXX_PEdi_SS_L9302_L9303_1:5:2203:6925:88074 TGAGGCTTTATACATAATAGTAATAAGATTTTATTCCAAGTACAAAAGAACCAATGATCTCATAAAAAAATT TAATATCACAGTTTTTAAATTAACAATTGTGGGGAAAA
N	3	>M_SN7001204_0131_BC0M3YACXX_PEdi_SS_L9302_L9303_2:2:2205:8390:20520 TTACGAGAAACACCCACAGGTGTGTAGGGGCAACCCACCCCTACAACAATAATATTGATACCAACATAT CAGTGACAAGTGCTATAGAAGGAGCAAAACAAAGGGCTAGAGAAAGTAGGATAGTTT

		Ne5
N	5	>M_SN928_0068_BB022WACXX:5:2202:19259:167871 TCCTTTTCCCCACATAGACATGTAAATATTCCTAAAGGTTGGAAATAAAAATAGATTCCAGAAGTTAGAGA GAAAACATATAGACCCGGCATATGTTGCTTTTAAAAAATAAAATAGCTTCATGGGATTTTT
N	3	>M_SN928_0068_BB022WACXX:8:2105:7982:115764 AGAGGAATGTGCTTGCTAGCCCTGAAGGAGATATGGCCAAAGGTAGAGTCTTGGACTCAAACTTTGATG TTCACACTCCTTATAGACTGTAGGGGTGGGTTGCCCCTAAACATCTGTGGGTGTTTCTCGTAAGGTGGG ACGAGAGATTTGG

		Ne6
N	5	>M_SN928_0073_BD0J78ACXX:3:1208:10790:12853 CGCTTTTCCCCACAACCACAATGAGATATCACCTCACACCTGTCAGAATGGCAATTAACAACAACAACAA AAAGTGTTGGAAAGGATGTGGAGAAACTGGAATCCTTACACACTACTGATGAGAATAAAAAATGGTGCT G
N	3	>M_SN928_0073_BD0J78ACXX:8:1106:13073:37043 TGATCAGTCAATTCTTGAGAATAAAGTCACCAAAAGAAAGAACATGAAAATCAAAACCACTGTAGGGGTG GGTTGCCCCTACACACCTGTGGGTGTTTCTT

		Ne7
N	5	>M_SN7001204_0130_AC0M6HACXX_PEdi_SS_L9302_L9303_1:5:2216:11632:4673 GAAAACTATGAGTAGTAAACAGGCTGGGATGAAAAGAAAGCTGCTTGAGATTTTAAAACATTTTAAAATA AAAAACGTAATAGTAGACGTGTGTGGGGAAAAAGCAAGA
N	3	>M_SN928_0068_BB022WACXX:1:1308:8132:95135 TGTGTAGGGGCAACCCACCCCTACAGACGTGACATATGCATCACAGCCCCAAAAGTGTACAATTGACA TTCTTGATAAC

		De13
D	3	>ERR141695_191710687 CACGAAGAAGCCGAATCCCTGAATAGACCAATAACAGGCTCTGAAATTGTAGCTGTGGGGAAAAGCAA GAGAGATCAGATTGTTACTGT

		De14
D	5	>ERR141695_199156238 ATTCAGTCTCCTAAAGCATGATGTTTTTGAATTAAATGGCTACTGGGAATTAAGTGGTGTGGGGAAAAG GAAGAGAGATCAGATTGTTACT
D	3	>ERR145624_1_80736435 CAACCCACCCCTACAAGTGGGAAGGAGGATAAAGCTGAAATATTTTGCCAGTGCTGCATGAAGGCTAA AGAAGTAA

3.3.3 Fuzzy-Matching Retroviral Screen of High-Coverage Denisovan Genome: Using the fuzzy-matching scripts, a further five novel HERV-K(HML-2) loci were identified (unpublished; Figure 3.4; Table 3.2). All of these were present in Denisova, four were present in Neanderthal, and all were absent from the hg38 human reference sequence. Two of these HERVs were represented at both 5′ and 3′ ends of the element, while the remaining three were single ends, represented by multiple unique sequence reads. Two of these were 5′ ends, and one, a 3′ end. Nomenclature follows the same system as before, with shared HERVs being named onwards from –DeNe4 and the single Denisovan HERV following onwards from –De14.

A

HERV Insertion	Neanderthal		Denisova		hg38		Lee *et al.* / Marchi *et al.*		Breakpoint Locus (Mapped to hg38)
	5′	3′	5′	3′	5′	3′	5′	3′	
HERV-K-DeNe4	●	●	●	●	○	○	○	○	chr11:55316157
HERV-K-DeNe5	●	●	●	●	○	○	○	○	chrX:25773640 (L1)
HERV-K-De15	N.D.	N.D.	●	N.D.	○	○	○	○	chr9:29333033
HERV-K-DeNe6	●	N.D.	●	N.D.	○	○	○	○	chr11:48625792
HERV-K-DeNe7	N.D.	●	N.D.	●	○	○	○	○	chr22:11242076

B

HERV Insertion	+/-	5′ Flanking Sequence	Endogenous Retrovirus 5′	Endogenous Retrovirus 3′	3′ Flanking Sequence
HERV-K-DeNe4	+	~GTAAATGGAATCATCTT**GTCTAC**	TGTGGGGA - - →	CCCCTACA	**GTCTAC**GTTTAAGGTTTAGTCAT~
HERV-K-DeNe5	+	~TATAATCTATCATATATA**CAGTAC**	TGTGGGGA - - →	CCCCTACA	**CAGAAC**CAATGACAAAAACCAC~
HERV-K-De15	+	~ACTCAGAATGGCACACAACTGTC	TGTGGGGA - - →		
HERV-K-DeNe6	+	~GTAAATGGAATCATCTTGTCTAC	TGTGGGGA - - →		
HERV-K-DeNe7	-		- - →	CCCCTACA	TTTCCAAACTGCTCAATGAAAAG~

Figure 3.4: **Overview of novel proviral loci** obtained from the fuzzy-matching screen. A). This table shows the distribution of the archaic hominid HERVs within Neanderthals and Denisovans. Preintegration site loci are given, and were determined by mapping the flanks to the hg38 assembly of the human genome. If insertions are flanked by part of a SGE, these are indicated besides the first matching locus, in parentheses. Elements were also compared against HERVs obtained from published screens of 401 high-coverage, modern human genomes, and this is shown in a dedicated column (Lee *et al*., 2012; Marchi *et al*., 2013; Marchi *et al*., 2014). Filled circles indicate that an element is present, while open circles denote absence. N.D. indicates that there was no sequence data available. B). Sequence flanking the five identified HERVs. The 5′ and 3′ flanking regions are shown, together with the proximal and distal ends of the HERV-K113 LTR. Nomenclature follows the convention and numbering set by Agoni *et al.* (2012), and follows on from the published elements in Lee *et al*. (2014). Reads where flanking sequence was extensive were truncated and are provided, in full, in Table 3.2. + or − denotes the native orientation of the Denisovan read realitve to hg38. For successfully back-filtered HERVs, the matching TSD sequence is underlined and in boldface.

3.3.4 Fuzzy-Matching Loci: Comparisons with Other Insertionally Polymorphic HERVs

: Of the five identified loci, all five appeared to be novel HERVs. None of these were previously identified by screens conducted by either Lee *et al*. (2012) or Marchi *et al*. (2014).

Table 3.2: **Example HERV-K(HML-2)-containing reads** for the novel proviruses in the fuzzy-matching screen. The names of the reads, as given in the genome data for Denisovan and Neanderthal man, are given below. While many of these elements were represented by multiple reads, the list below comprises individual examples of reads containing both flank and retroviral LTR. Provirus loci found in Denisova, and subsequently shared in Neanderthal, are shaded pink and those unique to Denisova are shaded blue. Shaded regions include the name of the HERV insertion – following on from the nomenclature set by Agoni *et al*. (2012) and from the viruses published in Lee *et al*. (2014) – and two columns indicate the host genome ('D' for Denisova; 'N' for Neanderthal) and the end of the virus represented ('5' for 5′; '3' for 3′).

		HERV-K-DeNe4
D	5	>ERR141692_198431456 ACAGCTCAAACTCTGCCTGCATATGCGCTCTTCCACTGCCAGATCTATGTAAATGGAATCATCTTGTCTA CTGTGGGGAAAAGGAAGAGAGAACAGATTGTTA
D	3	>ERR141693_63620696 TGTGGAGGGGCAACCCACCCCTTCAGTCTACGTTTAAGGTTTAGTCATTCTGTAGGTTTCTTCAGCAAT GAATAATCAGAAAGTCGGGTTTGTATCT
N	5	>M_SN928_0073_BD0J78ACXX:8:2103:18043:26107 AGACAGCTCAAACTCTGCCTGCATATGCGCTCTTCCACTGCCAGATCTATGTAAATGGAATCATCTTGTC TACTGTGGGGAAAAGGAAGAGAGATCAGATTGTTAATGT
N	3	>M_SN928_0068_BB022WACXX:8:1105:9023:110174 TTCTTTTCCAAGTCTCTCGTCCCACCTAATGAGAAACACCCACAGGTGTGGAGGGGCAACCCACCCCTT CAGTCTACGTTTAAGGTTTAGTCATTCTGTAGGTTTCTTCAGCAATGAATAAT

		HERV-K-DeNe5
D	5	>ERR145627_2_63105816 TTTAACATATAAAAATCAATAAATATAATCTATCATATATACAGTACTGTGGGGAAAAGCAATAGAGATCA GGTTG
D	3	>ERR141697_3_37427729 GAGAAACACCCACAGGTGTGGAGGGGCAACCCACCCCTACACAGAACCAATGACAAAAACCACATGAT TATCTCAATAGATGCAGAAAAGGCCTTTGACAAAATTCAATAGCCATTAATGCTAAAAACTCTCA
N	5	>AltaiNea.hg19_1000g.5.dq_24214635 AACAATCTGATCTCTCTTGCTTTTCCCCACAGTTCTGTTTATATGATGGATTATGTTTATTGATTTGTGTAT GTTAAACAAGACTTGCATCCCAGGGATGAAG
N	3	>AltaiNea.hg19_1000g.5.dq_24203179 TTTGTCGAAGGCCTTTTCTGCATCTATTGAGATAATCATGTGGTTTTTGTCATTGGTTCTGTGTAGGGGT GGGTTGCCCCTCCACACCTGTGGGTGTTTCTCGTAAGGTGGAATGAGAGACTTA

		HERV-K-DeNe6
D	5	>ERR145622_1_74957044 TGCATATGCGCTCTTCCACTGCCAGATCTATGTAAATGGAATCATCTTGTCTACTGTGGGGAAAAGTAA GAGAGAT
N	5	>M_NIOBE_0139_A_D0B5GACXX:7:1301:13880:6939 AGACACATTAACAATCTGATCTCTCTTCCTTTTCCCCACAGTAGACAAGATGATTCCATTTACATAGATC TGGCAGTGGAAGAGCGCATATGCAGG

		HERV-K-DeNe7
D	3	>ERR141693_174655934 CACACAGTTGAATCTTTCTTTTCATTGAGCAGTTTGGAAATGAAGGGGTGGGTTGCCCCTCCACACTTGT GCGTGTTTCTCATTAAGTGGAATGTGAGACTTGGAAAAGAAATAAACACAGAGACAAAGT
N	3	>M_SN928_0068_BB022WACXX:4:1308:13173:67456 TGTGATGTATGCTTTCATCTCACACAGTTGAATCTTTCTTTTCATTGAGCGGTTTGGAAATGAAGGGGTG GGTTGCCCCTCCACACTTGTGCGTGTTTCTCATTAAGTGGAATGTGAGACTTGGAAAAGAAATAAACACA GAGACAAAGTGTAGAGAAAGAAATTAGGTTG

		HERV-K-De15
D	5	>ERR141699_119115074 TAAAGGAGAACAGCATGAGATTTCATCATAGTACTCAGAATGGCACACAACTGTCTGTGGGGGAAAGCA AGAGAGATCAGATTGTTACTGTGTCTGTGTAGAAAGAAGTAGA

3.3.5 Template Switching: One recurrent problem that was encountered frequently, was template switching. This is a sequencing artefact caused by DNA polymerase switching templates during sequencing. The resulting read comprises sequence made from two different DNA sequence templates. This is one of the many reasons that it was necessary to verify matches manually. An example of a template switch is shown in Figure 3.5.



Figure 3.5: An example of a template switch sequencing artefact occurring in a HERV flanking sequence. The emboldened, black sequence at the top is a high-coverage Denisovan genome read, containing viral LTR (underlined in blue). The green sequences below represent human genome sequences from assembly hg38, aligned to the Denisovan read using BLAST. Chromosomes are given on the side. This is an example of a candidate novel provirus, where there is no contiguous match spanning the length of the LTR and flank. The flank match is also incomplete, as interrupted by a template switch, with a completely different DNA sequence being appended, underlined in red.

3.4 Discussion

Observed declines in HERV activity over time, has led to the once common opinion that all HERV lineages are defective, and have been for millions of years (Lander *et al*., 2001). However, several studies across the last 10-15 years have revealed increasing evidence pointing towards the potential activity of the HERV-K(HML-2) retroviral group. The 11 novel HERV-K(HML-2) loci reported in this chapter are the result of the one of the most extensive retroviral screens conducted on archaic hominid genomes, to date. Further to these, an additional three loci were also shared with loci identified by Lee *et al.* (2012) and Marchi *et al*. (2014) in their analyses of human cancer genomes.

HERV-K113 is a well-described member of the HERV-K(HML-2) retroviral group. It is unique, in that its genome comprises full-length ORFs that code retroviral gene products. These ORFs lack non-synonomous substitutions in genomic regions that are highly conserved among the *Retroviridae*. It is for these reasons, that HERV-K113 was used as a probe in the screens reported here. Furthermore, early reports of HERV-K(HML-2) insertions in the human genome investigated both HERV-K113 and HERV-K115 insertions in a group of human individuals of different ethnic origins. It was found that HERV-K113 elements occurred at approximately double the frequency of HERV-K115 (Turner, 2001). Additionally, as the HERV-K113 lineage is implicated in present-day activity, these provide sufficient incentive to use it as a probe when profiling genomes for HERV-K(HML-2) insertions (Turner *et al*., 2001; Belshaw *et al*., 2004).

Another property of HERV-K113 that makes it ideal for screening, is the high sequence conservation between its LTRs – both of which are identical upon integration. This is reflected by the fact that the high-coverage genomic screens presented here identified 14 loci absent from the human reference sequence, using matching scripts that – at the most – allowed for a single base mismatch in the 20 bp LTR probe used. Nine of these 14 loci were retrieved based on 100% matches to the published HERV-K113 sequence.

The fact that there is next to no degeneracy in the LTRs, further supports the recent integration dates estimated for these viruses, and that they could still be reinfecting today.

Since HERVs have previously been assumed to be defective for many millions of years, this would imply that all HERVs observed in humans today have likely reached fixation in the germline (Gifford & Tristem, 2003). In this event, a profile of HERV insertions would be expected to be identical from one individual to the next. As demonstrated by the novel HERV loci reported in this chapter – and by those reported in several other studies – it is evident that this is not the case. These results show that HERV-K(HML-2) elements exhibit insertional polymorphism, where a HERV locus can exist in one individual, but not another. This is evidence that the virus has probably been active in the last 800,000 years, since this is the time it takes for a gene to reach fixation in humans (Kimura, 1984).

The results reported in section 3.3, present 11 HERV-K(HML-2) loci that are suggested to be unique to archaic hominids. This would imply that they reinfected the germline, less than 800,000 years ago, since they have not drifted to fixation in extant humans (Kimura, 1984). Since anatomically modern humans are estimated to have diverged from Denisovan and Neanderthal lineages approximately 553-589,000 years ago (Prüfer *et al*., 2014), it seems plausible that replication-competent HERV-K(HML-2) insertions reinfecting the germlines of these hominid subspecies would be unique. Yet this is not conclusive. It remains possible that these HERVs have persisted in the genomes of some modern humans as a result of ancestral polymorphism, albeit at low allele frequencies – such that the human genomes screened so far do not represent them.

Aside from these 11 novel loci, three archaic hominid loci – HERV-K-Ne6, HERV-K-Ne7 and HERV-K-De14 – were recovered by Lee *et al*. 2012 in their screen of 43 high-coverage modern human genomes. These datasets were derived from 5 cancer types. Despite the often-observed upregulation of HERV-K(HML-2) particle expression in certain cancers (Boller *et al*., 2008),

the three loci identified here are unlikely to be directly linked to cancers. Given that both Neanderthal and Denisovan phalanx samples are at least 50,000 years old, it is unlikely – albeit possible – that such a deleterious HERV allele would have persisted into the modern human lineage for this length of time (Ohta & Gillespie, 1995). Additionally, this comparison of identified loci against existing, high-coverage modern human genomes is paramount to suggesting that they are novel, and specific to archaic hominids. In 2012, Agoni *et al*. published 14 HERV loci that they concluded were novel, based only on their absence from the hg19 assembly of the human genome. However, Marchi *et al*. (2013) subsequently identified nearly all of these published loci in the genomes of modern humans, proving that they were, in fact, insertionally polymorphic. This reiterates the importance of setting putative novel HERVs into context with the results of other retroviral screens, since the possibility of these existing within the human population at low allele frequencies can never be conclusively disproved. The identification of insertionally polymorphic, unfixed HERVs using archaic hominid lineages does, however, augment the ongoing activity of the HERV-K(HML-2) lineage (Turner *et al*., 2001).

Seven of the loci reported in this chapter were also found to be shared in both Denisovans and Neanderthals, whilst being absent from the human reference sequence. Given that modern humans and the lineage leading to Denisovans and Neanderthals diverged from one another approximately 553-589,000 years ago (Prüfer *et al*., 2014), it would seem as though these proviruses infected the germline of a hominid ancestral to Denisovans and Neanderthals. Since Denisovans and Neanderthals diverged from one another 381,000 years ago (Prüfer *et al*., 2014), it is possible to conclude that all the HERV-K-DeNe- proviruses given in this chapter are between 381-589,000 years old. It is also possible to conclude that HERV-K-De- and HERV-K-Ne- elements are between 50-381,000 years old, based on this divergence date, and the approximate age of the phalanxes from which the genomes were sequenced. This would, however, assume that no inbreeding occurred between Neanderthals and Denisovans.

Previous archaic hominid HERV-K(HML-2) screens have only used low-coverage genomes (Agoni *et al*., 2012), which offer limited validating evidence of novel loci, due to many of these only being represented once, and by short reads. When conducting a screen on the low-coverage Denisovan genome (reported in appendix C), many loci were rejected, and back-filtering discarded even more. While this screen was aborted upon Agoni *et al*.'s (2012) publication in favour of screening higher quality data, it still identified an element that they did not detect (see Figure 1 in appendix C). This, together with all of their 14 published proviruses, was recovered in the subsequent screens presented here (the missed HERV was later named HERV-K-DeNe1 (Figure 3.3)). Agoni *et al*. (2012) screened for a minimum of 18 bp of LTR, together with 20 bp of flank. In the screen of the same genome reported in appendix C, a 25 bp LTR probe was used to BLAST the low-coverage Denisovan genome, using a word size of 11. In this genome, HERV-K-DeNe1 was represented by two sequence reads, although both contained only 11 bp of LTR – this was picked up by BLAST in this study, but missed by Agoni *et al*. (2012) who set a requirement of ≥18 bp of LTR. Overlooking novel retroviral insertions is an inevitability when screening an unassembled, 1.9-fold coverage genome, consisting of short reads. The Illumina technologies used to sequence the biological sample for this genome typically produce short reads, less than 100 bp in length. Irrespective of this, this missing HERV was subsequently recovered from the high-coverage genomic screen presented here, and was found also to be shared in Neanderthals (Figure 3.3). It was subsequently named HERV-K-DeNe1.

Aside from the computational demands of these screens, complications also arose from the genomic sequence data used. Many false positives were encountered prior to arriving at a novel HERV insertion. These manifested either as a HERV integration already existing within the human reference sequence (and thus assumed to be fixed), or as an anomaly resulting from sequencing artefacts. In order to verify that an apparent novel HERV is not a false positive, a combination of sufficient HERV-K113 LTR sequence and unique flanking sequence are required within any one read. While a seemingly logical approach would be to assemble the archaic hominid

genomes by mapping reads onto the already assembled human genome, this would likely discard reads containing unique, novel HERVs. For instance, HERV-K(HML-2) elements are over 9 Kb in length (Subramanian *et al*., 2011), and if an element of this size is missing from the human reference sequence, assembly would likely discount reads containing novel HERV sequence. Thus, for studies of this nature, the most appropriate way to decisively conclude that a HERV is present in an unassembled genome, is to use its flanking sequence to identify it – which requires a sequence read to contain sufficient flank to map to the human reference sequence. Low-coverage genomes provide limited potential in this regard. During the low-coverage Denisovan screen presented in the appendix C, UCSC's BLAT was used to search for the flanking sequence in the human genome, which meant the flanks had to be set to at least 20 bp long; as this is the minimum length that can be inputted into BLAT. Reads with shorter flanks were discarded, which inevitably led to many novel HERVs being lost.

In stark contrast to this, the high-coverage Denisovan and Neanderthal genomes used for the second set of screens provided such vast amounts of data, that reducing the volume of hits obtained became a concern, since it became too exhaustive to verify so many candidate elements. Accordingly, the first screen stringently matched LTR – even a single base difference was discounted. To further increase the stringency and reduce the number of hits, the size of the probe used was 20 bp. While this meant the entire genome could feasibly be screened for retroviruses, it also inevitably resulted in many retroviruses being missed. Thus, the parameters set and methods used in this study met a compromise between feasibly screening all the data, and conducting an exhaustive retroviral screen.

Reducing the volume of hits was not the only reason for using a stringent-matching perl script, however. While these screens were all ultimately performed on a powerful computing cluster, I first attempted to start the screen on the same high-specification personal computer referenced in section 3.2.2, which was used to screen the low-coverage Denisovan genome. In order to do this, BLAST was the proposed tool for retrieving

HERVs. BLAST can be used to directly align a probe against a target FASTA format sequence, however it will only display the matching region. Since this study requires flank, the only way to use BLAST to extract the sequence read in which a match is located, is to build a BLAST database. This can be done using the makeblastdb script provided by NCBI, however this consumes extensive computing resources. Not only does building these databases take a considerable amount of time and processing power on an ordinary personal computer, it also requires storage space that equals or exceeds the amount required by the genome in the first place. While BLAST searches were used as final verification, the goal was to avoid using BLAST for mining the genomes for provirus sequences. Employing a simple perl script that reads sequences line by line and handles them using regex pattern matching is a more efficient solution to this problem.

The back-filtering procedure described in section 3.2.4 was one area where these perl scripts were particularly beneficial. Ordinarily, this process would involve BLAST-searching the archaic hominid genomes to recover corresponding 3′ ends of candidate HERVs. However, by using perl scripts to filter out reads not containing viral LTR, it was possible to construct a library of HERV-containing reads, that would be more likely to contain the opposite end of the HERV being verified. The resulting library of reads was of a much more easily manageable size, and could be used in the preliminary stages of back-filtering. Since the luxury of a high-performance computing cluster was available, any reads that did not successfully back-filter were ultimately checked by BLAST, regardless. In spite of this, the perl-based back-filter was very successful at recovering the vast majority of opposing HERV ends. This was particularly true when creating the library using fuzzy-matching, as this returned a larger volume of HERV-containing reads.

Since the fossilised phalanxes from which the archaic hominid genomes were sequenced belong to extinct subspecies, they are very valuable (Reich *et al*., 2010). It is therefore not feasible to use PCR as a means of verifying a candidate novel HERV. This must instead be done *in silico*. While the early stages of the workflow are automated, many of the subsequent verification

steps (particularly the steps described in section 3.2.3) are done manually, by eyeballing the sequence. This is exhaustive, but would prove cumbersome to automate. While BLAST-searching HERV-containing reads against the hg38 human reference sequence, there are a multitude of sequence match patterns that can arise depending on the total read length, proportion of flank to LTR and whether or not the read is marred by sequencing artefacts, such as template-switching. The latter occurs during the amplification of sample DNA (as is required by the Illumina protocols (Metzker, 2010) used for all the genomes screened in this chapter), whereby the thermostable DNA polymerases used produce sequence reads from two separate DNA sequences, spliced together erroneously. This is caused by incomplete strand extension, and then subsequent jumping of the DNA polymerase onto another template strand, creating a read that is a hybrid of two independent DNA sequences (Chakravarti & Mailander, 2008).

There are several ways in which template switching can present problems during HERV verification. If a template switch occurs precisely at the virus-host breakpoint (where the virus inserts), it can lead to a LTR appearing adjacent to an erroneous flank. When aligned to the human reference genome, this template-switched flanking sequence may successfully map to an apparent empty preintegration site, resulting in a false positive. While this occurred in several of the reads, a template switch occurring precisely at the virus breakpoint is rare. A far more common occurrence is a template switch happening at any point within the flanking sequence or LTR, which could result in two possibilities. If a template switch occurs in the flank, it might have the effect of truncating the flanking sequence to a length that is insufficient to map to the human reference sequence for verification. Alternatively, template switches occurring in the LTR sequence may lead to the impression that the LTR does not match HERV-K113 entirely. However, the latter is seldom a problem, since the perl scripts used to mine the reads from the archaic hominid genomes are already set to detect 20 bp of LTR. These 20 bases are likely sufficient to conclude that the sequence is homologous to the LTR of HERV-K113. In poorer quality sequence data, it might be expected that template switches would be more frequent. This was certainly true of the

Neanderthal data that was left unmapped by the authors. Their main Altai Neanderthal genome dataset (Prüfer *et al*., 2014) went through stringent quality filtering protocols, and the unmapped reads were a mix of both high-quality and poorer-quality sequence (much of which contained template switches). However, the 52-fold coverage that this genome was sequenced at meant that this was largely not a problem. Template switching appears to be random, and so when it occurs, it usually affects a single read. If two or more unique reads sufficiently represent a candidate novel HERV, then it was considered that they accurately represented the sample DNA. Unique here refers to two or more reads largely representing the same sequence, but one read may be positionally offset from another. An example of how template switching can affect a BLAST match, is illustrated in Figure 3.5.

While the LTRs of HERV-K113 are very highly conserved (Lebedev *et al*., 2000), it is possible that stringent-matching 100% of the proximal and distal ends of these may be discarding reads containing insertionally polymorphic HERVs. This is an intrinsic limitation of using regex pattern matching in perl to detect biological sequence data. Therefore, using a fuzzy-matching perl module, it was possible to allow for a customisable degree of error in the target sequence. Screening the high-coverage Denisovan genome using the fuzzy-matching script in appendix G, returned approximately 100,000 matched reads. Since subsequent verification is an involved process, a filtering script was developed that was set to discard reads with less than 25 bp of flanking sequence. Since stringent-matching scripts relied on regexes for matching sequence to a specified probe, this allowed for a specified number of characters before or after the match to be returned as a condition for keeping that match. This proved useful in reducing the volume of hits, since many hits often do not contain sufficient flanking sequence to unambiguously determine their integration locus as a unique one. When using the fuzzy-matching script, however, this capability is lost, since matching doesn't rely on regex pattern matching. To circumvent this, a second script was developed, which detected the location of the fuzzy matched probe, and subsequently detected the number of characters before and after that probe (appendix I). While verifying 100,000 loci would have been the most

exhaustive option, it would be expected that the majority of novel loci would still be detected using these parameters, as the genome coverage is so high. Since the stringent matches had already been screened, the perl script in appendix H was used to discard 100% matches to the 20 bp LTR probe. This left approximately 22,000 HERV-containing reads with a single base pair mismatch, which were subsequently verified. Although only a small, one base pair change was allowed for, this fuzzy-matching screen uncovered a further five novel HERVs.

Fuzzy-matching is well-suited to dealing with nucleic acid sequences, since it can account for additions, deletions and substitutions (Hietaniemi, 2001-2005). The degree of error can also be customised (in percent), which could potentially be useful for detecting less homologous sequences, although it is likely better suited to sequences exhibiting higher levels of homology. Certainly, its application in screening large datasets for insertionally polymorphic HERV-K(HML-2) elements in this study, have proven that it is well-suited to such screens. Since fuzzy-matching is computationally far more efficient than BLAST, I propose that it could confer great advantage when screening high-coverage genomes for highly conserved ERVs and similar SGEs.

Chapter 4: ERV Nomenclature: Implementing a Revised System for Naming Endogenous Retroviruses

4.1 Introduction

The nomenclature of Endogenous Retroviruses has long been a subject of debate among retrovirologists. The number of described retroviral groups is continually increasing, with several hundred thousand ERV loci having already been identified in the available vertebrate genomes (Stoye, 2012). While existing classification systems are loosely based on the ICTV system developed for classifying exogenous retroviruses, the growing number of discoveries and interest in the field necessitates that ERVs be named and catalogued under a unified, conventional system to eliminate ambiguity.

At present, ERVs are named using a variety of different conventions. Some ERVs, for instance, are named according to their host species; human ERVs are named HERVs, and murine ERVs are often named MuERV. A hyphenated letter following these designations is also used to name ERVs according to the PBS-binding tRNA primer required for reverse transcription (Cohen & Larsson, 1988). The appended, hyphenated letters reflect the single amino acid code: ERV-L indicates a leucine primer, HERV-K indicates a lysine primer. While this is useful in detailing the primer required, an amino-acid appendage to the name does not necessarily account for the phylogenetic relationship between members of the named retroviral group. There are, however, subgroups which have parenthesised designations intended to confer phylogenetic similarity. Such an example would be HERV-K(HML-2), where HML- stands for *h*uman *M*MTV-*l*ike (Andersson *et al*., 1999). Furthermore, one group in particular – HERV-K(HML-5) – exhibits strong phylogenetic similarity to other HERV-K(HML-) elements, yet is technically not correctly assigned to the HERV-K lineage. The tRNA used to prime HERV-K(HML-5) reverse transcription is not a lysine primer, but instead a methionine primer (Lavie *et al*., 2004). Despite this, HERV-K(HML-5) is still named as belonging to the HERV-K retroviral group, when it should technically be classed as HERV-M.

In the event that the tRNA primer is not ascertainable, some ERVs have been named according to the nearest known host gene. An example of this is

HERV.ADP, named after its neighbouring ADP-ribosyltransferase gene (Lyn *et al*., 1993). ERVs can alse be named by cloning number and amino acid motifs, as for HERV.S71 and HERV.FRD, respectively (Seifarth *et al*., 1995). In addition, some researchers have deemed it important to confer the locus of the insertion in the name; a HERV located on the 11q arm of chromosome 22, for instance, is named HERV-K 11q22 to reflect this (Stoye, 2012). Other retroviruses have simply been assigned a unique number for identification (i.e. HERV-K113). Furthermore, since the ERV designation is not always unanimous – with human ERVs being named HERVs, for example – it can be difficult to correctly refer to orthologous elements. While all these naming conventions have their merits, a universal system for nomenclature is lacking. Considering that the number of ERV loci identified is in excess of several hundred thousand elements, there becomes a pressing requirement for a universal nomenclature system.

All endogenous and exogenous retroviruses are contained within the family *Retroviridae* (Doolitle *et al*., 1989). An ambiguous term, much of the literature has adopted 'family' to describe a lineage of phylogenetically related ERVs that are the progeny of an initial germline infection event (Lander *et al*., 2001). As an example, this would incorrectly assume that HERV-K is a 'family', when it would more correctly be referred to as either a retroviral group, or a lineage.

Based on both sequence similarity and replication strategies, the *Retroviridae* are divided into two distinct sub-families: *Spumaretrovirinae* and *Orthoretrovirinae*. The latter are divided into further genera, which mirror those of XRVs. These are *Alpharetrovirus*, *Betaretrovirus, Gammaretrovirus, Deltaretrovirus, Lentivirus* and *Epsilonretrovirus* (Griffiths, 2001).

As previously discussed in chapter 1 (section 1.3), ERVs are generally grouped into three phyletic classes: I-III. These are based on phylogenetic similarity of ERV sequences to those of their exogenous counterparts (Jern *et al*., 2005a). Class I is loosely based on similarity to exogenous gammaretroviruses, while class II and III are based on similarity to betaretroviruses and spumaretroviruses, respectively (Doolitle *et al*., 1989).

This classification system has a long history within the field, and it is important that revisions made to the existing system continue to reflect the relationship between ERVs and XRVs.

In this chapter, I have written a computer program designed to work as part of a pipeline for systematically naming ERVs, developed in collaboration with some of the field's leading experts. This program will effect a novel, tri-partite ERV nomenclature system to assign all existing ERVs a unique identifier. This system will allow for the unambiguous cataloguing of every identified ERV insertion, and is the new proposed standard for naming and referring to ERVs (Blomberg *et al.*, unpublished). The nomenclature system will incorporate the following criteria to assign every ERV a unique identifier:

- ERV phylogenetic relationship to other retroviruses.
- A unique numerical identifier.
- A four-letter host genus-species code.

An example of how ERVs are named using this system is illustrated in Figure 4.1.

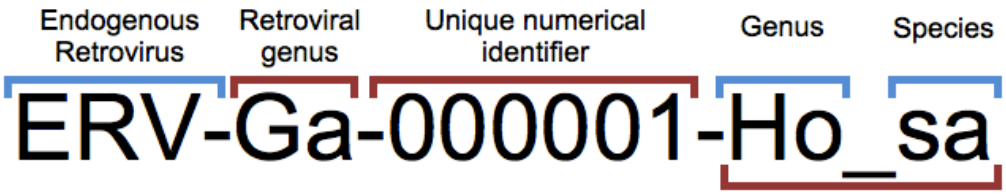


Figure 4.1: **ERV nomenclature schematic**, following the conventions of the unpublished, proposed nomenclature system (adapted from Blomberg *et al.* unpublished). Each component of the tri-partite system is bracketed in dark red; other components in blue.

While this naming system generates a long, unmemorable name, it is intended that ERVs that are already named, and well-studied in the literature, can be referred to using this existing name; in a similar fashion to how an organism might have a scientific, binomial name, in addition to a common name. There will be a dedicated online database that will provide a key for tri-partite and existing names, which will serve as a definitive reference. The new naming system will result in a definitive catalogue of every reported ERV

insertion, working to remove ambiguity when referring to ERV insertions in the future.

In order to implement this nomenclature system, Gifford *et al*. (unpublished) are building a pipeline, designed to mine vertebrate genomes for ERV insertions. This generates and tabulates several fields of data for each identified ERV. The pipeline is currently incomplete, as it lacks a final, automated step to assign ERVs their unique identifiers based on the screen-generated output. In this chapter, I present a program (under preparation for publication) – written in perl – that can be used to systematically parse the output of Gifford *et al*.'s existing pipeline, in order to generate individual, unique names for each ERV, according to the nomenclature system outlined in Figure 4.1. This program will form an integral part of the Gifford *et al*. (unpublished) pipeline for batch-naming large numbers of identified ERVs. The program provides a powerful, efficient and foolproof method for naming ERV insertions across multiple screened genomes, facilitating the cumbersome task of naming all the ERVs existing in the vertebrate genomes that have been sequenced, to date. While currently under publication, it is expected that this revised nomenclature standard will rename all existing retroviruses, providing a novel ERV nomenclature system to be implemented globally.

4.2 Methods and Results:

This chapter presents a computer program which, together with its output, constitutes both methods and results. Accordingly, methods and results sections have been combined.

4.2.1 Data Used: The program presented in this chapter is designed to work on the output of an existing ERV locus-mining pipeline created by Dr. Robert J. Gifford (unpublished). At the time of writing, this pipeline has been run on the catarrhine primates, and has identified 156,614 ERV loci across these species. Gifford *et al*.'s (unpublished) pipeline does not merely screen the genomes for ERVs, but also obtains information about each ERV insertion,

and outputs these as 11 tab-delimited columns. These columns, from left to right, are:

- Record ID – this is a numerical identifier.
- Assigned to – this is the retroviral group to which an ERV belongs.
- Assigned notes – this also contains the retroviral group.
- Extract start – the ERV start locus.
- Extract end – the ERV end locus
- Organism – the host organism.
- Orientation – the native orientation of the ERV in the host genome data.
- Chunk name – the name of the source file.
- Scaffold – the gi identifier used by NCBI to catalogue identified sequences, or the chromosome number.
- Genome structure – the structure of the identified ERV genome. If clear ORFs are detected, they are assigned their correct name (i.e. *gag*, *pol*, *env*), and are laid out in the order they occur in, from 5′-3′.
- Timestamp – the date and time an ERV was identified using the pipeline.

With the exception of the 'Record ID', 'Extract Start' and 'Extract End' columns, all of the values represented within each column are enclosed in double quotes. Most importantly, for every ERV identified, all this information is contained within a single line, separated at the end by a newline character that denotes the next ERV record. Within each line, individual columns are all separated by tabs.

4.2.2 Data Capture: The ERV Name Assigner program (written in perl) was designed to read all the aforementioned data, and use it to generate names for each ERV locus, according to the new proposed nomenclature system. Furthermore, ERVs are assigned their correct genus from the 'Assigned to' column, which contains the retroviral group to which each element belongs. The first program task is to read in each line, which uses a 'foreach' loop to do

this. All subsequent steps are then executed on a single line at a time, implemented using a while loop.

One main subroutine acts on these individual lines of data, with its main component being a regex pattern match. This splits all the data in each column by tabs, and assigns them to scalar variables, which can then be used to name each ERV integration. As mentioned in the previous section (4.2.1) the data comprises 11 columns, each separated by a tab. The regex line therefore splits each of these by tab, and by the type of data expected to occur within each column (i.e. numbers in the 'Record ID' column, a single word character in the 'Assigned to' column, small words separated by hyphens in the 'genome structure' column *etc*.). Each of these is then assigned to a scalar of the same name.

Since the newly devised naming system requires the first two characters of genus and species names, the 'organism' column was assigned to three different scalars. These were '$organism' (in case this is required at a later date, in its native form) and '$organism_genus' and '$organism_species'. The latter two scalars capture just the first two characters of the genus name, and the first two characters of the species name. For *Homo sapiens*, for example, '$organism_genus' would capture 'Ho', and 'organism_species' would capture 'sa'.

While not all the data in this set is required for generating a tri-partite ERV name, each column was assigned to its namesake scalar, in case this data would be required in the future, should the program be expanded for additional functionality. The regex line used to capture the data from the input file is illustrated in Figure 4.2, and can be seen in the script in Figure 4.5 (subunits B1 and C1).

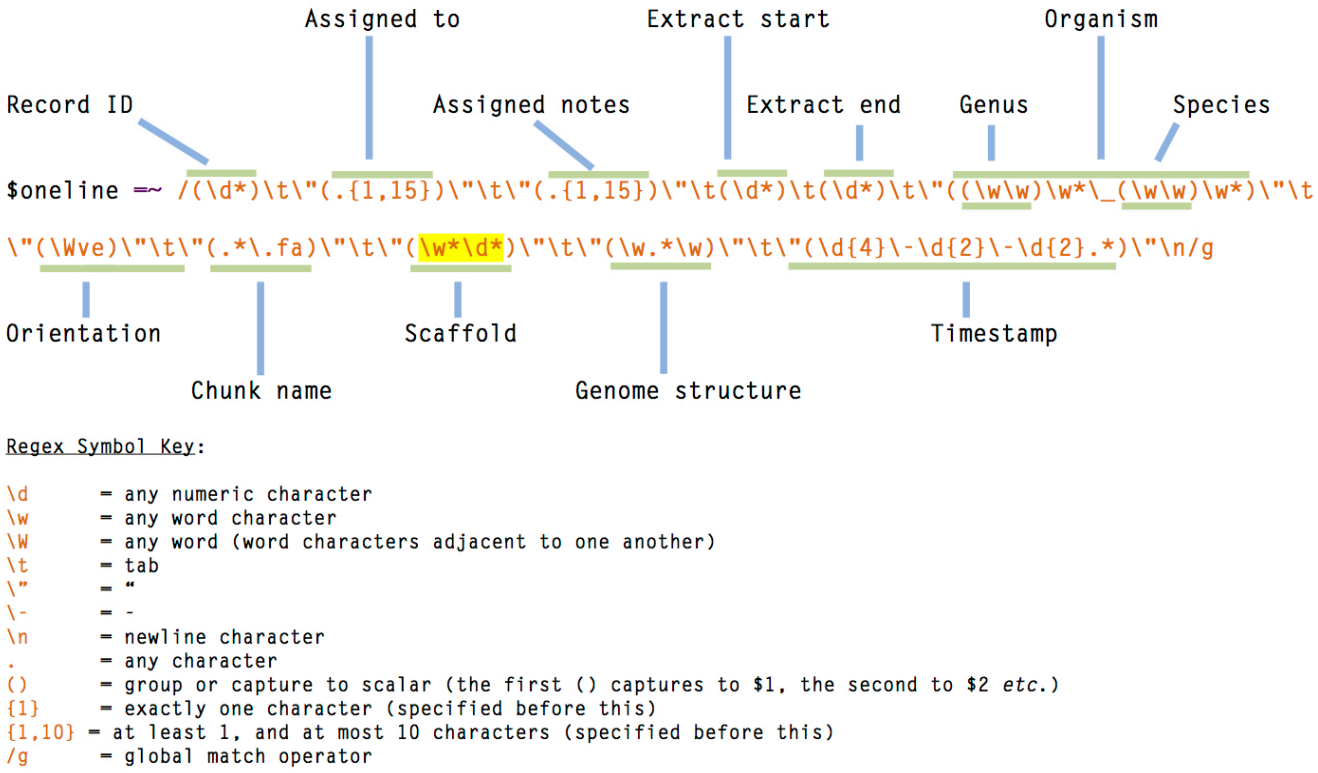


Figure 4.2: **Data capture schematic**. This shows one of the two main regex pattern matching conditionals that captures the data from the input file. Each of the parenthesised units (rounded brackets) indicates a captured column value. These are demarcated above the regex with light green lines. A key of all the symbols used in this regex is also provided. The yellow highlight indicates the scaffold column, which differs slightly in the second conditional, to pick up chromosome numbers, rather than a gi sequence identifier. The whole regex shown here acts on a single line of the input data from Gifford *et al.*'s ERV-mining pipeline (unpublished).

4.2.3 Leading Zeros: The number of digits in the 'Record ID' column varies. For ERV number 1, for example, this is just one digit; for ERV number 1000, this is four digits *etc*. Since the nomenclature and classification systems are still unpublished, the authors are still in debate regarding many of the specifics, and one such area is the number of leading zeros an ERV name should incorporate. A requirement of the naming system is to have ERV names contain the same total number of digits, irrespective of the 'Record ID' number. At present, since the dataset comprises 156,614 loci, I have set this to 6 digits, which would require up to 5 leading zeros to be inserted for single-digit 'Record ID' numbers (i.e. 1 would become 000001, and 156,000 would remain 156000). This works by ascertaining the number represented by the Record ID column (assigned to the $idnumber scalar), and assigning the

appropriate number of zeros to another scalar, $zeros, which is printed in the final ERV name. The described loops are shown in Figure 4.3, and can be seen within the program in Figure 4.5 (subunits B2 and C2).

```
#ADD LEADING ZEROS. CURRENTLY ASSUMES 6-DIGIT NUMERICAL IDENTIFIER.
if ($idnumber < 10 )         {$zeros = "00000";}
elsif ($idnumber < 100)      {$zeros = "0000";}
elsif ($idnumber < 1000)     {$zeros = "000";}
elsif ($idnumber < 10000)    {$zeros = "00";}
elsif ($idnumber < 100000)   {$zeros = "0";}
elsif ($idnumber < 1000000) {$zeros = "";}
```

Figure 4.3: **Loops for assigning leading zeros**, to create numerical identifiers totalling six digits across all ERVs.

4.2.4 Genus Assignment: A set of 'if' and 'elsif' conditional loops then operate on the '$Assigned_to' scalar. These are a set of conditionals, which assign ERVs to one of three retroviral genera: *Gammaretrovirus, Betaretrovirus* and *Spumaretrovirus*. These names are abbreviated to 'ga', 'be' and 'sp', respectively, and it is these abbreviations that are used in the final ERV names. These genus-assigning loops are shown in Figure 4.4, and can be seen within the main program in Figure 4.5 (subunits B3 and C3).

Since the catarrhine dataset comprises over 150,000 ERVs, belonging to a range of different retroviral groups, it is expected that some of these ERVs may have names that have not yet been assigned to genera within the main perl script. Therefore, the script incorporates a feedback loop, which counts and captures all of the ERVs to which genera have not yet been assigned. These are then written to a separate file called 'Unknown_Groups.txt'. The program segment effecting this can be seen in Figure 4.5 (subunits B4, C4 and D1).

As this file will often have many different ERVs from different retroviral groups, a second program in appendix J is designed to output just the retroviral group. For example, if 'Unknown_Groups.txt' contains 10,000 unassigned MPMV retroviruses, this second script will just output 'MPMV', removing duplicates. This makes it possible to easily add in an elsif loop assigning MPMV to the correct genus.

```
if ($assigned_to =~ /CERV-2/ )          { $group = 'Ga'; &name_printer2;}
elsif ($assigned_to =~ /HERV-T/ )       { $group = 'Ga'; &name_printer2;}
elsif ($assigned_to =~ /HERV-E/ )       { $group = 'Ga'; &name_printer2;}
elsif ($assigned_to =~ /HERV-Pb/ )      { $group = 'Ga'; &name_printer2;}
elsif ($assigned_to =~ /HERV-V/ )       { $group = 'Ga'; &name_printer2;}
elsif ($assigned_to =~ /RR-HERV-I/ )    { $group = 'Ga'; &name_printer2;}
elsif ($assigned_to =~ /HERV-Fb/ )      { $group = 'Ga'; &name_printer2;}
elsif ($assigned_to =~ /HERV-Fb-2/ )    { $group = 'Ga'; &name_printer2;}
elsif ($assigned_to =~ /HERV-XA/ )      { $group = 'Ga'; &name_printer2;}
elsif ($assigned_to =~ /ERV-Fc/ )       { $group = 'Ga'; &name_printer2;}
elsif ($assigned_to =~ /HERV-H/ )       { $group = 'Ga'; &name_printer2;}
elsif ($assigned_to =~ /HERV-I/ )       { $group = 'Ga'; &name_printer2;}
elsif ($assigned_to =~ /HERV-R/ )       { $group = 'Ga'; &name_printer2;}
elsif ($assigned_to =~ /HERV-W/ )       { $group = 'Ga'; &name_printer2;}
elsif ($assigned_to =~ /HERV-B7H6/ )    { $group = 'Ga'; &name_printer2;}
elsif ($assigned_to =~ /ERV-9/ )        { $group = 'Ga'; &name_printer2;}
elsif ($assigned_to =~ /HERV-30/ )      { $group = 'Ga'; &name_printer2;}
elsif ($assigned_to =~ /HERV-K-14C/ )   { $group = 'Be'; &name_printer2;}
elsif ($assigned_to =~ /HERV-K-HML2/ )  { $group = 'Be'; &name_printer2;}
elsif ($assigned_to =~ /HERV-K-HML4/ )  { $group = 'Be'; &name_printer2;}
elsif ($assigned_to =~ /HERV-K-HML5/ )  { $group = 'Be'; &name_printer2;}
elsif ($assigned_to =~ /HERV-K-HML6/ )  { $group = 'Be'; &name_printer2;}
elsif ($assigned_to =~ /HERV-K-HML7/ )  { $group = 'Be'; &name_printer2;}
elsif ($assigned_to =~ /HERV-K-HML8/ )  { $group = 'Be'; &name_printer2;}
elsif ($assigned_to =~ /HERV-K-HML9/ )  { $group = 'Be'; &name_printer2;}
elsif ($assigned_to =~ /RhERV-K-HML2/ ){ $group = 'Be'; &name_printer2;}
elsif ($assigned_to =~ /HERV-L/ )       { $group = 'Sp'; &name_printer2;}
elsif ($assigned_to =~ /HERV-S/ )       { $group = 'Sp'; &name_printer2;}

    #ADDITIONAL GENUS ASSIGNMENTS.
    elsif ($assigned_to =~ /MPMV/ )             { $group = 'Be'; &name_printer2;}
    elsif ($assigned_to =~ /RhERV-1b/ )         { $group = 'Ga'; &name_printer2;}
    elsif ($assigned_to =~ /BAEV/ )             { $group = 'Ga'; &name_printer2;}
    elsif ($assigned_to =~ /RhERV-d1/ )         { $group = 'Be'; &name_printer2;}
    elsif ($assigned_to =~ /CERV-1/ )           { $group = 'Ga'; &name_printer2;}
    elsif ($assigned_to =~ /RhERV-1a/ )         { $group = 'Ga'; &name_printer2;}
    elsif ($assigned_to =~ /SRV-2/ )            { $group = 'Be'; &name_printer2;}
    elsif ($assigned_to =~ /Nomascus-ERV-K/ )   { $group = 'Be'; &name_printer2;}
    elsif ($assigned_to =~ /Rh-HERV-K2/ )       { $group = 'Be'; &name_printer2;}
```

Figure 4.4: **Genus-assinging loops** used to assign genus. The bottom elsif loops show genera that were missed by the first run of the screen, but subsequently processed with the script in appendix J, and added back into the original naming program.

In order to assign an unknown retroviral group to the correct genus, each group was researched in the literature to retrieve this information, prior to being added back into the program within additional elsif conditionals (see Figure 4.4).

4.2.5 Name Printing: The final part of the script uses the captured data to construct and assign a unique, tri-partite alphanumeric identifier to each ERV insertion. This follows the convention outlined in Figure 4.1. The 'printer' subroutine compiles this name, and outputs this; both directly into the terminal (for monitoring purposes) and into a final, tab-delimited table. The segment of

the program responsible for creating the final ERV name can be seen in Figure 4.5 (subunits B5 and C5).

4.2.6 Output Files: Due to a difference in the data that occurs in the 'scaffold' column, the program contains two conditional loops that carry out all of the steps outlined in sections 4.2.1-5. Both these loops carry out exactly the same functions. However, one of these loops contains a regex that handles gi identifiers in the 'scaffold' column, and the other handles lines that contain a simple chromosome number in place of this. In order to capture and process lines containing both types of scaffold data, two regexes (and thus, two conditional loops) were required. Therefore, while it appears as though there are duplicated lines in the program, these are actually two separate loops for handling differences in the 'scaffold' column data.

As there are two loops, ERV names are outputted into two files: 'finaltable1.txt' and 'finaltable2.txt'. Once both of the main program conditional loops have processed the entire input data file, the program concatenates these two files into one: 'finaltable.txt'. This table is the same as the input data – maintaining all the information in its native form – except that it contains an additional appended column entitled 'ERV_name', which contains the new ERV identifiers, as specified by the 2014 nomenclature system that is currently under development (Blomberg *et al*., unpublished). The code handling the final program output can be seen in Figure 4.5 (subunit E1).

4.2.7 Feedback: The script is written to provide feedback to the user in the terminal itself, regarding the processes it is running, throughout. A loop counter in each loop registers the number of ERVs that have been named, and outputs a sentence in the terminal, once the main part of the program has finished running. These indicate to the user how many ERVs have been named successfully and how many have not been named due to unassigned genera. If the input data is not in the format outlined in section 4.2.1, the script will abort, and appropriate feedback is given in the terminal (see Figure 4.5, subunit E1).

4.2.8 Sorting: Finally, the final output table contains all the data in a tab-delimited format, such that it can be opened using a program such as Microsoft Excel. ERVs can then be sorted by column. For example, it may be desirable to sort them in ascending numerical order, which can be done by sorting the first 'Record ID' column in Excel.

The full ERV Name Assigner program is shown here, over the following pages (Figure 4.5).

```perl
#! usr/bin/perl
#*************************************************************************
#*************************************************************************
#*************************************************************************

#SET FILEPATH HERE. SCRIPT AND DATA FILE SHOULD BE IN THIS FILEPATH:
$filepath = "/Users/Adam/Desktop/PerlTest2014";

#AND DATA FILE NAME HERE:
@input = <$filepath/CatarrhineTestDB.txt>;

#*************************************************************************
#*************************************************************************
#*************************************************************************
print "\n*******************************************************\n";
print "ERV NAME ASSIGNER. REQUIRES INPUT DATA FROM R. J. GIFFORD PIPELINE. \n\n NAMING SCRIPT ", '(C)', " Adam Lee 2014.\n _________________ \n\nRENAMING ERVs IN ACCORDANCE WITH 2014 CLASSIFICATION SYSTEM.\n";
sleep (2);
print "\n\n- Your results will be saved in finaltable.txt, with ERV names in the final column.\n";
sleep (3);
print "- You will be able to monitor the naming loop iterations in the terminal.\nNOTE: The loop iteration number is two values behind the assigned number.\n";
sleep (4);
print "- Any sequences not picked up by the script are saved into finalfailseqs.txt.\n";
sleep (3);
print "Commencing in 2 seconds:\n";
sleep (2);

my $titlestrip = "Record_ID     Assigned_to     Assigned_notes  Extract_start   Extract_end     Organism        Orientation     Chunk_name      Scaffold        Genome_structure        Timestamp       ERV_Name\n";

open (MAKERESULTFILE1, ">$filepath/finaltable1.txt");
        print MAKERESULTFILE1 $titlestrip;
        close (MAKERESULTFILE1);

open (MAKERESULTFILE2, ">$filepath/finaltable2.txt");
        print MAKERESULTFILE2;
        close (MAKERESULTFILE2);

my $failtitlestrip = "Sequences that were not picked up by the script:\n\n";

open (FAILSEQS, ">$filepath/finalfailseqs.txt");
    print FAILSEQS $oneline;
    close (FAILSEQS);
#*************************************************************************

foreach $inputline (@input)
{
    #print $inputline . "\n";
    open (A,  $inputline) || die $!;
    my @splitinputfile = <A>;
    my @splitlines = split (/>/, "@splitinputfile");
    $splitdata = join (">", @splitlines);

}

$test = 0;
while ($splitdata =~ /(\d.*\n)/g)
{
    $oneline = $1;
    &checker;
    $counter = $test++;
```

A1

A2

```
61      #last;
62  }
63
64  $test1 = 1;
65  $test2 = 1;
66  $unknownlineageid1 = 1;
67  $unknownlineageid2 = 1;
68  #print $test;
69
70  #*********************************************************************************************
71  #*********************************************************************************************
72
73  sub checker {
74      #print "test value: ", $test, "\n";
75
76      #This first loop finds the majority of sequences. The thing that causes some to be missed is
    the 'scaffold'. This loop finds all those that are gi| identifiers.
77      #Elsif loop deals with the rest.
78  if ($oneline =~ /(\d*)\t\"(.{1,15})\"\t\"(.{1,15})\"\t(\d*)\t(\d*)\t\"((\w\w)\w*\_(\w\w)
\w*)\"\t\
    "(\Wve)\"\t\"(.*\.fa)\"\t\"(\w*\d*)\"\t\"(\w.*\w)\"\t\"(\d{4}\-\d{2}\-\d{2}.*)\"\n/g )
79      {
80          $idnumber = $1;
81          $assigned_to = $2;
82          $assigned_notes = $3;
83          $extract_start = $4;
84          $extract_end = $5;
85          $organism = $6;
86          $organism_genus = $7;
87          $organism_species = $8;
88          $orientation = $9;
89          $chunk_name = $10;
90          $scaffold = $11;
91          $genome_structure = $12;
92          $timestamp = $13;
93
94          #ADD LEADING ZEROS. CURRENTLY ASSUMES 6-DIGIT NUMERICAL IDENTIFIER.
95          if ($idnumber < 10 )         {$zeros = "00000";}
96          elsif ($idnumber < 100)     {$zeros = "0000";}
97          elsif ($idnumber < 1000)    {$zeros = "000";}
98          elsif ($idnumber < 10000)   {$zeros = "00";}
99          elsif ($idnumber < 100000)  {$zeros = "0";}
100         elsif ($idnumber < 1000000) {$zeros = "";}
101
102         #ASSIGN GENERA.
103         if ($assigned_to =~ /CERV-2/ )         { $group = 'Ga'; &name_printer1;}
104         elsif ($assigned_to =~ /HERV-T/ )      { $group = 'Ga'; &name_printer1;}
105         elsif ($assigned_to =~ /HERV-E/ )      { $group = 'Ga'; &name_printer1;}
106         elsif ($assigned_to =~ /HERV-Pb/ )     { $group = 'Ga'; &name_printer1;}
107         elsif ($assigned_to =~ /HERV-V/ )      { $group = 'Ga'; &name_printer1;}
108         elsif ($assigned_to =~ /RR-HERV-I/ )   { $group = 'Ga'; &name_printer1;}
109         elsif ($assigned_to =~ /HERV-Fb/ )     { $group = 'Ga'; &name_printer1;}
110         elsif ($assigned_to =~ /HERV-Fb-2/ )   { $group = 'Ga'; &name_printer1;}
111         elsif ($assigned_to =~ /HERV-XA/ )     { $group = 'Ga'; &name_printer1;}
112         elsif ($assigned_to =~ /ERV-Fc/ )      { $group = 'Ga'; &name_printer1;}
113         elsif ($assigned_to =~ /HERV-H/ )      { $group = 'Ga'; &name_printer1;}
114         elsif ($assigned_to =~ /HERV-I/ )      { $group = 'Ga'; &name_printer1;}
115         elsif ($assigned_to =~ /HERV-R/ )      { $group = 'Ga'; &name_printer1;}
116         elsif ($assigned_to =~ /HERV-W/ )      { $group = 'Ga'; &name_printer1;}
117         elsif ($assigned_to =~ /HERV-B7H6/ )   { $group = 'Ga'; &name_printer1;}
118         elsif ($assigned_to =~ /ERV-9/ )       { $group = 'Ga'; &name_printer1;}
119         elsif ($assigned_to =~ /HERV-30/ )     { $group = 'Ga'; &name_printer1;}
120         elsif ($assigned_to =~ /HERV-K-14C/ )  { $group = 'Be'; &name_printer1;}
121         elsif ($assigned_to =~ /HERV-K-HML2/ ) { $group = 'Be'; &name_printer1;}
122         elsif ($assigned_to =~ /HERV-K-HML4/ ) { $group = 'Be'; &name_printer1;}
123         elsif ($assigned_to =~ /HERV-K-HML5/ ) { $group = 'Be'; &name_printer1;}
124         elsif ($assigned_to =~ /HERV-K-HML6/ ) { $group = 'Be'; &name_printer1;}
125         elsif ($assigned_to =~ /HERV-K-HML7/ ) { $group = 'Be'; &name_printer1;}
```

B1

B2

B3

```
126         elsif ($assigned_to =~ /HERV-K-HML8/ ) { $group = 'Be'; &name_printer1;}
127         elsif ($assigned_to =~ /HERV-K-HML9/ ) { $group = 'Be'; &name_printer1;}
128         elsif ($assigned_to =~ /RhERV-K-HML2/ ){ $group = 'Be'; &name_printer1;}
129         elsif ($assigned_to =~ /HERV-L/ )      { $group = 'Sp'; &name_printer1;}
130         elsif ($assigned_to =~ /HERV-S/ )      { $group = 'Sp'; &name_printer1;}
131
132                 #ADDITIONAL GENOME ASSIGNMENTS.
133                 elsif ($assigned_to =~ /MPMV/ )            { $group = 'Be'; &name_printer1;}
134                 elsif ($assigned_to =~ /RhERV-1b/ )        { $group = 'Ga'; &name_printer1;}
135                 elsif ($assigned_to =~ /BAEV/ )            { $group = 'Ga'; &name_printer1;}
136                 elsif ($assigned_to =~ /RhERV-d1/ )        { $group = 'Be'; &name_printer1;}
137                 elsif ($assigned_to =~ /CERV-1/ )          { $group = 'Ga'; &name_printer1;}
138                 elsif ($assigned_to =~ /RhERV-1a/ )        { $group = 'Ga'; &name_printer1;}
139                 elsif ($assigned_to =~ /SRV-2/ )           { $group = 'Be'; &name_printer1;}
140                 elsif ($assigned_to =~ /Nomascus-ERV-K/ )  { $group = 'Be'; &name_printer1;}
141                 elsif ($assigned_to =~ /Rh-HERV-K2/ )      { $group = 'Be'; &name_printer1;}
142
143         else
144         {
145             $unknown_counter1 = $unknownlineageid1++;
146             print "\nUnkonwn counter is: ", $unknown_counter1, "\n";
147             @group1 = ($assigned_to, "-(UNKNOWN_LINEAGE)");
148             &unknown_lineage1;
149         }
150
151         sub name_printer1 {
152         if ($oneline =~ /(.*)\n/)
153         {
154             $newoneline = $1;
155         }
156
157         @ERVname = ("ERV-", $group, "-", $zeros, $idnumber, "-", $organism_genus, "_", $organism_
    species, "\n");
158         print "ERV-", $group, "-", $zeros, $idnumber, "-", $organism_genus, "_", $organism_specie
    s, "\n";
159
160         open (FIRSTLOOPERVS, ">>$filepath/finaltable1.txt");
161         print FIRSTLOOPERVS $newoneline, "\t", @ERVname;
162         close (FIRSTLOOPERVS);
163         #last;
164                         }
165
166
167         $counter1 = $test1++;
168         print "Loop iteration: ", $zeros, $counter, "\n";
169     }
170 #*****************************************************************************************************
    **********************************
171 #*****************************************************************************************************
    **********************************
172 #IMPORTANT NOTE: This is a REPEAT of the above regex and associated actions.
173 #It's for sequences where the 'scaffold' is not a gi| identifier, but a simple chromosome number.
     The regex is essentially the same, except for the
174 #scaffold capture at $11.
175 #*****************************************************************************************************
    **********************************
176 #*****************************************************************************************************
    **********************************
177
178 elsif ($oneline =~ /(\d*)\t\"(.{1,15})\"\t\"(.{1,15})\"\t(\d*)\t(\d*)\t\"((\w\w)\w*\_(\w\w)
\w*)\"
    \t\"(\Wve)\"\t\"(.*\.fa)\"\t\"(.*\.\d{1,3}\|)\"\t\"(\w.*\w)\"\t\"(\d{4}\-\d{2}\-\d{2}.*)\"\n/g)
179     {
180         $idnumber = $1;
181         $assigned_to = $2;
182         $assigned_notes = $3;
183         $extract_start = $4;
184         $extract_end = $5;
```

B4

B5

C1

```
185         $organism = $6;
186         $organism_genus = $7;
187         $organism_species = $8;
188         $orientation = $9;
189         $chunk_name = $10;
190         $scaffold = $11;
191         $genome_structure = $12;
192         $timestamp = $13;
193
194         #ADD LEADING ZEROS.
195         if ($idnumber < 10)         {$zeros2 = "00000";}
196         elsif ($idnumber < 100)     {$zeros2 = "0000";}
197         elsif ($idnumber < 1000)    {$zeros2 = "000";}
198         elsif ($idnumber < 10000)   {$zeros2 = "00";}
199         elsif ($idnumber < 100000)  {$zeros2 = "0";}
200         elsif ($idnumber < 1000000) {$zeros2 = "";}
201
202         if ($assigned_to =~ /CERV-2/ )          { $group = 'Ga'; &name_printer2;}
203         elsif ($assigned_to =~ /HERV-T/ )       { $group = 'Ga'; &name_printer2;}
204         elsif ($assigned_to =~ /HERV-E/ )       { $group = 'Ga'; &name_printer2;}
205         elsif ($assigned_to =~ /HERV-Pb/ )      { $group = 'Ga'; &name_printer2;}
206         elsif ($assigned_to =~ /HERV-V/ )       { $group = 'Ga'; &name_printer2;}
207         elsif ($assigned_to =~ /RR-HERV-I/ )    { $group = 'Ga'; &name_printer2;}
208         elsif ($assigned_to =~ /HERV-Fb/ )      { $group = 'Ga'; &name_printer2;}
209         elsif ($assigned_to =~ /HERV-Fb-2/ )    { $group = 'Ga'; &name_printer2;}
210         elsif ($assigned_to =~ /HERV-XA/ )      { $group = 'Ga'; &name_printer2;}
211         elsif ($assigned_to =~ /ERV-Fc/ )       { $group = 'Ga'; &name_printer2;}
212         elsif ($assigned_to =~ /HERV-H/ )       { $group = 'Ga'; &name_printer2;}
213         elsif ($assigned_to =~ /HERV-I/ )       { $group = 'Ga'; &name_printer2;}
214         elsif ($assigned_to =~ /HERV-R/ )       { $group = 'Ga'; &name_printer2;}
215         elsif ($assigned_to =~ /HERV-W/ )       { $group = 'Ga'; &name_printer2;}
216         elsif ($assigned_to =~ /HERV-B7H6/ )    { $group = 'Ga'; &name_printer2;}
217         elsif ($assigned_to =~ /ERV-9/ )        { $group = 'Ga'; &name_printer2;}
218         elsif ($assigned_to =~ /HERV-30/ )      { $group = 'Ga'; &name_printer2;}
219         elsif ($assigned_to =~ /HERV-K-14C/ )   { $group = 'Be'; &name_printer2;}
220         elsif ($assigned_to =~ /HERV-K-HML2/ )  { $group = 'Be'; &name_printer2;}
221         elsif ($assigned_to =~ /HERV-K-HML4/ )  { $group = 'Be'; &name_printer2;}
222         elsif ($assigned_to =~ /HERV-K-HML5/ )  { $group = 'Be'; &name_printer2;}
223         elsif ($assigned_to =~ /HERV-K-HML6/ )  { $group = 'Be'; &name_printer2;}
224         elsif ($assigned_to =~ /HERV-K-HML7/ )  { $group = 'Be'; &name_printer2;}
225         elsif ($assigned_to =~ /HERV-K-HML8/ )  { $group = 'Be'; &name_printer2;}
226         elsif ($assigned_to =~ /HERV-K-HML9/ )  { $group = 'Be'; &name_printer2;}
227         elsif ($assigned_to =~ /RhERV-K-HML2/ ){ $group = 'Be'; &name_printer2;}
228         elsif ($assigned_to =~ /HERV-L/ )       { $group = 'Sp'; &name_printer2;}
229         elsif ($assigned_to =~ /HERV-S/ )       { $group = 'Sp'; &name_printer2;}
230
231                 #ADDITIONAL GENUS ASSIGNMENTS.
232                 elsif ($assigned_to =~ /MPMV/ )             { $group = 'Be'; &name_printer2;}
233                 elsif ($assigned_to =~ /RhERV-1b/ )         { $group = 'Ga'; &name_printer2;}
234                 elsif ($assigned_to =~ /BAEV/ )             { $group = 'Ga'; &name_printer2;}
235                 elsif ($assigned_to =~ /RhERV-d1/ )         { $group = 'Be'; &name_printer2;}
236                 elsif ($assigned_to =~ /CERV-1/ )           { $group = 'Ga'; &name_printer2;}
237                 elsif ($assigned_to =~ /RhERV-1a/ )         { $group = 'Ga'; &name_printer2;}
238                 elsif ($assigned_to =~ /SRV-2/ )            { $group = 'Be'; &name_printer2;}
239                 elsif ($assigned_to =~ /Nomascus-ERV-K/ )   { $group = 'Be'; &name_printer2;}
240                 elsif ($assigned_to =~ /Rh-HERV-K2/ )       { $group = 'Be'; &name_printer2;}
241         else
242         {
243             $unknown_counter2 = $unknownlineageid2++;
244             print "\nUnkonwn counter is: ", $unknown_counter2, "\n";
245             @group2 = ($assigned_to, "-(UNKNOWN_LINEAGE)");
246             &unknown_lineage2;
247         }
248
249
250         sub name_printer2{
251         if ($oneline =~ /(.*)\n/)
252         {
```

C2

C3

C4

```
253             $newoneline2 = $1;
254         }
255
256         @ERVname2 = ("ERV-", $group, "-", $zeros2, $idnumber, "-", $organism_genus, "_", $organis
    m_species, "\n");
257         print "ERV-", $group, "-", $zeros2, $idnumber, "-", $organism_genus, "_", $organism_speci
    es, "\n";
258
259
260         #last;
261         open (SECONDLOOPERVS, ">>$filepath/finaltable2.txt");
262         print SECONDLOOPERVS $newoneline2, "\t", @ERVname2;
263         close (SECONDLOOPERVS);
264                         }
265
266
267         #$test2 = 1;
268         $counter2 = $test1++;
269         print "Loop iteration: ", $zeros2, $counter, "\n";
270
271     }
272
273 else
274 {
275     #print $oneline;
276     open (FAILSEQS, ">>$filepath/finalfailseqs.txt");
277     print FAILSEQS $oneline;
278     close (FAILSEQS);
279     $failcounter = $test2++;
280     #print "Fail loop iteration: ", $failcounter, "\n";
281 }
282
283 }
284
285 #************************************************************************************************
    ************************************
286 #************************************************************************************************
    ************************************
287 #************************************************************************************************
    ************************************
288 #************************************************************************************************
    ************************************
289
290         sub unknown_lineage1 {
291         my @unknown_lineage = ("ERV-", @group1, "-", $zeros1, $idnumber, "-", $organism_genus, "_
    ", $organism_species, "\n");
292         print @unknown_lineage;
293         push (@unknowngroups, $idnumber, "\t", $assigned_to, "\t", "ERV-", @group1, "-", $zeros1,
     $idnumber, "-", $organism_genus, "-", $organism_species, "\n");
294                             }
295
296 #************************************************************************************************
    ************************************
297
298         sub unknown_lineage2 {
299         my @unknown_lineage = ("ERV-", @group2, "-", $zeros2, $idnumber, "-", $organism_genus, "_
    ", $organism_species, "\n");
300         print @unknown_lineage;
301         push (@unknowngroups, $idnumber, "\t", $assigned_to, "\t", "ERV-", @group2, "-", $zeros2,
     $idnumber, "-", $organism_genus, "-", $organism_species, "\n");
302                             }
303
304 #************************************************************************************************
    ************************************
305
306 sleep (3);
307
308 open (UNKNOWNGROUPS, ">$filepath/Unknown_Groups.txt");
```

C5

D1

```
309         print UNKNOWNGROUPS @unknowngroups;
310         close (UNKNOWNGROUPS);
311
312     #*****************************************************************************************************
    ************************************
313     #*****************************************************************************************************
    ************************************
314     #*****************************************************************************************************
    ************************************
315     #*****************************************************************************************************
    ************************************
316
317     #IN THE CASE OF ANY LINES IN THE DATA NOT PICKED UP BY EITHER REGEX, THIS DUMPS THEM INTO A SEPER
    ATE FILE.
318     if ($failcounter <1)
319     {
320         $allpass = "No";
321     }
322
323     else
324     {
325         $failstatement = "These have been routed into failseqs.txt";
326     }
327
328
329     print "\n", $failcounter, $allpass, " ERVs were not named.", $failstatement, "\n";
330
331     #_____________________________________________________________
332
333     $counter_global = $counter1 + $counter2;
334     $unknown_counter_global = $unknown_counter1 + $unknown_counter2;
335
336     if ($unknown_counter_global > 0 )
337     {
338         print "\n*****\n", $unknown_counter_global, " ERVs were not assigned a genus, because the lin
    eage was unknown.\n\nThese are found in Unknown_Groups.txt.\n*****\n";
339     }
340
341     else
342     {
343         print "\n*****\nAll ERVs have been assigned the correct genus identifier.\n*****\n";
344     }
345
346     #_____________________________________________________________
347
348     $file1 = <$filepath/finaltable1.txt>;
349     $file2 = <$filepath/finaltable2.txt>;
350
351     print "\nConcatenating two results files into finaltable.txt\n";
352
353     if ($counter_global => 1) {sleep (3);
354                         my $cat_exit = system ("cat $file1 $file2 > finaltable.txt");
355                         if ($cat_exit==0) {
356                             print "\nSuccessfully concatenated files.\n";}
357                                 if ($counter_global == 1)
358                                 {
359                                     print "\n*****\n", $counter_global, " ERV was named successfully.
    \n*****\n";
360                                 }
361                                 elsif ($counter_global > 1)
362                                 {
363                                     print "\n*****\n", $counter_global, " ERVs were named successfull
    y. \n*****\n";
364                                 }
365                         }
366
367     elsif ($counter < 1)
368     {
```

E1

```
369         print "\nNo ERVs were named. Check your input file.\n";
370     }
371
372     print "\nFinal results table ready in finaltable.txt.\n\n";
373
374
```

Figure 4.5: **The full ERV-naming perl program** – ERV Name Assigner – is shown here (Lee *et al*., unpublished). As some of the lines are very long, they have been wrapped at the 105th character, in order to fit the whole script on the page. Each new line is numbered on the left, so as to clarify the wrapped lines. The various functional compartments of the program are demarcated by coloured boxes. Each functional unit is assigned a letter, and functional subunits within these are assigned a number. Blue A boxes show regions involved with reading in input files (A1), and splitting files into individual lines (A2). Orange B boxes indicate the first half of the main naming subroutine (entitled checker); green C boxes are a duplicate of these, with the slightly modified regex for handling chromosome numbers in the 'scaffold' column of the input data. B1 includes the main regex explained in Figure 4.2, and is the portion of the program involved in splitting an individual line by column, capturing each data value to a scalar of the appropriate name; C1 carries out the same function on columns containing chromosome numbers, using the modified regex. B2/C2 are responsible for adding leading zeros according to the ERV 'Record ID' number. B3/C3 contains the if and elsif conditionals involved in assigning ERVs to the correct retroviral genus. B4/C4 monitors for any ERVs that have not successfully been assigned genera in the previous conditionals; any that are detected are piped to another, later segment of the program (D1). B5/C5 is the program unit involved in compiling a tri-partite ERV name from the captured data, and this outputted into a tabulated, tab-delimited text file together with all of the relevant – one for the B unit, and one for the C unit of the program. If any ERVs were not assigned genera, these are fed into the D1 subunit, which outputs them into 'Unknown_Groups.txt' – a separate file for processing with UnknownLineage.pl (appendix J). The E1 unit of the script is responsible for providing a final output report to the user, as well as concatenating the B and C subunit files into 'finaltable.txt'.

4.2.9 ERV Names: The main focus of ERV Name Assigner is to name ERVs using the newly devised tri-partite nomenclature system (Blomberg *et al*., unpublished). When run on Gifford *et al*.'s (unpublished) catarrhine primate pipeline data file, the program successfully named all 156,614 viruses. These ERV names are outputted in the terminal, as the program runs, allowing the user to obtain an overview of the naming processes.

Since the dataset used here represents primate ERVs, only three of the six ERV genera are assigned; the other three occur in other host species. Within primates, the three most commonly encountered genera are *Betaretrovirus*, *Gammaretrovirus* and *Spumaretrovirus*. Table 4.1 shows a few examples of the ERVs that were successfully assigned tri-partite names from the catarrhine primate dataset.

Table 4.1: **Table showing a few sample ERV names** of the 156,614 successfully named loci. The table indicates the record ID – a randomly assigned number unique to every ERV identified by the pipeline – and used in the final ERV name. The retroviral group, that leads to the genus assigned, is also given, together with the host species in which the ERV was identified. All these components are used in the final, tri-partite ERV name (shaded red).

Record ID	Retroviral Group	Assigned Genus	Host Species	Tri-Partite ERV Name
71	RhERV-1b	*Gammaretrovirus*	*Gorilla gorilla*	**ERV-Ga-000071-Go_go**
173	RhERV-1b	*Gammaretrovirus*	*Gorilla gorilla*	**ERV-Ga-000173-Go_go**
995	HERV-K-HML8	*Betaretrovirus*	*Nomascus leucogenys*	**ERV-Be-000995-No_le**
1075	HERV-K-HML8	*Betaretrovirus*	*Macaca mulatta*	**ERV-Be-001075-Ma_mu**
12792	HERV-Fb-2	*Gammaretrovirus*	*Pan paniscus*	**ERV-Ga-012792-Pa_pa**
28349	HERV-Fb-2	*Gammaretrovirus*	*Papio hamadryas*	**ERV-Ga-028349-Pa_ha**
37535	ERV-9	*Gammaretrovirus*	*Homo sapiens*	**ERV-Ga-037535-Ho_sa**
117223	HERV-S	*Spumaretrovirus*	*Nomascus leucogenys*	**ERV-Sp-117223-No_le**
117339	HERV-S	*Spumaretrovirus*	*Pan troglodytes*	**ERV-Sp-117339-Pa_tr**

4.2.10 Unassigned Genera: Since not all the retroviral groups were included in the genera conditionals the first time the program was run, they were not all assigned genera, at first. As mentioned in section 4.2.4, the program outputs these unassigned ERVs into a dedicated file, and these were processed with another script, which outputted a list of each unique retroviral group which has not been assigned a genus (see added genera conditionals in Figure 4.4). In this case, there were nine unassigned groups. These were then researched in the literature, and added back into the program, and the second run of the script generated names for the entire dataset. Ultimately, the more ERV groups the program is run on, the more possible genus-identifying conditionals the program accumulates.

4.2.11 Final Tabulated Output: The final output of the program is a file containing each successful name appended onto each ERV data line. Since the catarrhine dataset used here resulted in all ERVs being named, there is a name for every new line. In the event that some genera are unassigned, or some ERVs are unnamed, the ERV name column of this table will be blank for that particular ERV.

4.3 Discussion

4.3.1 General Discussion: While the many classification and nomenclature conventions discussed in the chapter introduction (section 4.1) have a valid basis, many of these are in need of revision. Since the advent of bioinformatic analyses and genome sequencing, retrovirologists have been able to conduct vast analyses on ERVs. A growing plethora of publications and interest in the field reinforces the requirement for a revised nomenclature system, where all ERV names relay the same, categorised information about an ERV, at a glance, while simultaneously being unambiguous, eliminating confusion when referring to a particular element. The primary objective of the work presented in this chapter is to present an integral program as part of an automated pipeline to systematically identify and name ERVs in any given genome. The program ensures that all inputted ERVs are unambiguously named with a permanent, unique tri-partite identifier. Given that the number of genomic records for ERV hosts is continually increasing, identifying and naming all ERVs presents a considerable challenge; the total number of elements currently identified is vast and continually growing. This reiterates the pressing requirement for streamlined, systematic methods to reliably identify and correctly name ERV insertions. This is the main objective of Gifford *et al*.'s (unpublished) pipeline, of which ERV Name Assigner will be an integral part.

In conjunction with the pipeline developed by Gifford *et al*. (unpublished), the ERV Name Assigner (Figure 4.5) presented in this chapter provides a powerful, efficient way to name hundreds of thousands of loci in a short space of time. This program will form part of the final nomenclature pipeline, which can then be run on any genome to mine it of ERV loci, subsequently assigning them unique names. It is also proposed that an online database for all ERVs named using the system, is maintained and made freely available in the future.

Of the six ERV genera, only three are currently assigned within ERV Name Assigner, since Gifford *et al*.'s (unpublished) pipeline is still under development, and has currently only identified ERVs belonging to

Betaretrovirus, Gammaretrovirus and *Spumaretrovirus* genera. Upon expanding the species reach, it is expected that ERV Name Assigner will need to be run twice with each new host range provided. The first run is not expected to name all the ERVs, since their genera need to be assigned within the program, first. This is the function of the conditionals that output the 'Unknown_Group.txt' file, which can be processed with the script in appendix J to identify all the unassigned genera. This list can then be assigned a genus from the literature, and reincorporated into the program which can then be run a second time, consequently naming all the ERVs in that dataset. With every new host species inputted into ERV Name Assigner, it is expected that all retroviral groups for which known genera are documented in the literature, will be assigned to every ERV locus inputted into the program. Once the Gifford *et al*. (unpublished) pipeline has been run on more genomes, the ERV Name Assigner program will be able to handle all ERVs for which there are known genera. Any that are left unassigned subsequent to this, have yet to be classified into the correct genus.

4.3.2 Future Developments and Considerations: One important facet of this nomenclature system that is still under development is the assigning of ERV orthologues. While the final draft of the nomenclature system has yet to be agreed at the time of writing, it is likely that the described ERV pipeline will comprise additional scripts to determine the orthology of ERV insertions in a given dataset. Currently, the Gifford *et al*. (unpublished) pipeline outputs the start and end subranges of each identified ERV. A provisional script that is currently being designed to extract a specified amount of sequence flanking an ERV insertion, is provided in appendix K. Currently, this script works on the results of a BLAST search against a genome, and has not yet been directly tested on the Gifford *et al*. (unpublished) pipeline output, but this is an emerging development. The extracted flanking sequences are then outputted into a FASTA format file, and can be compared between species in order to resolve orthologous relationships.

I propose that, as outlined in chapter 3, fuzzy-matching (Hietaniemi, 2001-5) can be used to compare flanking sequences, customising the percentage

error rate using a user-controlled modifier, as would be appropriate for the evolutionary relatedness of the species pair being investigated. Considering there would be potentially hundreds of thousands of flanking sequences to compare, the computational advantage provided by using fuzzy-matching could prove advantageous in this application. While not complete, a further suggestion for this nomenclature system, would be to add an additional component to the tri-partite naming system. For elements that identify as orthologues, it is suggested that an orthologue group identifier be appended onto the existing ERV name (Tristem & Lee, unpublished). For instance, if a group of five primates were processed by the nomenclature pipeline, an identified ERV orthologue across all five elements would have an orthologue group assigned using an –o group identifier (i.e. –o1). A database of all orthologue groups could then be constructed and maintained, that contains a repository of all the host species belonging to each –o group, together with details of the flanking sequence used to verify this. The main advantage of appending an –o group to the end of the existing name, is that it does not interrupt the existing cataloguing system. Furthermore, if an ERV is assigned its unique tri-partite name, appending an –o group at a later date does not invalidate the original name. So, should any studies be published on an ERV, and that same ERV is later found to be orthologous in a species whose genome may not have been available at the time, appending an –o group identifier will serve to update the existing name, rather than replacing it.

As well as comparing flanking sequences, it may also be advantageous to directly compare the genomic sequence of each ERV element, against other identified ERVs in the same host genome, in order to ascertain paralogy. For retroviral groups that are known to be active, such as KoRV (Tarlinton *et al*., 2005), or implicated in activity (such as HERV-K(HML-2)) (Turner *et al*., 2001), it may be a valid suggestion to add a –p group identifier to the end of each tri-partite name. This could facilitate a method for tracking the copy number of an ERV lineage over time. As with the suggested orthology system, a –p group would require a frequently updated online database, where paralogous elements stemming from an original infection event, could be loosely monitored. This database could include additional information such as

genome build and assembly, and most importantly, timestamps for every entry. Thus, developments in the numbers of elements entering a –p group over time could potentially be helpful in identifying whether a group of paralogous ERVs is amplifying, as well as potentially relaying information about ERV transmission dynamics. As with the –o group identifier, these would be appended onto the original tri-partite name, so that updates can be made in the future without warranting renaming of a particular ERV element. Once again, the homology of ERV genomic sequences can be detected using fuzzy-matching (see script in appendix L), inputting a set of flanks from the same host, rather than two separate species.

Another suggestion for this system would be that solo-LTRs are detected and named accordingly. With the ERV Name Assigner program, I have intentionally parsed every column of the Gifford *et al*. (unpublished) dataset. The 'Genome Structure' column could be particularly important here, since this contains information regarding the nature of the insertion, whether it's a full-length element, partial internal region, solo-LTR, or otherwise. Since the estimated total of solo-LTRs outnumber full-length proviruses tenfold, it is probably an important suggestion to incorporate this into the ERV name, as well. This could be in the form of a simple appended one-letter code, such as –s. As it would be expected that some genomes contain solo-LTRs in the place of orthologous full-length proviruses in others, appending a single-letter identifier in this fashion would allow easy monitoring of recombination events within any one –o group. As with all of the suggestions presented here, the nomenclature system is still under development, and is yet to be finalised.

For the first time in this field, one of the most biologically important retroelement groups is under preparation for classification and naming under a novel, unified system, providing a rich catalogue of information, where each individual ERV is assigned a unique, informative name. It is quite possible that a successful implementation of this new nomenclature and classification system into endogenous retroviruses could lead to a new standard that could be used for naming other SGEs, as well; potentially providing an important update to the global genomic record.

CHAPTER 5: General Discussion and Future Developments

5.1 General Discussion and Future Developments

The study of endogenous retroviruses continues to develop, and more is being uncovered regarding the nature and history of these once little understood retroelements. The advent of the bioinformatic age, where global libraries of genomic data are becoming increasingly informative, has greatly facilitated research in this field, and will doubtless continue to do so. While advances in sequencing technologies are pivotal to this progression, the potential for ERVs as human pathogens or mediators for disease has spurred research into identifying whether germline ERV insertions are really defective. Moreover, ancient ancestral ERV infections may also provide valuable insight into the elusive origns of retroviruses and retrotransposons.

The results presented in chapters 2, 3 and 4 of this thesis all have the common thread of identifying ERVs, using bioinformatic analyses and providing suggestions for improved *in silico* workflows for tackling bioinformatic challenges. The research presented in chapter 2 (and Lee *et al*., 2013) provides the oldest definitively dated ERV, MIR3 and MER3 elements described, to date. Previous to this study, few orthologies have been used to date ancient ERV insertions. In this study, orthologies were reconstructed to assign minimum ages to the identified SGEs, as this gives a definitive, accurate upper age estimate that can be inferred from published phylogenies (Gilbert & Feschotte, 2010). Other endogenised viral elements, however, have been dated using orthologies; amongst the most ancient of these is an orthologue belonging to the *Bornaviridae*, proven to date back to at least 93 million years (Katzourakis & Gifford, 2010). This is still considerably more recent than the minimum age of 104 million years given for the ancient ERV-L presented in this thesis. Unlike other SGEs, ERVs and LTR-retrotransposons are flanked by LTRs, which are identical at the point of integration. This makes it possible to extrapolate the minimum age further, through estimating sequence divergence between the element's two LTRs over time (Dangel *et al*., 1995). In the case of the ERV-L element reported in chapter 2, this gave an age range from 128-140 million years old (Lee *et al*., 2013).

Furthermore, since it is expected that host genomes spanning the placental mammal superorders are highly divergent (Murphy *et al*., 2001a), it is often harder to identify orthologous elements when using just one bioinformatic method. Accordingly, in this study (chapter 2), it can be concluded that aligning ancient, divergent sequences should be approached using a synergistic combination of global and local alignment algorthims, together with manual alignment, in order to retrieve maximal lengths of homologous sequence. For example, in sequences which, when correctly aligned, share 65% homology, having a 1 Kb insertion in one could throw a local alignment algorithm like BLAST completely off track. Algorithms such as MUSCLE and ClustalO, however, are very efficient at resuming an alignment despite interruption from large indels. BLAST, however, is more useful for searching genomic records, whereas MUSCLE is better suited to aligning much smaller regions of DNA. Thus using BLAST with truncated probes to locate the rough locus of an element (starting with regions known to be more conserved in ERVs), and then extracting sequence around this match for alignment in MUSCLE, can be one way to use the two alignment algorithms to their best potential. Applying these techniques has aided reconstruction of the orthology presented in chapter 2 – particularly in the species with faster generation times, where the ERV-L homology was harder to resolve. Through implementing the outlined methodology, it may be possible to increase the lookback time for ERVs and other SGEs, with potential broader applications for the field of paleogenomics, in general.

In contrast to the ancient ERV-L reported in chapter 2, the results presented in chapter 3 describe a set of 6 (Lee *et al*., 2014) and a second set of 5 novel, insertionally polymorphic HERV-K(HML-2) loci (Lee *et al*., unpublished) in the high-coverage genomes of Denisova and Neanderthal man (Meyer *et al*., 2012; Prüfer *et al*., 2014). This study investigates a retroviral lineage representing the most recently active HERVs, which could indeed still be active today (Turner *et al*., 2001). These 11 novel HERVs have not been described in the literature previous to this study. An earlier screen of low-coverage Denisovan and Neanderthal genomes led to the publication of 14 HERV loci (Agoni *et al*., 2012), with the authors subsequently concluding that

absence from the human genome implied novel, archaic hominid-specific HERVs. However, in 2013, Marchi *et al*. proved that nearly all of these viruses could be recovered from the genomes of modern humans, and that they were all insertionally polymorphic. This further emphasises the importance of screening outside of the human reference sequence, before reaching conclusions; hence the comparison of the 11 novel loci identified in chapter 3 against those recovered from published screens of high-coverage modern human genomes (Lee *et al*., 2012; Marchi *et al*., 2014). While these 11 HERV loci were far more rigorously verified, it still remains possible that these elements exist at low allele frequencies within some extant human individuals.

Both archaic hominid genomes (Meyer *et al*., 2012; Prüfer *et al*., 2014) provided limited sequence data within individual reads, making assembly of unmapped regions – such as those containing insertionally polymorphic HERVs – very difficult to accomplish with any degree of accuracy. Thus, a methodical workflow was implemented, that was aimed at maximising the potential of a single read. This meant that sufficient flanking sequence was obtained to successfully map the preintegration sites onto the human genome, while simultaneously verifying reads with BLAST searches, back-filtering and corroboration *via* multiple, unique reads. Checking these against elements obtained from screens of a further 401 high-coverage modern human genomes (Lee *et al*., 2012; Marchi *et al*., 2014), makes this study one of the most extensive presentations of insertionally polymorphic HERV-K(HML-2) elements, to date.

Advances in genome sequencing technology are currently overtaking the technology required to compute and handle the voluminous data produced. Thus, in many bioinformatic applications, functionality and efficiency are of equal importance. In contrast to the archaic hominid HERV screen presented in chapter 3, the methodology involved in chapter 2 was far less computationaly demanding. The most intensive processes involved were sequence alignment with MUSCLE and phylogenetic analyses – these can be executed adequately on a well-specified personal computer. The archaic hominid screen, however, would have been unfeasible to attempt on a

personal computer; the genome sequence data for Denisova and Neanderthal alone comprised over 1.2 TB of plain text data. Using conventional algorithms – such as BLAST – to screen for HERVs was not an option on such a machine. In fact, BLAST searching genomic datasets of this scale was even challenging on a high-performance computer cluster, pressing for an alternative method for HERV screening. The perl scripts discussed in this chapter – both stringent- and fuzzy-matching – are far more efficient at recovering HERV loci than more sensitive algorithms, such as BLAST. Since HERV-K113 LTRs are highly homologous, the sensitivity advantages provided by BLAST were less relevant to this screen. Thus, a less selective method – such as the aforementioned perl scripts – could be employed, in favour of screening the whole genome efficiently. These perl scripts could even be optimised to make screens run on high-specification personal computers, in some cases; although subsequent verification steps using BLAST would not be a feasible option, given the technology available at the time of writing.

While the stringent-matching perl scripts recovered tens of thousands of HERV-containing reads, it was observed that fuzzy-matching recovered a greater number still. Writing perl code in order to match patterns – such as a fragment of the HERV-K113 LTR – does not typically allow for a straightforward solution to accommodate for mismatching biological sequence data. This was always the main limitation of HERV screens using regex pattern matching in perl. However, incorporating the fuzzy-matching String::Approx (Hietaniemi, 2001-2005) CPAN module provided great advantage, in this regard. In fact, its potential in *in silico* biological applications is limitless, with regards to screening large datasets.

Although many components of the archaic hominid screen were automated (namely the entire HERV-mining procedure), much of the verification of candidate HERVs was done manually, by eyeballing matches and sequences for visible patterns. Since the current workflow – from HERV mining to rigorous *in silico* locus verification – has been streamlined to its effective potential, it is important to consider the future development of a user-controlled, automated pipeline for effecting candidate novel HERV verification.

This is expected to be a complex task. Manual verification was largely unavoidable, due to the sheer number of possibilities of sequence matches that could occur within a read. For example, when verifying by BLAST searching candidate reads against the human reference sequence, it was found that some matches comprised more LTR than flank. Others comprised more flank, and some comprised equal amounts of both flank and LTR. Template switches can occur at any point within the read. Sometimes this occurs at the breakpoint and, at other times, in the flank or within the LTR sequence. Consequently, setting program thresholds is difficult, since an unknown, erroneous sequence can interrupt the read at any point. These template switches are not easy to identify, and thus automated programs may struggle to accurately detect their positions.

Conversely, careful manual observation can be used to identify template switches, which can subsequently be confirmed or refuted by a number of verification steps. For instance, candidate flanking sequences are manually selected and BLAT searched against the human genome, to see where the template switch occurs within the read. Several different BLAT results for each read can then be analysed manually to help infer the nature of a potential template switch. For example, the top BLAT match may match one half of the flank perfectly, and not match the second half at all. At first glance, this could appear to be a template switch within the archaic hominid sequence read. However, more careful observation may reveal that the same flank completely matches another locus in the human genome, albeit with lower homology. It could be that this inferior, full-length match reflects the locus represented in the archaic hominid read, and that sequence divergence between archaic and modern man could be one of many reasons behind the observed loss of homology between the two. This is one of many potential scenarios that can make automation diffiicult when verifying candidate novel HERVs, and would require complex programming to achieve. Another strategy for future consideration would be to construct libraries of all known HERVs present within the human reference sequence, and subtract these from the output of filtered fuzzy-matches. This could provide a way to reduce

the number of HERVs that would subsequently require verification, helping to streamline the process further.

Despite the many difficulties of completely automating HERV verification, it is suggested that a future pipeline of perl scripts – with adjustable parameters and thresholds – be developed to assist in the verification of mined HERVs as insertionally polymorphic, relative to the human reference sequence. Automation of the entire workflow into a software program – called PolyMorph – with a user interface is currently under discussion (Lee, Huntley & Tristem, unpublished), and would allow for a powerful pipeline for identifying insertionally polymorphic elements in the fastest manner possible. This could have important applications in monitoring active ERV families, and could even potentially be extended to LTR-retrotransposons, in the future.

As with the mining stages of the archaic hominid screens, the work presented in chapter 4 also involves perl-based automation. The ERV Name Assigner program was developed to assign ERV insertions unique, tri-partitle identifiers as part of a pipeline that will be used to rename all existing ERVs. This is a fundamental part of a worldwide initiative involving the most esteemed retrovirolgoists, to change the way ERVs are discussed and named, hereforth. As was previously discussed in this chapter, the sample dataset comprised the catarrhine primate ERV loci, totalling 156,614. This puts into perspective the size of the task at hand, since the ultimate goal is to screen and name all the vertebrate genomes sequenced, to date. It is expected that the total loci generated will number in the millions, stressing the requirement of an efficient computer program, when handling such data. The ERV Name Assigner program presented in this thesis has successfully named all the currently extracted 156,614 catarrhine loci, and will be developed to incorporate additional information into each ERV identifier, which could prove invaluable to researchers. At a glance, an ERV name under the newly proposed nomenclature system could reveal information about orthology, paralogy, ERV genome structure and host species (Tristem & Lee, unpublished). Perhaps through the mining and characterisation of ERVs and their relationships relative to one another, it may be possible to observe patterns that could

reveal the population dynamics and potential activity of ERV lineages. Once this system has been successfully implemented, it may even be possible to devise a similar system for naming other classes of SGE.

Overall, the findings presented in this thesis highlight the importance of using different bioinformatic techniques synergistically in order to achieve a particular research objective. From reporting a retrovirus that is approximately 128-140 million years old, through to identifying 11 novel, insertionally polymorphic loci belonging to the most recently active HERV lineage, as well as providing a systematic solution to implementing a revised ERV nomenclature system, the rapid advances in sequencing and computing technology spell a bright future for the study of endogenous retroviruses. Further research carried out in the same vein as presented here, could bring virologists closer to understanding not only the origin of retroviruses, but also their evolution, transmission dynamics, host ranges and – perhaps most importantly – their potential to cause disease.

References

1. Abbink, T. E., & Berkhout, B. (2008). RNA structure modulates splicing efficiency at the human immunodeficiency virus type 1 major splice donor. *Journal of virology*, *82*(6), 3090-3098.
2. Agoni, L., Golden, A., Guha, C., & Lenz, J. (2012). Neandertal and Denisovan retroviruses. *Current Biology*, *22*(11), R437-R438.
3. Aiewsakun, P. & Brockman, A. (2011) 'Identification of Novel Human Endogenous Retroviruses in an Extinct Paleo-Eskimo Genome and Khoisan Genomes and their Implications', final year thesis, Imperial College London.
4. Altschul, S. F., Gish, W., Miller, W., Myers, E. W., & Lipman, D. J. (1990). Basic local alignment search tool. *Journal of molecular biology*, *215*(3), 403-410.
5. Andersson, M. L., Lindeskog, M., Medstrand, P., Westley, B., May, F., & Blomberg, J. (1999). Diversity of human endogenous retrovirus class II-like sequences. *Journal of general virology*, *80*(1), 255-260.
6. Barbulescu, M., Turner, G., and Su, M., Kim, R., Jensen- Seaman, M. L., Deinhard, A. S., Kidd, K. K., Lenz, J., (2001). A HERV-K provirus in chimpanzees, bonobos and gorillas, but not humans. *Curr. Biol.* 11: 779-783.
7. Barbulescu, M., Turner, G., Seaman, M. I., Deinard, A. S., Kidd, K. K., & Lenz, J. (1999). Many human endogenous retrovirus K (HERV-K) proviruses are unique to humans. *Current Biology*, *9*(16), 861-S1.
8. Belshaw, R., Dawson, A. L. A., Woolven-Allen, J., Redding, J., Burt, A. & Tristem, M. (2005) Genomewide Screening Reveals High Levels of Insertional Polymorphism in the Human Endogenous Retrovirus Family HERV-K(HML-2): Implications for Present-Day Activity. *Journal of Virology* 79, (19), 12507-12514.
9. Belshaw, R., Pereira, V., Katzourakis, A., Talbot, G. Paces, J., Burt, A. & Tristem, M. 2004 Long-term reinfection of the human genome by endogenous retroviruses. PNAS 101, (14): 4894-4899.
10. Belyi, V. A., Levine, A. J. & Skalka, A. M. (2010) Unexpected Inheritance: Multiple Integrations of Ancient Bornavirus and Ebolavirus/Marburgvirus Sequences in Vertebrate Genomes. *PLoS Pathogens*. 6 (7): e1001030.
11. Benachenhou, F., Jern, P., Oja, M., Sperber, G., Blikstad, V., Somervuo, P., ... & Blomberg, J. (2009). Evolutionary conservation of orthoretroviral long terminal repeats (LTRs) and ab initio detection of single LTRs in genomic data. *PloS one*, *4*(4), e5179.
12. Bénit, L., Lallemand, J. B., Casella, J. F., Philippe, H., & Heidmann, T. (1999). ERV-L elements: a family of endogenous retrovirus-like elements active throughout the evolution of mammals. *Journal of virology*, *73*(4), 3301-3308.
13. Bénit, L., Lallemand, J. B., Casella, J. F., Philippe, H., & Heidmann, T. (1999). ERV-L elements: a family of endogenous retrovirus-like elements active throughout the evolution of mammals. *Journal of virology*, *73*(4), 3301-3308.
14. Bieda, K., Hoffmann, A. & Boller, K. (2001) Phenotypic heterogeneity of human endogenous retrovirus particles produced by teratocarcinoma cell-lines. *J. Gen. Virol.* 82, 591-596.
15. Bininda-Emonds, O. R. P., Cardillo, M., Jones, K. E., MacPhee, R. D. E., Beck, R. M. D., Grenyer, R., Price, S. A., Vos., R. A., Gittleman, J. L. & Purvis, A. 2007 The delayed rise of present-day mammals. Nature 446, (29): 507-51.
16. Bjerregaard, B., Holck, S., Christensen, I. J. & Larsson, L. I. (2006) Syncytin is involved in breast cancer-endothelial cell fusions. *Cell Mol. Life Sci.*, 63, 1906-1911.
17. Blaise, S., De Parseval, N., & Heidmann, T. (2005). Functional characterization of two newly identified Human Endogenous Retrovirus coding envelope genes. *Retrovirology*, *2*(1), 19.
18. Blomberg, J., Benachenhou, F., Blikstad, V., Sperber, G., & Mayer, J. (2009). Classification and nomenclature of endogenous retroviral sequences (ERVs): problems and recommendations. *Gene*, *448*(2), 115-123.
19. Blomberg, J., Coffin, J., Fan, H., Gifford, R., Heidmann, T., Johnson, W., Mayer, J., Stoye, J. & Tristem, M. (unbpublished) Proposal to Revise the Nomenclature for Endogenous Retrovirus Loci.
20. Boeke, J. D. & Stoye, J. P. (1997). Retrotransposons, Endogenous Retroviruses, and the Evolution of Retroelements. In: Coffin, J., Hughes, S. & Varmus, H. E. (1997) *Retroviruses*. pp.343-435. New York, Cold Spring Harbor Laboratory Press.
21. Boese, A., Sauter, M., Galli, U., Best, B., Herbst, H., Mayer, J., Kremmer, E., Roemer, K. & Mueller-Lantzsch N. (2000) Human endogenous retrovirus protein cORF supports cell transformation and associates with the promyelocytic leukemia zinc finger protein. *Oncogene*. 19, 4328-4336.

22. Boese, A., Sauter, M., Galli, U., Best, B., Herbst, H., Mayer, J., ... & Mueller-Lantzsch, N. (2000). Human endogenous retrovirus protein cORF supports cell transformation and associates with the promyelocytic leukemia zinc finger protein. *Oncogene*, *19*(38), 4328-4336.
23. Boller, K., Janssen, O., Schuldes, H., Tonjes, R. R. & Kurth, R. (1997) Characterization of the antibody response specific for the human endogenous retrovirus HTDV/HERV-K. *J. Virol.* 4581-4588.
24. Boller, K., Schönfeld, K., Lischer, S., Fischer, N., Hoffmann, A., Kurth, R. & Tönjes, R. R. (2008) Human endogenous retrovirus HERV-K113 is capable of producing intact viral particles. *J. Gen. Virol.* 89 (2) 567-572.
25. Buzdin, A., Ustyugova, S., Khodosevich, K., Mamedov, I., Lebedev, Y., Hunsmann, G., & Sverdlov, E. (2003). Human-specific subfamilies of HERV-K (HML-2) long terminal repeats: three master genes were active simultaneously during branching of hominoid lineages. *Genomics*, *81*(2), 149-156.
26. Chakravarti, D., & Mailander, P. C. (2008). Formation of Template-Switching Artifacts by Linear Amplification. *Journal of biomolecular techniques: JBT*, *19*(3), 184.
27. Coffin, J., Hughes, S. & Varmus H. E. (1997). *Retroviruses*. New York, Cold Spring Harbor Laboratory Press.
28. Cohen, C. J., Lock, W. M., & Mager, D. L. (2009). Endogenous retroviral LTRs as promoters for human genes: a critical assessment. *Gene*, *448*(2), 105-114.
29. Cohen, J. C., & Varmus, H. E. (1979). Endogenous mammary tumour virus DNA varies among wild mice and segregates during inbreeding. *Nature*. 278, 418-423.
30. Cohen, M., & Larsson, E. (1988). Human endogenous retroviruses. *Bioessays*, *9*(6), 191-196.
31. Cook, J. M. & Tristem, M. (1997) 'SINEs of the times' -transposable elements as clade markers for their hosts. *Tree*. 12, (8), 295-297.
32. Cordonnier, A., Casella, J. F., & Heidmann, T. (1995). Isolation of novel human endogenous retrovirus-like elements with foamy virus-related pol sequence. *Journal of virology*, *69*(9), 5890-5897.
33. Dangel, A. W., Baker, B. J., Mendoza, A. R. & Yu, C. Y. (1995) Complement component C4 gene intron 9 as a phylogenetic marker for primates: long terminal repeats of the endogenous retrovirus ERV-K(C4) are a molecular clock of evolution. *Immunogenetics*. 42 (1): 41-52.
34. Delsuc, F., Scally, M., Madsen, O., Stanhope, M. J., De Jong, W. W., Catzeflis, F. M., ... & Douzery, E. J. (2002). Molecular phylogeny of living xenarthrans and the impact of character and taxon sampling on the placental tree rooting. *Molecular Biology and Evolution*, *19*(10), 1656-1671.
35. Dewannieux, M., Harper, F., Richaud, A., Letzelter, C., Ribet, D., Pierron, G., & Heidmann, T. (2006). Identification of an infectious progenitor for the multiple-copy HERV-K human endogenous retroelements. *Genome research*, *16*(12), 1548-1556.
36. Doolittle, R. F., Feng, D. F., Johnson, M. S., & McClure, M. A. (1989). Origins and evolutionary relationships of retroviruses. *Quarterly Review of Biology*, 1-30.
37. Dressman, D., Yan, H., Traverso, G., Kinzler, K. W., & Vogelstein, B. (2003). Transforming single DNA molecules into fluorescent magnetic particles for detection and enumeration of genetic variations. *Proceedings of the National Academy of Sciences*, *100*(15), 8817-8822.
38. Duffy, S., Shackelton, L. A. & Holmes, E. C. (2008) Rates of evolutionary change in viruses: paterns and determinants. *Nature Reviews Genetics*. 9: 267-276.
39. Dunn, B. M., Goodenow, M. M., Gustchina, A., & Wlodawer, A. (2002). Retroviral proteases. *Genome Biol*, *3*(4), 3006-1.
40. Edgar, R. C. 2004 MUSCLE: multiple sequence alignment with high accuracy and high throughput. Nucleic Acids Research 32, (5): 1792-97.
41. Eid, J., Fehr, A., Gray, J., Luong, K., Lyle, J., Otto, G., ... & Zhao, P. (2009). Real-time DNA sequencing from single polymerase molecules. *Science*, *323*(5910), 133-138.
42. Eizirik, E., Murphy, W. J., & O'brien, S. J. (2001). Molecular dating and biogeography of the early placental mammal radiation. *Journal of Heredity*, *92*(2), 212-219.
43. Elder, J. H., Lerner, D. L., Hasselkus-Light, C. S., Fontenot, D. J., Hunter, E., Luciw, P. A., ... & Phillips, T. R. (1992). Distinct subsets of retroviruses encode dUTPase. *Journal of virology*, *66*(3), 1791-1794.
44. Esnault, C., Priet, S., Ribet, D., Vernochet, C., Bruls, T., Lavialle, C., ... & Heidmann, T. (2008). A placenta-specific receptor for the fusogenic, endogenous retrovirus-derived, human syncytin-2. *Proceedings of the National Academy of Sciences*, *105*(45), 17532-17537.
45. Fan, J. B., Chee, M. S., & Gunderson, K. L. (2006). Highly parallel genomic assays. *Nature Reviews Genetics*, *7*(8), 632-644.

46. Fedurco, M., Romieu, A., Williams, S., Lawrence, I., & Turcatti, G. (2006). BTA, a novel reagent for DNA attachment on glass and efficient generation of solid-phase amplified DNA colonies. *Nucleic acids research*, *34*(3), e22-e22.
47. Feschotte, C. & Gilbert, C. (2012) Endogenous viruses: insights into viral evolution and impact on host biology. *Nature Reviews Genetics*. 13, 283-296.
48. Flockerzi, A., Ruggieri, A., Frank, O., Sauter, M., Maldener, E., Kopper, B., ... & Mayer, J. (2008). Expression patterns of transcribed human endogenous retrovirus HERV-K (HML-2) loci in human tissues and the need for a HERV Transcriptome Project. *BMC genomics*, *9*(1), 354.
49. Gifford, R. & Tristem, M. (2003). The Evolution, Distribution and Diversity of Endogenous Retroviruses. *Virus Genes*. 26: (3) 291-315.
50. Gifford, R. J. *et al*., (unpublished) Computer pipeline to automate and effect a revised nomenclature standard for endogenous retroviruses.
51. Gilbert, C. & Feschotte, C. (2010) Genomic fossils calibrate the long-term evolution of hepadnaviruses. *PLoS Biol.* 8, e1000495.
52. Gimenez, J., Montgiraud, C., Pichon, J. P., Bonnaud, B., Arsac, M., Ruel, K., ... & Mallet, F. (2010). Custom human endogenous retroviruses dedicated microarray identifies self-induced HERV-W family elements reactivated in testicular cancer upon methylation control. *Nucleic acids research*, *38*(7), 2229-2246.
53. Gogvadze, E., & Buzdin, A. (2009). Retroelements and their impact on genome evolution and functioning. *Cellular and molecular life sciences*, *66*(23), 3727-3742.
54. Goodier, J. L., & Kazazian Jr, H. H. (2008). Retrotransposons revisited: the restraint and rehabilitation of parasites. *Cell*, *135*(1), 23-35.
55. Goodman, M., Sterner, K. N., Islam, M., Uddin, M., Sherwood, C. C., Hof, P. R., ... & Wildman, D. E. (2009). Phylogenomic analyses reveal convergent patterns of adaptive evolution in elephant and human ancestries. *Proceedings of the National Academy of Sciences*, *106*(49), 20824-20829.
56. Green, R. E., Krause, J., Briggs, A. W., Maricic, T., Stenzel, U., Kircher, M., ... & Mullikin, J. C. (2010). A draft sequence of the Neandertal genome. *science*, *328*(5979), 710-722.
57. Griffiths, D. J. (2001) Endogenous retroviruses in the human genome sequence. *Genome Biology*. 2 (6) 1017.1-1017.5.
58. Gude, N. M., Roberts, C. T., Kalionis, B., & King, R. G. (2004). Growth and function of the normal human placenta. *Thrombosis research*, *114*(5), 397-407.
59. Gudleski, N., Flanagan, J. M., Ryan, E. P., Bewley, M. C., & Parent, L. J. (2010). Directionality of nucleocytoplasmic transport of the retroviral gag protein depends on sequential binding of karyopherins and viral RNA. *Proceedings of the National Academy of Sciences*, *107*(20), 9358-9363.
60. Harris, L. K. (2010). Review: trophoblast-vascular cell interactions in early pregnancy: how to remodel a vessel. *Placenta*, *31*, S93-S98.
61. Harris, T. D., Buzby, P. R., Babcock, H., Beer, E., Bowers, J., Braslavsky, I., ... & Xie, Z. (2008). Single-molecule DNA sequencing of a viral genome. *Science*, *320*(5872), 106-109.
62. Havecker, E. R., Gao, X., & Voytas, D. F. (2004). The diversity of LTR retrotransposons. *GENOME BIOLOGY.*, *5*(6), 225-225.
63. Hietaniemi, J. (2001-2005) String::Approx CPAN Module. Available online: http://search.cpan.org/~jhi/String-Approx/Approx.pm
64. Hughes, J. F., & Coffin, J. M. (2004). Human endogenous retrovirus K solo-LTR formation and insertional polymorphisms: implications for human and viral evolution. *Proceedings of the National Academy of Sciences of the United States of America*, *101*(6), 1668-1672.
65. Hunter, E. (2008) Retroviruses: General Features. In: Mahy, B. W. J. & van Regenmortel, M. H. V. (eds.) *Encyclopedia of Virology (Third Edition).* Academic Press, Oxford, pp. 459-467.
66. Jern, P., Sperber, G. O. & Blomberg, J. (2005a). Use of Endogenous Retroviral Sequences (ERVs) and structural markers for retroviral phylogenetic inference and taxonomy. *Retrovirology*. 2: (50) 1-12.
67. Jern, P., Sperber, G. O., Ahlsén, G., & Blomberg, J. (2005b). Sequence variability, gene structure, and expression of full-length human endogenous retrovirus H. *Journal of virology*, *79*(10), 6325-6337.
68. Johnson, W. E. (2008) Endogenous Retroviruses. *Elsevier* Ltd., USA.
69. Johnson, W. E., & Coffin, J. M. (1999). Constructing primate phylogenies from ancient retrovirus sequences. *Proceedings of the National Academy of Sciences*, *96*(18), 10254-10260.
70. Jones, K. E., Bininda-Emonds, O. R., & Gittleman, J. L. (2005). Bats, clocks, and rocks: diversification patterns in Chiroptera. *Evolution*, *59*(10), 2243-2255.

71. Jurka, J. 1990 Novel families of interspersed repetitive elements from the human genome. Nucl. Acids Res. 18, (1): 137-141.
72. Jurka, J., Zietkiewicz, E., & Labuda, D. (1995). Ubiquitous mammalian-wide interspersed repeats (MIRs) are molecular fossils from the mesozoic era. *Nucleic acids research*, *23*(1), 170-175.
73. Karlsson, H., Bachman, S., Schroder, J., McArthur, J., Torrey, E. G. & Yolken, R. H. (2001) Retroviral RNA identified in the cerebrospinal fluids and brains of individuals with schizophrenia. *Proc. Natl. Acad. Sci.* USA 98, 4634-4639.
74. Katz, R. A., & Skalka, A. M. (1990). Generation of diversity in retroviruses. *Annual review of genetics*, *24*(1), 409-443.
75. Katz, R. A., & Skalka, A. M. (1994). The retroviral enzymes. *Annual review of biochemistry*, *63*(1), 133-173.
76. Katzourakis, A. & Gifford, R. J. 2010 Endogenous Viral Elements in Animal Genomes. PloS Genetics 6, (11): e 1001191.
77. Katzourakis, A. Tristem, M., Pybus, O. G. & Gifford, R. J. 2007 Discovery and analysis of the first endogenous lentivirus. PNAS 104, (15): 6261-6265.
78. Keckesova, Z., Ylinen, L. M. J., Towers, G. J., Gifford, R. J., & Katzourakis, A. (2009). Identification of a RELIK orthologue in the European hare (*Lepus europaeus*) reveals a minimum age of 12 million years for the lagomorph lentiviruses. *Virology*, *384*(1), 7-11.
79. Kijima, T. E., & Innan, H. (2010). On the estimation of the insertion time of LTR retrotransposable elements. *Molecular biology and evolution*, *27*(4), 896-904.
80. Kim, J. B., Porreca, G. J., Song, L., Greenway, S. C., Gorham, J. M., Church, G. M., ... & Seidman, J. G. (2007). Polony multiplex analysis of gene expression (PMAGE) in mouse hypertrophic cardiomyopathy. *Science*, *316*(5830), 1481-1484.
81. Kimura, M. (1984). *The neutral theory of molecular evolution*. Cambridge University Press.
82. Kosakovsky Pond, S. L. & Frost, S. D. W. 2005 Not So Different After All: A Comparison of Methods for Detecting Amino Acid Sites Under Selection. Mol. Biol. Evol. 22, 1208-1222.
83. Labuda, D., & Striker, G. (1989). Sequence conservation in Alu evolution. *Nucleic acids research*, *17*(7), 2477-2491.
84. Laird, C. D., McCONAUGHY, B. L., & McCARTHY, B. J. (1969). Rate of fixation of nucleotide substitutions in evolution.
85. Lander, E. S., Linton, L. M., Birren, B., Nusbaum, C., Zody, M. C., Baldwin, J., Devon, K., Dewar, K., Doyle, M., FitzHugh, W. et al (2001). Initial sequencing and analysis of the human genome. *Nature*. 409, 860-921.
86. Lavialle, C., Cornelis, G., Dupressoir, A., Esnault, C., Heidmann, O., Vernochet, C., & Heidmann, T. (2013). Paleovirology of 'syncytins', retroviral env genes exapted for a role in placentation. *Philosophical Transactions of the Royal Society B: Biological Sciences*, *368*(1626), 20120507.
87. Lavie, L., Medstrand, P., Schempp, W., Meese, E., & Mayer, J. (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), 8788-8798.
88. Lebedev, Y. B., Belonovitch, O. S., Zybrova, N. V., Khil, P. P., Kurdyukov, S. G., Vinogradova, T. V., ... & Sverdlov, E. D. (2000). Differences in HERV-K LTR insertions in orthologous loci of humans and great apes. *Gene*, *247*(1), 265-277.
89. Lee, A., Huntley, D. & Tristem, M. (unpublished) PolyMorph: a computer program for the detection of insertionally polymorphic endogenous retroviruses.
90. Lee, A., Huntley, D., Aiewsakun, P., Kanda, R. K., Lynn, C., & Tristem, M. (2014). Novel Denisovan and Neanderthal Retroviruses. *Journal of virology*, *88*(21), 12907-12909.
91. Lee, A., Huntley, D., Aiewsakun, P., Kanda, R., Lynn, C. & Tristem, M. (unpublished). A fuzzy-matching HERV-mining program uncovers novel, insertionally polymorphic HERV-K(HML-2) retroviruses in the genomes of two archaic hominids.
92. Lee, A., Nolan, A., Watson, J., & Tristem, M. (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), 20120503.
93. Lee, E., Iskow, R., Yang, L., Gokcumen, O., Haseley, P., Luquette, L. J., ... & Park, P. J. (2012). Landscape of somatic retrotransposition in human cancers. *Science*, *337*(6097), 967-971.
94. Lee, Y. N., & Bieniasz, P. D. (2007). Reconstitution of an infectious human endogenous retrovirus. *PLoS pathogens*, *3*(1), e10.
95. Llorens, C., Futami, R., Covelli, L., Domínguez-Escribá, L., Viu, J. M., Tamarit, D., ... & Moya, A. (2010). The Gypsy Database (GyDB) of mobile genetic elements: release 2.0. *Nucleic acids research*, gkq1061.
96. Löwer, R., Löwer, J., & Kurth, R. (1996). The viruses in all of us: characteristics and biological

significance of human endogenous retrovirus sequences. *Proceedings of the National Academy of Sciences*, *93*(11), 5177-5184.
97. Lyn, D., Deaven, L. L., Istock, N. L., & Smulson, M. (1993). The Polymorphic ADP-Ribosyltransferase (NAD+) Pseudogene 1 in Humans Interrupts an Endogenous pol-like Element on 13q34. *Genomics*, *18*(2), 206-211.
98. Lynch, C., & Tristem, M. (2003). A Co-opted *gypsy*-type LTR-Retrotransposon Is Conserved in the Genomes of Humans, Sheep, Mice, and Rats. *Current biology*, *13*(17), 1518-1523.
99. Macfarlane, C., & Simmonds, P. (2004). Allelic variation of HERV-K (HML-2) endogenous retroviral elements in human populations. *Journal of molecular evolution*, *59*(5), 642-656.
100. Mager, D. L., & Medstrand, P. (2005). Retroviral repeat sequences. *eLS*.
101. Mallet, F., Bouton, O., Prudhomme, S., Chenyet, V., Oriol, G., Bonnaud, B., Lucotte, G., Duret, L. & Mandrand, B. 2004. The endogenous retroviral locus ERVWE1 is a bona fide gene involved in hominoid placental physiology. Proc. Natl. Acad. Sci. USA, 101, 1731-1736.
102. Marchi, E., Kanapin, A., Byott, M., Magiorkinis, G., & Belshaw, R. (2013). Neanderthal and Denisovan retroviruses in modern humans. *Current Biology*, *23*(22), R994-R995.
103. Marchi, E., Kanapin, A., Magiorkinis, G., & Belshaw, R. (2014). Unfixed endogenous retroviral insertions in the human population. *Journal of virology*, *88*(17), 9529-9537.
104. Martins, H., & Villesen, P. (2011). Improved integration time estimation of endogenous retroviruses with phylogenetic data. *PloS one*, *6*(3), e14745.
105. Mayer, J., & Meese, E. U. (2003). Presence of dUTPase in the various human endogenous retrovirus K (HERV-K) families. *Journal of molecular evolution*, *57*(6), 642-649.
106. McCarthy, E. M., & McDonald, J. F. (2004). Long terminal repeat retrotransposons of Mus musculus. *Genome biology*, *5*(3), R14-R14.
107. McClintock, B. (1956, January). Controlling elements and the gene. In *Cold Spring Harbor Symposia on Quantitative Biology* (Vol. 21, pp. 197-216). Cold Spring Harbor Laboratory Press.
108. Medstrand, P. & Mager, D. L. (1998) Human-specific integrations of the HERV-K endogenous retrovirus family. *Journal of Virology* 1998, 72: 9782-9787.
109. Metzker, M. L. (2005). Emerging technologies in DNA sequencing. *Genome research*, *15*(12), 1767-1776.
110. Metzker, M. L. (2009). Sequencing technologies—the next generation. *Nature Reviews Genetics*, *11*(1), 31-46.
111. Meyer, M., Kircher, M., Gansauge, M. T., Li, H., Racimo, F., Mallick, S., ... & Pääbo, S. (2012). A high-coverage genome sequence from an archaic Denisovan individual. *Science*, *338*(6104), 222-226.
112. Mugal, C. F., Wolf, J. B., & Kaj, I. (2013). Why time matters: Codon evolution and the temporal dynamics of dN/dS. *Molecular biology and evolution*, mst192.
113. Murata, S., Takasaki, N., Saitoh, M., Tachida, H. & Okada, N. (1996) Details of retropositional genome dynamics that provide a rationale for a generic revision: the distinct branching of all the Pacific salmon and trout (Oncorhynchus) from the Atlantic salmon and trout (Salmo). *Genetics.* 142, 915-926.
114. Murphy, J. E., & Goff, S. P. (1992). A mutation at one end of Moloney murine leukemia virus DNA blocks cleavage of both ends by the viral integrase in vivo. *Journal of virology*, *66*(8), 5092-5095.
115. Murphy, W. J., Eizirik, E., O'Brien, S. J., Madsen, O., Scally, M., Douady, C. J., Teeling, E., Ryder, O. A., Stanhope, M. J., de Jong, W. W. & Springer, M. S. (2001a) Resolution of the Early Placental Mammal Radiation Using Bayesian Phylogenetics. *Science*. 294 (5550): 2348-2351.
116. Murphy, W. J., Elzirik, E., Johnson, W. E., Ping Zhang, Y. P., Ryder, O. A. & O'Brien, S. J. (2001b) Molecular phylogenetics and the origins of placental mammals. *Nature*. 409, 614-618.
117. Muster, T., Waltenberger, A., Grassauer, A., Hirschi, S., Caucig, P., Romirer, I., Födinger, D., Seppele, H., Schanab, O., Magin-Lachmann, C., Löwer, R., Jansen, B., Pehamberger, H. & Wolff, K. (2003) An endogenous retrovirus derived from human melanoma cells. *Cancer Res.* 63 (24): 8735-8741.
118. Naas, T. P., DeBerardinis, R. J., Moran, J. V., Ostertag, E. M., Kingsmore, S. F., Seldin, M. F., ... & Kazazian, H. H. (1998). An actively retrotransposing, novel subfamily of mouse L1 elements. *The EMBO journal*, *17*(2), 590-597.
119. Nelson, P. N., Carnegie, P. R., Martin, J., Ejtehadi, H. D., Hooley, P., Roden, D., Rowland-Jones, S., Warren, P., Astley, J. & Murray, P. G. (2003). Demystified... human endogenous retroviruses. *Molecular Pathology*, *56*(1), 11.
120. Ohta, T., & Gillespie, J. H. (1996). Development of neutral and nearly neutral theories. *Theoretical population biology*, *49*(2), 128-142.

121. Oroszlan, S., and R. B. Luftig. "Retroviral proteinases." *Retroviruses*. Springer Berlin Heidelberg, 1990. 153-185.
122. Paces, J., Pavlicek, A. & Paces, V. (2002) HERVd: database of human endogenous retroviruses. *Nucleic Acid Res.* 30, 205-206.
123. Parent, L. J. (2011). New insights into the nuclear localization of retroviral Gag proteins. *Nucleus*, *2*(2), 92.
124. Perelman, P., Johnson, W. E., Roos, C., Seuánez, H. N., Horvath, J. E., Moreira, M. A., ... & Pecon-Slattery, J. (2011). A molecular phylogeny of living primates. *PLoS genetics*, *7*(3), e1001342.
125. Perron, H., Garson, J. A., Bedin, F., Beseme, F., Paranhos-Baccala, G., Komurian-Pradel, F., Mallet, F., Tuke, P. W., Voisset, C., Blond, J. L., Lalande, B., Seigneurin, J. M. & Mandrand, B. (1997) Molecular identification of a novel retrovirus repeatedly isolated from patients with multiple sclerosis. *Proc. Natl. Acad. Sci.* USA 94, 7583-7588.
126. Pertsemlidis, A., Fondon, J. W., & John, W. (2001). Having a BLAST with bioinformatics (and avoiding BLASTphemy). *Genome Biol*, *2*(10), 1-10.
127. Prüfer, K., Racimo, F., Patterson, N., Jay, F., Sankararaman, S., Sawyer, S., ... & Pääbo, S. (2014). The complete genome sequence of a Neanderthal from the Altai Mountains. *Nature*, *505*(7481), 43-49.
128. Reich, D., Green, R. E., Kircher, M., Krause, J., Patterson, N., Durand, E. Y., Viola, B., Briggs, A. W., Stenzel, U., Johnson, P. L. F., Maricic, T., Good, J. M., Marques-Bonet, T., Alkan, C., Fu, Q., Mallick, S., Li, H., Meyer, M., Eichler, E. E., Stoneking, M., Richards, M., Talamo, S., Shunkov, M. V., Derevianko, A. P., Hublin, J., Kelso, J., Slatkin, M. & Paabo, S. (2010) Genetic history of an archaic hominin group from Denisova Cave in Siberia. *Nature*. 468, 1053-1060.
129. Renoux-Elbé, C., Cheynier, R. & Wain-Hobson, S. (2002) Phylogeny Derived from Coding Retroviral Genome Organization. *Journal of Molecular Evolution* 54 (3): 376-385.
130. Rice, P., Longden, I., & Bleasby, A. (2000). EMBOSS: the European molecular biology open software suite. *Trends in genetics*, *16*(6), 276-277.
131. Roca, A. L. & O'Brien, J. O. (2005) Genomic inferences from Afrotheria and the evolution of elephants. *Current Opinion in Genetics & Development*. 15: 652-659.
132. Sassaman, D. M., Dombroski, B. A., Moran, J. V., Kimberland, M. L., Naas, T. P., DeBerardinis, R. J., ... & Kazazian, H. H. (1997). Many human L1 elements are capable of retrotransposition. *Nature genetics*, *16*(1), 37-43.
133. Seifarth, W., Skladny, H., Krieg-Schneider, F., Reichert, A., Hehlmann, R., & Leib-Mösch, C. (1995). Retrovirus-like particles released from the human breast cancer cell line T47-D display type B-and C-related endogenous retroviral sequences. *Journal of virology*, *69*(10), 6408-6416.
134. Shendure, J., Porreca, G. J., Reppas, N. B., Lin, X., McCutcheon, J. P., Rosenbaum, A. M., ... & Church, G. M. (2005). Accurate multiplex polony sequencing of an evolved bacterial genome. *Science*, *309*(5741), 1728-1732.
135. Skripkin, E., Paillart, J. C., Marquet, R., Ehresmann, B., & Ehresmann, C. (1994). Identification of the primary site of the human immunodeficiency virus type 1 RNA dimerization in vitro. *Proceedings of the National Academy of Sciences*, *91*(11), 4945-4949.
136. Smit, A. F. A. 1993 Identification of a new, abundant superfamily of mammalian LTR-transposons. Nucl. Acids Res. 21, (8): 1863-1872.
137. Smit, A. F. A., Hubley, R. & Green, P. 1996-2010 RepeatMasker Open-3.0. <http://www.repeatmasker.org>.
138. Springer, M. S., Cleven, G. C., Madsen, O., de Jong, W. W., Waddell, V. G., Amrine, H. M. & Stanhope, M. J. (1997) Endemic African mammals shake the phylogenetic tree. *Nature*. 388, 61-64.
139. Springer, M. S., Murphy, W. J., Eizirik, E. & O'Brien, S. J. (2003) Placental mammal diversification and the Cretaceous-Tertiary boundary. *PNAS*. 100, (3), 1056-1061.
140. Stocking, C., & Kozak, C. A. (2008). Endogenous retroviruses. *Cellular and molecular life sciences*, *65*(21), 3383-3398.
141. Stoye, J. P. (2012). Studies of endogenous retroviruses reveal a continuing evolutionary saga. *Nature reviews Microbiology*, *10*(6), 395-406.
142. Strauss, E. G., & Strauss, J. H. (2007). *Viruses and human disease*. Academic Press.
143. Subramanian, R. P., Wildschutte, J. H., Russo, C., & Coffin, J. M. (2011). Identification, characterization, and comparative genomic distribution of the HERV-K (HML-2) group of human endogenous retroviruses. *Retrovirology*, *8*(1).
144. Suzuki, Y. & Gojobori, T. 1999 A method for detecting positive selection at single amino acid sites. Mol. Biol. Evol. 16, 1315-1328.

145. Sverdlov ED (2000) Retroviruses and primate evolution. BioEssays 22: 161–171.
146. Tamura K, Peterson D, Peterson N, Stecher G, Nei M, and Kumar S (2011) MEGA5: Molecular Evolutionary Genetics Analysis using Maximum Likelihood, Evolutionary Distance, and Maximum Parsimony Methods. Molecular Biology and Evolution 28: 2731-2739.
147. Tamura, K., & Nei, M. (1993). Estimation of the number of nucleotide substitutions in the control region of mitochondrial DNA in humans and chimpanzees. *Molecular biology and evolution*, *10*(3), 512-526.
148. Tarlinton, R. E., Meers, J., & Young, P. R. (2006). Retroviral invasion of the koala genome. *Nature*, *442*(7098), 79-81.
149. Tarlinton, R., Meers, J., Hanger, J., & Young, P. (2005). Real-time reverse transcriptase PCR for the endogenous koala retrovirus reveals an association between plasma viral load and neoplastic disease in koalas. *Journal of general virology*, *86*(3), 783-787.
150. Temin, H. M. (1964). Nature of the provirus of Rous sarcoma. *Nat. Cancer Inst. Monogr*, *17*, 557-570.
151. Temin, H. M. (1985). Reverse transcription in the eukaryotic genome: retroviruses, pararetroviruses, retrotransposons, and retrotranscripts. *Molecular biology and evolution*, *2*(6), 455-468.
152. Temin, H. M., & Rubin, H. (1958). Characteristics of an assay for Rous sarcoma virus and Rous sarcoma cells in tissue culture. *Virology*, *6*(3), 669-688.
153. Thomas, J. A., Welch, J. J., Lanfear, R., & Bromham, L. (2010). A generation time effect on the rate of molecular evolution in invertebrates. *Molecular biology and evolution*, msq009.
154. Towers, G., Bock, M., Martin, S., Takeuchi, Y., Stoye, J. P., & Danos, O. (2000). A conserved mechanism of retrovirus restriction in mammals. *Proceedings of the National Academy of Sciences*, *97*(22), 12295-12299.
155. Tristem, M. (2000) Identification and characterisation of novel human endogenous retrovirus families by phylogenetic screening of the human genome mapping project database. *J. Virol.* 74, 3715-3730.
156. Tristem, M. & Lee, A. (unpublished) An *in silico* system for detecting orthology and paralogy using fuzzy-matching.
157. Turelli, P., Pétursson, G., Guiguen, F., Mornex, J. F., Vigne, R., & Quérat, G. (1996). Replication properties of dUTPase-deficient mutants of caprine and ovine lentiviruses. *Journal of virology*, *70*(2), 1213-1217.
158. Turner, G., Barbulescu, M., Su, M., Jensen-Seaman, M. I., Kidd, K. K. & Lenz, J. (2001) Insertional polymorphisms of full-length endogenous retroviruses in humans. *Curr. Biol.* 11, 1531-1535.
159. van Regenmortel, M. H. V. et al. (2000) *Virus Taxonomy*. Academic Press, San Diego, 2000, p. 1167.
160. Volff, J., Körting, C., Altschmied, J., Duschl, J., Sweeney, K., Wichert, K., Froschauer, A. & Schartl, M. 2001 *Jule* from the Fish *Xiphophorus* Is the First Complete Vertebrate *Ty3/Gypsy* Retrotransposon from the *Mag* Family. Mol. Biol. Evol. 18, (2): 101-111.
161. Weiss, R. A., & Vogt, P. K. (2011). 100 years of Rous sarcoma virus. *The Journal of experimental medicine*, *208*(12), 2351-2355.
162. Wilkinson, D. A., Mager, D. L., & Leong, J. A. C. (1994). Endogenous human retroviruses. In *The retroviridae* (pp. 465-535). Springer US.
163. Xiong, Y., & Eickbush, T. H. (1990). Origin and evolution of retroelements based upon their reverse transcriptase sequences. *The EMBO journal*, *9*(10), 3353.
164. Zehr, S. M., Nedbal, M. A. & Flynn J. J. 2001 Tempo and Mode of Evolution in an Orthologous *Can* SINE. Mammalian Genome 12, 38-44.

APPENDIX A

Published manuscript corresponding to the work presented in chapter 2.

Lee, A., Nolan, A., Watson, J. & Tristem, M. (2014) Identification of an ancient endogenous retrovirus, predating the divergence of placental mammals. *Phil. Trans. R. Soc. B* doi: 10.1098/rstb.2012.0503

Downloaded from http://rstb.royalsocietypublishing.org/ on November 29, 2014

PHILOSOPHICAL TRANSACTIONS OF THE ROYAL SOCIETY B

rstb.royalsocietypublishing.org

Research

Cite this article: Lee A, Nolan A, Watson J, Tristem M. 2013 Identification of an ancient endogenous retrovirus, predating the divergence of the placental mammals. Phil Trans R Soc B 368: 20120503. http://dx.doi.org/10.1098/rstb.2012.0503

One contribution of 13 to a Theme Issue 'Paleovirology: insights from the genomic fossil record'.

Subject Areas:
evolution, bioinformatics, genomics, paleontology, computational biology, molecular biology

Keywords:
retrovirus evolution, mammalian evolution, selfish genetic elements, mammalian-wide interspersed repeat, genomics, paleovirology

Author for correspondence:
Adam Lee
e-mail: adam.lee07@imperial.ac.uk

Electronic supplementary material is available at http://dx.doi.org/10.1098/rstb.2012.0503 or via http://rstb.royalsocietypublishing.org.

Royal Society Publishing
Informing the science of the future

Identification of an ancient endogenous retrovirus, predating the divergence of the placental mammals

Adam Lee[1], Alison Nolan[2], Jason Watson[2] and Michael Tristem[2]

[1]Imperial College London, South Kensington Campus, London SW7 2AZ, UK
[2]Imperial College London, Silwood Park Campus, Buckhurst Road, Ascot, Berkshire SL5 7PY, UK

The evolutionary arms race between mammals and retroviruses has long been recognized as one of the oldest host–parasite interactions. Rapid evolution rates in exogenous retroviruses have often made accurate viral age estimations highly problematic. Endogenous retroviruses (ERVs), however, integrate into the germline of their hosts, and are subjected to their evolutionary rates. This study describes, for the first time, a retroviral orthologue predating the divergence of placental mammals, giving it a minimum age of 104–110 Myr. Simultaneously, other orthologous selfish genetic elements (SGEs), inserted into the ERV sequence, provide evidence for the oldest individual mammalian-wide interspersed repeat and medium-reiteration frequency interspersed repeat mammalian repeats, with the same minimum age. The combined use of shared SGEs and reconstruction of viral orthologies defines new limits and increases maximum 'lookback' times, with subsequent implications for the field of paleovirology.

1. Introduction

Endogenous retroviruses (ERVs) are proviral DNA elements that insert into the germline cells of the host organism, being vertically transmitted from one generation to the next [1]. Intact elements usually comprise retroviral structural genes, an encoded polymerase and viral envelope, flanked on either side by long terminal repeats (LTRs). Over time, germline ERV insertions accrue an increasing number of irreversible mutations in their DNA, often leading to inactive retroviral elements, defective in their ability to produce infectious virions [2]. In some cases, ERV insertions remain in the genome as 'molecular fossils' of ancient infection events, and can be used to better understand and study the evolution of both retroviruses and their hosts [3]. Upon integration, ERVs likely retain greater genomic conservation than their exogenous counterparts, being subjected only to the host's rate of genome evolution [4].

ERVs are classified into families, each of which contains viruses thought to have derived from a single, initial infection event. Many human endogenous retroviruses (HERVs), for instance, stem from an infection event of an ancestral mammalian genome [5], formerly suggested to have occurred as much as 70 Ma in one particular family [6]. This family is HERV-L (HERV with leucine tRNA primer), discovered in 1995, with related sequences having been found at copy numbers between 1 and 200 in the genomes of many mammals [7]. As a family, ERV-L thus appears to have been active for much of mammalian evolution. Unusually for a retrovirus, ERV-L does not contain an envelope gene.

Viral genomes are typically prone to high rates of nucleotide substitutions, often resulting in many lineages becoming extinct relatively quickly [8]. However, unlike other viruses, retroviruses must integrate into the genome of their host as part of their replication cycle [9]. If this integration event occurs in a germ cell, then the retroviral genome becomes endogenized, and its mutation rate is generally confined to the host organism's neutral rate of evolution (with the exception of rare cases in which the retroviral element becomes

© 2013 The Authors. Published by the Royal Society under the terms of the Creative Commons Attribution License http://creativecommons.org/licenses/by/3.0/, which permits unrestricted use, provided the original author and source are credited.

Downloaded from http://rstb.royalsocietypublishing.org/ on November 29, 2014

co-opted) [10,11]. ERVs can thus provide a long-term record of host and virus evolution.

In order to date these integrated retroviruses, sequence divergence between the two LTRs can be measured [5]. While useful in estimating the age of recent insertions, this method is not as easily applicable to estimating the age of ancient ERVs, since these are often highly mutated. A more efficient procedure to obtain a minimum age of certain ancient ERVs is to identify orthologous elements in related taxa. The presence of orthologous ERVs in two or more host species implies that the initial infection event predates the divergence of those species [12]. Furthermore, co-divergence or co-speciation studies of viruses and their hosts can also indicate ancient associations. This approach, based largely on the study of extant viruses, and not a panel of ancient ERVs, has been used to indicate that some foamy viruses originated more than 100 Ma [13].

While there have been few studies investigating the age of viral insertions using orthologous elements, one study involved dating orthologous retroviruses in the European hare. The retrovirus, termed rabbit endogenous lentivirus type K, was found inserted at the same genomic locus across the order Lagomorpha, giving the element a minimum age of 12 Myr [14]. While orthologies were not used for dating, cophylogenetic analysis of a retroviral element discovered in the genome of a coelacanth gave an estimated age of 407 Myr for foamy-like viruses [15]. Another study involved reconstructing orthologies for several non-retroviral endogenous virus elements, one of which was a virus of the family *Bornaviridae*, found to have a minimum age of 93 Myr; members of the *Parvoviridae* and *Circoviridae* were also examined, with minimum ages of 30 and 60 Myr, respectively [16].

In addition to ERVs, host genomic DNA can also be subjected to the insertion of other selfish genetic elements (SGEs) such as short interspersed elements (SINEs) and long interspersed elements. These SGEs can amplify themselves within the genome, often to the detriment of the host organism, which may lead to their eventual deactivation or deletion via selection. However, if SGEs are neutral to the host, then they can reach fixation in the genome and, much like ERVs, can be passed down vertically across many generations, often making ancient SGE insertions useful in evolution and dating studies [17]. Among the most prevalent SGE families in the Mammalia, are mammalian-wide interspersed repeats (MIRs). Estimated to have started amplifying within mammalian genomes during the Mesozoic era, this family of repeats can be found at copy numbers as high as 10^7. SGE orthologies have also been reconstructed across several mammalian superorders. Namely, an orthologous MIR insertion was demonstrated occurring at nine loci within humans and non-primate mammalian sequences dating back 65 Myr which, together with the application of a molecular clock, suggested an age of approximately 130 Myr for this family of repeats [18].

Mammalian genomes also contain large numbers of medium-reiteration frequency interspersed repeats (MERs). In particular, the MER3 family exhibits the highest diversity, with poor conservation found between elements [19]. While orthologies have not been reconstructed to date, MER3 elements are thought to be at least 55 Myr old [20]. Another family of SGEs, with a distribution across the placental mammals, is mammalian apparent LTR-retrotransposons. An individual element found in both humans and mice, suggested an age of 80–100 Myr for this family, although, as of yet, no definitive orthology has confirmed this [21]. Likely older still are the *Ty3/Gypsy* family of LTR-retrotransposons, where related sequences were observed in both vertebrates and invertebrates. Phylogenies of these sequences suggested a potential age of 600 Myr for this family [22].

In this investigation, we examined a member of the ERV-L family and, via orthology reconstruction, confirmed that the element probably integrated before the divergence of placental mammals, making this the oldest individual virus described to date. Furthermore, three SGE insertions within the ERV-L were found across all, or nearly all, of the studied taxa, making these the oldest individual MIR and MER3 elements described. We further discuss some of the difficulties in identifying ancient retroviral insertions, while estimating possible maximum 'lookback' times for ERVs, and the corresponding implications for the field of paleovirology, in general.

2. Material and methods

(a) Element retrieval and sequence alignment

The published HERV-L reference sequence (X89211; [7]) was retrieved from NCBI's Nucleotide database, and used to perform BLAST searches against the genomes of various mammals. The algorithm selected was BLASTN. Initially, one particular ERV-L sequence was identified in several mammalian superorders; preliminary analysis of LTRs and sequence flanking the viral insertion in these species suggested possible orthology. This particular ERV-L element was therefore chosen as the subject of further characterization.

We used this newly identified human ERV-L element as a probe, *in silico*. A further 10 homologous retroviral sequences were retrieved from different mammalian hosts through various BLAST searches against NCBI's WGS, HTGS and refseq_genomic databases [23]. It should be noted that many of the taxa selected for investigation have low rates of neutral evolution. The viral sequences were extracted, together with sequences flanking the 5′ and 3′ LTRs of the elements, and were run through NCBI's open reading frame (ORF) finder, to screen for any intact retroviral ORFs. The mammals included in this study span superorders I–IV, and include the pig (*Sus scrofa*), horse (*Equus caballus*), Seba's short-tailed bat (*Carollia perspicillata*), large flying fox (*Pteropus vampyrus*), human (*Homo sapiens*), chimpanzee (*Pan troglodytes*), nine-banded armadillo (*Dasypus novemcinctus*), African elephant (*Loxodonta africana*), West Indian manatee (*Trichechus manatus*), aardvark (*Orycteropus afer*) and golden mole (*Chrysochloris asiatica*). In addition to this, BLAST searches of the available marsupial genomes were also conducted.

An initial alignment was constructed from approximately 1.2 kilobase (kb) of 5′ flanking sequence and 70 basepair (bp) of the 5′ LTR from the start of the element. Sequences were then aligned using EBI's multiple sequence comparison by log-expectation (MUSCLE) algorithm [24], with resulting alignments being adjusted and corrected manually. Sequence conservation was also recorded at each individual site, and plotted as a histogram beneath the alignment. This was repeated for approximately 2 kb of 3′ flanking sequence and 70 bp of the associated 3′ LTR. Sequences demonstrating orthology were retained, and the corresponding viral insertion in each species was loaded into a third alignment. The full-length alignments are available in the electronic supplementary material.

The flanking sequence was trimmed, and the complete viral elements, from the 5′ end of the 5′ LTR through to the 3′ end of the 3′ LTR, were individually run through RepeatMasker [25] with the appropriate species-specific repeat settings applied, in

Downloaded from http://rstb.royalsocietypublishing.org/ on November 29, 2014

order to identify SGEs, including regions of the ERV-L itself. Identified elements and repeats were then annotated onto the sequence using CLC's main workbench. Elements that appeared to be anomalously assigned by RepeatMasker were investigated at the nucleotide level, and manually corrected. Once annotated, sequences were aligned again using MUSCLE, and the alignments were carefully edited and adjusted manually, using NCBI's bl2seq and ClustalW to guide manual alignment of small fragments of the viral element. Following alignment, any insertions in two species or fewer that amounted to 10 bp or more were deleted (deleted regions are specified in the electronic supplementary material).

Subsequent to the annotation of the sequences, an additional LTR structure was identified in the Afrotheria, approximately 2300 bp into the viral alignment. In order to verify that this was a solo-LTR structure, approximately 400 bp of both 5′ and 3′ LTRs paralogous to the published HERV-L sequence were retrieved from the human genome. In combination with the Afrotherian LTR, these were used to construct a maximum-likelihood phylogeny, in order to confirm whether this was in fact a solo-LTR (A. Lee & M. Tristem 2013, unpublished data).

(b) Sequence identities

In order to numerically evaluate and represent homology between the orthologues, two sets of pairwise comparisons were made between species, to give observed percentage identities across all pairs of taxa. First, a 997 bp region from the start of the element was extracted, and all gaps were removed. This was repeated for a second, 830 bp region starting at position 4417 (see the electronic supplementary material). Shared, pairwise sequence identities were then computed using CLC main workbench.

(c) Phylogenetic analysis

A 1 kb aligned section of ERV internal region downstream of the 5′ LTR was used as the basis of phylogenetic analysis. The region used is designated region 1. All species were included, apart from Seba's short-tailed bat and the flying fox, both of which had insufficient sequence data at this region. A maximum-likelihood phylogenetic tree was built from this region using the TN93 substitution model, performed in MEGA [26]. Branch support was in the form of 10 000 bootstrap replicates. A host phylogeny was derived from the literature in order to construct a cophylogeny. Minimum divergence dates for the various host taxa were also obtained from the literature.

A second phylogeny was built from a segment of ERV internal region (from positions approx. 3220–3820; designated region 2), in order to confirm observed phylogenetic relationships resolved by the first tree. The same substitution model and bootstrapping were used. This was then plotted onto a cophylogeny, together with the mammalian tree. The Afrotheria were excluded from this phylogeny, as they did not align sufficiently. This second cophylogeny has been made available in the electronic supplementary material.

(d) Estimating initial integration date

To determine how long before the radiation of placental mammals the initial integration event occurred (assuming Eutherian divergence at 104 Ma), we calculated the percentage of non-identical residues between the LTRs of three combinations of species, representing each of the four mammalian superorders. These included human (superorder III) against elephant (superorder I), human against armadillo (superorder II) and human against horse (superorder IV).

The orthologous 5′ LTRs of the human and elephant ERV-L element were run through the EMBOSS Needle sequence alignment algorithm [27], creating a pairwise alignment. Elephant 5′ and human 3′ LTRs were also subsequently aligned using Needle, and the percentage of non-identical bases was obtained. Subtracting the former value from the latter value gives an approximate estimate for the amount of LTR divergence occurring prior to the initial placental mammal speciation events. This value was then divided by the neutral mutation rate of the mouse genome, which then allows an estimated date of the initial germline integration event to be calculated. These calculations assume that the LTRs are subjected solely to a steady rate of host genome evolution over time. The calculations were then repeated for the remaining species combinations.

(e) d_N/d_S ratios

The complete MuERV-L sequence (accession number Y12713) was obtained from NCBI's Nucleotide database, and a BLAST search of this sequence against the investigated ERV-L element was performed, to retrieve the areas with highest homology in the human sequence. These individual matching regions were compiled into a separate alignment, and all leading and trailing MuERV-L sequence between the blocks demonstrating homology were deleted (while maintaining an ORF in the MuERV-L sequence, and therefore the presumed ORFs in the original ERV-L sequences). Species included in the alignment comprised armadillo, human, Seba's short-tailed bat, flying fox, chimpanzee, pig and horse. The regions aligning to MuERV-L fell within the approximately 2.5 kb of poorly aligning sequence in the Afrotheria, so the corresponding four species were omitted.

The alignment was then used to calculate codonwise d_N/d_S ratios using HyPhy [28] in MEGA [26], together with p-values; if significant (less than 0.05), then these could be used to reject a null hypothesis assuming the sequences are evolving under neutrality [28,29]. A total of 108 amino acid sites were deduced from the alignment, and both p-values and d_N/d_S ratios were averaged to give individual values.

(f) Synteny analysis

Analysis of chromosome-wide syntenic genes surrounding the ERV-L integration was performed using ENSEMBL's SyntenyView, with available species for analysis, including human, pig, horse and chimpanzee. The human ERV-L element was found to be located on chromosome 17; synteny across this chromosome was visualized together with the corresponding chromosomes in the pig, horse and chimpanzee. Other host taxa were not available for analysis in SyntenyView.

3. Results

(a) Flanking sequence similarity between ERV-L elements

We identified an ERV-L sequence in 11 placental mammals by BLAST searches of genomic databases. While the available genomes of several marsupials were also screened, these returned negative results. Most of the examined host species were specifically selected for their relatively low neutral mutation rates, because this would be expected to best preserve sequence similarity between any orthologous viruses and other SGEs.

When aligned, the 5′ and 3′ flanking regions generally showed considerable levels of homology across all species (figure 1). However, approximately 450 bp of sequence just upstream of the 5′ LTR was found to contain a large insertion in the Afrotheria (which was subsequently excised from figure 1). Upstream of this 450 bp region, sequence

Downloaded from http://rstb.royalsocietypublishing.org/ on November 29, 2014

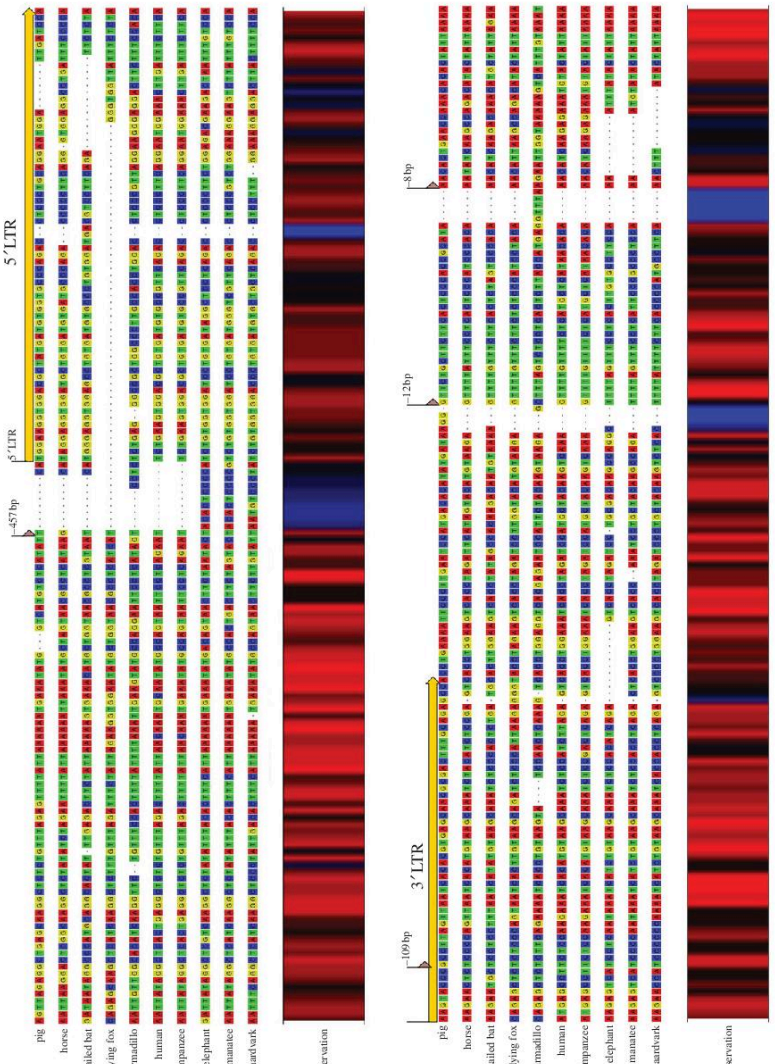


Figure 1. ERV-L integration site: (*a*) 5′ integration site, showing the extreme end of the 5′ LTR of the ERV-L insertion in 10 mammalian species, together with flanking sequence. (*b*) Corresponding 30 integration site, showing the extreme end of the 3′ LTR, together with flanking sequence. A conservation colour gradient beneath the nucleotides indicates the level of sequence homology across all 10 mammals, with red showing highest conservation. Triangular flags indicate segments of DNA deleted from a particular species for illustration purposes. Because alignment was poor at these regions, the golden mole was omitted (full flanking region alignments are given in the electronic supplementary material).

Downloaded from http://rstb.royalsocietypublishing.org/ on November 29, 2014

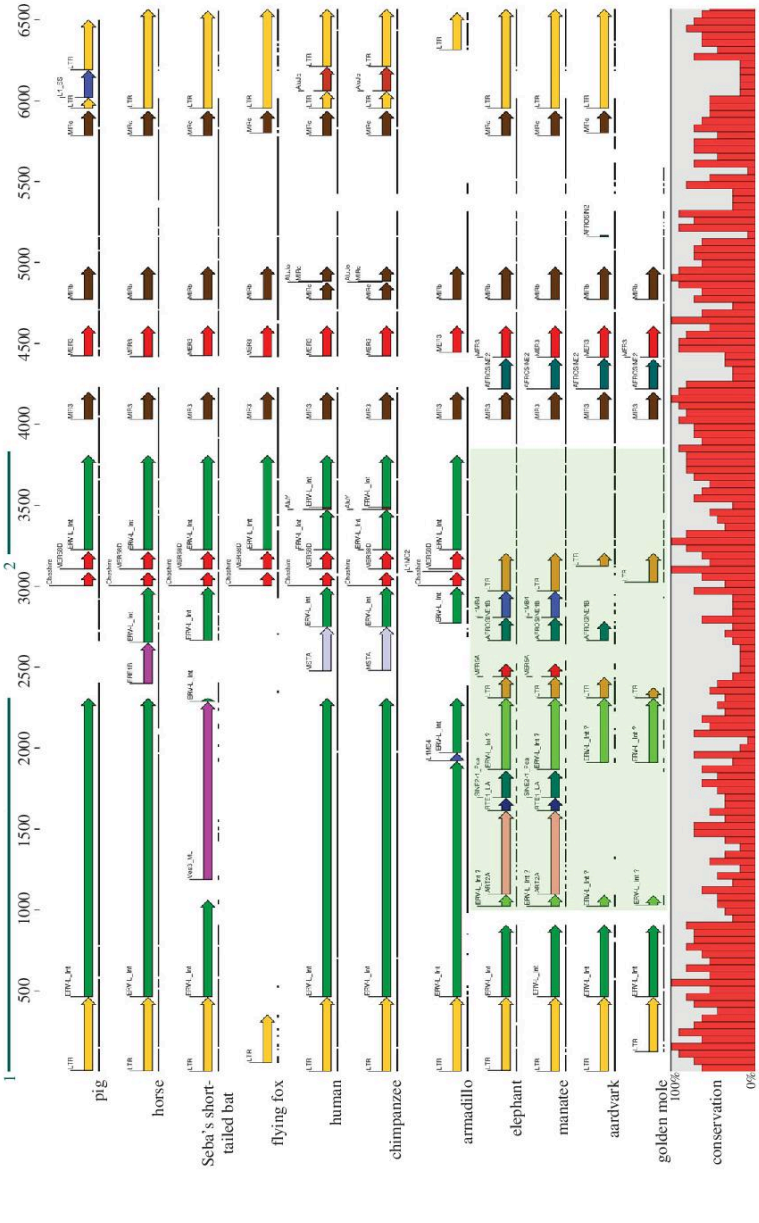


Figure 2. Schematic of ERV-L element and identified repeats: SGEs, identified by RepeatMasker, are individually flagged, with the repeat families being colour-coded. A solo-LTR, present only in the Afrotheria, is highlighted in beige, whereas a region of low homology in the same mammalian superorder is shaded light green. The presence of nucleotide sequence is indicated by horizontal black lines, to make obvious where gaps have been inserted in order to maintain the alignment. Sequence conservation is indicated by a histogram running underneath the alignment. Region 1 and region 2 are demarcated by dark lines running above the alignment, and correspond to the regions used for phylogenetic analysis (region 1 corresponds to figure 5; region 2 corresponds to a second cophylogeny given in the electronic supplementary material).

Downloaded from http://rstb.royalsocietypublishing.org/ on November 29, 2014

conservation was high across all species (see additional flanking sequence alignment in the electronic supplementary material). Figure 1 indicates that the ERV-L sequence from each species is therefore likely to have integrated at the same locus in the genomes of multiple placental mammals (the golden mole sequence is, however, missing from the 5′ flank alignment, owing to the absence of this region from the genome database). Because retroviral integration is generally random [30], this provides strong evidence for the orthology of the element.

(b) Full-length viral sequence alignment and identification of shared selfish genetic elements

Prior to alignment, the element was verified as an ERV-L element, although phylogenetic analyses demonstrated the element to be basal to other HERV-L elements selected from the human genome (A. Lee & M. Tristem 2013, unpublished data). The ERV elements were found to be highly mutated, because they did not encode any intact retroviral ORFs. To further demonstrate orthology, the proviral DNA of each species was aligned and annotated with mammalian repeats identified by RepeatMasker [25]. This revealed that many SGEs were shared across all the investigated species, at similar positions. Figure 2 shows high conservation at both 5′ and 3′ LTRs across all species and the internal region was shown to be well conserved at several points, most notably between approximately 500 and 1000 bp (excepting the Armadillo, in which there have been deletions), and between approximately 3200 and 3800 in the Boreoeutherian mammals. Two separate MIR elements, at positions approximately 4000 (figure 3) and approximately 4750, and an MER3 at approximately 4450 (figure 4) are shared across all 11 mammals. Additionally, another MIR element (at position approx. 5800) was shared across nine mammals; the corresponding sequence is missing from the armadillo and golden mole for this MIR (probably as the result of subsequent deletion of genomic DNA in these host species).

The green shaded area in figure 2 demarcates a region in which the Afrotheria shared very little sequence homology with the Boreoeutherian mammals. This region represents approximately 2.5 kb of the alignment. An additional LTR structure was also identified within this region, at positions approximately 2300–3200 bp in the Afrotheria. This LTR was itself interrupted by the insertion of a MER5A and L1MB4 repeat in the sirenians, and an AFROSINE1B in both sirenians and the aardvark. Phylogenetic analysis confirmed this to be a solo-LTR structure (A. Lee & M. Tristem 2013, unpublished data). This was most likely the result of a secondary viral integration within the existing ERV-L element, and its subsequent recombinational deletion, leaving behind a single 5′ solo-LTR structure [1].

In addition to the region of poor similarity in the Afrotheria, there was an approximately 800 bp region where RepeatMasker failed to identify any repeats (or similarity to other ERV-L sequences); this is present in all 11 mammalian sequences, starting approximately 5000 bp into the alignment (figure 2). As all the sequences exhibit significant homology to each other throughout this region, it is unclear whether this is a heavily mutated internal region, or whether it is an unexplained insertion. Regardless, this change in the ERV-L genome is expected to have occurred prior to the Afrotherian–Boreoeutherian divergence, because it is evident in all 11

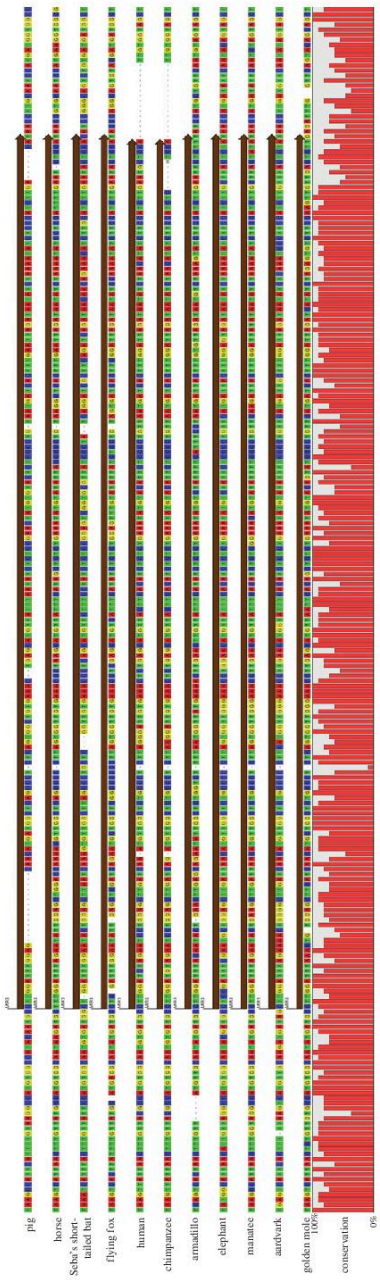


Figure 3. Orthologous MIR3 retrotransposon: an MIR3 retrotransposon, and flanking DNA, shared across all 11 mammalian species. The nucleotide sequence beneath each annotation is shown, with a graph running beneath the alignment to indicate site-specific sequence conservation across the taxa.

Downloaded from http://rstb.royalsocietypublishing.org/ on November 29, 2014

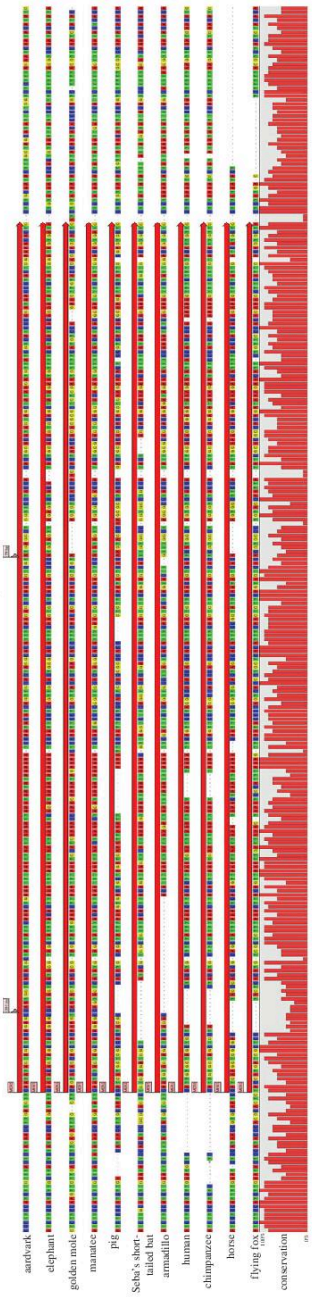


Figure 4. Orthologous MER3 DNA transposon: a non-autonomous MER3 hAT-Charlie DNA transposon, and flanking DNA, shared across all 11 mammals. The nucleotide sequence beneath each annotation is shown, with a graph to indicate site-specific sequence conservation across the taxa. Flags indicate deletions that were made for illustrative purposes, in regions that contained unique insertions.

placental mammals. This region is flanked on either side by two conserved MIR repeats.

(c) Sequence identities

We calculated the percentage shared sequence identities between ERV-L orthologues in pairs of taxa across two regions, from positions 1–997 (region A) to 4417–5247 (region B). Table 1 indicates relatively good sequence conservation between the most distant two superorders, I and IV, with up to 65% identities shared between species such as Seba's short-tailed bat against the manatee, in region A. Similarities for region B were generally equivalent. Sequence identities for some taxa, especially in region A, however, are far lower for the golden mole, ranging from 25 to 50%, relative to superorders II-IV; this would indicate a more divergent 5′ LTR in these taxa. Similarly, pairs of armadillo against superorders III and IV, show a lower range of shared identities, between 34 and 39%.

(d) Phylogenetic analysis

We constructed a cophylogeny of the ERV-L sequences together with a mammalian phylogeny adapted from the literature (figure 5). This demonstrated that the retroviral phylogenetic relationships resolve well to those shown in the existing mammalian phylogeny [37], as would be expected with orthologous elements. There are, however, topological differences between the Xenarthran and Afrotherian mammals. This is not surprising given the previous difficulties of reconstructing mammalian phylogenies from single genes [38].

(e) Estimating the initial integration date

The orthologous nature of the ERV-L sequences indicates that the virus first integrated into an ancestor of all extant placental mammals, which started diverging between 104 and 110 Ma [31,36]. Thus, the virus has the same minimum age. However, it could also be considerably older, as is suggested by multiple shared SGEs, which presumably integrated into the ERV-L element itself prior to the divergence of the placental mammals. To investigate this, we examined the similarity between various combinations of the ERV-L 5′ and 3′ LTRs.

Upon integration, both viral LTRs generally comprise identical sequences. Thus, a comparison of the percentage identity between any 5′ LTRs (or any 3′ LTRs) indicates how much LTR sequence divergence has occurred since the earliest divergence of the placental mammals. By contrast, a calculation of the percentage divergence between a 5′ LTR and a 3′ LTR indicates how much sequence divergence has occurred since the initial integration event. Using Needle, a pairwise comparison was made between the elephant and human 5′ LTR sequences, and a divergence of 31% was given between them. We then compared the 5′ LTR of the elephant ERV-L with the 3′ LTR of the human element (and vice versa), and obtained an average divergence of 53%. Thus, $53 - 31\% = 22\%$, meaning that 22% of the sequence divergence between the 5′ and 3′ LTRs likely occurred prior to divergence of the placental mammals, 104–110 Ma [31,36,39].

It is problematic to turn percentage 5′ versus 3′ LTR divergence data into timescales, owing to the lack of data regarding the neutral mutation rates of ancient mammals. Gene conversion and recombination between LTRs can also make these

Downloaded from http://rstb.royalsocietypublishing.org/ on November 29, 2014

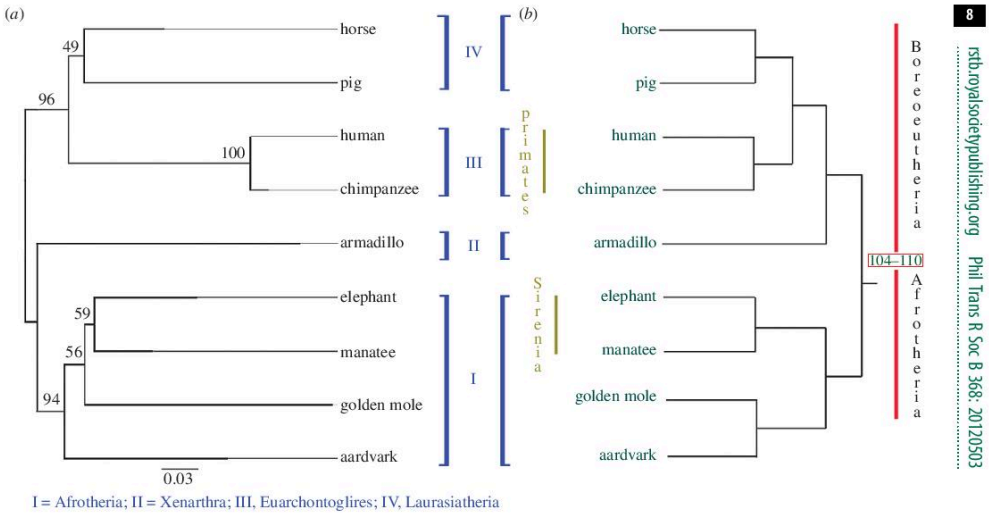


Figure 5. Retroviral-mammalian cophylogeny: cophylogeny comprising an unrooted maximum-likelihood phylogenetic tree constructed from the described ERV-L element (*a*) and an adapted schematic rooted mammalian phylogeny (*b*) [31–34]. Species falling into each of the four mammalian superorders are bracketed in blue, while the Afrotherian and Boreoeutherian clades are indicated in red; primates and Sirenia are indicated in beige. Branch lengths on the retroviral tree (*a*) are scaled to indicate the number of substitutions per site, and bootstrap values are shown. The initial divergence of the placental mammals into Afrotheria and Boreoeutheria is indicated in a box [31,36].

Table 1. Pairwise observed sequence identities (%) for two regions: region A is shaded blue, and region B is red. Region A covers the first 997 bp of the element, including the 5′ LTR and ERV internal region. Because the flying fox was missing sequence data in this region, it has been omitted. Region B covers a 830 bp segment of the alignment, starting at position 4417. The full alignment is available in the electronic supplementary material. Blue, region A; red, region B; grey, omitted data.

observed identities (%)		1	2	3	4	5	6	7	8	9	10	11
pig	1		68	64	62	65	65	57	62	65	59	51
horse	2	75		72	69	72	72	66	69	73	65	59
Seba's bat	3	62	67		64	65	65	60	63	67	59	55
flying fox	4					63	63	58	60	62	58	53
human	5	68	72	59			98	62	67	69	64	55
chimpanzee	6	67	72	59		98		62	67	69	64	55
armadillo	7	34	39	35		36	35		63	66	60	55
elephant	8	58	63	63		59	59	37		83	73	65
manatee	9	60	64	65		62	61	38	83		71	65
aardvark	10	56	60	61		59	58	39	70	73		59
golden mole	11	45	48	50		46	46	25	58	61	55	

estimates unreliable. However, if we assume the ancestral placental mammal was small, with a relatively fast generation time, and therefore apply a conservative neutral mutation rate based on the mouse genome, of 4.5×10^{-9} substitutions per site per year [40], we find an initial integration date of approximately 128 Ma (22 divided by $(0.45 \times 2) + 104$).

These calculations were then repeated for human against armadillo ERV-L LTRs, and human against horse ERV-L LTRs, giving initial integration date estimates of approximately 140 and approximately 139 Ma, respectively. We therefore suggest that the initial integration date could have preceded the divergence of the placental mammals by at least 24 Myr, and perhaps by up to 36 Myr.

(f) d_N/d_S ratios

HyPhy was used to calculate individual values for d_N and d_S, per codon, across the seven Boreoeutherian mammals. We

Downloaded from http://rstb.royalsocietypublishing.org/ on November 29, 2014

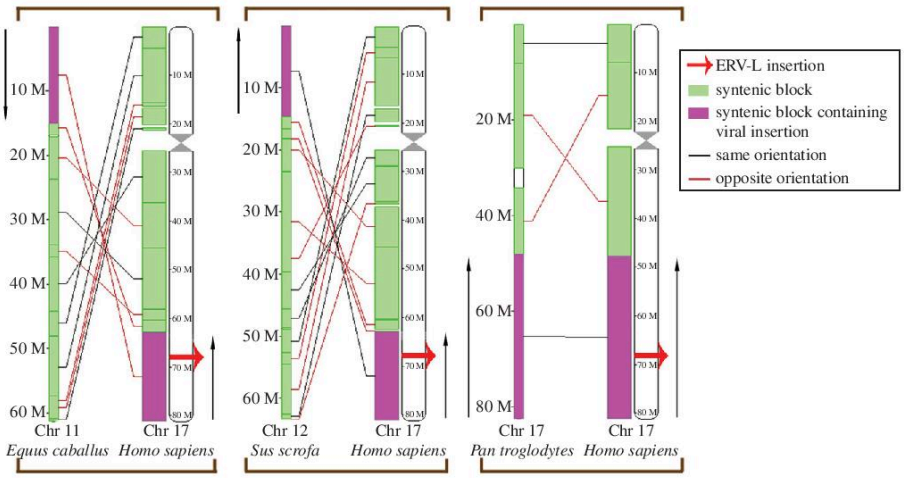


Figure 6. Syntenic relationship between human chromosome 17 and three selected mammals: a chromosome synteny map adapted from ENSEMBL's SyntenyView, showing syntenic blocks (shaded green) shared between human chromosome 17, where the ERV-L insertion occurs (highlighted in red), and corresponding chromosomes in the horse, pig and chimpanzee. Syntenic blocks containing the ERV-L insertion are shaded pink.

were unable to reject the null hypothesis of neutrality ($p = 0.71$), as would be expected in the case of uncoopted, orthologous insertions, but which is often not the case for paralogous insertions [41].

(g) Synteny

The human ERV-L insertion is present on chromosome 17. Analysis of synteny across this entire chromosome was performed relative to the horse, pig and chimpanzee. While the order is scrambled, as would be expected between distantly related species, figure 6 shows a high number of syntenic blocks shared between chromosome 17 of the human genome and corresponding chromosomes in each of the remaining three species. Furthermore, the syntenic block containing the virus was shown to be present in the same orientation in all species, except for the horse where it is inverted. This analysis is not yet available for any of the Xenarthran or Afrotherian genomes.

4. Discussion

(a) Orthology of the endogenous retrovirus sequences

This is the first time an ERV, or any other type of virus, has been demonstrated to exhibit orthology across all four Eutherian superorders, making this the oldest individual viral sequence examined to date. Within the Afrotheria, the ERV-L element was recovered from the genomes of the African elephant, manatee, aardvark and golden mole. Furthermore, the presence of additional orthologues in multiple Boreoeutherian hosts indicates that the initial infection event predates the earliest divergence of the placental mammals, estimated to have occurred approximately 104–110 Ma [31,36]. Prior to this study, ERV-L elements have previously been found across the placental mammals at low copy numbers, strongly implying a minimum age of 70 Myr for this retroviral family [6]. These findings, however, did not attempt to identify any orthologous retroviral elements. While several other studies have examined orthologous retroviral families between primate species [42], none have extensively described retroviral insertions covering the breadth of mammalian species studied here.

In this investigation, multiple lines of evidence pointing to the orthologous nature of the ERV-L element were found. Because retroviruses generally insert at random within host genomes, only orthologues, but not paralogues, would be expected to share similar host flanking sites. The investigated ERV-L sequences from the different hosts all contained flanking sites with strong sequence similarity and, where synteny data were available, were found to have inserted into the same region of their respective genomes.

We estimated that the initial integration event occurred tens of millions of years before the first divergence of the placental mammals studied. Consistent with this, we found that several SGEs are present in the same locations in all the ERV-L sequences. These would therefore have first inserted within an ancestral mammal after the initial integraton event of the ERV-L, but before the placental mammals started radiating. In particular, three SGEs were identified by RepeatMasker [25] to be conserved across most or all of the host taxa. In each case, the SGEs belong to the same family, and are in the same orientation with respect to the ERV-L. Furthermore, sequence alignments demonstrated that the same region of each SGE is present at almost exactly the same location within each of the ERV-L sequences.

These additional SGEs are therefore also some of the oldest discovered to date. Previous to this study, MIR elements were estimated to be 130 Myr old. However, orthology was only demonstrated to 65 Ma, with the remaining age estimated through the aid of a molecular clock [18]. Other studies that have previously investigated similar orthologues identified a *Can* SINE element with a minimum age of 40 Myr [43]. Moreover, SINE elements have also been used to

Downloaded from http://rstb.royalsocietypublishing.org/ on November 29, 2014

reconstruct ancient orthologies; such an example is the use of *Hpa*I SINEs as clade markers for the study of salmonoid fish evolution [44,45]. Other studies have recovered the complete chronology of Alu elements in primates, using specifically developed computational algorithms [46].

Once a retrovirus inserts into a host genome it remains at that location. For this reason, selection often operates differently on orthologous versus paralogous ERVs. When a retrovirus integrates into the germline and is inherited by host progeny, it is no longer under selection, because that copy remains at the same genomic locus, and the provirus mutates as a result of the host's DNA replication error rate (unless it becomes co-opted). However, when virus particles are produced (either extra- or intracellularly), the replicating retrovirus pool remains under negative selection until one of the viruses successfully inserts into, and becomes fixed within, the germline again. Thus, purifying selection operates on the viral genome between germline integration events [41]. There may be many hundreds of replication cycles within somatic cells between each germline integration event [8]. Therefore, considering the ERV-L element being studied as orthologous, a d_N/d_S ratio of unity would be expected. A greater number of synonymous substitutions over non-synonymous substitutions indicates negative selection, resulting in a ratio of less than unity. We analysed the d_N/d_S ratios between the ERV-Ls in this study and were unable to reject the null hypothesis of neutrality, further supporting the orthologous nature of the elements. This also suggests that there has been no recombination with different ERV-L sequences (while maintaining the same flanking sequences) within a particular host genome, which would also be expected to reduce d_N/d_S ratios.

Individual ERV-L elements do not change location within the host genome, and evolve neutrally. They would therefore be expected to track a phylogeny of their host organisms. Thus, phylogenies of the virus and its hosts should be congruent with one another. We constructed phylogenies of parts of the ERV-L sequences and found that they did indeed almost match the host phylogeny. Taken together, all these lines of evidence clearly indicate that the ERV-L sequences investigated here are orthologous in nature.

(b) Sequence conservation over time

As expected, observed sequence identities shared between species served to highlight the importance of selecting taxa with slower generation times when reconstructing ancient orthologies. For instance, it was observed that the range of percentage identities for pairwise comparisons containing smaller mammals was generally lower than with larger mammals. Such examples included pairs containing armadillo and golden mole, when compared against superorders III and IV. These differences are most likely the result of a higher neutral evolution rate in these smaller mammals. Conversely, the larger mammals with slower generation times, such as the elephant and manatee, contained ERV-L sequences accruing mutations at a comparatively slower rate. The viral insertions in these mammals have a higher range of shared identities when compared with other mammals in superorders II–IV. On some occasions ERVs, or parts of ERVs, can persist for long periods because they have been co-opted [47,48]. We can rule out this possibility here because of the multiply defective nature of the ERVs; all of them contained numerous in-frame stop codons and frameshifting indels. Co-opted elements are usually under selection to maintain an intact ORF.

(c) Difficulties encountered when reconstructing ancient orthologies

Global genomic databases are constantly growing, amassing into a widely varied library of genomic sequences. One of the many obstacles encountered in reconstructing ancient orthologies is the nature of the required sequence information. First, the full-length viral insertion must be retrieved together with sufficient 5′ and 3′ flanking sequence for demonstrating the orthology of the element. This can often mean that a single, intact DNA sequence upwards of 10 kb is required. While some genomes have been sequenced to completion, many others comprise unassembled reads and, depending on the sequencing technology used, individual contigs and reads may not be long enough to prove that a virus is orthologous. This limitation resulted in the elimination of certain species considered for this study. It is also important to study genomes from taxa that have relatively low neutral mutation rates to ensure a higher likelihood of orthologue conservation across multiple taxa.

Other problems revolve around the specificity and stringency of the alignment algorithms used to retrieve corresponding viral elements from genomic sequence libraries. NCBI's BLAST search, for instance, while accurate, can often be too stringent to uncover matches in highly divergent sequences. This can potentially lead to viral orthologues remaining undetected. A number of alternative alignment algorithms are available, but many of the most sensitive programs, in terms of detecting weak sequence similarity, cannot compute sufficient amounts of data in the same way that BLAST can search an entire genomic record in a relatively short space of time. Another of BLAST's limitations is its low sensitivity to the presence of large insertions or deletions. In such cases, alternative algorithms such as BlastAlign [49] are superior in retrieving homologous sequences. Thus, during this investigation, combinations of algorithms were used: BLAST, with its ability to perform fast, genome-wide screens, together with MUSCLE and Needle, with their heightened sensitivity and accuracy [24,27]. We also found that RepeatMasker was generally very efficient at identifying ERV and other SGE sequences.

This study has presented a pan-Eutherian viral orthologue, demonstrating the oldest single endogenized virus identified to date. It has also been possible to identify some of the oldest described individual SGEs, setting a new benchmark for studying orthologous elements within the field of molecular paleovirology. While it is possible that older orthologues may be identified, the problems discussed above suggest that we may be approaching the maximum achievable look-back time—at least as far as the mammals are concerned. It may be possible, although very difficult, to identify orthologous insertions between placental mammals and marsupials (both owing to the divergence date of 147 Ma [37], and also to relatively fast rates of neutral evolution in the extant marsupials) but the placental and monotreme divergence date (167 Ma [37]) may remain beyond reach.

Acknowledgements. The authors thank the Wellcome Trust for their support, and Mona Lee for her insight and meaningful conversation, throughout.

Downloaded from http://rstb.royalsocietypublishing.org/ on November 29, 2014

References

1. Boeke JD, Stoye JP. 1997 Retrotransposons, endogenous retroviruses, and the evolution of retroelements. In *Retroviruses* (eds JM Coffin, SH Hughes, HE Varmus), pp. 343–435. New York, NY: CSHL Press.
2. Stoye JP. 2001 Endogenous retroviruses: still active after all these years? *Curr. Biol.* **11**, R914–R916. (doi:10.1016/S0960-9822(01)00553-X)
3. Mayer J, Meese EU. 2002 The human endogenous retrovirus family HERV-K (HML-3). *Genomics* **80**, 331–343. (doi:10.1006/geno.2002.6839)
4. Kijima TE, Innan H. 2009 On the estimation of insertion time of LTR retrotransposable elements. *Mol. Biol. Evol.* **37**, 896–904. (doi:10.1093/molbev/msp295)
5. Tristem M. 2000 Identification and characterisation of novel human endogenous retrovirus families by phylogenetic screening of the human genome mapping project database. *J. Virol.* **74**, 3715–3730. (doi:10.1128/JVI.74.8.3715-3730.2000)
6. Benit L, Lallemand J, Casella J, Philippe H, Heidmann T. 1999 ERV-I elements: a family of endogenous retrovirus-like elements active throughout the evolution of mammals. *J. Virol.* **73**, 3301–3308.
7. Cordonnier A *et al*. 1995 Isolation of novel human endogenous retrovirus-like elements with foamy virus-related *pol* sequence. *J. Virol.* **69**, 5890–5897.
8. Gifford R, Tristem M. 2003 The evolution, distribution and diversity of endogenous retroviruses. *Virus Genes* **26**, 291–315. (doi:10.1023/A:1024455415443)
9. Tang KF, Lightner DV. 2006 Infectious hypodermal and hematopoietic necrosis virus (IHHNV)-related sequences in the genome of the black tiger prawn *Penaeus monodon* from Africa and Australia. *Virus Res.* **118**, 185–191. (doi:10.1016/j.virusres.2006.01.003)
10. Mallet F, Bouton O, Prudhomme S, Chenyet V, Oriol G, Bonnaud B, Lucotte G, Duret L, Mandrand B. 2004 The endogenous retroviral locus ERVWE1 is a bona fide gene involved in hominoid placental physiology. *Proc. Natl Acad. Sci. USA* **101**, 1731–1736. (doi:10.1073/pnas.0305763101)
11. Feschotte C, Gilbert C. 2012 Endogenous viruses: insights into viral evolution and impact on host biology. *Nat. Rev. Genet.* **14**, 283–296. (doi:10.1038/nrg3199)
12. Gilbert C, Feschotte C. 2010 Genomic fossils calibrate the long-term evolution of hepadnaviruses. *PLoS Biol.* **8**, e1000495. (doi:10.1371/journal.pbio.1000495)
13. Katzourakis A, Gifford RJ, Tristem M, Gilbert TP, Pybus OG. 2009 Macroevolution of complex retroviruses. *Science* **325**, 1512. (doi:10.1126/science.1174149)
14. Keckesova Z, Ylinen LMJ, Towers GJ, Gifford RJ, Katzourakis A. 2009 Identification of a RELIK orthologue in the European hare (*Lepus europaeus*) reveals a minimum age of 12 million years for the lagomorph lentiviruses. *Virology* **384**, 5. (doi:10.1016/j.virol.2008.10.045)
15. Han G, Worobey M. 2012 An endogenous foamy-like viral element in the coelacanth genome. *PLoS Pathog.* **8**, e1002790. (doi:10.1371/journal.ppat.1002790)
16. Katzourakis A, Gifford RJ. 2010 Endogenous viral elements in animal genomes. *PloS Genet.* **6**, e1001191. (doi:10.1371/journal.pgen.1001191)
17. Gershenson S. 1928 A new sex-ratio abnormality in *Drosophila obscura*. *Genetics* **13**, 488–507.
18. Jurka J, Zietkiewicz E, Labuda D. 1995 Ubiquitous mammalian-wide interspersed repeats (MIRs) are molecular fossils from the Mesozoic era. *Nucleic Acids Res.* **23**, 170–175. (doi:10.1093/nar/23.1.170)
19. Jurka J. 1990 Novel families of interspersed repetitive elements from the human genome. *Nucleic Acids Res.* **18**, 137–141. (doi:10.1093/nar/18.1.137)
20. Labuda D, Striker G. 1989 Sequence conservation in Alu evolution. *Nucleic Acids Res.* **17**, 2477–2491. (doi:10.1093/nar/17.7.2477)
21. Smit AFA. 1993 Identification of a new, abundant superfamily of mammalian LTR-transposons. *Nucleic Acids Res.* **21**, 1863–1872. (doi:10.1093/nar/21.8.1863)
22. Volff J, Körting C, Altschmied J, Duschl J, Sweeney K, Wichert K, Froschauer A, Schartl M. 2001 Jule from the fish *Xiphophorus* is the first complete vertebrate *Ty3/Gypsy* retrotransposon from the Mag family. *Mol. Biol. Evol.* **18**, 101–111. (doi:10.1093/oxfordjournals.molbev.a003784)
23. McEntyre J, Ostell J (eds) 2002 The NCBI handbook (Internet). Bethesda, MD: National Center for Biotechnology Information. See http://www.ncbi.nlm.nih.gov/books/NBK21101/.
24. Edgar RC. 2004 MUSCLE: multiple sequence alignment with high accuracy and high throughput. *Nucleic Acids Res.* **32**, 1792–1797. (doi:10.1093/nar/gkh340)
25. Smit AFA, Hubley R, Green P. 1996–2010 RepeatMasker Open-3.0. See http://www.repeatmasker.org.
26. Tamura K, Peterson D, Peterson N, Stecher G, Nei M, Kumar S. 2011 MEGA5: molecular evolutionary genetics analysis using maximum likelihood, evolutionary distance, and maximum parsimony methods. *Mol. Biol. Evol.* **28**, 2731–2739. (doi:10.1093/molbev/msr121)
27. Rice P, Longden I, Bleasby A. 2000 EMBOSS: the European molecular biology open software suite. *Trends Genet.* **16**, 276–277. (doi:10.1016/S0168-9525(00)02024-2)
28. Kosakovsky Pond SL, Frost SDW. 2005 Not so different after all: a comparison of methods for detecting amino acid sites under selection. *Mol. Biol. Evol.* **22**, 1208–1222. (doi:10.1093/molbev/msi105)
29. Suzuki Y, Gojobori T. 1999 A method for detecting positive selection at single amino acid sites. *Mol. Biol. Evol.* **16**, 1315–1328. (doi:10.1093/oxfordjournals.molbev.a026042)
30. Brown PO, Bowerman B, Varmus HE, Bishop JM. 1987 Correct integration of retroviral DNA *in vitro*. *Cell* **49**, 347–356. (doi:10.1016/0092-8674(87)90287-X)
31. Eizirik E, Murphy WJ, O'Brien J. 2001 Molecular dating and biogeography of the early placental mammal radiation. *Am. Genet. Assoc.* **92**, 212–219.
32. Delsuc F, Scally M, Madsen O, Stanhope MJ, de Jong Catzeflis FM, Springer MS, Douzery EJP. 2002 Molecular phylogeny of living Xenarthrans and the impact character and taxon sampling on the placental tree rooting. *Mol. Biol. Evol.* **19**, 1656–1671. (doi:10.1093/oxfordjournals.molbev.a003989)
33. Jones KE, Bininda-Emonds ORP, Gittleman JL. 2005 Bats, clocks and rocks: diversification patterns in Chiroptera. *Evolution* **59**, 2243–2255.
34. Perelman P *et al*. 2011 A molecular phylogeny of living primates. *PLoS Genet.* **7**, e1001342. (doi:10.1371/journal.pgen.1001342)
35. Bininda-Emonds ORP *et al*. 2007 The delayed rise of present-day mammals. *Nature* **446**, 507–512. (doi:10.1038/nature05634)
36. Springer MS, Cleven GC, Madsen O, de Jong WW, Waddell VG, Amrine HM, Stanhope MJ. 1997 Endemic African mammals shake the phylogenetic tree. *Nature* **388**, 61–64. (doi:10.1038/40386)
37. Bininda-Emonds ORP *et al*. 2007 The delayed rise of present-day mammals. *Nature* **446**, 507–551. (doi:10.1038/nature05634)
38. Tobe SS, Kitchener AC, Linacre AMT. 2010 Reconstructing mammalian phylogenies: a detailed comparison of the cytochrome *b* and cytochrome oxidase subunit I mitochondrial genes. *PLoS ONE* **5**, e14156. (doi:10.1371/journal.pone.0014156)
39. Katzourakis A, Tristem M, Pybus OG, Gifford RJ. 2007 Discovery and analysis of the first endogenous lentivirus. *Proc. Natl Acad. Sci. USA* **104**, 6261–6265. (doi:10.1073/pnas.0700471104)
40. Waterston RH *et al*. 2002 Initial sequencing and comparative analysis of the mouse genome. *Nature* **420**, 520–562. (doi:10.1038/nature01262)
41. Belshaw R, Pereira V, Katzourakis A, Talbot G, Paces J, Burt A, Tristem M. 2004 Long-term reinfection of the human genome by endogenous retroviruses. *Proc. Natl Acad. Sci. USA* **101**, 4894–4899. (doi:10.1073/pnas.0307800101)
42. Johnson WE, Coffin JM. 1999 Constructing primate phylogenies from ancient retrovirus sequences. *Proc.*

Downloaded from http://rstb.royalsocietypublishing.org/ on November 29, 2014

Natl Acad. Sci. USA **96**, 10 254–10 260. (doi:10.1073/pnas.96.18.10254)

43. Zehr SM, Nedbal MA, Flynn JJ. 2001 Tempo and mode of evolution in an orthologous *Can* SINE. *Mammalian Genome* **12**, 38–44. (doi:10.1007/s003350010224)
44. Murata S, Takasaki N, Saitoh M, Tachida H, Okada N. 1996 Details of retropositional genome dynamics that provide a rationale for a generic division: the distinct branching of all the Pacific salmon and trout (*Oncorhynchus*) from the Atlantic salmon and trout (*Salmo*). *Genetics* **142**, 915–926.
45. Cook JM, Tristem M. 1997 'SINEs of the times' — transposable elements as clade markers for their hosts. *Tree* **12**, 295–297. (doi:10.1016/s0169-5347(97)01121-X)
46. Churakov G, Grundmann N, Kuritzin A, Brosius J, Makalowski W, Schmitz J. 2010 A novel web-based TinT application and the chronology of the Primate *Alu* retroposon activity. *BMC Evol. Biol.* **10**, 376. (doi:10.1186/1471-2148-10-376)
47. Lynch C, Tristem M. 2003 A co-opted *gypsy*-type LTR-retrotransposon is conserved in the genomes of humans, sheep, mice, and rats. *Curr. Biol.* **13**, 1518–1523. (doi:10.1016/S0960-9822(03)00618-3)
48. Blaise S, Parseval N, Heidmann T. 2005 Functional characterization of two newly identified human endogenous retrovirus coding envelope genes. *Retrovirology* **2,** 19**.** (doi:10.1186/1742-4690-2-19)
49. Belshaw R, Katzourakis A. 2005 BlastAlign: a program that uses blast to align problematic nucleotide sequences. *Bioinformatics* **21**, 122–123. (doi:10.1093/bioinformatics/bth459)

APPENDIX B

The electronic supplementary information for the publication in appendix A is available online, at the following address:

http://rstb.royalsocietypublishing.org/content/368/1626/20120503.figures-only

Publication details:

DOI: 10.1098/rstb.2012.0503.

The alignments cited in this thesis are titled as follows:

- Flagged, shortened alignment: Main Alignment with Dels.png
- Unshortened alignment: Alignment_w_insertions.fasta
- 5′ flank alignment: 5′ flanks + LTRs.fasta
- 3′ flank alignment: 3′ flanks + LTRs.fasta

APPENDIX C

Low-Coverage Denisova Genome Screen.

1. Materials and Methods:

1.1 Data Preparation and Formatting: The genome sequence data (Reich *et al*., 2010) used for the low-coverage Denisovan screen can be found deposited in the European Nucleotide Archive (ENA) under the study accession number ERP000318. Data was compressed in Sequence Read Archive (SRA) format, which is not directly suitable for sequence analyses. These SRA files firstly needed to undergo conversion to a more conventional FASTQ format using the 'Fastq.dump' script as part of NCBI's SRA toolkit. FASTQ files were then converted to FASTA format, which is the native format handled by BLAST and many other bioinformatic tools. FASTQ to FASTA conversions were then perfomed using executable scripts as part of the FAST-X Toolkit.

The 1.9-fold Denisovan genome contains 47 files, comprising in excess of 80,000,000 sequences. Sequence reads were short, averaging 58-74 bp, and were built into a BLAST database using CLC's Main Workbench on a high specification personal desktop computer comprising an Intel Core i7 quad-core 3.4 gigahertz (GHz) processor, 16 GB of random access memory (RAM) and 2.25 terabytes (TB) of local storage.

1.2 BLAST: In order to identify HERV-K(HML2) group retroviral insertions, a probe sequence comprising the first 25 bp of the 5′ HERV-K113 LTR was used. BLAST searches were performed using BLASTN within CLC Main WorKbench, set to a default word size of 11 bp. In order to extract full-length reads from small matches, a BLAST database was built using CLC Main Workbench. Matches were then filtered by the size of the matching region, and by percentage identities (ID). While examining results, only hits with 100% IDs were analysed. Using a perl script), reads were filtered further, based on the number of bases of flanking sequence upstream of start of the viral LTR. Reads that contained 20 bp of flanking sequence or more were kept; those with fewer than 20 bp of flank were discarded. This allowed for manual verification (see 3.2A.3) in UCSC's BLAT, which requires a minimum probe size of 20 bp.

1.3 BLAT Verification: The following verification steps were used to identify proviruses in Denisova that are absent from the human reference sequence. To do this, reads containing viral LTR were verified using UCSC's BLAT alignment tool, with individual Denisovan reads being searched against the hg19 assembly of the human genome. Since reads contain the virus-host junctions, with both hominid genomic sequence and retroviral LTR, they could be BLAT-searched in full against hg19 using default parameters. BLAT was chosen, as its user interface allows easy viewing of the DNA surrounding each match. It was therefore possible to manually inspect the matches for retroviral LTR sequence at the terminal end of the flank sequences. Reads

which did not match hg19 across the virus-host junction were considered as candidate novel HERVs. All reads were manually examined in this manner.

1.4 Read Count: In order to confirm that candidate novel HERVs did not arise as a result of sequencing artefacts, they were checked for multiple representation in genome sequence reads. This was done by BLAST-searching the candidate novel HERVs against the Denisovan genome. Since the reads are already very short, a truncated 22 bp region surrounding the virus-host DNA junction was used as a probe. This comprised 11 bp of flank, and 11 bp of LTR sequence. Multiply represented candidate HERVs were recorded.

1.5 Back-Filtering: Candidate 5′ reads that pass the initial BLAT verification are further confirmed by retrieving the corresponding 3′ end of the viral insertion. This process is termed back-filtering (Aiewsakun & Brockman, 2011), and involves creating artificial constructs from a fragment of sequence obtained from the human genome (hg19), and appending this to a fragment of the furthermost 3′ end of the HERV-K113 LTR.

The first step in this procedure was to BLAT search the human genome using just the 5′ flank from a candidate Denisovan read. This located the corresponding empty preintegration site in the human genome. The sequence directly downstream of the end of this match would, in the event of a HERV insertion, flank the 3′ end of the virus. Therefore, 11 bp of this was extracted, and was appended downstream of the last 11 bp of the 3′ end of the HERV-K113 LTR. The resulting construct is similar to the expected sequence of the 3′ HERV provirus-host DNA junction. Given the limitation of small sequence reads in the 1.9-fold Denisovan genome, a 22 bp construct best satisfied a compromise between being able to retrieve viruses and missing them altogether.

The artificial 22 bp constructs were then BLAST searched against the Denisovan genome, using a small word size of 5 bp. Given that a virus creates a target site duplication (TSD) upon integration, where a 4-7 bp fragment of the 5′ flank is repeated at the 3′ end, a word size of 5 bp would allow BLAST to successfully match most of the read, allowing for a potential region of poorer homology at the TSD.

Reads that match across the virus-host junction are considered to have passed the back-filter. If more than 11 bp of 3′ flanking sequence is available, this can be further checked by BLAT searching against hg19.

1.6 Matching TSDs: As a final verification step for HERV insertions for which both ends of the element were obtained, the TSDs directly adjacent to the LTR can be manually checked for similarity. They are normally visibly homologous, and are often identical, but sometimes acquire point mutations.

2. Results:

2.1 Low-Coverage Denisova Screen: The low-coverage Denisovan genome was screened by using BLAST to identify HERV sequences, and filtering to retain only 100% ID-scored matches. This resulted in 1,702 hits, all of which were rigorously verified *in silico*. Matches that weren't back-filtered successfully, were not included in these results, which are only representative of elements identified at both ends. This identified a total of 4 unique, novel HERV loci, further verified by the observation of matching TSDs in 5′ and 3′ flanks.

Table 1: **A table of the four novel proviruses** identified in the low-coverage Denisovan screen. 5′ flanking sequences are given, together with their length, orientation in the Denisovan genome, and the BLAST match ID scores. The first three proviruses in this table were published by Agoni *et al*. (2012) before I had completed my screen. The fourth HERV was missed from their study, however. It was subsequently identified in the high-coverage genomes for Denisova and Neanderthal, and published in Lee *et al*. (2014).

Match No.	Flank Sequence (forward orientation)	Flank Length (bp)	Original Match Orientation	BLAST match ID (%)
1	TTGACCAGGCTGGTCTTGAACTC CTGACCTCATGTGATCCACCTGT CTCCGCCTCCC	56	Forward	100
2	TACTTAGATTTCTATTTCTTCTTGA GTCAGTTTCAGTCATGTCTGTATT TCTAGAAATTTGTCCATTTT	69	Forward	100
3	CTGGGATTATAGGTGTGTGCCAC	23	Forward	100
4	AACACACAGAAAATATGACAGTG AACACAGAGAAAAGATGATGGTA CACCCAGAGAAAAGATGATGGTA CGCC	73	Reverse	100

Of these 4 proviruses, the first three were published by Agoni *et al*. (2012). They used very similar procedures, verifying with BLAT and using a minimum of 20 bp of flank, yet they missed the fourth provirus presented in Table 1 (two represenative reads are also shown in Figure 1). In my results, I verified both ends of this virus using back-filtering.

Since Agoni *et al*. (2012) published their data before my screen had finished, I decided to resume this work by screening the higher-coverage Denisovan genome that was made available shortly afterwards. Subsequently, provirus 4 was identified in this more in-depth study.

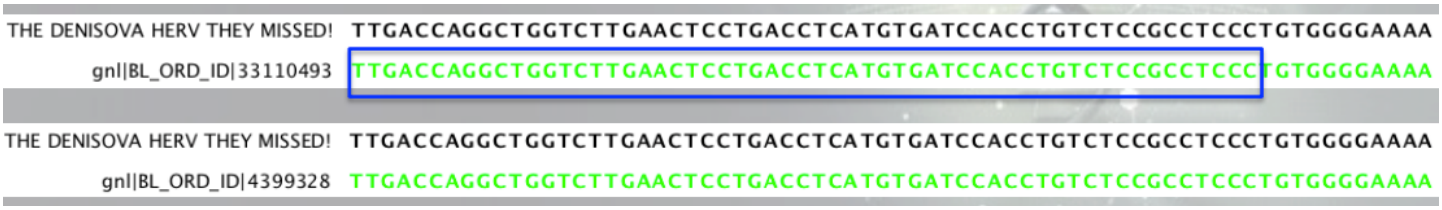


Figure 1: **The missing HERV** identified in the low-coverage Denisovan retroviral screen that was not picked up by Agoni *et al*. (2012). This was subsequently identified in my second high-coverage genome screen.

APPENDIX D

Published manuscript corresponding to the work presented in chapter 3.

Lee, A., Huntley, D., Aiewsakun, P., Kanda, R. K., Lynn, C. & Tristem, M. (2014) Novel Denisovan and Neanderthal Retroviruses. *J. Virol.* 88 (21), 12907-12909.

JVI
Journals.ASM.org

Novel Denisovan and Neanderthal Retroviruses

Adam Lee,[a] Derek Huntley,[a] Pakorn Aiewsakun,[b]* Ravinder K. Kanda,[b]* Claire Lynn,[a] Michael Tristem[b]

Imperial College London, South Kensington Campus, London, United Kingdom[a]; Imperial College London, Silwood Park Campus, Ascot, Berkshire, United Kingdom[b]

Following the recent availability of high-coverage genomes for Denisovan and Neanderthal hominids, we conducted a screen for endogenized retroviruses, identifying six novel, previously unreported HERV-K(HML2) elements (HERV-K is human endogenous retrovirus K). These elements are absent from the human genome (hg38) and appear to be unique to archaic hominids. These findings provide further evidence supporting the recent activity of the HERV-K(HML2) group, which has been implicated in human disease. They will also provide insights into the evolution of archaic hominids.

In 2008, an archaeological dig at a cave in the Siberian Altai mountain range led to the discovery of a finger bone belonging to a female hominid, dating to at least ~50,000 years ago (1, 2). From this, the DNA of a subspecies of *Homo sapiens*, designated Denisovans, was sequenced (1). Similarly, a draft Neanderthal genome—a sister group to Denisovans—was sequenced from three individuals in 2010 (3, 4). Using this data, Agoni et al. (5) identified 14 novel human endogenous retrovirus K (HERV-K) proviruses, which were absent from the human genome sequence (assembly hg19). The authors suggested that these HERVs were unique to archaic hominids and that no orthologous insertions would be found in modern humans (5). Subsequently, however, Marchi et al. reported that all of these sequences were actually present, or likely to be present, in some modern humans (6).

In this study, we screened the most recently available high-coverage genomes for a Denisovan (4) and an Altai Neanderthal (2) for HERV-K proviruses. We present six novel, endogenized retroviruses, absent from the hg38 human genome, 43 modern-human genomes reported by Lee et al. (7), and a further 358 reported by Marchi et al. (8). These may therefore represent the first proviruses unique to Neanderthal and Denisovan hominids.

While endogenized retroviral DNA makes up ~8% of the human genome, only one group—HERV-K(HML2)—appears to have been active within the past million years. This has been demonstrated by the observation that some members of this group, but not others, are insertionally polymorphic, having been identified in some modern humans (9). Although no active, replication-competent HERV-K(HML2) elements have been identified to date, it remains possible that such elements exist and may cause disease in some modern humans.

The high-coverage Neanderthal and Denisovan genomes screened in this study were both derived from fossils found in Denisova Cave (2, 4). These genome sequences consist of short, unassembled DNA reads averaging ~70 to 200 bp and were sequenced to 52- and 30-fold coverage, respectively (versus 1.3-fold [3] and 1.9-fold [1] coverage for the genomes screened previously by Agoni et al. [5]). They therefore likely represent almost-complete genome coverage.

Novel retroviral insertions in archaic hominids can be recognized when orthologous flanking DNA in modern humans is not interrupted by a HERV insertion, manifesting instead as an empty preintegration site. We obtained reads containing 5′ host-virus junctions using a perl script that stringently detected the first 20 bp of the start of the HERV-K(HML2) long terminal repeat (LTR), allowing us to build libraries of reads containing hominid and viral DNA. Flanking sequences were then extracted and BLAST searched against the human genome (hg38), using blastn and a word size of 11. Apparent novel HERV-K(HML2) insertions were identified by a lack of retroviral sequence downstream of a matching flank in the modern-human genome.

To confirm that putative novel insertions were not a result of sequencing artifacts, such as template switching, we used three approaches. First, we confirmed that each retrovirus was represented by multiple reads, as this would be unlikely to occur in the event of sequencing error. Second, we attempted to identify the corresponding 3′ flanks for each candidate provirus. This involved obtaining the modern-human sequence directly downstream of the flank-virus breakpoint and locating matching sequence in the Denisovan and Neanderthal genomes. Matching reads containing LTR sequence directly upstream of the 3′ flank were extracted using BLAST, utilizing word sizes of 5 to 7. This enabled matches to be returned, despite small mismatches occurring due to target site duplications (TSDs) at the virus-host junction. Lastly, the presence of matching TSDs was considered additional verification of the virus. Conversely, we then repeated each of these steps with the 3′ end of the HERV-K(HML2) LTR to identify 3′ host-virus junctions. Sequence reads are given in Table S1 in the supplemental material.

Using this approach, we identified a total of nine HERV-K(HML2) proviruses present within the Neanderthal and Denisovan genomes—while also absent from hg38—that were not reported by Agoni et al. (5). However, their absence from the hg38 sequence does not necessarily imply absence from all modern humans; such elements could be insertionally polymorphic, as demonstrated by Marchi et al. (6). While we did not directly screen further modern-human genomes, we compared our nine ele-

Received 22 June 2014 **Accepted** 18 August 2014
Published ahead of print 20 August 2014
Editor: K. L. Beemon
Address correspondence to Adam Lee, adam.lee07@imperial.ac.uk.
* Present address: Pakorn Aiewsakun and Ravinder K. Kanda, Department of Zoology, University of Oxford, Oxford, United Kingdom.
Supplemental material for this article may be found at http://dx.doi.org/10.1128/JVI.01825-14.
Copyright © 2014 Lee et al. This is an open-access article distributed under the terms of the Creative Commons Attribution 3.0 Unported license.
doi:10.1128/JVI.01825-14

Downloaded from http://jvi.asm.org/ on November 29, 2014 by Imperial College London Library

A

HERV Insertion	Neanderthal		Denisovan		hg38		Lee *et al.* / Marchi *et al.*		Breakpoint Locus (Mapped to hg38)
	5′	3′	5′	3′	5′	3′	5′	3′	
HERV-K-DeNe1	●	●	●	●	○	○	○	○	chr19:2890621 (Alu)
HERV-K-DeNe2	●	●	●	●	○	○	○	○	chr18:30138227
HERV-K-DeNe3	N.D.	●	N.D.	●	○	○	○	○	chrUn_K1270749v1:63347
HERV-K-De13	○	N.D.	●	N.D.	○	○	○	○	chr4:117182160
HERV-K-Ne4	●	●	○	○	○	○	○	○	chr8:74432832
HERV-K-Ne5	●	●	○	○	○	○	○	○	chr19:57485571
HERV-K-Ne6**	●	●	○	○	○	○	●	●	chr12:43919858 (L1)
HERV-K-Ne7**	●	●	○	○	○	○	●	●	chr5:65092618
HERV-K-De14*	○	○	●	●	○	○	●	●	chr20:12421693

B

HERV Insertion	+/-	5′ Flanking Sequence	Endogenous Retrovirus 5′ 3′	3′ Flanking Sequence
HERV-K-DeNe1	+	~GATCCACCTGTCTCCGC**CTCCC**	TGTGGGGA - - - CCCCTACA	**CTCCC**AAAGTGCTGGGATGACA~
HERV-K-DeNe2	+	~GTATAATTCTACTAACC**CAAGTA**	TGTGGGGA - - - CCCCTACA	**CAAGTA**AAACCAAAAATTATTA~
HERV-K-DeNe3	+		- - - CCCCTACA	AAGAAGTGTGAAAAAAGAAATA~
HERV-K-De13	+	~TAACAGGCTCTGAAATTGTAGC	TGTGGGGA - - -	
HERV-K-Ne4	+	~ATCACAGTTTTTAAATTA**ACAAT**	TGTGGGGA - - - CCCCTACA	**ACAAT**AATATTGATACCAACAT~
HERV-K-Ne5	-	~CTTTAGGAATATTTACAT**GTCTA**	TGTGGGGA - - - CCCCTACA	**GTCTA**TAAGGAGTGTGAACATC~
HERV-K-Ne6**	+	~TGAGGTGATATCTCATT**GTGGT**	TGTGGGGA - - - CCCCTACA	**GTGGT**TTTGATTTTCATGTTCTT~
HERV-K-Ne7**	+	~AAAAAACGTAATAGTA**GACGTG**	TGTGGGGA - - - CCCCTACA	**GACGTG**ACATATGCATCACAGC~
HERV-K-De14*	-	~ATGGCTACTGGGAATTA**AGTGG**	TGTGGGGA - - - CCCCTACA	**AGTGG**GAAGGAGGATAAAGCT~

FIG 1 (A) Distribution of the archaic hominid insertions within Neanderthals and Denisovans, compared to hg38 and the HERV-K elements recovered in previous studies (6–8). Filled circles denote that an element is present, while open circles denote absence. N.D. indicates that there were no sequence data available. The loci of corresponding empty preintegration sites, mapped to hg38, are also given. If this occurred within a repeat, the class of repeat is listed in parentheses. Single asterisks adjacent to the last three viruses represent those elements that were also recovered in some modern humans by Lee et al. (7), while double asterisks represent those viruses also recovered by Marchi et al. (8). The distribution of these three elements is likely explained by ancestral polymorphism. (B) HERV-K(HML2) flanking sequences for the nine endogenized retroviruses identified here. The 5′ and 3′ flanking regions are shown, together with the proximal and distal ends of the HERV LTR. Nomenclature follows the convention and numbering set by Agoni et al. (5). Reads where flanking sequence was extensive were truncated and are provided in full in Table S1 in the supplemental material. + or − denotes the native orientation of the read against hg38. For proviruses for which both 5′ and 3′ flanks were obtained, the matching TSD sequence is underlined and in boldface.

ments against those recovered by Lee et al. (7) and Marchi et al. (8) from their analyses of high-coverage modern-human genomes. This revealed that three of these viruses were present within the data of Lee et al. (7), of which two were also identified by Marchi et al. (8).

The remaining six elements therefore appear to be absent from both hg38 and the additional 401 modern-human genomes investigated in previous reports (7, 8). For four of these elements, both the 5′ and 3′ virus-host junctions were identified, while the remaining two—which were both represented by multiple sequence reads—were derived from single ends. Of the six proviruses, three were shared by both Denisovans and Neanderthals, while two were unique to Neanderthals and one to Denisovans (Fig. 1). We note that one of these proviruses, De13, is located approximately 1 kb upstream of an existing HERV-K(HML2) solo LTR in hg38; it also appears to share the same TSD, as well as a similar flanking sequence. This would normally suggest that it is a sequencing artifact. However, it is represented by multiple sequence reads in the Denisovan genome and lies within a highly repetitive region. Its flanking sequence exactly matches the region 1 kb upstream of the known solo LTR in hg38.

As a result of genetic drift, neutral HERV insertions can become fixed in a population within a time frame dependent on population size and generation time. It is estimated that the average time taken to fixation in humans is ~800,000 years (10). Since modern humans are estimated to have diverged from Denisovan and Neanderthal lineages approximately 553,000 to 589,000 years ago (2), we would expect that some—but not all—of the novel HERV-K(HML2) elements in these archaic hominids would be absent in modern humans. This is consistent with our results; six elements appear to be absent from all of the modern-human genomes investigated to date, whereas others (identified in this and previous reports [5, 7, 8]) are present within some of them. However, it remains possible that these six elements are also present in modern humans, albeit at very low allele frequencies.

We suggest that at least some of the six proviruses identified in this study inserted into archaic hominids after their divergence from modern humans; however, it is also possible that they inserted before the divergence of archaic hominids and modern humans, with these ancestral polymorphisms being subsequently lost from modern humans by genetic drift. These findings will help improve our understanding of archaic hominid evolution and provide additional insight into the recent activity of the HERV-K(HML2) retroviral group.

ACKNOWLEDGMENTS

We thank the Wellcome Trust for supporting this work and the Max Planck Institute for the provision of additional unmapped Neanderthal genome sequence data.

We also thank Steve Cook, Magda Charalambous, Andrew Drake-

Downloaded from http://jvi.asm.org/ on November 29, 2014 by Imperial College London Library

Brockman, Dominic John Bennett, and Mona Lee for their valuable insight and Suhail Islam for his assistance with data preparation and formatting.

REFERENCES

1. **Reich D, Green RE, Kircher M, Krause J, Patterson N, Durand EY, Viola B, Briggs AW, Stenzel U, Johnson PLF, Maricic T, Good JM, Marques-Bonet T, Alkan C, Fu Q, Mallick S, Li H, Meyer M, Eichler EE, Stoneking M, Richards M, Talamo S, Shunkov MV, Derevianko AP, Hublin J, Kelso J, Slatkin M, Pääbo S.** 2010. Genetic history of an archaic hominin group from Denisova Cave in Siberia. Nature **468:**1053–1060. http://dx.doi.org/10.1038/nature09710.
2. **Prüfer K, Racimo F, Patterson N, Jay F, Sankararaman S, Sawyer S, Heinze A, Renaud G, Sudmant PH, Filippo C, Li H, Mallick S, Dannemann M, Fu Q, Kircher M, Kuhlwilm M, Lachmann M, Meyer M, Ongyerth M, Siebauer M, Theunert C, Tandon A, Moorjani P, Pickrell J, Mullikin JC, Vohr SH, Green RE, Hellman I, Johnson PLF, Blanche H, Cann H, Kitzman JO, Shendure J, Eichler EE, Lein ES, Bakken TE, Golovanova LV, Doronichev VB, Shunkov MV, Derevianko AP, Viola B, Slatkin M, Reich D, Kelso J, Pääbo S.** 2014. The complete genome sequence of a Neanderthal from the Altai Mountains. Nature **505:**43–49. http://dx.doi.org/10.1038/nature12886.
3. **Green RE, Krause J, Briggs AW, Maricic T, Stenzel U, Kricher M, Patterson N, Heng L, Zhai W, Fritz MH, Hansen NF, Durand EY, Malaspinas A, Jensen JD, Marques-Bonet T, Alkan C, Prüger K, Meyer M, Burbano HA, Good JM, Schultz R, Aximu-Petri A, Butthof A, Höber B, Höffner B, Siegemund M, Weihmann A, Nusbaum C, Lander ES, Russ C, Novod N, Affourtit J, Egholm M, Verna C, Rudan P, Brajkovic D, Kucan Z, Gusic I, Doronichev VB, Golovanova LV, Lalueza-Fox C, Rasilla M, Fortea J, Rosas A, Schmitz RW, Johnson PLF, Eichler EE, Falush D, Birney E, Mullikin JC, Slatkin M, Nielsen R, Kelso J, Lachmann M, Reich D, Pääbo S.** 2010. A draft sequence of the Neandertal genome. Science **328:**710–722. http://dx.doi.org/10.1126/science.1188021.
4. **Meyer M, Kircher M, Gansauge M, Li H, Racimo F, Mallick S, Schraiber JG, Jay F, Prüfer K, Filippo C, Sudmant PH, Alkan C, Fu Q, Do R, Rohland N, Tandom A, Siebauer M, Green RE, Bryc K, Briggs AW, Stenzel U, Dabney J, Shendure J, Kitzman J, Hammer MF, Shunkov MV, Derevianko AP, Patterson N, Andrés AM, Eichler EE, Slatkin M, Reich D, Kelso J, Pääbo S.** 2012. A high-coverage genome sequence from an archaic Denisovan individual. Science **338:**222–226. http://dx.doi.org/10.1126/science.1224344.
5. **Agoni L, Golden A, Guha C, Lenz J.** 2012. Neandertal and Denisovan retroviruses. Curr. Biol. **22:**R437–R438. http://dx.doi.org/10.1016/j.cub.2012.04.049.
6. **Marchi E, Kanapin A, Byott M, Magiorkinis G, Belshaw R.** 2013. Neanderthal and Denisovan retroviruses in modern humans. Curr. Biol. **23:**R994–R995. http://dx.doi.org/10.1016/j.cub.2013.10.028.
7. **Lee E, Iskow R, Yang L, Gokcumen O, Haseley P, Luquette LJ, III, Lohr JG, Harris CC, Ding L, Wilson RK, Wheeler DA, Gibbs RA, Kucherlapati R, Lee C, Kharchenko PV, Park PJ, The Cancer Genome Atlas Research Network.** 2012. Landscape of somatic retrotransposition in human cancers. Science **337:**967–971. http://dx.doi.org/10.1126/science.1222077.
8. **Marchi E, Kanapin A, Magiorkinis G, Belshaw R.** 2014. Unfixed endogenous retroviral insertions in the human population. J. Virol. **88:**9529–9537. http://dx.doi.org/10.1128/JVI.00919-14.
9. **Belshaw R, Dawson ALA, Woolven-Allen J, Redding J, Burt A.** 2005. Genomewide screening reveals high levels of insertional polymorphism in the human endogenous retrovrius family HERV-K(HML2): implications for present-day activity. J. Virol. **79:**12507–12514. http://dx.doi.org/10.1128/JVI.79.19.12507-12514.2005.
10. **Kimura M.** 1984. The neutral theory of molecular evolution. Cambridge University Press, Cambridge, United Kingdom.

Downloaded from http://jvi.asm.org/ on November 29, 2014 by Imperial College London Library

APPENDIX E

This presents the stringent-matching perl script.

```
#! usr/bin/perl -w

#Stringent_Matcher.pl

# (C) Adam Lee 2014

#This is a stringent matcher, designed to catch 100% matches to LTR probes.
Make sure you specify the genomic FASTA filepath in the input
#of the script itself, below on line 8. Same for the output.
#You can interchange the amount of K113 LTR you want to use, by pasting it in to the
#splitseqs regex in each of the four for loops.

#Script can be used to filter existing 100% ID results by flank amount, or to mine genomes
#directly.

#_____________________________________________________________________________________________
_____
mkdir "Hits";

@files = <NEANDERTAL_DIRECTORY_FILEPATH_HERE/*>;

open STDOUT, '>>', "FILEPATH_HERE/3_Prime_LTR_Hits.txt";

foreach $file (@files)
        {
       #print $file . "\n";
       open (A,  $file) || die $!;
       @array1 = <A>;
       @split = split (/>/, "@array1");
       $splitseqs = join (">", @split);

       for ( $splitseqs =~ / (.{20,}TGTGGGGAAAAGCAAGAGAG.*)/g )
               {
                 print ">FORWARD_5'_LTR_match ". $nr++ . "\n";
                 print $_, "\n";
               }


       for ( $splitseqs =~ / (.*CTCTCTTGCTTTTCCCCACA.{20,})/g )

               {
                 print ">REVERSE_3'_LTR_match ". $nr++ . "\n";
                 print $_, "\n";
               }

       for ( $splitseqs =~ / (.{20,}TGTAGGGGTGGGTTGCCCCT.*)/g )
               {
                 print ">REVERSE_3'_LTR_match ". $nr++ . "\n";
                 print $_, "\n";
               }


       for ( $splitseqs =~ / (.*AGGGGCAACCCACCCCTACA.{20,})/g )

               {
                 print ">FORWARD_3'_LTR_match ". $nr++ . "\n";
                 print $_, "\n";
               }

       }
```

APPENDIX F

This is the script that strips LTR away from a sequence read, leaving behind just flanking sequence. Currently set to prepare 5′ flanks. Enter 3′ LTR to prepare 3′ flanks, as required.

```
#! usr/bin/perl -w

#Script for preparing flanks, and stripping away LTR. Set thresholds as desired for application.

#_____________________________________________________________________________________________
_____
mkdir "Flanks";

@files = </Users/Adam/Desktop/Final_Pipeline/Hits/*.txt>;

open STDOUT, '>>', "/Users/Adam/Desktop/Final_Pipeline/Flanks/SampleFlanks.txt";

foreach $file (@files)

        {
        open STDOUT, '>', "/Users/Adam/Desktop/TestPerl/Flanks/SampleResult.txt";
        open (A, $file) || die $!;
        @seqs = <A>;

        @split = split (/>/, "@seqs");
        $splitseqs = join (">", @split);

#Note line below. The '{4,}' includes newline character. Always make this 1 interger bigger than
desired threshold
        while ( $splitseqs =~ /.FORWARD 5' LTR match ([0-9])\n(.{8,})TGTGGGGAAAA.*/g )

                {
                    my $numbering = $1;
                    my $flankingsequence = $2;
                    print ">FORWARD 5' Flank ",$numbering,"\n",$flankingsequence, "\n";
                }

        while ( $splitseqs =~ /.REVERSE 5' LTR match ([0-9])\n.*TTTTCCCCACA(.{8,})/g )

                {
                    my $revnumbering = $1;
                    my $revflankingsequence = $2;
                    print ">REVERSE 5' Flank ",$revnumbering,"\n",$revflankingsequence,"\n";
                }

        }

```

APPENDIX G

Perl script for HERV mining using the fuzzy-matching algorithm. This script was amended by Dr. Steve Cook.

```
#!/usr/bin/perl

#Script amended by Dr. Steve Cook for more efficient operation.

#Script for fuzzy matching LTR, using String::Approx CPAN module. Set as required.

use strict;
use warnings;

use String::Approx 'amatch';

# To use this, type into the terminal:
# sudo perl RevisedPipeline2.pl DIRECTORY
# _____________________________PROBES_____________________________

my ( $directory )   = @ARGV;
$directory        ||= '/Users/Adam/Desktop/PerlTest2014/Data';
my @files           = glob( "$directory/*.txt" );

my %probes = (
                LTRfor5PRIME => "TGTGGGGAAAAGCAAGAGAG",
                LTRrev5PRIME => "CTCTCTTGCTTTTCCCCACA",
                LTRfor3PRIME => "AGGGGCAACCCACCCCTACA",
                LTRrev3PRIME => "TGTAGGGGTGGGTTGCCCCT",
);

while ( my ( $probe_name, $probe_sequence ) = each %probes ) {
        my %fuzzy_matches;
        foreach my $file ( @files ) {
                open my $input, '<', $file
                                                or die "Can't open '$file' for reading: $!\n";
                local $/ = undef;
                my $fasta_data = <$input>;
                close $input;
                my @sequences = grep { $_ !~ /^\s*$/ } split />/, $fasta_data;
                for my $sequence ( @sequences ) {
                        my ( $id, $dna ) = $sequence =~ /(.+)\n(.+)/;
                        $dna =~ s/\s//g;
                        my ( $has_matched ) = amatch( $probe_sequence, $dna ); #modifier can be a
dded to change error rate here
                        if ( $has_matched ) {
                                $fuzzy_matches{ ">$probe_name-$id" } = $has_matched;
                        }
                }
        }
        if ( -e "$probe_name.txt" ) {
                die "The output file name already exists. Please manually delete it!";
        }
        open my $output, '>', "$probe_name.txt"
                                or die "Can't open '$probe_name.txt' for writing: $!\n";
        while ( my ( $id, $dna ) = each %fuzzy_matches ) {
                print { $output } "$id\n$dna\n";
        }
        close $output;
}
```

APPENDIX H

Perl script used to strip away 100% matches from raw fuzzy-match output, leaving only reads where mismatching occurs in the LTR.

```
#! usr/bin/perl -w

#After fuzzy matching has been performed, this script removes 100% matches, just to leave fuzzy matches behind.

#IMPORTANT. You need to change the output file name at the bottom of the script to match that of the input.
#To run, replace the LTR in the regex with the appropriate 5'/3' rev/for LTR probe, for the file you're using / generating.

@input = </Users/Adam/Desktop/FuzzyDenisovaMatches/LTRrev5PRIME.txt>;

$filepath = "/Users/Adam/Desktop/FuzzyDenisovaMatches/";

$test = 0;

foreach $inputline (@input)
{
    #print $inputline . "\n";
    open (A,  $inputline) || die $!;
    my @splitinputfile = <A>;
    my @splitlines = split (/>/, "@splitinputfile");
    $splitdata = join (">", @splitlines);
}

while ($splitdata =~ /(\>.*)\n(.*)\n/g)
{
    $gi = $1;
    $sequence = $2;
    #$tomod = "$gi\n$sequence\n";
    #push (@tomodinput, $tomod);

if ($sequence !~ /.*CTCTCTTGCTTTTCCCCACA.*/g)
    {
        $sequence =~ s/\s//;
        $deletion = "$gi\n$sequence\n";
        push (@fuzzymatches, $deletion);
        ##@tomodlinput = grep {!/$deletion/} @tomodinput;
        $counter = $test++;
    }
}

print @fuzzymatches, "\n\nLoops run: $counter.\n\n";

open (A, ">$filepath/LTRrev5FUZZY.txt");
print A @fuzzymatches;
close (A);
```

APPENDIX I

This is the fuzzy-match filtering script. This allows the user to filter away hits whose flank doesn't correspond to the inputted amount, vastly reducing the number of obtained hits, facilitating a more concise set of candidates for subsequent verification.

```
#! usr/bin/perl -w

#SCRIPT TO OBTAIN FUZZY MATCHES WITH ≥8 BP OF FLANK.

##############################################################################################
##############################################################################################
#SET INPUT AND OUTPUT FILEPATHS HERE:

@FuzzyMatchFile = </Users/Adam/Desktop/FuzzyDenisovaMatches/LTRrev3FUZZY.txt>;
$outputpath = "/Users/Adam/Desktop/FuzzyDenisovaMatches";

#YOU MUST SET THESE THREE SCALARS AS FOLLOWS:
$execute = "auto"; #IF YOU USED MY EARLIER SCRIPTS TO GENERATE FUZZY FILES, SELECT "auto". OTHERW
ISE "MANUAL", and fill out $virusend and $orientation manually.
$flankamount = 25;                  #type the number of base pairs of flank you want ahead of you
r hit. NO quotes.
$virusend = "5 PRIME";              #IF SELECTING MANUAL ONLY: type "5 PRIME" or "3 PRIME".
$orientation = "forward";           #IF SELECTING MANUAL ONLY: type "forward" or "reverse".
#NOTE: If in manual mode, you match $virusend and $orientation incorrectly to the @FuzzyMatchFile
 input, the script will STALL.
##############################################################################################
##############################################################################################

push (@starter, $virusend, $orientation);
#print @starter[1];

@FOR5 = ("TGTGGGGAAAAGCAAGAGAG\n","GGTGGGGAAAAGCAAGAGAG\n","AGTGGGGAAAAGCAAGAGAG
\n","CGTGGGGAAAAG
        CAAGAGAG\n","TTTGGGGAAAAGCAAGAGAG\n","TCTGGGGAAAAGCAAGAGAG\n","TATGGGGAAAAGCAAGAGAG
\n","TGAGGGGAA
        AAGCAAGAGAG\n","TGGGGGGAAAAGCAAGAGAG\n","TGCGGGGAAAAGCAAGAGAG\n","TGTAGGGAAAAGCAAGAGAG
\n","TGTCGG
        GAAAAGCAAGAGAG\n","TGTTGGGAAAAGCAAGAGAG\n","TGTGTGGAAAAGCAAGAGAG\n","TGTGCGGAAAAGCAAGAGAG
\n","TGT
        GAGGAAAAGCAAGAGAG\n","TGTGGTGAAAAGCAAGAGAG\n","TGTGGCGAAAAGCAAGAGAG
\n","TGTGGAGAAAAGCAAGAGAG\n","
        TGTGGGTAAAAGCAAGAGAG\n","TGTGGGCAAAAGCAAGAGAG\n","TGTGGGAAAAAGCAAGAGAG
\n","TGTGGGGCAAAGCAAGAGAG\n
        ","TGTGGGGTAAAGCAAGAGAG\n","TGTGGGGGAAAGCAAGAGAG\n","TGTGGGGACAAGCAAGAGAG
\n","TGTGGGGATAAGCAAGAGA
        G\n","TGTGGGGAGAAGCAAGAGAG\n","TGTGGGGAACAGCAAGAGAG\n","TGTGGGGAATAGCAAGAGAG
\n","TGTGGGGAAGAGCAAG
        AGAG\n","TGTGGGGAAACGCAAGAGAG\n","TGTGGGGAAATGCAAGAGAG\n","TGTGGGGAAAGGCAAGAGAG
\n","TGTGGGGAAAAAC
        AAGAGAG\n","TGTGGGGAAAACCAAGAGAG\n","TGTGGGGAAAATCAAGAGAG\n","TGTGGGGAAAAGTAAGAGAG
\n","TGTGGGGAAA
        AGAAAGAGAG\n","TGTGGGGAAAAGGAAGAGAG\n","TGTGGGGAAAAGCTAGAGAG\n","TGTGGGGAAAAGCCAGAGAG
\n","TGTGGGG
        AAAAGCGAGAGAG\n","TGTGGGGAAAAGCATGAGAG\n","TGTGGGGAAAAGCACGAGAG\n","TGTGGGGAAAAGCAGGAGAG
\n","TGTG
        GGGAAAAGCAAAAGAG\n","TGTGGGGAAAAGCAACAGAG\n","TGTGGGGAAAAGCAATAGAG\n","TGTGGGGAAAAGCAAGTGAG
\n","T
        GTGGGGAAAAGCAAGGGAG\n","TGTGGGGAAAAGCAAGCGAG\n","TGTGGGGAAAAGCAAGACAG
\n","TGTGGGGAAAAGCAAGAAAG\n"
        ,"TGTGGGGAAAAGCAAGATAG\n","TGTGGGGAAAAGCAAGAGTG\n","TGTGGGGAAAAGCAAGAGCG
\n","TGTGGGGAAAAGCAAGAGGG
        \n","TGTGGGGAAAAGCAAGAGAT\n","TGTGGGGAAAAGCAAGAGAC\n","TGTGGGGAAAAGCAAGAGAA\n");
@REV5 = ("CTCTCTTGCTTTTCCCCACA\n","TTCTCTTGCTTTTCCCCACA\n","ATCTCTTGCTTTTCCCCACA
\n","GTCTCTTGCTTT
        TCCCCACA\n","CCCTCTTGCTTTTCCCCACA\n","CACTCTTGCTTTTCCCCACA\n","CGCTCTTGCTTTTCCCCACA
\n","CTGTCTTGC
        TTTTCCCCACA\n","CTTTCTTGCTTTTCCCCACA\n","CTATCTTGCTTTTCCCCACA\n","CTCCCTTGCTTTTCCCCACA
\n","CTCACT
        TGCTTTTCCCCACA\n","CTCGCTTGCTTTTCCCCACA\n","CTCTGTTGCTTTTCCCCACA\n","CTCTTTTGCTTTTCCCCACA
\n","CTC
        TATTGCTTTTCCCCACA\n","CTCTCCTGCTTTTCCCCACA\n","CTCTCATGCTTTTCCCCACA
```

```
\n","CTCTCGTGCTTTTCCCCACA\n","
        CTCTCTCGCTTTTCCCCACA\n","CTCTCTAGCTTTTCCCCACA\n","CTCTCTGGCTTTTCCCCACA
\n","CTCTCTTTCTTTTCCCCACA\n
        ","CTCTCTTCCTTTTCCCCACA\n","CTCTCTTACTTTTCCCCACA\n","CTCTCTTGGTTTTCCCCACA
\n","CTCTCTTGATTTTCCCCAC
        A\n","CTCTCTTGTTTTTCCCCACA\n","CTCTCTTGCATTTCCCCACA\n","CTCTCTTGCGTTTCCCCACA
\n","CTCTCTTGCCTTTCCC
        CACA\n","CTCTCTTGCTATTCCCCACA\n","CTCTCTTGCTGTTCCCCACA\n","CTCTCTTGCTCTTCCCCACA
\n","CTCTCTTGCTTGT
        CCCCACA\n","CTCTCTTGCTTATCCCCACA\n","CTCTCTTGCTTCTCCCCACA\n","CTCTCTTGCTTTGCCCCACA
\n","CTCTCTTGCT
        TTACCCCACA\n","CTCTCTTGCTTTCCCCCACA\n","CTCTCTTGCTTTTTCCCACA\n","CTCTCTTGCTTTTGCCCACA
\n","CTCTCTT
        GCTTTTACCCACA\n","CTCTCTTGCTTTTCTCCACA\n","CTCTCTTGCTTTTCGCCACA\n","CTCTCTTGCTTTTCACCACA
\n","CTCT
        CTTGCTTTTCCTCACA\n","CTCTCTTGCTTTTCCGCACA\n","CTCTCTTGCTTTTCCACACA\n","CTCTCTTGCTTTTCCCTACA
\n","C
        TCTCTTGCTTTTCCCGACA\n","CTCTCTTGCTTTTCCCAACA\n","CTCTCTTGCTTTTCCCCTCA
\n","CTCTCTTGCTTTTCCCCCCA\n"
        ,"CTCTCTTGCTTTTCCCCGCA\n","CTCTCTTGCTTTTCCCCATA\n","CTCTCTTGCTTTTCCCCAGA
\n","CTCTCTTGCTTTTCCCCAAA
        \n","CTCTCTTGCTTTTCCCCACT\n","CTCTCTTGCTTTTCCCCACC\n","CTCTCTTGCTTTTCCCCACG\n");
  26    @FOR3 = ("AGGGGCAACCCACCCCTACA\n","TGGGGCAACCCACCCCTACA\n","GGGGGCAACCCACCCCTACA
\n","CGGGGCAACCCA
        CCCCTACA\n","ATGGGCAACCCACCCCTACA\n","ACGGGCAACCCACCCCTACA\n","AAGGGCAACCCACCCCTACA
\n","AGTGGCAAC
        CCACCCCTACA\n","AGCGGCAACCCACCCCTACA\n","AGAGGCAACCCACCCCTACA\n","AGGTGCAACCCACCCCTACA
\n","AGGCGC
        AACCCACCCCTACA\n","AGGAGCAACCCACCCCTACA\n","AGGGTCAACCCACCCCTACA\n","AGGGCCAACCCACCCCTACA
\n","AGG
        GACAACCCACCCCTACA\n","AGGGGTAACCCACCCCTACA\n","AGGGGGAACCCACCCCTACA
\n","AGGGGAAACCCACCCCTACA\n","
        AGGGGCTACCCACCCCTACA\n","AGGGGCCACCCACCCCTACA\n","AGGGGCGACCCACCCCTACA
\n","AGGGGCATCCCACCCCTACA\n
        ","AGGGGCACCCCACCCCTACA\n","AGGGGCAGCCCACCCCTACA\n","AGGGGCAATCCACCCCTACA
\n","AGGGGCAAACCACCCCTAC
        A\n","AGGGGCAAGCCACCCCTACA\n","AGGGGCAACTCACCCCTACA\n","AGGGGCAACACACCCCTACA
\n","AGGGGCAACGCACCCC
        TACA\n","AGGGGCAACCTACCCCTACA\n","AGGGGCAACCAACCCCTACA\n","AGGGGCAACCGACCCCTACA
\n","AGGGGCAACCCCC
        CCCTACA\n","AGGGGCAACCCGCCCCTACA\n","AGGGGCAACCCTCCCCTACA\n","AGGGGCAACCCAGCCCTACA
\n","AGGGGCAACC
        CAACCCTACA\n","AGGGGCAACCCATCCCTACA\n","AGGGGCAACCCACGCCTACA\n","AGGGGCAACCCACACCTACA
\n","AGGGGCA
        ACCCACTCCTACA\n","AGGGGCAACCCACCTCTACA\n","AGGGGCAACCCACCACTACA\n","AGGGGCAACCCACCGCTACA
\n","AGGG
        GCAACCCACCCTTACA\n","AGGGGCAACCCACCCATACA\n","AGGGGCAACCCACCCGTACA\n","AGGGGCAACCCACCCCGACA
\n","A
        GGGGCAACCCACCCCCACA\n","AGGGGCAACCCACCCCAACA\n","AGGGGCAACCCACCCCTCCA
\n","AGGGGCAACCCACCCCTGCA\n"
        ,"AGGGGCAACCCACCCCTTCA\n","AGGGGCAACCCACCCCTATA\n","AGGGGCAACCCACCCCTAAA
\n","AGGGGCAACCCACCCCTAGA
        \n","AGGGGCAACCCACCCCTACT\n","AGGGGCAACCCACCCCTACC\n","AGGGGCAACCCACCCCTACG\n");
  27    @REV3 = ("TGTAGGGGTGGGTTGCCCCT\n","CGTAGGGGTGGGTTGCCCCT\n","AGTAGGGGTGGGTTGCCCCT
\n","GGTAGGGGTGGG
        TTGCCCCT\n","TCTAGGGGTGGGTTGCCCCT\n","TATAGGGGTGGGTTGCCCCT\n","TTTAGGGGTGGGTTGCCCCT
\n","TGAAGGGGT
        GGGTTGCCCCT\n","TGGAGGGGTGGGTTGCCCCT\n","TGCAGGGGTGGGTTGCCCCT\n","TGTTGGGGTGGGTTGCCCCT
\n","TGTCGG
        GGTGGGTTGCCCCT\n","TGTGGGGGTGGGTTGCCCCT\n","TGTATGGGTGGGTTGCCCCT\n","TGTAAGGGTGGGTTGCCCCT
\n","TGT
        ACGGGTGGGTTGCCCCT\n","TGTAGCGGTGGGTTGCCCCT\n","TGTAGAGGTGGGTTGCCCCT
\n","TGTAGTGGTGGGTTGCCCCT\n","
        TGTAGGCGTGGGTTGCCCCT\n","TGTAGGAGTGGGTTGCCCCT\n","TGTAGGTGTGGGTTGCCCCT
\n","TGTAGGGCTGGGTTGCCCCT\n
        ","TGTAGGGATGGGTTGCCCCT\n","TGTAGGGTTGGGTTGCCCCT\n","TGTAGGGGAGGGTTGCCCCT
\n","TGTAGGGGCGGGTTGCCCC
        T\n","TGTAGGGGGGGGTTGCCCCT\n","TGTAGGGGTCGGTTGCCCCT\n","TGTAGGGGTAGGTTGCCCCT
```

```
\n","TGTAGGGGTTGGTTGC
        CCCT\n","TGTAGGGGTGCGTTGCCCCT\n","TGTAGGGGTGAGTTGCCCCT\n","TGTAGGGGTGTGTTGCCCCT
\n","TGTAGGGGTGGCT
        TGCCCCT\n","TGTAGGGGTGGATTGCCCCT\n","TGTAGGGGTGGTTTGCCCCT\n","TGTAGGGGTGGGATGCCCCT
\n","TGTAGGGGTG
        GGCTGCCCCT\n","TGTAGGGGTGGGGTGCCCCT\n","TGTAGGGGTGGGTAGCCCCT\n","TGTAGGGGTGGGTCGCCCCT
\n","TGTAGGG
        GTGGGTGGCCCCT\n","TGTAGGGGTGGGTTCCCCCT\n","TGTAGGGGTGGGTTACCCCT\n","TGTAGGGGTGGGTTTCCCCT
\n","TGTA
        GGGGTGGGTTGGCCCT\n","TGTAGGGGTGGGTTGACCCT\n","TGTAGGGGTGGGTTGTCCCT\n","TGTAGGGGTGGGTTGCGCCT
\n","T
        GTAGGGGTGGGTTGCACCT\n","TGTAGGGGTGGGTTGCTCCT\n","TGTAGGGGTGGGTTGCCGCT
\n","TGTAGGGGTGGGTTGCCACT\n"
        ,"TGTAGGGGTGGGTTGCCTCT\n","TGTAGGGGTGGGTTGCCCAT\n","TGTAGGGGTGGGTTGCCCGT
\n","TGTAGGGGTGGGTTGCCCTT
        \n","TGTAGGGGTGGGTTGCCCCA\n","TGTAGGGGTGGGTTGCCCCC\n","TGTAGGGGTGGGTTGCCCCG\n");
 28
 29    if ($execute =~/au/)
 30    {
 31        for $firstline (@FuzzyMatchFile)
 32        {
 33            #print "\n\nYO YO YO\n\n$firstline\n\n YO YO \n\n";
 34            if ($firstline =~/for5/) {$flag = "FOR5"; $outputname = "FilteredFOR5FuzzyMatches"; &LTRD
       B1; &action; print "\nInput file detected: 5' forward hits.\n";}
 35            if ($firstline =~/rev5/) {$flag = "REV5"; $outputname = "FilteredREV5FuzzyMatches"; &LTRD
       B2; &action; print "\nInput file detected: 5' reverse hits.\n";}
 36            if ($firstline =~/for3/) {$flag = "FOR3"; $outputname = "FilteredFOR3FuzzyMatches"; &LTRD
       B3; &action; print "\nInput file detected: 3' forward hits.\n";}
 37            if ($firstline =~/rev3/) {$flag = "REV3"; $outputname = "FilteredREV3FuzzyMatches"; &LTRD
       B4; &action; print "\nInput file detected: 3' reverse hits.\n";}
 38
 39        }
 40    }
 41
 42    elsif ($execute =~ /man/)
 43    {
 44        if      ($starter[0] =~ /5 / and $starter[1] =~ /for/)       {$flag = "FOR5"; $outputname =
       "FilteredFOR5FuzzyMatches"; &LTRDB1; &action; print "\nFile was set as containing: 5' forward hit
       s.\n";}
 45        elsif   ($starter[0] =~ /5 / and $starter[1] =~ /rev/)       {$flag = "REV5"; $outputname =
       "FilteredREV5FuzzyMatches"; &LTRDB2; &action; print "\nFile was set as containing: 5' reverse hit
       s.\n";}
 46        elsif   ($starter[0] =~ /3 / and $starter[1] =~ /for/)       {$flag = "FOR3"; $outputname =
       "FilteredFOR3FuzzyMatches"; &LTRDB3; &action; print "\nFile was set as containing: 3' forward hit
       s.\n";}
 47        elsif   ($starter[0] =~ /3 / and $starter[1] =~ /rev/)       {$flag = "REV3"; $outputname =
       "FilteredREV3FuzzyMatches"; &LTRDB4; &action; print "\nFile was set as containing: 3' reverse hit
       s.\n";}
 48        else    {die || 'Cannot run script. Check orientation and virus end settings';}
 49    }
 50
 51    #############
 52
 53    sub LTRDB1 {@input = @FOR5;};
 54    sub LTRDB2 {@input = @REV5;};
 55    sub LTRDB3 {@input = @FOR3;};
 56    sub LTRDB4 {@input = @REV3;};
 57
 58    sub action{
 59
 60
 61    $test = 1;
 62    $test2 = 1;
 63    $test3 = 1;
 64    $test4 = 1;
 65    $test5 = 1;
 66
 67    foreach $inputline (@input)
```

```
68    {
69        @LTRposs = @input;
70        my @splitlines = split (/\n/, "@splitinputfile");
71    }
72
73
74    foreach $new (@LTRposs)
75    {
76        $new =~ s/\R//;
77    }
78
79    #print @LTRposs2;
80
81    #@input = </Users/Adam/Desktop/FuzzyDenisovaMatches/LTR_possibilities
/5\'_For_LTR_possibilities.t
      xt>;
82
83    $test2 = 1;
84
85    foreach $hit (@FuzzyMatchFile)
86    {
87        #print $inputline . "\n";
88        open (B,  $hit) || die $!;
89        my @splithits = <B>;
90        my @splitlinehit = split (/\>/, "@splithits");
91        $splithitdata = join (">", @splitlinehit);
92    }
93
94    #print $splithitdata;
95
96    foreach $testLTR (@LTRposs)
97    {
98        my $foreachcount = $test++;
99        #print "\nHI $foreachcount";
100   }
101
102   while ($splithitdata =~ /(\>.*)\n\s(.*\n)/g)
103   {
104       #print $1, "\n";
105       $hit = "$1\n$2";
106       #print $hit;
107       my $whilecount = $test2++;
108       #print "\nTHING $whilecount\n";
109       #push (@hits, $hit);
110
111   foreach $LTR (@LTRposs)
112   {
113       if    ($flag =~ /FOR5/)     {&FOR5REV3;}
114       elsif ($flag =~ /REV5/)     {&FOR3REV5;}
115       elsif ($flag =~ /FOR3/)     {&FOR3REV5;}
116       elsif ($flag =~ /REV3/)     {&FOR5REV3;}
117
118       my $foreachcount = $test3++;
119       ############################################################
120       sub FOR5REV3 {
121       if ($hit =~ /.{$flankamount,}$LTR/)
122       {
123           #print "FOR5REV3";
124           $ifcount = $test4++;
125           #print $hit;
126           push (@result, $hit);
127           #print "\nSUCCESS! $ifcount";
128       }
129       }
130       ############################################################
131
132       ############################################################
133       sub FOR3REV5 {
```

```
134         if ($hit =~ /$LTR.{$flankamount,}/)
135         {
136             #print "FOR3REV5";
137             $ifcount = $test4++;
138             #print $hit;
139             push (@result, $hit);
140             #print "\nSUCCESS! $ifcount";
141         }
142         }
143         ############################################################
144
145
146     }
147
148     #last;
149     }
150     print @result, "\nSequences passed through: $ifcount\n";
151
152     open (OUTPUT, ">$outputpath/$outputname.txt");
153     print OUTPUT @result;
154     close (OUTPUT);
155     }
156
```

APPENDIX J

This is a perl script for stripping away duplicate ERV names form the 'Unknown_Group.txt' output of ERV Name Assigner, in the case of genera that have not yet been assigned. This allows the user to easily visualise which retroviral groups need to be assigned genera within the ERV Name Assigner program.

```
#! usr/bin/perl -w

#A QUICK SCRIPT TO CHECK THROUGH THE FILE CONTAINING ERV GROUPS UNASSIGNED TO GENERA. IT OUTPUT
ONE OF EACH TYPE OF GROUP IN THE FILE, TO MAKE IT EASY TO DISCERN
#WHICH GROUPS NEED ASSIGNING.

$filepath = "/Users/Adam/Desktop/PerlTest2014";

@unknown_lineage = <$filepath/Unknown_Groups.txt>;

foreach $unknownline (@unknown_lineage)
{
    #print $inputline . "\n";
    open (A,  $unknownline) || die $!;
    my @splitunknownfile = <A>;
    my @splitunknownlines = split (/\n/, "@splitunknownfile");
    $splitunknowndata = join ("\n", @splitunknownlines);
}

#print $splitunknowndata, "\n";

$test = 1;

while ($splitunknowndata =~ /(\d.*\n)/g)
{
    $oneline = $1;
    &printer;
    #print $test++, " HI\n";
    #print $oneline, "\n";
    #$counter = $test++;
    #last;
    #EXIT;
}

sub printer{

    if ($oneline =~ /\d*\t(.{1,15})\tERV\W.*\n/g)
    {
        $group_name = $1;
        push (@grouparray, $group_name, "\n");
        #print $group_name, "\n";
    }

    $counter = $test++;
    #print "Loop iteration: ", $counter, "\n";
    #last;
}

print "\nDONE!\n";

#print @grouparray;

#This removes duplicates, so you can clearly see which individual groups were left unassigned.

sub uniq {
    my %seen;
    grep !$seen{$_}++, @_;
}

#my @array = qw(one two three two three);
@duplicates_removed = uniq(@grouparray), "\n\n";

print "\n\n@duplicates_removed\n\n";
```

APPENDIX K

This is a program that is currently under development, where the aim is to confer the main operating strategy to the nomenclature pipeline, for extracting the flanking sequence from around an ERV insertion. This is currently designed to work on the output of a BLAST search, but will effectively work on the three columns of data provided by the output of the Gifford *et al*. (unpublished) pipeline, as well. These are: a gi identifier, sequence start and sequence end subranges. As with all the previous perl scripts, thresholds are customisable. There are a number of feedback loops included in this program as well, should the selected flank range fall outside of the sequence read limit.

```perl
#! /usr/bin/perl -w

$individualseqdirectory = "/Users/Adam/Desktop/PerlTest2014/ExtractedIndiSeqs";
$FinalResultFilepath = "/Users/Adam/Desktop/PerlTest2014";

#INPUT THE RANGE AROUND YOUR BLAST MATCH YOU WISH TO EXTRACT, HERE:

$sstartrange = 1000;
$sendrange = 1000;

$REVsstartrange = $sstartrange;
$REVsendrange = 2*$sendrange;

print "REVsstartrange: ", $REVsstartrange, "\nREVsendrange: ", $REVsendrange, "\n";

#MAY WANT TO INSERT A QUESTION HERE WARNING WHETHER DELETION IS DESIRED WITH Y = PROCEED, N = DIE
.

#This deletes all files where INDIVIDUAL sequences will be outputted. Important so that
#resulting concatenated file contains ONLY the desired sequences from this script instance.
print "\nDeleting all files within sequence extraction repository directory...\n";
sleep (1);
unlink glob "$individualseqdirectory/*.*";
sleep (2);

#SEQUENCE EXTRACTOR SCRIPT.
#RUN THIS AFTER TESTBLASTERE2.PL

#THESE SHOULD POINT TO YOUR GI NUMBER FILE, SUBJECT START FILE and SUBJECT END FILE:

@file1 = "/Users/Adam/Desktop/PerlTest2014/Seq_Extraction_Info/GInumbers.txt";
@file2 = "/Users/Adam/Desktop/PerlTest2014/Seq_Extraction_Info/SubjectStarts.txt";
@file3 = "/Users/Adam/Desktop/PerlTest2014/Seq_Extraction_Info/SubjectEnds.txt";

mkdir ExtractedIndiSeqs ;

foreach $inputline (@file1)
{
    #print $inputline . "\n";
    open (A,  $inputline) || die $!;
    my @splitinputfile = <A>;
    my @splitlines = split (/\n/, "@splitinputfile");
    $splitgis = join ("\n", @splitlines);

}

foreach $inputline2 (@file2)
{
    #print $inputline . "\n";
    open (B,  $inputline2) || die $!;
    my @splitinputfile = <B>;
    my @splitlines = split (/\n/, "@splitinputfile");
    $splitsstarts = join ("\n", @splitlines);

}

foreach $inputline3 (@file3)
{
    #print $inputline . "\n";
    open (C,  $inputline3) || die $!;
    my @splitinputfile = <C>;
    my @splitlines = split (/\n/, "@splitinputfile");
    $splitsends = join ("\n", @splitlines);

}

```

```
66
67
68
69  $number = () = $splitgis =~ /gi.*/gi;
70  $number2 = $number*3;
71  print "Total number of sequences to extract = ", $number, "\n";
72  print "Total number of expected loop iterations = ", $number2, "\n\n";
73  $test = 1;
74  sleep (6);
75
76  sub loop {
77  while ( $test <= $number2 )
78  {
79      sub gi{
80      while ($splitgis =~ /(gi.*)/g )
81      {
82          $gi = $1;
83          if ($gi =~ /(gi\|.*)/g ) {
84              $ginumber = $1;
85              print "\nGI NUMBER HERE>> ", $ginumber, "\n";
86              }
87          last;
88      }
89      }
90
91      while ($splitsstarts =~ /(\d*)/g )
92      {
93          $ss = $1;
94
95        if ($ss =~ /(\d)/g ) {
96              if ($ss =~ /([1-9]*)/g)
97                      {
98
99                          $sscheck = ($ss -= $sstartrange);
100                         if ($sscheck < 1)
101                         {
102                             print "\n EMERGENCY! SUBJECT START IS LESS THAN ZERO! Subject start r
    eset to 1. \n";
103                             $sstart = 1;
104                             #$emergencycounter = $test++;
105                         }
106                         else
107                         {
108                             $sstart = $sscheck;
109                         }
110                     }
111
112                         &gi;
113                         print $ss+=$sstartrange, "\n";
114                         print "SUBJECT START HERE>> ", $sstart, "\n";
115                         }
116
117         last;
118     }
119
120     while ($splitsends =~ /(\d*)/g )
121     {
122         $se = $1;
123         print $se, "\n";
124         if ($se =~ /(\d)/g ) {
125             if ($se =~ /([1-9]*)/g)
126                     {
127                         $send = $se += $sendrange;
128                         print "SUBJECT END HERE>> ", $send, "\n";
129                         &blastdbcmd;}
130                     }
131         last;
132      #   redo;
```

```
133         }
134
135         #sleep (2);
136
137             sub blastdbcmd{
138             #$one = 1;
139             #$filenumber = $one++;
140             #$counter = $test++;
141
142             #FOR REVERSE MATCHES
143
144             if ($sstart > $send)
145             {
146                 $sendchecker = ($send -= $REVsstartrange);
147
148                         if ( $sendchecker < 1)
149                             {
150                                 print "\n EMERGENCY! SUBJECT START IS LESS THAN ZERO! Subject start r
        eset to 1. \n";
151                                 $REVsstart = 1;
152                                 $emergencycounter = $test++;
153                             }
154                             else
155                             {
156                                 $REVsstart = ($send -= $REVsstartrange);
157                             }
158                 #print "THIS IS YOUR REVSSTART: ", $REVsstart, "\n";
159
160                 #NOTE: Despite the above if statement, some $REVsstart values are coming out negative
        . This next if resets them to 1 again.
161                 if ($REVsstart < 1)
162                 {
163                     $REVsstart = 1;
164                     $emergencycounter = $test++;
165                 }
166
167                 $REVsend = ($sstart += $REVsendrange);
168                 &REV_Extractor;
169             }
170             else
171             {
172                 &FOR_Extractor;
173             }
174
175             sub FOR_Extractor {
176             print "\nSequence in 'FORExtractedSeq$counter.txt'\n";
177             $exit_value = system("blastdbcmd -db /Users/Adam/Desktop/Gorilla_Genome/BLAST_db
/BLAST_Da
        tabase -entry '$ginumber' -range $sstart-$send -out $individualseqdirectory/FORExtractedSeq$count
        er.txt");
178             if ($exit_value == 0)
179             {
180                 print "SUCCESS! Sequence $counter extracted. Your exit code is $exit_value. \n";
181             #    print $exit_value, "\n";
182              #   &loop;
183             }
184             else
185             {
186                 print "FAILED. Sequence $counter NOT extracted. Your exit code is $exit_value. \n";
187             }
188             }
189
190             sub REV_Extractor {
191             print "\nSequence in 'REVExtractedSeq$counter.txt'\n";
192             $exit_value = system("blastdbcmd -db /Users/Adam/Desktop/Gorilla_Genome/BLAST_db
/BLAST_Da
        tabase -entry '$ginumber' -range $REVsstart-$REVsend -out $individualseqdirectory/REVExtractedSeq
        $counter.txt");
```

```
193             if ($exit_value == 0)
194             {
195                 print "SUCCESS! Sequence $counter extracted. Your exit code is $exit_value. \n";
196             #    print $exit_value, "\n";
197              #   &loop;
198             }
199             else
200             {
201                 print "FAILED. Sequence $counter NOT extracted. Your exit code is $exit_value. \n";
202             }
203             }
204
205             }
206         #COUNTER MUST BE HERE. REGARDLESS OF -w TRYING TO TELL YOU IT'S UNINITIATED.
207         $counter = $test++;
208         print "\nLoop iteration: ", $counter, "\n", "__________________\n";
209         if ($sscheck < 1 or $REVsstart < 1)
210             {
211                     $one = 1;
212                     $emergencycounter = $one++;
213             }
214         &relooper;
215         last;
216     }
217     }
218
219     #$exit_value = 0;
220     #print "EXIT VALUE = ", $exit_value, "\n";
221
222     &loop;
223
224     sub relooper{
225     if ($exit_value==0)
226     {
227         print "\nRe-Running While Loop\n";
228         #sleep (2);
229         &loop;
230     }
231     }
232
233     if ($counter == $number2)
234     {
235         print "\n You have completed $counter out of $number2 iterations. \n";
236         my $second_exit_value = `cat /Users/Adam/Desktop/PerlTest2014
/ExtractedIndiSeqs/* > $FinalRes
        ultFilepath/ExtractedSequence.txt`;
237         print "\nConcatenating sequences...\n";
238
239         if ($second_exit_value =!0)
240         {
241             print "\nSUCCESS! Your extracted sequence awaits in ExtractedSequence.txt Exit code $seco
        nd_exit_value\n";
242                 if ($emergencycounter => 1)
243                 {
244                 print "\nSame contig(s) started AFTER your specified start range. Maximum sequence wa
        s extracted by resetting sequence start to 1.\n";
245                 }
246         }
247         else
248         {
249             print "\nFAILED. You did not manage to concatenate sequences. Exit code: $second_exit_val
        ue.\n";
250         }
251
252     }
253
254     #print "BLASTDBCMD exit code = ", $exit_value, "\n";
```

APPENDIX L

This is the main fuzzy-matching unit that will be developed into a system for detecting orthologues, and subsequently naming them as part of the nomenclature pipeline. This script is a collaboration with Dr. Derek Huntley, who has revised the following script.

```
#! /usr/bin/perl

use String::Approx 'amatch';

{
open(IN, "Sample5ORTHset.txt"); #CHANGE THIS TO 5 PRIME FLANKS FILEPATH.

$/ = ">";
@seqs = <IN>;
shift @seqs;
}

foreach(@seqs){
        ($id) = $_ =~ /^(.*)\n/;
        #print "$id\n";
        s/^.*\n//;
        s/\n//g;
        s/>//;

        push(@seqs2, $_);

        $ids{$_} = $id;
        #print "$_\n";

        }

#print "SEQS2: @seqs2\n";

foreach $key(keys %ids){
        print "$key Matches\n";
        @fuzzymatches = amatch($key, ["10%"], @seqs2); # ["20%"],

        foreach $m(@fuzzymatches){
                print "$m\n";
                }
        #print @fuzzymatches;
        print "\n";
        }

#@fuzzymatches = amatch ($querymess, @genomespaced, ["10%"]);
```