

ENDOGENOUS RETROVIRUSES IN PRIMATES

Katherine Brown BSc, MSc

Thesis submitted to the University of Nottingham for the
degree of Doctor of Philosophy

July 2015

Abstract

Numerous endogenous retroviruses (ERVs) are found in all mammalian genomes, for example, they are the source of approximately 8% of all human and chimpanzee genetic material. These insertions represent retroviruses which have, by chance, integrated into the germline and so are transmitted vertically from parents to offspring. The human genome is rich in ERVs, which have been characterised in some detail. However, in many non-human primates these insertions have not been well-studied.

ERVs are subject to the mutation rate of their host, rather than the faster retrovirus mutation rate, so they change much more slowly than exogenous retroviruses. This means ERVs provide a snapshot of the retroviruses a host has been exposed to during its evolutionary history, including retroviruses which are no longer circulating and for which sequence information would otherwise be lost. ERVs have many effects on their hosts; they can be co-opted for functional roles, they provide regions of sequence similarity where mispairing can occur, their insertion can disrupt genes and they provide regulatory elements for existing genes. Accurate annotation and characterisation of these regions is an important step in interpreting the huge amount of genetic information available for increasing numbers of organisms.

This project represents an extensive study into the diversity of ERVs in the genomes of primates and related ERVs in rodents. Lagomorphs (rabbits and hares) and tree shrews are also analysed, as the closest relatives of primates and rodents. The focus is on groups of ERVs for which previous analyses are patchy or outdated, particularly in terms of their evolutionary history and possible transmission routes. A pipeline has been developed to comprehensively and rapidly screen genomes for ERVs and phylogenetic analysis has been performed in order to characterise these ERVs.

Almost 200,000 ERV fragments, many of which have not previously been characterised, were identified using this pipeline, distributed across six retroviral genera and 33 vertebrate genomes. These fragments were used to investigate several areas of interest: the potential origin of primate ERVs, in rodents or other hosts; the ERV content of the less well-studied primates; the endogenous lentiviruses; mammalian endogenous epsilonretroviruses and the origin of pathogenic gibbon ape

leukaemia virus. Laboratory study was used to complement the bioinformatics analysis where appropriate.

This analysis had several interesting outcomes. First, a novel endogenous member of the lentivirus genus of retroviruses, which are rarely found in an endogenous form, was identified in the bushbaby *Galago moholi*. This ERV may represent an ancient ancestor of modern human immunodeficiency virus (HIV), as it is the oldest member of the lentivirus genus (the genus which HIV belongs to) that has been identified in a primate living on the African mainland, alongside the primate hosts where the HIV pandemic originated. This ERV appears to have been transmitted between *G. moholi* and two species of Malagasy primate in the last five million years, many millions of years after these species have had any contact, suggesting that the virus has been transmitted from one host to another via a third, vector species. (Hart et al., 1996)

Gibbon ape leukaemia virus was responsible for leukaemia and lymphoma in several gibbon colonies during the 1970s and has since then been thought of as a circulating pathogen in this species. Using a combination of techniques we have established that this virus is not a common pathogen of modern gibbons and identified a route through which a single cross species transmission event from a rodent may have resulted in all known cases of this disease worldwide.

We have also identified endogenous epsilonretroviruses, usually considered to be viruses of fish and amphibians, in all screened species of primates. Based on these results, there is an ancient evolutionary relationship between epsilonretroviruses and primates. As these viruses once had the potential to infect primates and are currently widespread in fish, this result raises questions about the pathogenic potential of these viruses.

Many other ERVs were identified in primates, rodents and related species and we propose a classification scheme for these viruses and use this scheme as a basis to explore the ERV content of these hosts. Using this technique, previously unknown ERVs which are recombinant and which have the potential to produce active viral particles have been identified.

List of Published Papers

BROWN, K., EMES, R. D. & TARLINTON, R. E. 2014. Multiple Groups of Endogenous Epsilon-Like Retroviruses Conserved across Primates. *Journal of Virology*, 88, 12464-12471.

BROWN, K., MORETON, J., MALLA, S., ABOOBAKER, A. A., EMES, R. D. & TARLINTON, R. E. 2012. Characterisation of retroviruses in the horse genome and their transcriptional activity via transcriptome sequencing. *Virology*, 433, 55-63.

TARLINTON, R. E., BARFOOT, H. K. R., ALLEN, C. E., BROWN, K., GIFFORD, R. J. & EMES, R. D. 2013. Characterisation of a group of endogenous gammaretroviruses in the canine genome. *The Veterinary Journal*, 196, 28-33.

Acknowledgements

I would like to thank the following people for their contribution to this work:

- My supervisor Dr. Rachael Tarlinton for her guidance, support, friendship and constant reassurance over the last four years.
- Prof. Ed. Louis for supervision during the early stages of this project and advice throughout.
- Dr. Li Li for prior work on the prosimian lentiviruses and Dr. Liz Bailes, who was involved in the planning stages of the project and sadly passed away before the work was underway.
- Dr. Richard Emes for development of the Exonerate pipeline used to generate the majority of these results and general bioinformatics advice.
- Prof. John Brookfield, Dr. Beth Hellen and the University of Nottingham HPC team for advice on computing, bioinformatics and data analysis.
- Dr. Christian Roos at the German Primate Centre, Dada Gottelli at the Zoological Society of London and Mads Frost Bertelsen at Copenhagen Zoo for kindly providing primate samples.
- BIAZA for providing a letter of support for this project.
- Frank Wessely for answering a huge number of bioinformatics questions on a daily basis in the early days of my PhD.
- My friends at the University of Nottingham Vet School for keeping me sane and making me laugh throughout this PhD, especially Mansi, Frank, Donna, Jasmine, James, Tim and Laura. I will never hear the words robot or funnel again without laughing. Isaac Malter for entertainment during my thesis pending period.
- My mum, dad and brother for emotional (and occasionally financial!) support and practical help. And Dad and Olly for proofreading while sending me photos of monkeys.
- The University of Nottingham School of Veterinary Medicine and Science and School of Biology for funding, plus travel funding from the University of Nottingham Graduate School, the Genetics Society and the Society for General Microbiology.

Table of Contents

Scope of Thesis	1
Chapter 1. Introduction	5
1. 1. Classification, Structure and Life Cycle of Exogenous Retroviruses	5
1.1.1. Classification	6
1.1.2. Structure	10
1.1.2.1. Viral Particles	10
1.1.2.2. Genome Structure, Genes and Proteins	11
1.1.2.3. Long Terminal Repeats	12
1.1.3. Life Cycle	15
1.1.3.1. Receptor Binding	15
1.1.3.2. Entry	17
1.1.3.3. Transport to the Nucleus and Reverse Transcription	17
1.1.3.4. Reverse Transcription	18
1.1.3.5. Nuclear Import	21
1.1.3.6. Integration	21
1.1.3.7. RNA Synthesis	23
1.1.3.8. Translation	23
1.1.3.9. Assembly, Packaging and Release	25
1.1.3.10. Maturation	26
1.1.4. Accessory Proteins	26
1.1.4.1. Betaretroviruses	27
1.1.4.2. Lentiviruses	27
1.1.4.3. Deltaretroviruses	29
1.1.4.4. Epsilonretroviruses	29
1.1.4.5. Spumaviruses	30
1. 2. Exogenous Retroviruses, Disease and Host Defences	32
1.2.1. HIV, SIV and FIV	32
1.2.1.1. Naturally Infected Hosts	32
1.2.1.2. Progression to AIDS	33

1.2.1.3. Treatment	36
1.2.2. Oncogenic Retroviruses	37
1.2.2.1. Transducing Retroviruses	38
1.2.2.2. Cis-Activating Retroviruses	38
1.2.2.3. Trans-Activating Retroviruses	39
1.2.3. Restriction Factors	39
1.2.3.1. Uncoating and Reverse Transcription	40
1.2.3.2. Translation	44
1.2.3.3. Release	45
1. 3. Endogenous Retroviruses	46
1.3.1. Life Cycle and Evolution	47
1.3.1.1. Integration	47
1.3.1.2. Proliferation	47
1.3.1.3. Degeneration	50
1.3.2. Host Control of ERVs	51
1.3.2.1. Control of Transcription	51
1.3.2.2. Innate Immunity	53
1.3.2.3. Restriction Factors	54
1.3.3. Benefits of ERVs	55
1.3.3.1. Capture of ERV Genes by the Host	56
1.3.3.2. Protection Against Other Retroviruses	57
1.3.3.3. Gene Regulation	58
1.3.4. ERVs and Disease	59
1.3.4.1. Active ERVs	59
1.3.4.2. Transcription and Expression	60
1.3.4.3. Chromosome Disruption	64
1. 4. Endogenous retroviruses in vertebrate genomes.	64
1.4.1. Overview	64
1.4.2. Vertebrate Evolution	66
1.4.3. Gammaretroviruses	71
1.4.3.1. HERV-I Group	72
1.4.3.2. HERV-F/H Group and HERV-W	73

1.4.3.3. Crocodile Group	75
1.4.3.4. REV Group	75
1.4.3.5. GALV Group	78
1.4.3.6. MLV Group	82
1.4.3.7. HERV-E Group	85
1.4.3.8. Rabbit ERV H Group	85
1.4.3.9. Syncytins	86
1.4.4. Epsilonretroviruses	89
1.4.4.1. Fish epsilonretroviruses	89
1.4.4.2. Amphibian epsilonretroviruses	91
1.4.4.3. Reptile Epsilonretroviruses	92
1.4.5. Spumaviruses	94
1.4.5.1. ERV-L Elements	94
1.4.5.2. Foamy viruses	94
1.4.6. Alpharetroviruses	95
1.4.7. Betaretroviruses	96
1.4.7.1. HERV-K and $\beta 1$	97
1.4.7.2. SERV, SRV, BaEV, RD114, MusD and TvERV	99
1.4.7.3. Mouse Mammary Tumour Virus	102
1.4.7.4. Jaagsiekte Sheep Retrovirus and Enzootic Nasal Tumour Virus	103
1.4.8. Lentiviruses	104
1.4.8.1. RELIK	104
1.4.8.2. pSIVgml and pSIVfdl	104
1.4.8.3. MELVs	106
1.4.9. Gypsy Elements and Errantiviruses	107
Chapter 2. Materials and Methods	109
2. 1. Genome Screening for ERVs	109
2.1.1. Genome Screening: Techniques	109
2.1.1.1. BLAST and BLAT	109
2.1.1.2. Retroector	110
2.1.1.3. LTR_STRUC	111

2.1.1.4. RepeatMasker and Repbase	111
2.1.1.5. Comparison of Techniques	112
2.1.2. Exonerate	114
2.1.3. Exonerate Pipeline	115
2.1.4. Input Dataset	119
2.1.4.1. Categorisation of Reference Sequences	122
2.1.5. Input Genomes	123
2.1.5.1. Pre-processing	126
2.1.6. Screening	128
2. 2. Parsing Output	128
2.2.1. Quality Control	129
2.2.2. Clustering	129
2.2.2.1. Clustering Sequences: Techniques	129
2.2.2.2. Clustering Analysis	131
2.2.3. Identification of Output Sequences	135
2. 3. Phylogenetic Analysis	136
2.3.1. Aligning Retroviral Sequences: Techniques	136
2.3.2. Building Trees: Techniques	137
2.3.2.1. Choosing a Gene	137
2.3.2.2. Model Selection	138
2.3.2.3. Tree Building Algorithm	139
2.3.3. Phylogenetic Analysis of Exonerate Output	142
2.3.3.1. Phylogenetic Test Datasets	142
2.3.3.2. Model Selection	143
2.3.3.3. Alignment and Phylogenetic Analysis	146
2. 4. Characterisation of ERVs	146
2.4.1. Determining Presence or Absence	147
2.4.2. Determining Copy Number	147
2.4.3. ERVs with Multiple Genes	147
2.4.4. LTRs	148
2.4.5. Identification of Open Reading Frames	149

2.4.6. Determining Age	149
2.4.6.1. LTRs and Degeneration	149
2.4.6.2. Host Tracking and Locus-by-Locus Analysis	150
2.4.6.3. Locus-by-locus Analysis	151
2.4.7. Identifying Selection	152
2. 5. Host Phylogeny	152
Chapter 3. Overview of Results	155
3. 1. Raw Output and Quality Control	155
3. 2. Clustering	160
3. 3. Intact ERVs	162
3. 4. Host Phylogeny	164
Chapter 4. Genus-by-genus Analysis	165
4. 1. Overview	165
4. 2. Gammaretroviruses	167
4.2.1. HERV-I Group	173
4.2.2. HERV-F Group	182
4.2.3. HERV-E Group and HERV-R Group	195
4.2.4. REV-Like Group	198
4.2.5. MLV-Like Group	208
4. 3. Epsilonretroviruses	209
4. 4. Spumaviruses	210
4. 5. Alpharetroviruses	211
4. 6. Betaretroviruses	212
4.6.1. HERV-K-Like Group	214
4.6.2. SERV-Like Group	224
4.6.3. IAP-Like Group	235
4.6.4. JSRV-Like Group	236
4. 7. Lentiviruses	239

Chapter 5. Endogenous lentiviruses in mainland African bushbabies provide insight into the origin of SIV.	240
5. 1. Introduction	240
5. 2. Materials and Methods	241
5. 3. Results	243
5. 4. Discussion	248
Chapter 6. The origin and proliferation of gibbon ape leukaemia virus	251
6. 1. Introduction	251
6.1.1. History of GALV	251
6.1.2. GALV Phylogeny	253
6. 2. Materials and Methods	255
6. 3. Results	257
6. 4. Discussion	262
Chapter 7. Endogenous Epsilon-Like Retroviruses in Primates	266
7. 1. Introduction	267
7. 2. Materials and Methods	269
7.2.1. Genome Screening	269
7.2.2. Comparison between Primate Genomes	270
7.2.3. Genome Characterisation	271
7.2.4. Comparison with Other Mammals	273
7. 3. Results	273
7. 4. Discussion	282
Chapter 8. General Discussion	287
8. 1. Vector Species and Cross-Species Transmissions	287
8. 2. Host Range and Recombination	289
8. 3. Potentially Active ERVs	290

8. 4. Comparison of Genomes	292
8. 5. Relationship between ERVs and XRVs	293
8. 6. Defining an ERV Group	295
8. 7. Predicting ERV Diversity	299
8. 8. Future Work	307
8. 9. Conclusions	311
Chapter 9. Appendices	312
Appendix A. Prior publications	312
A.1 Characterisation of retroviruses in the horse genome and their transcriptional activity via transcriptome sequencing.	312
A.2 Characterisation of a group of endogenous gammaretroviruses in the canine genome.	312
Appendix B. Supplementary Tables	313
B.1 Figure_sequences.xlsx	313
B.2 Full_ERV_Nucleotide_Database.xlsx	313
B.3 Refseq_Retrovirus_Sequences.xlsx	313
B.4 GROUPED_PREVKNOWN_groups.xlsx	313
B.5 Test_datasets.xlsx	313
B.6 Euarchontoglires_accessions.xlsx	314
B.7 Output_db.xls	314
B.8 ERV_Regions.xlsx	314
B.9 Lemur_Tree.xlsx	315
B.10 Epsilonretrovirus_Input.xlsx	315
B.11 Epsilonretrovirus_loci.xlsx	315
B.12 Epsilonretrovirus_positions.xlsx	315

Appendix C.	Fasta files	316
C.1	FULL_PREVKNOWN.fas	316
C.2	PARSED_UT_PREVKNOWN.fas	316
C.3	PARSED_T_PREVKNOWN.fas	316
C.4	RAW_EXO_OUT.fas	316
C.5	PARSED_EXO_OUT.fas	316
C.6	CLU_EXO_OUT.fas	317
Appendix D.	Python and R scripts	318
D.1	make_chromosomes.py	318
D.2	reciprocal_blast.py	318
D.3	distance.R	319
D.4	make_cons.py	319
D.5	classify_sets.py	320
Appendix E.	Documents concerning gibbon transportation	321
E.1	1974BANGKO17800.pdf	321
E.2	1975BANGKO15111_b.pdf	321
E.3	1974BANGKO19028_b.pdf	321
E.4	1974STATE244644_b.pdf	321
E.5	1974TAIPEI06749_b.pdf	322
E.6	1974STATE260770_b.pdf	322
E.7	1974BANGKO17734_b.pdf	322
E.8	1974STATE260768_b.pdf	322
References		323

Table of Figures

Figure 1: The phylogenetic relationships of common exogenous retroviruses.....	7
Figure 2: Schematic diagram showing the basic structure of a retroviral particle.	10
Figure 3: The genome structure of a simple provirus.....	11
Figure 4: The structure of the 5' and 3' LTRs of an integrated complex retrovirus.	14
Figure 5: Stages of the retroviral life cycle.....	16
Figure 6: Stages of reverse transcription.	20
Figure 7: Stages of integration.....	22
Figure 8: The impact of APOBEC3G on cells without Vif.....	42
Figure 9: Pol gene phylogeny of the retroviruses showing the seven retroviral genera and their hosts.....	65
Figure 10: Overview of vertebrate phylogeny.....	66
Figure 11: Overview of mammalian phylogeny.	67
Figure 12: Phylogeny of the major groups of primates.	69
Figure 13: Phylogeny of the major groups of rodents and lagomorphs.....	70
Figure 14: The relationship between the <i>pol</i> genes of the major groups of gammaretroviruses.....	71
Figure 15: The phylogeny of HERV-I-like elements from birds, reptiles, cartilaginous fish and humans.....	73
Figure 16: The relationship between the <i>pol</i> genes of HERV-H and HERV-F lineages of HERV.	73
Figure 17: The proposed evolutionary history of SNV, DIAV and RSV.	77
Figure 18: The relationship between the <i>pol</i> genes of members of the GALV group of gammaretroviruses.....	78
Figure 19: The species of <i>Mus</i> in which MuRRS, MuRV-Y, GLN and MmERV have been detected.....	80
Figure 20: The relationship between the <i>pol</i> genes of members of the MLV group of gammaretroviruses.....	82
Figure 21: The relationship between syncytins and other gammaretroviral <i>env</i> genes.....	86
Figure 22: Two phylogenetic trees showing the amphibian ERVs and their relationship with other epsilonretroviruses.....	92

Figure 23: Phylogenetic tree showing reptile ERVs and their relationship with other epsilonretroviruses.	93
Figure 24: <i>Pol</i> gene phylogeny of primate and rodent betaretroviruses and their relatives. Adapted from Baillie et al. 2004.....	97
Figure 25: The relationship between the <i>gag</i> , <i>pol</i> and <i>env</i> genes of SERVs, SRVs, BaEV and RD114.	101
Figure 26: Comparison of the ERV sequences identified in the cow genome using LTR_STRUC, Retrotector and BLAST.	113
Figure 27: Flow chart representing the Exonerate pipeline used to detect candidate ERVs.	116
Figure 28: Scatter graph showing the proportion of ERVs identified on each chromosome of the horse genome using Exonerate compared to BLAST, LTR_STRUC and Exonerate and the Pearson's correlation coefficient for each comparison.	118
Figure 29: The process used to cluster candidate ERV sequences.	132
Figure 30: Results of model testing using the JModelTest software.....	145
Figure 31: Graph showing the total number of candidate <i>gag</i> , <i>pol</i> and <i>env</i> ERV fragments identified and the proportions of these fragments which were both verified by BLAST and appeared to represent unique genes.....	156
Figure 32: Graph showing the number of unique, BLAST-verified <i>gag</i> , <i>pol</i> and <i>env</i> genes identified using Exonerate in each host.	157
Figure 33: The relationship between number of scaffolds and number of ERV fragments identified using Exonerate, including and excluding results from the tarsier genome.	159
Figure 34: The distribution of the similarity of sequences in PARSED_EXO_OUT to the group consensus sequence by which they were represented in CLU_EXO_OUT, in cases where the group consensus and the original sequence are not identical.	161
Figure 35: The number of regions with each possible combination of multiple ERV gene fragments across all hosts.	163
Figure 36: The number of <i>gag-pol-env</i> regions identified in each host.	163
Figure 37: Phylogenetic tree based on 15 nuclear genes showing the relationships between the sequences of primates, rodents, lagomorphs and tree shrews.	164

Figure 38: The distribution of <i>gag</i> , <i>pol</i> and <i>env</i> insertions identified in each host between genera.....	166
Figure 39: The distribution of ERV fragments between genomes for the gammaretrovirus genus.	167
Figure 40: The proportion of gammaretroviruses identified in each host here (green) and by Hayward et al. (2013a) (pink).	168
Figure 41: <i>Pol</i> gene phylogeny showing the seven proposed groups of gammaretroviruses in the Euarchontoglires.	171
Figure 42: <i>Pol</i> gene phylogeny showing the HERV sequences classified as class ii (red) and class iii (blue) by Hayward et al. (2013a).	172
Figure 43: Phylogenetic comparison between the nucleotide sequences of the <i>env</i> ORFs identified in HERV-I related insertions and of other known gammaretroviral syncytin proteins and <i>env</i> genes.	177
Figure 44: <i>Pol</i> gene phylogeny showing the relationship between the HERV-I like insertions identified in prosimians and rodents with those found in mammals.....	179
Figure 45: <i>Pol</i> gene phylogeny showing the relationship between the HERV-I like insertions identified in birds and those found in mammals.	181
Figure 46: The <i>pol</i> gene phylogenetic relationships between each lineage of HERV-H, HERV-F, HERV-W and HERV-P and the closest matching sequence from each host species, where a sequence was present in the host.	184
Figure 47: The copy number and estimated integration dates of the HERV-F/H/W family of ERVs.....	185
Figure 48: <i>Env</i> gene phylogeny showing the relationship between known carnivore, new world rodent and ruminant syncytins and the HERV-P <i>env</i> genes identified here.	188
Figure 49: <i>Pol</i> gene phylogeny showing the relationship between the bushbaby HERV-W-like insertions and those of other primates. Details of previously known sequences are provided in Appendix B.2.	190
Figure 50: <i>Pol</i> gene phylogeny showing the relationship between HERV-H like clusters identified in humans (green), bonobos (pink) and chimpanzees (blue).	191
Figure 51: <i>Env</i> gene phylogeny showing the regions identified with ORFs corresponding to HERV Fc1 and HERV Fc2.	194

Figure 52: <i>Pol</i> gene phylogeny for the closest sequence identified in each host to HERV E (left) and HERV R (right).	198
Figure 53: The number of REV/HERV-T like <i>pol</i> gene insertions identified in each host.	200
Figure 54: Phylogeny showing the relationships between the <i>pol</i> gene at selected REV-like non-recombinant loci and known gammaretroviruses.	202
Figure 55: The structure of the two potentially intact gammaretroviral loci identified in the guinea pig genome.	205
Figure 56: <i>Gag</i> , <i>pol</i> and <i>env</i> gene phylogenies of the REV/HERV-T like full length recombinant insertions identified in the guinea pig and chinchilla genomes.	207
Figure 57 The distribution of ERV fragments between genomes for the epsilonretrovirus genus.	209
Figure 58: The distribution of ERV fragments between genomes for the spumavirus genus.	210
Figure 59: The distribution of ERV fragments between genomes for the alpharetrovirus genus.	211
Figure 60: The distribution of ERV fragments between genomes for the betaretrovirus genus.	212
Figure 61: <i>Pol</i> gene phylogeny showing the four proposed groups of betaretroviruses in the Euarchontoglires.	213
Figure 62: <i>Pol</i> gene phylogenies of the most similar insertion to each HML type sequence identified in each host where an insertion clustering with the type sequence was identified.	217
Figure 63: The copy number and estimated integration dates of the HERV-K family of ERVs in primates.	218
Figure 64: <i>Pol</i> gene phylogenetic trees for the two HML-5 loci which appear to predate the divergence of new and old world primates.	220
Figure 65: <i>Gag</i> , <i>pol</i> and <i>env</i> gene phylogenies of the SERV-like betaretroviruses identified in old world monkeys and reference betaretroviruses.	226
Figure 66: <i>Gag</i> , <i>pol</i> and <i>env</i> gene phylogenies of the SRV-like recombinants with betaretrovirus-like <i>gag</i> and <i>pol</i> genes and gammaretrovirus-like <i>env</i> genes with reference gamma- and betaretroviruses.	227

Figure 67: The structure of the most intact SRV-like insertions indentified with betaretrovirus-like <i>gag</i> and <i>pol</i> genes and gammaretrovirus-like <i>env</i> genes.	230
Figure 68: <i>Gag</i> , <i>pol</i> and <i>env</i> gene phylogenies of the BaEV-like recombinants with gammaretrovirus-like <i>gag</i> and <i>pol</i> genes and betaretrovirus-like <i>env</i> genes with reference gamma- and betaretroviruses.	234
Figure 69: <i>Pol</i> gene phylogenies showing the relationships between the JSRV-like ERVs in the marmoset (blue, left) and the tarsier (green, right) and known members of this group.	238
Figure 70: The distribution of ERV fragments between genomes for the lentivirus genus.	239
Figure 71: The phylogenetic relationships between the prosimian primates.	245
Figure 72: Gel electrophoresis photograph showing the 300bp band identified using the FR- RR primer pair in <i>M. murinus</i> , <i>C. medius</i> and <i>G. moholi</i>	245
Figure 73: Maximum likelihood phylogenetic tree showing the phylogenetic relationship between pSIVmb (marked in yellow) and other lentiviruses.	247
Figure 74: The geographical distribution of <i>Galago moholi</i> (yellow) and <i>Microcebus murinus</i> / <i>Cheirogaleus medius</i> (blue).	249
Figure 75: The relationship between the <i>pol</i> genes of GALV and the sequences described as its closest genetic relatives.	254
Figure 76: <i>Pol</i> gene phylogenetic tree showing the phylogenetic relationship between GALV and its closest relative in each host genome screened.	258
Figure 77: <i>Env</i> gene phylogeny showing the relationship between the GALV strains and related sequences from rodents, primates and tree shrews.	259
Figure 78: PhyML phylogenetic tree based on a 7426 nucleotide multiple alignment of the consensus sequences for 87 epsilon-like <i>pol</i> gene fragments found in primates, showing the clustering of primate epsilonretroviral loci into three major phylogenetic groups.....	276
Figure 79: PhyML phylogenetic tree based on a 5510 base pair multiple alignment of the consensus sequences of three phylogenetic groups of primate epsilon-like <i>pol</i> gene fragments and known epsilon and epsilon-like retroviruses.	278
Figure 80: A comparison of identified regions of the PE genome (A) and the reference genome of WDSV (GenBank Accession NC_001867) with orf-a, orf-b and orf-c excluded (B) and included (C) in the genome length and gene position calculations.	282

Figure 81: The classification scheme proposed for newly identified ERVs, using the 16 HERV-Fc1 like loci as an example.	298
Figure 82: Scatterplot showing the relationship between the estimated integration date of a subgroup of ERVs and the number of ERVs in the subgroup, based on the HERV-F-Like and HERV-K-Like subgroups in chimpanzees, bonobos, humans and gorillas.	300
Figure 83: Ten example runs of Katzourakis et al.'s stochastic model of ERV integration rate showing the estimated number of active integrated ERVs (blue) and inactive integrated ERVs (red) over time.	302
Figure 84: Ten example runs of Katzourakis et al.'s stochastic model of ERV integration rate with varying values of parameter s (top-left of each graph) showing the estimated number of active integrated ERVs (blue) and inactive integrated ERVs (red) over time.	303

List of Tables

Table 1: Core morphologies for each genus of retrovirus and electron micrographs of mature particles from example species.....	9
Table 2: Table showing the translational strategies for gag,pro and pol used by different retroviruses.....	25
Table 3: The accessory genes and genome structure of an example of each genus of retrovirus.	31
Table 4: Common types of cancer with a reported significant increase in transcription of at least one HERV family.	61
Table 5: Diseases of fish either confirmed to be or provisionally associated with retroviruses.....	90
Table 6: Comparison of the number of pol genes and of Class I pol genes identified in the horse genome with Exonerate, Retrorector, LTR_STRUC and TBLASTN.	117
Table 7: Taxonomic details of the 33 vertebrate genomes screened using Exonerate.	124
Table 8: Assemblies and source databases for the 33 vertebrate genomes screened using Exonerate.....	125
Table 9: The probability that an average ERV (10,000 bp) and an average Exonerate ERV fragment (1100 bp) would span more than one contig or scaffold in each screened genome based on mean segment (contig or scaffold) length and on N50.	127
Table 10: Genes used for host phylogeny.	154
Table 11: Statistical comparison between the number of fragments and mean fragment length for each host in PARSED_EXO_OUT and various genome metrics.	159
Table 12: The gammaretrovirus groups identified by Hayward et al. (2013a)	169
Table 13: The number of HERV-I-like ERV fragments identified in each host type.	174
Table 14: The position of the HERV-I <i>env</i> ORF in various simian hosts.	176
Table 15: The number of HERV-F-like ERV fragments identified in each host type.....	183
Table 16: The number of HERV-E-like ERV fragments identified in each host type.	195
Table 17: The number of HERV-R-like ERV fragments identified in each host type.	196
Table 18: The number of REV-like ERV fragments identified in each host type.....	199
Table 19: Table showing the details of the potentially intact loci in the guinea pig genome.	205

Table 20: The number of MLV-like ERV fragments identified in each host type.	208
Table 21: The number of HERV-K-Like ERV fragments identified in each host type.....	215
Table 22: The presence or absence of each HERV-K lineage (HML-1 to HML-8) in each primate group.	216
Table 23: The number of SERV-like ERV fragments identified in each host type.	225
Table 24: The four subcategories of SERV-like insertions with betaretroviral <i>gag-pol</i> regions and gammaretroviral <i>env</i> regions.	228
Table 25: The number of SRV-like insertions of each type identified in each host with <i>gag</i> and <i>pol</i> genes clustering with the betaretroviruses and <i>env</i> genes with the gammaretroviruses, any previous references to these insertions, the estimated age of the insertions and the length of the longest ORF.	229
Table 26: The position of the most intact SRV-like insertions with betaretroviral <i>gag</i> and <i>pol</i> genes and gammaretroviral <i>env</i> genes in their host genomes. Scaffold numbers are from the genome builds listed in Table 8.	231
Table 27: The number of IAP-like ERV fragments identified in each host type.....	235
Table 28: The number of JSRV-like ERV fragments identified in each host type.....	236
Table 29: The gibbon colonies in which GALV was identified, the date the colony was started (where available) and the source and import date of the gibbons in which GALV was isolated.	261
Table 30: The number of epsilon-like ERVs of each type (primate epsilon 1 to primate epsilon 3, PE1 to PE3) identified in each host species.....	274
Table 31: The phylogenetic group, LTR_type, proportion of sites at which LTRs are not identical to each other and median age of each of the 11 epsilon-like ERV loci flanked by two recognisable LTRs.	279
Table 32: Spearman's rank correlation coefficient for the relationship between various life history traits and number of ERV fragments per host.	306
Table 33: Spearman's rank correlation coefficient for the relationship between various life history traits and number of ERV fragments per host, excluding primates.	306

List of Abbreviations

Abbreviation	Term
A	adenine
AIDS	acquired immunodeficiency syndrome
aLTR	approximate likelihood ratio test
ALV	avian leukaemia virus
APOBEC	apolipoprotein B-editing catalytic polypeptide family
ARKS	animal record keeping software
ART	anti-retroviral therapy
ATL	adult T-cell leukaemia
BaEV	baboon endogenous retrovirus
BLAST	basic local alignment search tool
BLAT	BLAST-like alignment tool
bp	bp
C	cytosine
C6P	Compara six primate alignment
CA	capsid
CFS	chronic fatigue syndrome
COL	University of California at Davis Comparative Oncology Laboratory
DIAV	duck infectious anaemia virus
EIAV	equine infectious anaemia virus
ENTV	enzootic nasal tumour virus
<i>env</i>	envelope (gene)
Env	envelope (protein)
ERV	Endogenous retrovirus
<i>Fgf</i>	fibroblast growth factor
FIV	feline immunodeficiency virus
FWPV	fowlpox virus
G	guanine
<i>gag</i>	group-specific antigen (gene)
Gag	group-specific antigen (protein)
GALV	gibbon ape leukaemia virus
GALV-Br	gibbon ape leukaemia virus brain
GALV-H	gibbon ape leukaemia virus Hall's island

GALV-SEATO	gibbon ape leukaemia virus Southeast Asia Treaty Organization
GALV-SF	gibbon ape leukaemia virus San Francisco
GHV	gallid herpesvirus
GLN	murine retrovirus using tRNA ^{Gln}
GTR	general time reversible
HERV	human endogenous retrovirus
HIV	human immunodeficiency virus
HML	human mouse mammary tumour virus like
HTLV	human T-cell lymphotropic virus
IAP	intracisternal A-type particle
II	integrase inhibitor
IN	integrase
JC	Jukes-Cantor
JSRV	Jaagsiekte sheep retrovirus
K80	Kimura 80
KoRV	koala retrovirus
KWERV	killer whale ERV
LTR	long terminal repeat
MbERV	<i>Melomys burtoni</i> ERV
MDERV	<i>Mus dunni</i> ERV
MLV	murine leukaemia virus
MmERV	<i>Mus musculus</i> ERV
MMTV	mouse mammary tumour virus
mRNA	messenger RNA
MS	multiple sclerosis
MSRV	multiple sclerosis associated retrovirus
MuERV-C	murine ERV C
MuRRS	mouse retrovirus related sequence
MuRV-Y	murine repeated virus on the Y-chromosome
NC	nucleocapsid
NCBI	National Center for Biotechnology Information
NCI	National Cancer Institute
Nef	negative factor
NIH	National Institute of Health
NNRTI	nucleoside RT inhibitor
ORF	open reading frame

PBS	primer binding site
PcEV	Papio cynocephalus endogenous retrovirus
PCR	polymerase chain reaction
PE	primate epsilon
PERV	porcine endogenous retrovirus
PI	protease inhibitor
PIC	pre-integration complex
<i>pol</i>	polymerase (gene)
Pol	polymerase (protein)
PPT	polypurine tract
PR	protease (protein)
<i>pro</i>	protease (gene)
Psi	packaging signal
R	repeat
RELK	rabbit endogenous lentivirus type K
Rem	regulatory protein of MMTV
Rev	regulator of expression of virion proteins
REV	reticuloendotheliosis virus
RNAPII	RNA polymerase II
RT	reverse transcriptase
SAg	superantigen
SAMHD1	SAM domain and HD domain containing protein 1
SEATO	Southeast Asia Treaty Organization
SERV	simian endogenous retrovirus
SFMC	San Francisco Medical Center
SIV	simian immunodeficiency virus
SIVcpz	SIV chimpanzee
SIVgor	SIV gorilla
SIVmac	SIV macaque
SIVsm	SIV sooty mangabey
SMRV	squirrel monkey retrovirus
SNRV	snakehead retrovirus
SNV	spleen necrosis virus
SRV	simian retrovirus
SSSV	salmon swim bladder sarcoma virus
STLV	human T-cell lymphotropic virus

SU	surface unit
T	thymine
TAR	trans-acting responsive element
Tas	transactivator of spumaviruses
Tat	transactivator of transcription
TF	translational frameshift
TLR	Toll-like receptor
TM	transmembrane unit
TR	translational readthrough
TRIM5 α	tripartite motif containing protein 5 alpha
tRNA	transfer RNA
TvERV	<i>Trichosurus vulpecula</i> (possum) endogenous retrovirus
TVM	transversion model
U	uracil
U3	untranslated 3'
U5	untranslated 5'
UC Davis	University of California at Davis
UCSC	University of California Santa Cruz
UTR	untranslated region
Vif	viral infectivity factor
Vpr	viral protein R
Vpu	viral protein U
Vpx	viral protein X
WDSV	walleye dermal sarcoma virus
WEHV	walleye epidermal hyperplasia virus
WGS	whole genome shotgun
WMSV	woolly monkey sarcoma virus
XMRV	xenotropic murine leukaemia related virus
XRV	exogenous retrovirus
ZFERV	zebrafish ERV

List of Datasets

The datasets created and used for this analysis are referred to using the following abbreviated names and are available as Appendices.

Dataset Name	FASTA Appendix	Table Appendix	Description
FULL_PREVKNOWN (full previously known retrovirus)	C.1	B.1	The full unparsed untranslated dataset of 4124 previously known retrovirus sequences
PARSED_UT_PREVKNOWN (parsed untranslated previously known retrovirus)	C.2	B.1	The parsed, untranslated version of the database of previously known retroviruses, consisting of 1590 nucleotide sequences
PARSED_T_PREVKNOWN (parsed translated previously known retrovirus)	C.3	B.1	The parsed, translated version of the database of previously known retroviruses used as an input to the Exonerate pipeline,
GROUPED_PREVKNOWN (grouped previously known retrovirus)	N/A	B.4	The grouped dataset of previously known sequences from PARSED_UT_PREVKNOWN
RAW_EXO_OUT (raw Exonerate output)	C.4		The raw output from the Exonerate algorithm for all host genomes, containing 190,196 candidate ERV fragments
PARSED_EXO_OUT (parsed Exonerate output)	C.5	B.4	The parsed output from the Exonerate algorithm for all host genomes, containing only the 169,424 sequences which passed the quality control step
CLU_EXO_OUT (clustered Exonerate output)	C.6	B.4	The clustered output from the Exonerate algorithm for all host genomes, consisting of 47,896 sequences with consensus sequences representing highly similar sequences from the same host.
GROUPED_EXO (grouped Exonerate output)	N/A	B.4	The grouped dataset of sequences from CLU_EXO_OUT.

Scope of Thesis

Chapter 1. Introduction

1. 1. Classification, Structure and Life Cycle of Exogenous Retroviruses
1. 2. Exogenous Retroviruses, Disease and Host Defences
1. 3. Endogenous Retroviruses

Endogenous retroviruses (ERVs) are retroviruses which have, by chance, at some point in their evolutionary history, integrated into a germline cell and therefore become an inherited part of the host genome. ERVs usually lose their pathogenicity over time but fragments of ERVs can remain visible in the genome for many millions of years. Chapter 1 provides a review of the current literature on retroviruses, with a particular focus on endogenous retroviruses, and an outline of the aims of this thesis.

Chapter 2. Materials and Methods

2. 1. Genome Screening for ERVs
2. 2. Parsing Output
2. 3. Phylogenetic Analysis
2. 4. Characterisation of ERVs
2. 5. Host Phylogeny

The first focus of this project was the development of a pipeline able to quickly and comprehensively characterise the ERV content of mammalian genomes. Various techniques used for genome-wide screening, clustering large numbers of sequences, identifying degenerate genetic material and characterising the evolutionary history of this material are reviewed in Chapter 2 and appropriate methodology is selected, tested and refined.

Chapter 3. Overview of Results

- 3. 1. Raw Output and Quality Control
- 3. 2. Clustering
- 3. 3. Intact ERVs

Using this pipeline, the 15 primate genomes which have been sequenced to date were screened for ERVs and their ERV content identified and characterised. Rodents have abundant ERVs and cross-species transmissions from these hosts are common, therefore rodents were also screened in order to identify insertions which they may have transmitted to primates. Rodents and primates form a phylogenetic group with the Lagomorphs (rabbits and hares) and the tree shrews, which were also screened to allow comparisons between host and ERV phylogenies. The overall distribution of ERVs in these hosts was then examined and is discussed in Chapter 3.

Chapter 4. Genus-by-genus Analysis

- 4. 1. Host Phylogeny
- 4. 2. Overview
- 4. 3. Gammaretroviruses
- 4. 4. Epsilonretroviruses
- 4. 5. Spumaviruses
- 4. 6. Alpharetroviruses
- 4. 7. Betaretroviruses
- 4. 8. Lentiviruses

Of the large number of ERVs identified in this study, several groups were of particular interest. These were previously unidentified ERVs, ERVs with distributions or phylogenetic relationships which are inconsistent with the literature, ERVs of less well-studied primates (new world monkeys and prosimians) and insertions which may perform a role in the host or have the potential to produce active retroviral particles. These groups are discussed in each retroviral genus in Chapter 4.

Chapter 5. Endogenous lentiviruses in mainland African bushbabies provide insight into the origin of SIV.

- 5. 1. Introduction
- 5. 2. Materials and Methods
- 5. 3. Results
- 5. 4. Discussion

Chapter 5 is presented as a research article and is an analysis of the endogenous lentiviruses of prosimian primates. The known endogenous lentiviruses were considered to be unlikely to represent all of the endogenous lentiviruses present in mammalian genomes, as they are found in species with a patchy geographic and phylogenetic distribution. Therefore, this study aimed to investigate the distribution of these ERVs in other hosts. Firstly, as endogenous lentiviruses are known in two species of lemur, a laboratory based approach was used to screen samples from other prosimians (where a genome sequence is not available) and determine the presence or absence of endogenous lentiviruses in these hosts. Secondly, the pipeline discussed in Chapter 2 was used to screen the available Euarchontoglires genomes for these insertions.

Chapter 6. The origin and proliferation of gibbon ape leukaemia virus

- 6. 1. Introduction
- 6. 2. Materials and Methods
- 6. 3. Results
- 6. 4. Discussion

Chapter 6 is also presented as a research article and concentrates upon the origin of the GALV pathogen of gibbons. This virus is widely considered to have originated in rodents and been transmitted to gibbons shortly before the GALV outbreak in the 1970s. However, little is known about this outbreak and the virus has not been analysed in depth since the advent of modern laboratory and sequencing techniques. The prevalence of this pathogen in contemporary gibbons and the risk it poses to these primates are poorly understood.

Therefore, several analyses were performed with the objective of establishing where this virus originated, how it spread and its current prevalence.

Chapter 7. Endogenous Epsilon-Like Retroviruses in Primates

- 7. 1. Introduction
- 7. 2. Materials and Methods
- 7. 3. Results
- 7. 4. Discussion

Chapter 7 is identical to an article accepted for publication in the Journal of Virology. Prior analysis of the horse genome has identified epsilon-like ERVs in mammalian hosts (Brown et al., 2012) (Appendix A.1), despite these ERVs being generally considered to be exogenous fish pathogens. A small number of endogenous epsilon-like insertions have also previously been described in the human genome (Katzourakis and Tristem, 2005, Tristem, 2000, Oja et al., 2005). However, little detailed work has been carried out to find out how widespread mammalian epsilonretroviruses may be. Therefore, this chapter provides the first detailed analysis of epsilonretrovirus-related fragments in primate genomes.

Chapter 8. General Discussion

- 8. 1 Vector Species and Cross-Species Transmissions
- 8. 2 Host Range and Recombination
- 8. 3 Potentially Active ERVs
- 8. 4 Comparison of Genomes
- 8. 5 Relationship between ERVs and XRVs
- 8. 6 Defining an ERV Group
- 8. 7 Predicting ERV Diversity
- 8. 8 Future Work
- 8. 9 Conclusions

Chapter 8 is a general discussion of our results in the context of the literature and of potential future work.

Chapter 1. Introduction

This chapter provides a review of the current literature on exogenous and endogenous retroviruses.

Section 1. 1 provides an overview of the XRVs in terms of their classification, structure and life cycle.

Section 1. 2 provides a brief summary of the effect of some of the major pathogenic XRVs.

Section 1. 3 provides an introduction to ERVs, their life cycle and their interactions with their hosts.

Section 1. 4 provides a more detailed review of some of the ERVs identified in mammalian genomes to date

1. 1. Classification, Structure and Life Cycle of Exogenous Retroviruses

The Retroviridae, or retroviruses, are a family of related viruses with shared characteristics in terms of life cycle, morphology and genetics. Retroviruses are enveloped viruses with a positive sense, single stranded RNA genome, 7,000 to 11,000 nucleotides in length (International Committee on Taxonomy of Viruses, 2002). All retroviral genomes include four major protein coding genes: group specific antigen (*gag*), protease (*pro*), polymerase (*pol*) and envelope (*env*) and have flanking 5' and 3' long terminal repeats (LTRs) (Goff, 2007). Some also code for further accessory proteins. The life cycle of the retroviruses involves reverse transcription of viral RNA into double stranded

DNA, which is integrated into the genome of the host and transcribed by cellular factors (Gifford and Tristem, 2003). This section provides a general introduction to the exogenous retroviruses and their life cycle.

1.1.1. Classification

The family Retroviridae is divided into two subfamilies and seven genera. The Orthoretrovirinae subfamily consists of the genera alpharetrovirus, betaretrovirus, gammaretrovirus, deltaretrovirus, lentivirus and epsilonretrovirus (International Committee on Taxonomy of Viruses, 2002).

The Spumaretrovirinae subfamily contains one genus, spumavirus (International Committee on Taxonomy of Viruses, 2002). These classifications are based on morphological and structural characteristics (section 1.1.2), life cycle (section 1.1.3), accessory genes (section 1.1.4) and genetic similarity (Figure 1).

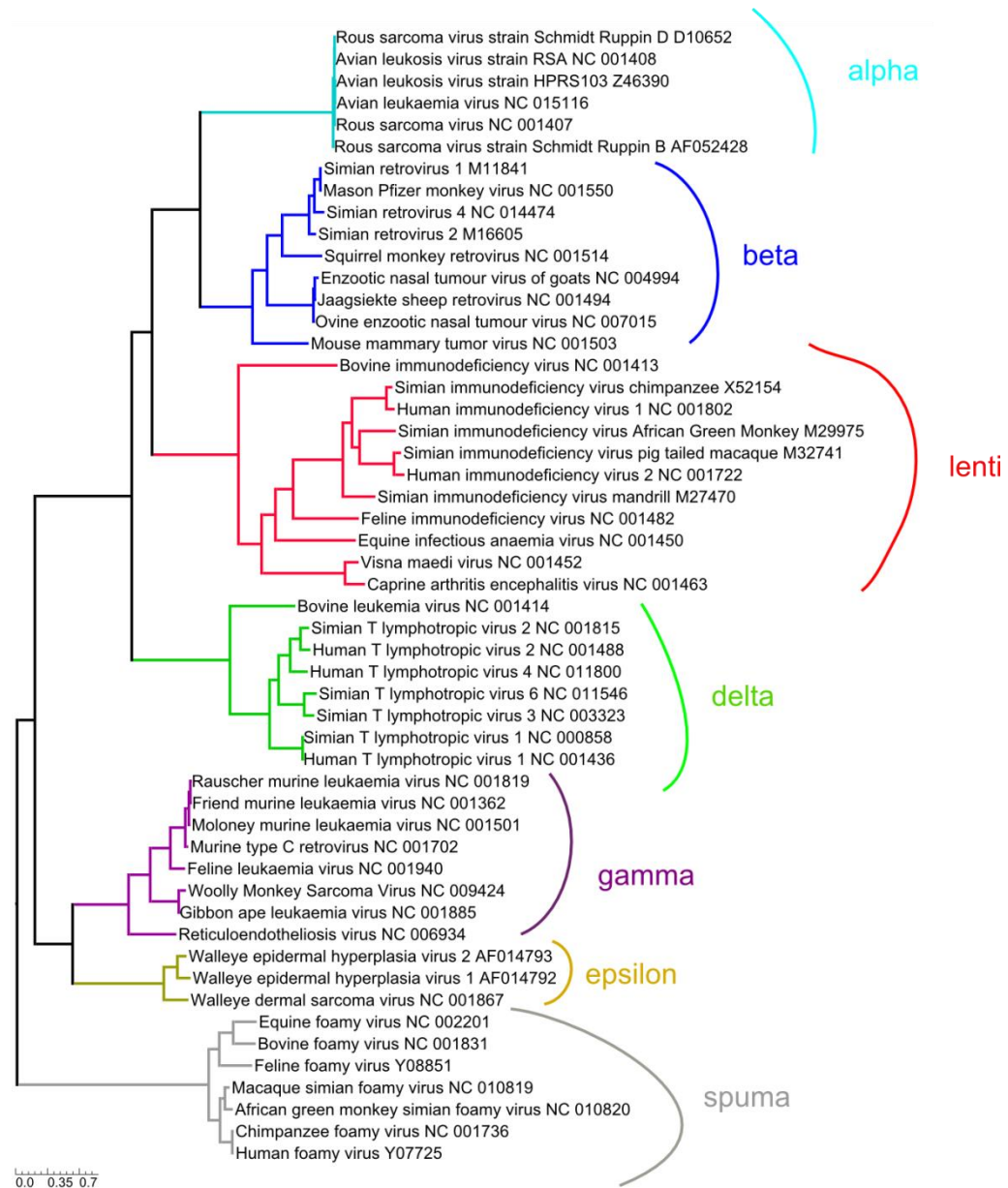

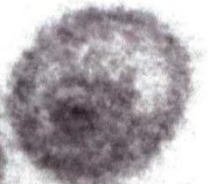
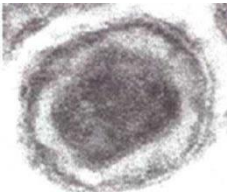

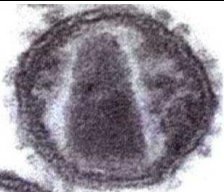



Figure 1: The phylogenetic relationships of common exogenous retroviruses. Phylogenetic tree rooted on the spumaviruses showing the relationship between the *pol* gene amino acid sequences of the major exogenous retroviruses listed in RefSeq (Pruitt et al., 2012b) and the international committee on the taxonomy of viruses database (ICTVdb) (International Committee on Taxonomy of Viruses, 2002). This tree was generated here to incorporate specific sequences but is consistent with the literature [for example (Weiss, 2006, Jern et al., 2005, Han and Worobey, 2012a)]. Amino acid sequences were aligned under the default settings in MUSCLE (Edgar, 2004) and the tree generated under the RtRev model in PhyML (Guindon and Gascuel, 2003).

Retroviruses were originally classified into four types – types A to D – based on shape of the core, as seen via electron microscopy(Goff, 2007). Briefly, A-type particles, now considered to be immature capsids, appear as thick shelled, hollow intracellular structures(Goff, 2007). B-type particles have a round, non-central inner core(Goff, 2007). C-type particles assemble at the plasma membrane and have a central, spherical core(Goff, 2007). D-type particles assemble in the cytoplasm and have a cylindrical core(Goff, 2007). Type A is no longer considered to be a separate morphological type. Type C particles are typical of alpha-, gamma-, epsilon- and deltaretroviruses and types B and D are seen in betaretroviruses (Table 1). Lentiviruses and spumaviruses have their own unique core types, lentiviruses have cylindrical or conical cores while spumaviruses have a spiked surface and a central, uncondensed core (Goff, 2007) .

Table 1: Core morphologies for each genus of retrovirus and electron micrographs of mature particles from example species.

All images from Bannert et al. (2010).

Genus	Particle Type	Example	
Alpha	C		Avian leukosis virus
Beta	B or D		Mouse mammary tumour virus (B type)
Gamma	C		Murine leukaemia virus
Delta	C		Bovine leukaemia virus
Epsilon	C		
Lenti	Lenti		Simian immunodeficiency virus
Spuma	Spuma		Simian foamy virus

1.1.2. Structure

1.1.2.1. Viral Particles

All mature exogenous retroviruses form 100 to 150nm enveloped particles (Bannert et al., 2010). The viral core contains two copies of the RNA genome of the retrovirus, which is protected from degradation by nucleocapsid (NC) proteins (Figure 2). The core also encloses the viral enzymes protease (PR), reverse transcriptase (RT) and integrase (IN) (Goff, 2007)(Figure 2). The core is surrounded by a protein capsid (CA) (Goff, 2007). The viral particle is enclosed in a host-derived lipid bilayer envelope, studded with viral glycoproteins with two subunits, the transmembrane (TM) and surface (SU) units (Bannert et al., 2010) (Figure 2).

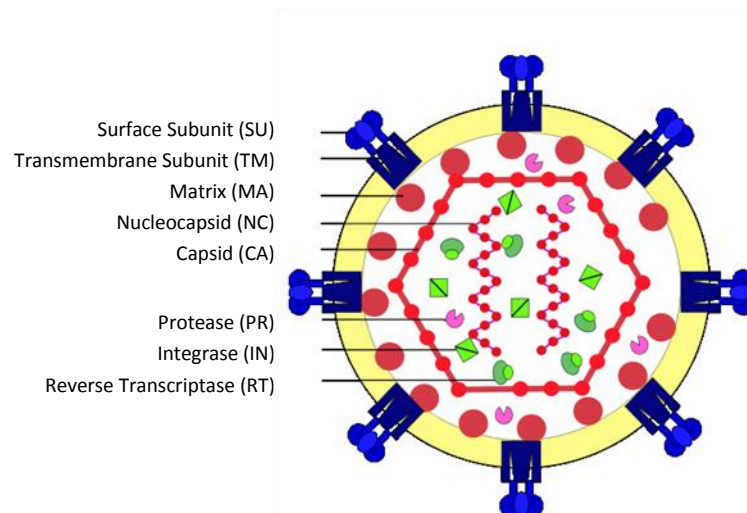


Figure 2: Schematic diagram showing the basic structure of a retroviral particle. Proteins encoded by *gag* are shown in red, *pol* green and *env* blue. Adapted from Voisset and Andrawiss (2000).

1.1.2.2. Genome Structure, Genes and Proteins

The four major retroviral genes each generate proteins which play specific roles in either the structure of the retrovirus or its life cycle. The genome order 5' LTR-*gag*-*pro*-*pol*-*env*-LTR 3' is completely conserved amongst known retroviruses (Jern and Coffin, 2008) (Figure 3).

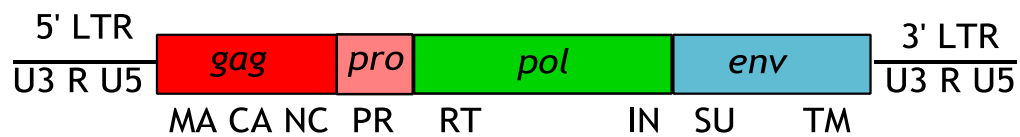


Figure 3: The genome structure of a simple provirus.

Excluding *pro*, these major genes encode polyproteins which are later cleaved into smaller subunits (Figure 2). The *gag* gene encodes the CA, MA and NC proteins (Goff, 2007). As well as forming structural components of the virion, these proteins are involved in assembly and packaging of newly formed retroviral particles (Goff, 2007). *Gag* ranges in length from less than 1200 bp (bp) to almost 2000 bp (Bannert et al., 2010). *Pro* encodes the viral enzyme PR, which is involved in cleaving viral polyproteins into their separate subunits and is approximately 700 bp in length (Goff, 2007). *Pol* encodes two enzymes, RT and IN. RT catalyses the transcription of viral RNA into DNA and IN the integration of viral cDNA into host DNA. *Pol* genes range in length from 2500 to 3500 bp. *Env* encodes the SU and TM glycoproteins of the retroviral envelope and is approximately 1500 to 3000 bp in length (Goff, 2007). SU is involved in receptor binding and TM in membrane fusion (Goff, 2007). Complex retroviruses also have various accessory genes coding for additional proteins, the function which will be discussed in section 1.1.4.

The strategy used to generate multiple proteins from short retroviral genomes depends on the genus. *Gag*, *pro* and *pol*, or subsets of this group, are often translated as single fusion proteins and later cleaved (Goff, 2007) (section

1.1.3.8). Multiple proteins, especially accessory proteins, are often the product of subsections of the same DNA sequence in different reading frames (section 1.1.4).

1.1.2.3. Long Terminal Repeats

Retroviral genomes are flanked by LTRs. On integration, these regions are identical to each other and each LTR consists of three regions – untranslated 3' (U3), repeat (R) and untranslated 5' (U5) (Lenasi et al., 2010). In an unintegrated RNA retrovirus U5 is found only at the 5' end and U3 at the 3' end, as these regions are duplicated during reverse transcription (section 1.1.3.4). Many elements regulating transcription of integrated retroviruses are in the LTRs. The same regulatory structures are present in both LTRs but the majority of retroviruses use the 5' LTR for transcription initiation and the 3' LTR for termination (Bannert et al., 2010).

After integration, the 5' LTR has the structure 5'-U3-R-U5-3'. The U3 region begins with a highly conserved dinucleotide, the *att* site, used as an attachment site during integration, at the far 5' end (Bannert et al., 2010). U3 is the promoter region for the retrovirus, so its main role is transcription initiation. It incorporates the TATA box, to which cellular RNA polymerase II (RNAPII) binds, plus several other transcription factor binding sites (Bannert et al., 2010). When retroviral DNA is transcribed by the host, a 5' cap is added to a specific site at the 3' end of U3, while the remainder of the 5' U3 is not transcribed. The 5' untranslated region (UTR) runs from this point, through R and U5 to the AUG start codon at the 5' end of *gag* (Bannert et al., 2010). The TATA box marks the boundary between the U3 and R regions (Bannert et al., 2010). The region between R and the start codon of *gag* is known as the 5' leader region and contains the majority of regulatory elements involved in transcription, reverse transcription and packaging (Bannert et al., 2010). The trans-acting responsive (TAR) element binds to accessory proteins, increasing the speed of RNAPII and preventing premature termination (Bannert et al.,

2010). The polyadenylation site, usually adenylated only in the 3' LTR, marks the boundary of R and U5 (Bannert et al., 2010). U5 includes the primer binding site (PBS), which binds to a complementary cellular transfer RNA (tRNA) during reverse transcription (Bannert et al., 2010) (section 1.1.3.4) and the packaging signal (Psi) sites, which increase the efficiency of RNA packaging into virions (Bannert et al., 2010) (section 1.1.3.9).

The 3' LTR is predominantly involved in transcription termination and has the same 5'-U3-R-U5-3' structure as the 5' LTR after integration. The transcript terminates with the poly(A) tail, which is added at the poly(A) loop downstream of the polyadenylation signal, a conserved sequence in the 3' R region (Bannert et al., 2010). The terminus of the 3' LTR is an *att* site, mirroring the 5' LTR (Bannert et al., 2010).

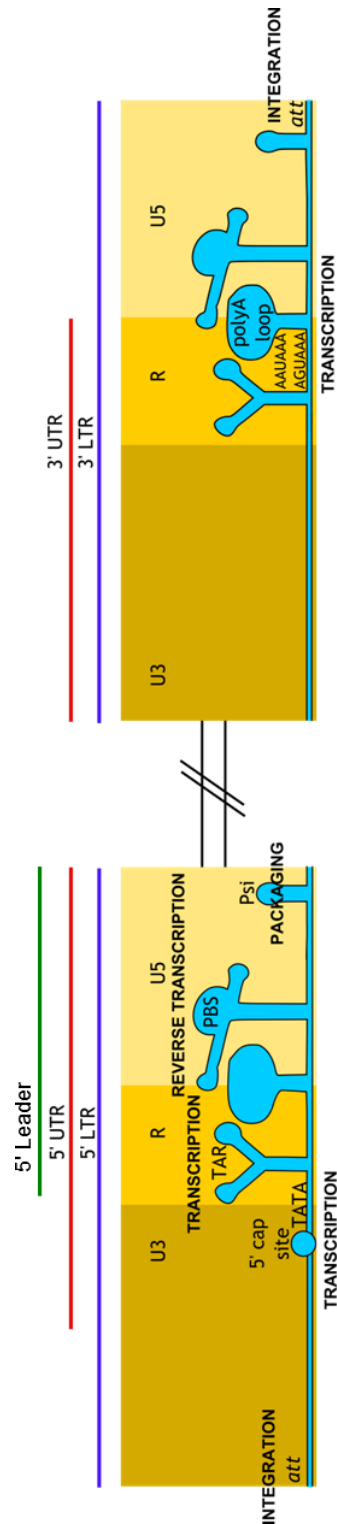


Figure 4: The structure of the 5' and 3' LTRs of an integrated complex retrovirus. Structural components are labelled with the stage of the life cycle they are involved in: integration, transcription, reverse transcription or packaging. The green line represents the 5' leader region, the red line the untranslated region and the blue line the LTR. Adapted from Balvay et al. (2007) and Bannert et al. (2010).

1.1.3. Life Cycle

1.1.3.1. Receptor Binding

The life cycle of all retroviruses begins when the SU subunit of the Env protein interacts with a receptor on the surface of a cell (Goff, 2007) (Figure 5A). In order for this to occur, the virus needs to reach the receptor. For murine leukaemia virus (MLV), HIV and avian leukaemia virus (ALV) this has been shown to involve protrusions from the cell surface called filopodia, which “pick” viral particles and pull them towards the cell body for fusion (Lehmann et al., 2005). Receptor specificity is determined by a region of the N-terminus of the SU. A wide range of cell surface molecules can be used as receptors (Goff, 2007). (Sommerfelt, 1999)

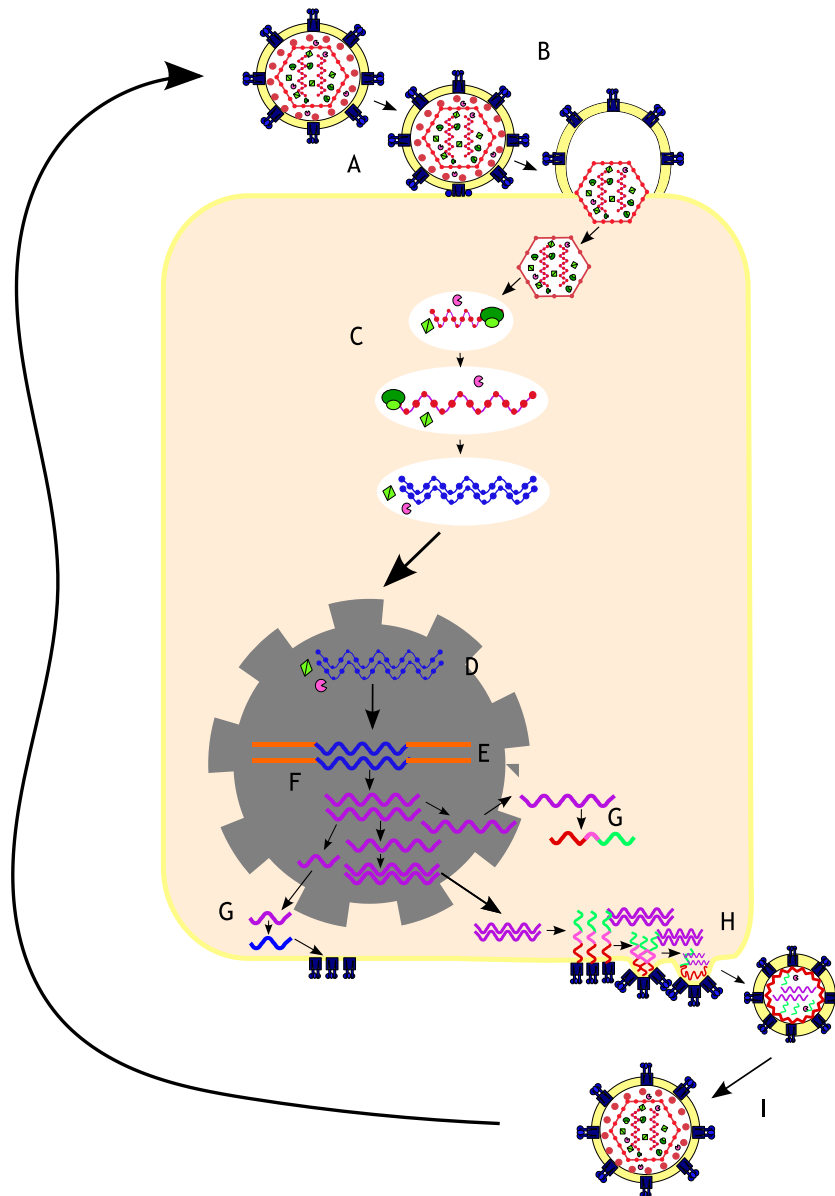


Figure 5: Stages of the retroviral life cycle.

A) Receptor binding, the retrovirus particle binds to a receptor on the host cell surface; B) Entry, the viral and cellular membrane fuse and the retrovirus enters the cell; C) Transport to the nucleus and reverse transcription, the viral core is uncoated and its RNA genome (red) is reverse transcribed to form double stranded DNA (blue); D) nuclear import, the retrovirus enters the nucleus via the nuclear pore; E) integration, the retrovirus is incorporated into the host genome (orange); F) RNA synthesis, mRNA copies of the retroviral genome (purple) are generated by cellular factors and selectively spliced; G) translation, mRNA transcripts are translated into polyproteins: Gag (red), PR (pink), Pol (green) and Env (blue). Env is localised to the plasma membrane; H) assembly, budding and release, Gag, PR and Pol accumulate at the cell membrane and co-opt a host pathway to bud out as immature virions; I) maturation, polyproteins are cleaved by PR and the virion undergoes various structural changes to form a mature virus. Adapted from Goff (2007), Bannert et al. (2010), Engelman et al. (2010), Göttlinger and Weissenhorn (2010).

1.1.3.2. Entry

In order to infect a cell, the retrovirus needs to transfer its genome across its own membrane and that of the cell (Blumenthal et al., 2012). This is achieved either by fusion of the viral and cell surface membranes as a response to receptor binding (the pH independent pathway), or via receptor mediated endocytosis (pH dependent pathway) (Goff, 2007, Mothes and Uchil, 2010) (Figure 5B). Receptor binding triggers disassociation of the SU and TM subunits of Env, bringing a region of TM known as the fusion peptide into contact with the host cell membrane. This peptide induces changes in the host cell membrane which allow it to enter the target cell (Shchelokovskyy et al., 2011).

Alternatively, a number of retroviruses enter the host cell by receptor-mediated endocytosis (Mothes and Uchil, 2010). Viruses are taken into the cell by endocytosis and the acidic environment in the endosome triggers fusion between the virus and the endosome, releasing the capsid into the cell cytoplasm (Miyachi et al., 2011).

1.1.3.3. Transport to the Nucleus and Reverse Transcription

Once the virus has fused with the host cell, a step-by-step uncoating of the viral core occurs and the retroviral RNA genome is reverse transcribed into DNA (Figure 5C). When the viral core enters the cell, it consists of at least CA, NC, RT, IN, PR, some MA, and, in complex retroviruses, accessory proteins (Warrilow et al., 2009). CA and most of RT are usually lost before the virus enters the nucleus. This uncoating is not well understood, but in HIV it seems to involve phosphorylation of MA (Peterlin, 2002). Uncoating does not seem to be necessary prior to nuclear import for MLV, which enters the nucleus with its capsid intact (Arhel, 2010).

1.1.3.4. Reverse Transcription

For most retroviruses, the majority of reverse transcription occurs between infection and nuclear import (Hu and Hughes, 2012). Reverse transcription depends on reverse transcriptase and RNase H, both of which are parts of the RT protein (Hu and Hughes, 2012).

Reverse transcription begins at a cellular tRNA bound to the PBS (Figure 4) close to the 5' end of the viral genome (Hu and Hughes, 2012) (Figure 6A). The RT enzyme then proceeds towards the 5' end, synthesising minus strand DNA of the U5 and R sequences (Goff, 2007) (Figure 6B). This creates a short RNA-DNA duplex, from which RNase H degrades the RNA, leaving only DNA (Hu and Hughes, 2012). The product of this process is known as the minus strand strong stop DNA (Goff, 2007) (Figure 6C). The exposed R region DNA in this intermediate is complementary to the 3' R region of the minus strand of RNA, and therefore acts as a bridge to transfer the minus strand strong stop DNA to the 3' end of the viral RNA (Hu and Hughes, 2012) (Figure 6D). This is known as the first strand transfer (Engelman, 2010). DNA synthesis then continues in the 3' to 5' direction until a complete minus strand is generated (Hu and Hughes, 2012) (Figure 6E). The majority of the RNA genome is then degraded by RNase H (Engelman, 2010).

The PPT, found just upstream of the 3' LTR, is not degraded and acts as a primer for plus strand DNA synthesis (Engelman, 2010). 3' to 5' elongation proceeds from here until 18 nucleotides into the tRNA primer, where a modified adenine residue terminates elongation (Engelman, 2010) (Figure 6F). A second RNA-DNA duplex is produced and the RNA is degraded, leaving a plus strand strong stop DNA (Hu and Hughes, 2012) (Figure 6G). The nucleotides from the tRNA primer are complementary to the nucleotides in the minus strand DNA which were copied from the PBS, so these regions anneal together in the second strand transfer (Hu and Hughes, 2012) (Figure 6F). On annealing, the plus strand is elongated from the 3' to 5' end of the genome and the minus strand is elongated to incorporate the sequence of the plus strand

strong stop DNA (Figure 6I). This produces a double stranded DNA copy of the viral genome (Hu and Hughes, 2012). This DNA copy is longer than the original retrovirus because it contains the U3-R-U5 strong stop sequence at either end (Figure 6J).

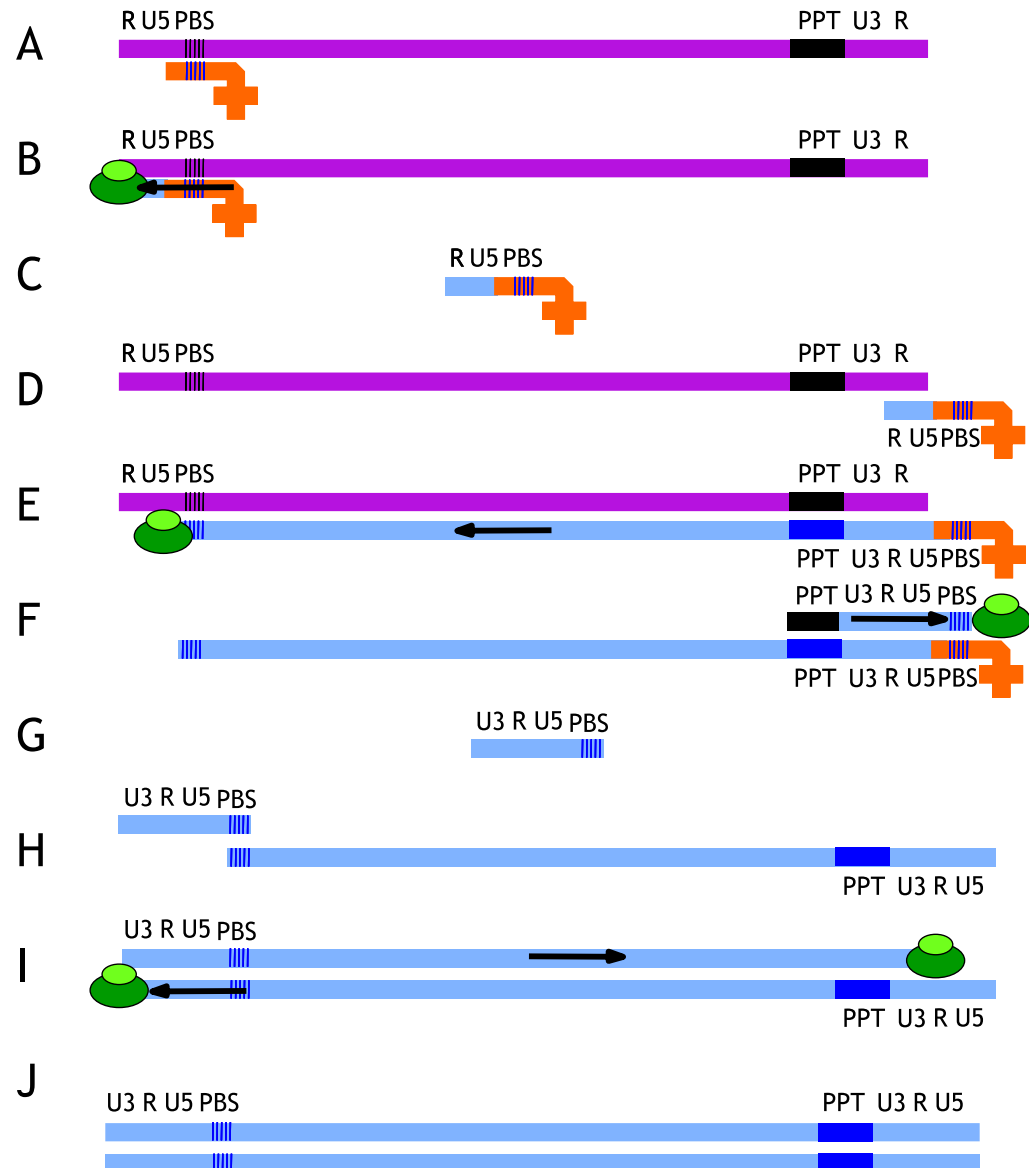


Figure 6: Stages of reverse transcription.

RNA is shown in purple and DNA in blue. A) the cellular tRNA (orange) binds to the primer binding site (PBS) near the 5' end of the viral RNA genome; B) the RT enzyme (green) elongates towards the 5' end of the minus strand DNA; C) this generates the minus strand strong stop DNA; D) the strong stop DNA is transferred to the 3' end of viral RNA, this is the first strand transfer; E) DNA synthesis 3' to 5' generates a complete DNA minus strand (blue); F) complementary sequences in the polypurine tract (PPT) act as a primer for plus strand DNA synthesis towards the 5' end; G) this forms the plus strand strong stop DNA; H) this is complementary to and anneals to the 3' PBS on the plus strand, this is the second strand transfer; I) elongation of the plus strand occurs 3' to 5' and the minus strand is elongated to incorporate the plus strand strong stop DNA; J) a double stranded DNA copy of the viral genome is produced. Adapted from Engelman (2010), Hu and Hughes (2012).

1.1.3.5. Nuclear Import

Prior to integration, the DNA copy of the virus forms a complex, known as the pre-integration complex (PIC), with the uncoated viral core and some specific host proteins (Bannert et al., 2010). This complex is then imported into the nucleus (Bannert et al., 2010). Simple retroviruses, for example MLV, enter the nucleus during mitosis, when the nuclear envelope breaks down, in a process directed by viral proteins (Suzuki and Craigie, 2007). HIV is able to integrate in non-dividing cells, such as macrophages (Bukrinsky, 2004). It is thought that HIV virus uses the cell's nuclear import processes to enter the nucleus via the nuclear pore (Bukrinsky, 2004). This may be because a component of the PIC contains targeting signals, known as nuclear localisation signals, which engage with cellular transport proteins (Bukrinsky, 2004).

1.1.3.6. Integration

Once a retrovirus is in the nucleus, the next step is to integrate into the DNA sequence of the host (Figure 5G).

Integration into the host genome has three steps, end processing, joining and gap repair. In the end processing step, a dinucleotide is removed from both 3' ends of the double stranded viral cDNA, exposing hydroxyl groups (Hindmarsh and Leis, 1999) (Figure 7A). This hydroxyl ion then hydrolyses the phosphodiester bond on the 3' side of a conserved CA dinucleotide (the *att* site) near the 3' end (Figure 4), releasing the adjacent dinucleotide and exposing a hydroxyl group (Engelman, 2010) (Figure 7B). In the joining step, this hydroxyl group is used by integrase to cut the host DNA and join it to the viral DNA (Engelman, 2010) (Figure 7C). The hydroxyl group attacks phosphodiester bonds in the host DNA and a new phosphodiester bond forms between the 3' end of the virus and the host DNA, displacing one of the bonds in the host (Goff, 2007)(Figure 7D). The gap repair step closes the gap between the *att* site at the 5' end of the viral DNA and the host DNA (Goff, 2007). This

is performed by host enzymes (Engelman, 2010) (Figure 7E). The integrated virus is known as a provirus (Bannert et al., 2010). The CA *att* sites on either end are conserved in newly integrated proviruses (Bannert et al., 2010).

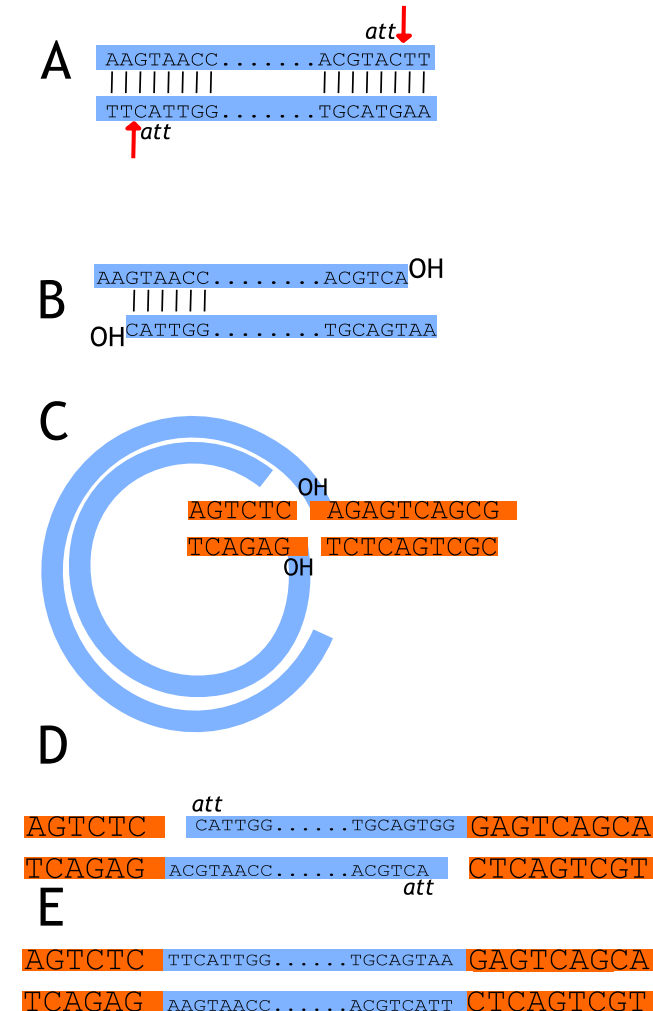


Figure 7: Stages of integration.

A) End processing. A dinucleotide is removed from both 3' ends of the viral cDNA (blue) through hydrolysis of the phosphodiester bond 3' to a CA dinucleotide, by a water molecule from which a proton is removed by an Mg^{2+} ion, catalysed by IN; B) this leaves an exposed hydroxyl group at the 3' end of each strand; C) Joining. The exposed hydroxyl group is used by IN to cut host DNA (orange) and join it to viral DNA; D) new phosphodiester bonds form between the 3' ends of the virus and the host DNA; E) Gap repair. Host enzymes close the gap between the 5' end of viral DNA and the host DNA. Adapted from Goff (2007), Hindmarsh and Leis (1999), Engelman (2010).

1.1.3.7. RNA Synthesis

After integration, the life cycle of the virus reaches the “late” stage, which is mediated by host rather than viral enzymes (Goff, 2007). Viral DNA in the host genome is transcribed to produce full length viral RNA genomes (Figure 5F).

Transcription is initiated by factors in the promoter and enhancer regions of the U3 region of the LTR of the provirus (Lenasi et al., 2010). First, the host TATA-binding protein identifies the TATA box in the promoter and recruits proteins, forming a transcription factor complex which associates with RNAPII. The promoter also binds to transcription factors involved in the activation of host cells (Lenasi et al., 2010). Elongation occurs until transcription is terminated at the polyadenylation signal (Guntaka, 1993).

This process generates a single pre messenger RNA (mRNA) transcript for the whole length of the retroviral genome(Goff, 2007). There are three possible fates for this transcript. First, a proportion of transcripts are directly exported from the nucleus and serve as the genome for progeny virions(Goff, 2007). Others are exported to the cytoplasm but are then translated to form Gag or Gag-Pol polyproteins(Goff, 2007). The remainder are spliced in the nucleus and translated to form Env and, in complex retroviruses, accessory proteins (Goff, 2007).

1.1.3.8. Translation

With the exception of the unspliced mRNA incorporated into new virions, retroviral mRNAs are translated into proteins (Figure 5G).

The Gag precursor protein is translated from full length mRNA transcripts in the cell cytoplasm and later cleaved by protease (Goff, 2007). Some retroviruses, for example MoMLV, also express a second, longer Gag protein, a modification which is thought to be involved in budding (Bannert et al., 2010).

Pro and Pol are also translated from full-length transcripts in the cytoplasm (except in spumaviruses, which use a spliced mRNA).

The abundance of Gag, Pro and Pol is determined by either translational read-through, translational frameshifting or splicing depending on the genus of retrovirus (Table 2). For translational read-through, *gag* and *pro-pol* are in the same reading frame, separated by a stop codon (Goff, 2007). Translation usually stops after Gag is produced, however in 5-10% of cases an amino acid is inserted at the position of the stop codon, allowing translation to continue through the *pro-pol* open reading frame (ORF) and form a Gag-Pro-Pol polyprotein. (Goff, 2007). In translational frameshifting, the genes are expressed separately in different reading frames. In alpharetroviruses and lentiviruses, translation normally results in a Gag protein, but in 10% of cases the ribosome slips back one nucleotide at a specific site near the end of the *gag* ORF, meaning it passes through the stop codon out of frame and synthesises the Pol protein (Goff, 2007). In beta and deltaretroviruses there are two frameshift sites, one between *gag* and *pro* and one between *pro* and *pol* (Goff, 2007). Each frameshift occurs approximately 30% of the time (Goff, 2007). In spumaviruses, *gag* and *pro-pol* are produced from separate spliced transcripts (Goff, 2007).

Unlike the other major genes, *env* is always expressed from a separate, spliced mRNA, from which *gag* and *pol* are removed as an intron (Goff, 2007).

Translation begins at a start codon at the 5' end of the gene, except in alpharetroviruses, where the start codon from the *gag* gene is used, after being brought close to the *env* sequence by splicing (Goff, 2007).

Table 2: Table showing the translational strategies for gag,pro and pol used by different retroviruses.

Abbreviations: TF, translational frameshift; TR, translational readthrough.

Genus	Strategy	Details
Alpha	TF	<i>Gag</i> and <i>pro</i> expressed as a single protein, <i>pol</i> in a separate reading frame.
Beta, Delta	TF	<i>Gag</i> , <i>pro</i> and <i>pol</i> in separate reading frames, separated by successive frameshifts.
Gamma, Epsilon	TR	<i>Gag</i> and <i>pro-pol</i> fusion in the same reading frame, separated by a stop codon.
Lenti	TF	<i>Gag</i> and <i>pro-pol</i> fusion in different reading frames, separated by a frameshift.
Spuma	Spliced	<i>Pro-pol</i> fusion expressed without <i>gag</i> in a separate, spliced mRNA.

1.1.3.9. Assembly, Packaging and Release

The next step in the viral life cycle is assembly of the newly synthesised proteins and the RNA genome and release of immature viral particles (Figure 5H). Assembly of new virions is co-ordinated by subunits of the Gag polyprotein, and the Gag polyprotein precursor is sufficient to assemble immature virus-like particles (Göttlinger and Weissenhorn, 2010).

The viral particle is then released by budding through the plasma membrane of the cell. There are two types of retroviral budding. In gammaretroviruses, epsilonretroviruses, alpharetroviruses and lentiviruses the viral capsid assembles during budding, so is not apparent while the virion is in the cell. In betaretroviruses and spumaviruses the capsid assembles in the cytoplasm of the host cell prior to budding (Strauss and Strauss, 2008). *Env* is recruited during budding by both groups (Strauss and Strauss, 2008). The incorporation of the RNA genome into the retroviral particle is known as packaging and is co-ordinated by interactions between the Psi element in the 5' UTR and the NC subunit of Gag (Goff, 2007). These interactions are highly specific and allow Gag to identify full length retroviral RNA when spliced transcripts are also present (Jouvenet et al., 2011).

Release from the plasma membrane is also co-ordinated by Gag (Pincetic and Leis, 2009). Specific motifs in Gag recruit host cell factors involved in producing vesicles in normal cells and co-opt this for viral budding (Pincetic and Leis, 2009).

1.1.3.10. Maturation

Shortly after their release from the host cell, immature retroviral particles undergo conformational changes to produce mature particles (Bannert et al., 2010) (Figure 5I). First, PR is activated and cleaves Gag-Pro-Pol polyproteins to release mature PR protein(Goff, 2007). Gag is then split by PR into MA, CA and NC. MA binds to the inside of the plasma membrane and the cytoplasmic tail of the Env protein(Goff, 2007). CA forms a shell around the inner core of the virus by assembling into higher order structures(Goff, 2007). NC associates with the viral RNA and protects it from degradation(Goff, 2007). Pol is also cleaved by PR into the IN and RT subunits(Goff, 2007). Env is already cleaved into its SU and TM subunits while it is being transported to the plasma membrane(Goff, 2007). The mature virus is structurally distinct and more stable than the immature form(Goff, 2007). After maturation, the virus is able to infect new cells (Goff, 2007).

1.1.4. Accessory Proteins

Simple retroviruses contain only *gag*, *pro*, *pol* and *env* ORFs, while complex retroviruses code for further accessory proteins (Goff, 2007).

Gammaretroviruses and alpharetroviruses have no known accessory genes. Accessory genes for each genus of retrovirus are shown in Table 3. Proteins encoded by these genes can have essential or advantageous effects on the retroviral life cycle and many are involved in counteracting host factors which otherwise limit retroviral infectivity.

1.1.4.1. **Betaretroviruses**

Betaretroviruses are generally classified as simple retroviruses, although two of the most widely studied families, mouse mammary tumour virus (MMTV) and human endogenous retrovirus (HERV) K, encode accessory genes.

Hayward et al. (2013b) propose that ancestral betaretroviruses were complex and that the simple betaretroviruses are a subgroup which later lost their accessory genes.

MMTV generates the superantigen (SAg) protein via alternative splicing of the *env* gene (Holt et al., 2013a). MMTV targets B-cells and SAg expressed on the surface of infected B-cells stimulates specific T-cells, activating them and leading to recruitment of further B and T cells for MMTV to infect (Holt et al., 2013a). Another type of alternative splicing of MMTV *env* produces the protein regulatory protein of MMTV (Rem), which is related to the HIV-1 protein “regulator of expression of virion proteins” (Rev) and is involved in nuclear export (Holt et al., 2013a). HERV-K can produce two alternative proteins from its *env* gene, the Rec protein, again functioning similarly to HIV-1 Rev and the Np9 protein via a 292 base pair deletion in Rec, which interacts with host pathways and is involved in tumorigenesis (Ruprecht et al., 2008).

1.1.4.2. **Lentiviruses**

Lentiviruses have a particularly high number of accessory genes. All lentiviruses encode the regulatory proteins transactivator of transcription (Tat) and Rev. All except equine infectious anaemia virus (EIAV) encode viral infectivity factor (Vif) (Bannert et al., 2010). The simian immunodeficiency viruses (SIVs) (including HIVs) encode two further proteins, viral protein R (Vpr) and negative factor (Nef) and sometimes additionally either viral protein U (Vpu) or viral protein X (Vpx) (Bannert et al., 2010) (Table 3). All lentiviral accessory genes are translated from separate, spliced mRNAs (Bannert et al., 2010) (Table 3).

The primary role of *vif* appears to be in counteracting the cellular apolipoprotein B-editing catalytic polypeptide family 3 (APOBEC3) retroviral restriction factors (section 1.2.3.1). APOBEC3 restriction factors cause a large decrease in viral reverse transcription if the viral *vif* is knocked out but the infectivity of wild-type HIV-1 is unaffected by APOBEC3G expression (Sheehy et al., 2002).

Vpr seems to enhance infection of macrophages in SIVs by facilitating nuclear import of the virus, although the extent of this effect varies between viruses and hosts (Ayinde et al., 2010). Vpx is only found in certain SIVs and is thought to be the result of a duplication event of Vpr (Ayinde et al., 2010). Vpx has a better understood role in degradation of the restriction factor SAM domain and HD domain-containing protein 1 (SAMHD1) expressed in dendritic cells, in which only Vpx positive SIVs replicate (section 1.2.3.1.) (Ayinde et al., 2010). In some species, Vpr may also perform this function (Lim et al., 2012).

Vpu, which is specific to certain subgroups of SIVs, counteracts another restriction factor, tetherin (Poli and Erfle, 2010). Nef may also be involved in this effect in some SIVs. Nef also has other known roles, for example in suppression of the host immune response and in apoptosis (Poli and Erfle, 2010).

Tat and Rev are both regulatory proteins and neither has yet been implicated in interaction with a restriction factor. Both are essential for viral infectivity (Poli and Erfle, 2010). Tat activates transcription of integrated proviruses via the LTR and also seems to be involved in interactions with cell surface receptors and in T-cell apoptosis (Poli and Erfle, 2010). Rev is involved in nuclear export of viral RNA, preventing excessive splicing (Poli and Erfle, 2010, Nakano and Watanabe, 2012).

1.1.4.3. **Deltaretroviruses**

Deltaretroviruses encode two accessory proteins, Tax and Rex, both of which are regulatory proteins (McGirr and Buehning, 2006) (Table 3). As with the lentiviral accessory genes, both are produced from separate, spliced mRNAs (Nakano and Watanabe, 2012). Tax has a similar role to lentiviral Tat, in that it activates transcription of the integrated virus and disrupts the cell cycle (Nakano and Watanabe, 2012). Rex is related to lentiviral Rev in that it is involved in nuclear export (Nakano and Watanabe, 2012). Although Tax and Tat perform similar roles, they are not homologous and work via different mechanisms, so cannot be replaced by each other (Nakano and Watanabe, 2012). Rev and Rex share a minimal amount of homology but have similar mechanisms, and HIV-1 Rev can be functionally replaced by human T-cell lymphotropic virus (HTLV) Rex (Nakano and Watanabe, 2012).

1.1.4.4. **Epsilonretroviruses**

The exogenous fish epsilonretrovirus walleye dermal sarcoma virus (WDSV) encodes three accessory proteins, Rv-cyclin (encoded by ORF *a*), Orf-B and Orf-C (Joel and Sandra, 2010) (Table 3). Rv-cyclin and Orf-B are expressed from spliced transcripts while Orf-C is cleaved from full length viral mRNAs (Joel and Sandra, 2010). Rv-cyclin and Orf-B are involved in tumour development while Orf-C is involved in apoptosis tumour development and tumour regression. These three proteins are essential for WDSV proliferation and dissemination (Rovnak and Quackenbush, 2010). Snakehead retrovirus, another fish epsilonretrovirus, also contains three potential ORFs, but these have not been well characterised (Hart et al., 1996).

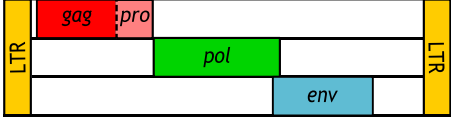
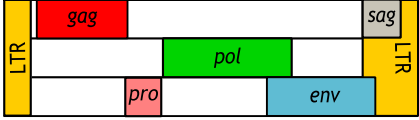
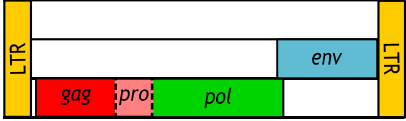
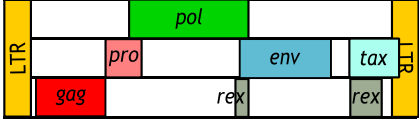
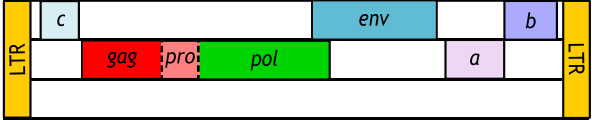
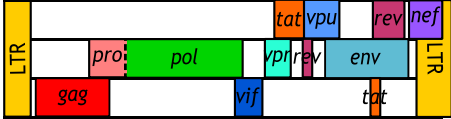

1.1.4.5. **Spumaviruses**

Spumaviruses encode two accessory proteins, transactivator of spumaviruses (Tas) and Bet (Table 3). Tas acts to increase its own production and, on

reaching a critical level, acts on the 5' LTR, where it activates the promoter to express the main retroviral genes (Bannert et al., 2010). Tat activates Bet, which is not well characterised but seems to encourage viral persistence in the host (Bannert et al., 2010).

Table 3: The accessory genes and genome structure of an example of each genus of retrovirus.

Each row of the genome structure diagrams represents a reading frame, dotted lines represent the division between genes which are transcribed together. Genome diagrams adapted from Bannert et al. (2010).

Genus	Type	Accessory Genes	Example Genome Structure
alpha	simple		 <p>Rous sarcoma virus</p>
beta	simple / complex	<i>sag, rem</i>	 <p>Mouse mammary tumour virus</p>
gamma	simple		 <p>Murine leukaemia virus</p>
delta	complex	<i>tax, rex</i>	 <p>Human T-Cell Lymphotropic Virus 1</p>
epsilon	complex	<i>a, b, c</i>	 <p>Walleye Dermal Sarcoma Virus</p>
lenti	complex	<i>vif, vpr, tat, vpu, rev, nef</i>	 <p>HIV-1</p>
spuma	complex	<i>bet, tas</i>	 <p>Simian Foamy Virus 1</p>

1. 2. Exogenous Retroviruses, Disease and Host Defences

Several major pathogens infecting humans and animals are retroviruses. Most notably, HIV affects over 30 million people worldwide and is responsible for approximately 1.8 million deaths per year (Central Intelligence Agency, 2013). A similar virus, feline immunodeficiency virus (FIV), is a major cause of disease in domestic cats, affecting approximately 11% of cats worldwide and usually progressing to feline acquired immunodeficiency syndrome (feline AIDS), which is fatal (Richards, 2005). Many oncogenic retroviruses are also known. This section provides a general introduction to some of the major retroviral pathogens.

1.2.1. HIV, SIV and FIV

SIVs (including HIV) and FIVs are major pathogens infecting old world primates and cats respectively. Both viruses cause disease through progressive immunodeficiency leading to vulnerability to opportunistic infections (Vogel et al., 2010).

1.2.1.1. Naturally Infected Hosts

SIVs tend not to result in disease in their natural primate hosts. SIVs causing known immunodeficiency syndromes, such as HIV-1 and HIV-2, SIV chimpanzee (SIVcpz) and SIV macaque (SIVmac) are all the result of cross-species transmission events (Mansfield et al., 1995, Khan et al., 1991, Hahn et al., 2000, Bailes et al., 2003). For FIVs this is less well characterised but not all cat species with FIV infections progress to feline AIDS (Pecon-Slaterry et al., 2008).

Several factors are thought to be responsible for the lack of progression to AIDS of natural SIV infections (Pandrea and Apetrei, 2010). The old world monkeys naturally infected with SIVs have been exposed to the viruses for at

least four million years, giving the hosts a lot of time to adapt and avoid the deleterious effects of the viruses (Gilbert et al., 2009, VandeWoude and Apetrei, 2006). Many cellular restriction factors have been identified in these species which reduce or prevent the replication of lentiviruses in certain cell types and certain hosts (section 1.2.3). These restriction factors are a major anti-lentiviral adaptation and are likely to be largely responsible for the differences in SIV response in different hosts. The disease-causing cross-species transmissions of SIV which we are currently aware of are recent, for example the HIV-1 outbreak is estimated to have originated in around 1920 and SIVmac outbreaks in the 1970s (Hahn et al., 2000). Therefore, these hosts may not have had time to develop the kind of defences seen in the original host species.

1.2.1.2. Progression to AIDS

HIV, FIV, SIVcpz and SIVmac infections cause gradual deterioration of immune function, eventually leading to death of the host through opportunistic infections (Murphy et al., 2008, Poli and Erfle, 2010, Sellon and Hartmann, 2006). HIV-1 and FIV are particularly widespread within their host populations, have serious detrimental effects and share many characteristics. HIV and FIV mutate rapidly while in the host, so many different variants can result from a single infection (Murphy et al., 2008). This allows different cell types to be infected and different co-receptors to be used within the same infection (Murphy et al., 2008).

In humans, HIV is usually transmitted through sexual contact, as the virus is present in semen and mucosal surfaces (Moir et al., 2011). It can also be transmitted through use of needles contaminated with infected blood, blood transfusions from an infected to an uninfected individual, from an infected mother to her foetus or to young children via breast milk (Moir et al., 2011). In cats, FIV transmission usually occurs through bites (Voevodin and Marx, 2009).

The progression of a typical, untreated HIV or FIV infection can be broadly divided into four stages. First, during early infection, the virus enters the body and makes its way to its target cells (Moir et al., 2011). HIV-1, HIV-2 and SIVs can only bind to cells bearing the receptor CD4 and a co-receptor, usually a CXCR4 or CCR5 chemokine receptor (Reitz and Gallo, 2010). CD4 receptors are found on immune cells, including T-cells, macrophages and dendritic cells (Stevenson, 2003). FIV does not use feline CD4 as a receptor, but rather CXCR4, which is expressed in a large number of cell types susceptible to FIV infection (Sellon and Hartmann, 2006). With transmission into the bloodstream the HIV or FIV virus comes directly into contact with immune cells, while with sexual transmission it crosses the mucosal surface of the genital or rectal tract and enters the lamina propria layer beneath the surface, which is rich in resting CD4+ T-cells which have not yet been activated by exposure to a pathogen (Haase, 2005). These cells support viral replication, so the virus population slowly begins to increase (Haase, 2005). This stage tends to last only a few days and the virus population generally remains too low to activate the host immune system (Haase, 2005). The host does not display clinical signs of the disease during this time (Haase, 2005).

The virus replicates in these CD4+ T-cells and spreads through the lymphatic system until the viral population is large enough to form a reservoir in the lymphoid tissue (Haase, 2005). This corresponds with the acute stage of the disease and begins two to four weeks after infection (Moir et al., 2011).

Lymphoid tissue is dense in CD4+ T-cells, so once this tissue is accessed the virus can infect many cells in a short time and replicate very quickly (Haase, 2005, Moir et al., 2011). It is at this stage that viraemia reaches its peak (Moir et al., 2011). Approximately 70% of humans infected with HIV suffer from an influenza-like illness during this acute phase but are rarely diagnosed with HIV (Fauci, 1993, Murphy et al., 2008). In FIV infected cats the acute phase lasts from several days to a few weeks and causes mild physical symptoms (Sellon and Hartmann, 2006).

The immune system is activated by the high viral load at this stage and the host generates a cellular and humoral immune response (Sellon and Hartmann, 2006, Fauci, 1993). The humoral immune response involves generation of antibodies against viral proteins (Sellon and Hartmann, 2006). HIV-specific CD8⁺ T-cell mediated immunity is also activated in the host (Sellon and Hartmann, 2006, Murphy et al., 2008). However, this also supplies new CD4⁺ T-cells for the virus to infect, which suppresses the immune response to some extent (Haase, 2005). The virus evolves quickly to evade the specific CD8⁺ T-cell mediated immune response by accumulation of mutations in the regions recognized by the CD8⁺ cells, resulting in the development of a population of viral “quasi-species” (Moir et al., 2011). The virus forms large reservoirs in the gut-associated lymphoid tissue and lymph nodes consisting of latently infected cells and cells with low level viral propagation.

The presence of these reservoirs, the rapid evolution of the virus and suppression of the immune system by the virus mean that the virus is never fully cleared by the host immune system (Moir et al., 2011). However, activation of the immune system does decrease the viral load and most leukocytes return to their normal levels, with the exception of CD4⁺ T-cells, which remain moderately reduced (Fauci, 1993). This occurs over several months, until the virus reaches its set point, an equilibrium at which low level virus replication occurs while the immune system is chronically activated, with a low, relatively stable virus population (Moir et al., 2011). At this equilibrium, the host enters the asymptomatic phase of infection and generally appears healthy. This phase can last between six months and twenty years in humans (Murphy et al., 2008, Sellon and Hartmann, 2006).

Although the host appears healthy during this phase, it is accompanied by gradual progressive disruption of normal immune function (Sellon and Hartmann, 2006, Murphy et al., 2008). This is partly due to the ongoing depletion of CD4⁺ T-cells by the virus. The combination of gradually

decreasing CD4+ T-cell populations and general immune disruption continue until the host has few remaining CD4+ T-cells.

At this point, the host reaches the terminal phase of infection, referred to as AIDS (Murphy et al., 2008, Sellon and Hartmann, 2006). At this stage, various opportunistic infections can occur, these can be viral, bacterial, protozoal and fungal (Sellon and Hartmann, 2006). These result in a wide range of clinical symptoms (Sellon and Hartmann, 2006).

1.2.1.3. Treatment

HIV infection is currently generally considered to be irreversible but can be managed with anti-retroviral therapy (ART). There are currently five main categories of ART: nucleoside/nucleotide RT inhibitors (NRTIs), non-nucleoside RT inhibitors (NNRTIs), protease inhibitors (PIs), integrase inhibitors (IIs) and entry inhibitors (CCR5 agents and fusion inhibitors). (Kanters et al., 2014). Very briefly, NRTIs are incorporated into the viral DNA during reverse transcription, preventing further transcription. NNRTIs, PIs and IIs bind to RT, protease and integrase respectively and inhibit their activities in the retroviral life cycle (Michael and Moore, 1999). CCR5 inhibitors bind CCR5 HIV coreceptors on the host cell and fusion inhibitors bind to Env on the HIV surface, both of which prevent the interactions necessary for membrane fusion (Eggink et al., 2010, Michael and Moore, 1999). Combinations of classes of ART are often used to effectively treat HIV (Kanters et al., 2014). ART is currently in use by more than nine million people worldwide (Kanters et al., 2014). In developed countries the life expectancy of HIV infected patients on combined ART is near-normal (Samji et al., 2013). However, the high mutation rate of HIV means that drug resistance mutations can occur (Tang and Shafer, 2012).

Research into ART tends to concentrate on treating HIV and the majority of ART has been shown to be ineffective against FIV, due to differences in

enzyme structure between the two viruses (Schwartz et al., 2014). However, NRTIs can be effective against FIV (Schwartz et al., 2014). Due to cost, side effects and difficulty of administration these drugs are rarely used in cats and instead treatment concentrates on managing secondary infections (Schwartz et al., 2014).

No treatment has been developed which can completely remove HIV from an infected patient, but, in a few cases, functional cures (meaning the virus is not completely eradicated but the patient no longer needs ART) have been observed. Persaud et al. (2013) treated an infant with ART from 30 hours until 18 months of age, at which point treatment was stopped. When the child was 30 months old HIV-1 RNA, DNA and antibodies were undetectable in his system (Persaud et al., 2013). An adult patient in Germany who was HIV-1 positive and suffered from acute myeloid leukaemia was treated with a stem-cell transplant from an individual with a specific CCR5 mutation which confers resistance to HIV-1 and had undetectable levels of HIV 20 months later without ART (Hütter et al., 2009). Finally, Sáez-Cirión et al. (2013) identified 14 patients who started ART very soon after initial infection and continued for an average of three years then retained very low levels of viraemia despite stopping therapy (Sáez-Cirión et al., 2013). The mechanism for this is not clear.

1.2.2. Oncogenic Retroviruses

Many retroviruses are oncogenic, causing cell transformation and leading to excessive cell proliferation and tumours. There are three main mechanisms through which retroviruses can cause cellular transformation – transduction, *cis*-activation of host genes and *trans*-activation of host genes.

1.2.2.1. Transducing Retroviruses

Transduction occurs when errors in recombination lead to the replacement of some retroviral RNA with cellular RNA, probably as a result of errors in packaging and reverse transcription (Pedersen and Sørensen, 2010). If the acquired cellular RNA includes a proto-oncogene (a normal cellular gene which has the potential to transform cells when overexpressed or altered) it can be mutated, or overexpressed by strong viral promoters, leading to transformation of cells with modified virus (Burmeister, 2001, Pedersen and Sørensen, 2010). Transducing retroviruses are also known as “rapidly transforming viruses” because they only require a short incubation period (Burmeister, 2001).

There are many transducing retroviruses. The first to be discovered was Rous sarcoma virus, in which the *src* gene acts as a proto-oncogene (Weiss and Vogt, 2011). *src* encodes a tyrosine kinase which affects cellular signalling and increases cell division, leading to transformation. In primates, woolly monkey sarcoma virus (WMSV) includes the transforming gene *v-sis*, a growth factor, which is constitutively expressed in its proviral form but only transiently expressed in normal cells and therefore leads to transformation (Doolittle et al., 1983). Acquired cellular genes sometimes replace an essential part of the viral RNA, meaning the virus is defective and co-infection with a second retrovirus is needed for propagation (Weiss and Vogt, 2011).

1.2.2.2. Cis-Activating Retroviruses

Cis-activation of host genes occurs when a retrovirus is inserted close to a cellular proto-oncogene, which can therefore be activated by viral promoters or enhancers (Burmeister, 2001). This is usually the route through which simple retroviruses become oncogenic (Burmeister, 2001). *Cis*-activating retroviruses are also known as non-acutely transforming retroviruses (Pedersen and Sørensen, 2010).

MMTV is an important example of a *cis*-acting retrovirus (Burmeister, 2001). It acts via activation of proto-oncogenes, particularly the fibroblast growth factor (*Fgf*) family and *Wnt1*, which show significant overexpression in MMTV tumours (Theodorou et al., 2007). Typically, both *Fgf* and *Wnt1* are found close to MMTV integration sites in MMTV-infected mice (Ross, 2010). Enhancer sequences in the LTRs of MMTV act on the promoters of *Wnt1* and *Fgf* genes, disrupting the regulatory controls normally in place during development and allowing transformation to occur (Callahan and Smith, 2000, Pedersen and Sørensen, 2010).

1.2.2.3. Trans-Activating Retroviruses

Trans-activation of host genes occurs when complex retroviruses encode viral proteins which act oncogenically. One important group of *trans*-activating viruses is the deltaretroviral HTLVs (Burmeister, 2001). HTLV-1 causes adult T-cell leukaemia (ATL) in humans (Matsuoka and Jeang, 2007). Only 6.7% of male and 2.1% of female carriers of HTLV-1 develop the disease (Matsuoka and Jeang, 2007). HTLV-1 transforms cells via transcription of virally encoded Tax proteins, which prevent apoptosis in infected cells and disrupt cell cycle checkpoints which would otherwise detect damaged ATL cells, triggering cell proliferation (Matsuoka and Jeang, 2007).

1.2.3. Restriction Factors

As retroviruses evolve they are constantly working against host mechanisms to minimise the damage caused by retroviral infection. One key route through which this is achieved is via retroviral restriction factors: proteins encoded by host genes which block or slow the spread of retroviral infection (Luban, 2010). Restriction factors have been identified which act to disrupt almost every stage of the retroviral life cycle. The majority of research into restriction factor function has concentrated on HIV-1, however, it is likely that in many

cases these restriction factors are the product of past selection pressure from historical infection with other retroviruses and possibly also non-retroviruses.

1.2.3.1. Uncoating and Reverse Transcription

The most well-characterised and specific retroviral restriction factors appear to act once the virus has entered the cell but before integration can occur, during the uncoating and reverse transcription stages of the life cycle. Three of these restriction factors – the APOBECs, tripartite motif containing protein 5 alpha (TRIM5 α) and SAMHD1 appear to have played a particularly major role in the host specificity of retroviruses and in the evolution of viral accessory genes.

The APOBEC family of genes, APOBEC1, APOBEC2, APOBEC3A to APOBEC3H and APOBEC4, code for proteins catalysing the deamination of cytosine (C) to uracil (U) in DNA and RNA (Sawyer et al., 2004, Jarmuz et al., 2002). Several of these proteins have been shown to reduce the infectivity of HIV-1 and several other retroviruses when the viral *vif* accessory gene is either removed or is not present. Notably, only one human protein, APOBEC3B, restricts wild-type HIV-1, however this protein is not expressed in the T-cells and macrophages targeted by HIV-1 (Chiu and Greene, 2008). When exposed to HIV-1 strains lacking *vif* (Δvif HIV-1), there is strong restriction of infectivity in the presence of two human proteins, APOBEC3G and APOBEC3F, plus moderate restriction in the presence of APOBEC3B, APOBEC3C and APOBEC3DE. APOBEC3G and APOBEC3F are expressed in T-cells and macrophages and are thought to be the major APOBECs affecting HIV-1 host range and cell tropism (Chiu and Greene, 2008).

APOBEC3G and APOBEC3F use the minus strand of retroviral DNA as a substrate for C to U deamination. Accordingly, MLV produced in cells expressing APOBEC3G has been shown to have a much higher level of plus strand guanine (G) to adenine (A) mutations than MLV produced in cells without APOBEC3G, the result of deamination from C to U in the minus

strand (Conticello et al., 2003) . The restriction of reverse transcription by these proteins has two stages (Chiu and Greene, 2008). Mariani et al. (2003) demonstrated that in cells expressing APOBEC3G, newly synthesised reverse transcripts are degraded as their APOBEC3G induced mutations reduce stability, which greatly reduces integration (Mariani et al., 2003). This instability seems to be the result of N-glycosylase activity, which removes uracil residues from the DNA (Zhang and Webb, 2004) (Figure 8). Secondly, any uracil residues which evade this degradation become G to A mutations in the plus strand, often resulting in inability of these transcripts to produce functional proteins (Zhang and Webb, 2004, Chiu and Greene, 2008) (Figure 8).

Given that APOBEC3 proteins cause such a severe decrease in HIV-1 infectivity and are expressed in T-cells and macrophages, the primary targets of HIV-1, it is surprising that HIV-1 is able to infect humans so effectively. This appears to be the result of the *vif* accessory protein (Mariani et al., 2003) (see section 1.1.4.2). Cells expressing APOBEC3G infected with Δvif HIV-1 show minimal levels of infection, while the infectivity of wild-type HIV-1 is unaffected by APOBEC3G expression (Sheehy et al., 2002). Mariani et al. (2003) hypothesised that Vif binds to the APOBEC3G protein, forming a complex which prevents APOBEC3G incorporation into virions, where it would normally induce cytidine deamination. They found that APOBEC3G incorporation in virions is greatly decreased in the presence of Vif (Figure 8). Vif also appears to counteract APOBEC3G by triggering its degradation through a proteasome-dependent pathway (Conticello et al., 2003).

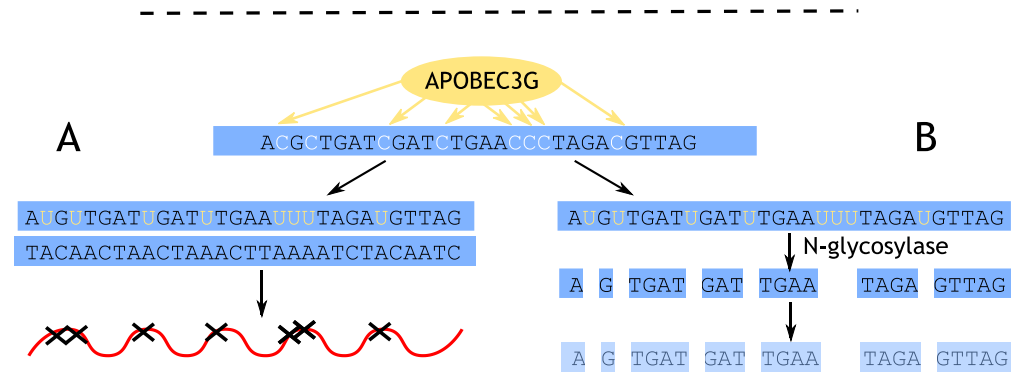


Figure 8: The impact of APOBEC3G on cells without Vif.

Without Vif (Δvif) APOBEC3 induces C to U hypermutation in the minus strand of retroviral DNA. This leads to A) G to A hypermutation in the plus strand, leading to translation of non-functional viral proteins or B) N-glycosylase degradation of U residues leading to instability and degradation of reverse transcripts. In the presence of Vif, Vif binds to APOBEC3G and prevents its incorporation into virions, preventing its action on viral DNA.

SAMHD1 is a phosphohydrolase enzyme responsible for converting deoxynucleoside triphosphates (dNTPs) into deoxynucleosides and inorganic triphosphate (Goldstone et al., 2011). Deficiencies in this gene can result in Aicardi-Goutieres syndrome, which results in inappropriate activation of the immune system (Laguetta et al., 2011). The gene is highly expressed in dendritic and other myeloid cells, which are known to be resistant to HIV-1 replication (Laguetta et al., 2011). If SAMHD1 is silenced in these resistant cell lines they become susceptible to HIV-1 and accumulate viral DNA (Laguetta et al., 2011).

The restrictive ability of SAMHD1 appears to be partly the result of its ability to hydrolyse cellular dNTPs (Goldstone et al., 2011). In the presence of SAMHD1, the pool of dNTPs available in a cell is greatly reduced (Goldstone et al., 2011). As dNTPs are required for reverse transcriptase to convert RNA to DNA, this provides a feasible route for SAMHD1 to limit reverse transcription (Goldstone et al., 2011). However, SAMHD1 is only able to restrict HIV-1 if it is unphosphorylated at a specific residue, T592, but phosphorylation of this residue does not affect dNTP depletion by SAMHD1 (White et al., 2013).

Therefore, another mechanism must be limiting reverse transcription in this case, possibly direct targeting of viral RNA by SAMHD1 (White et al., 2013).

Although HIV-1 is effectively restricted by SAMHD1, SIVs with the *vpx* accessory gene are able to replicate in dendritic cells and primary macrophages (White et al., 2013). This appears to be the result of the binding of Vpx protein to SAMHD1, which targets the protein for proteasomic degradation (Goldstone et al., 2011). If a dendritic cell is infected with HIV-1 and *vpx* positive SIV sooty mangabey (SIVsm) the HIV-1 becomes replication competent, presumably as a result of the presence of Vpx (Hrecka et al., 2011). In some SIV strains lacking *vpx*, the product of the *vpr* gene is able to target SAMHD1 for degradation, however this is not the case for HIV-1 *vpr* (Lim et al., 2012). SAMHD1 is found in all primates and the proteins from different primates are degraded with various degrees of specificity by Vpx and Vpr proteins from different SIV strains (Lim et al., 2012).

TRIM5 α is a member of the tripartite motif family, a group of more than 50 proteins containing RING domains, zinc binding domains which are typically involved in protein-protein interactions (Towers, 2007). The macaque version of this protein is thought to be responsible for a 100-fold reduction in HIV-1 replication in macaque cells compared to human cells (Himathongkham and Luciw, 1996). Introduction of rhesus TRIM5 α into human cells also leads to resistance to SIV and MLV (Stremlau et al., 2004). TRIM5 α proteins seem to be a major determinant of the ability of the host species to restrict certain retroviruses and prevent their replication (Keckesova et al., 2004).

TRIM5 α acts after the virus has entered the cell but before integration (Towers, 2007, Gong et al., 2011). The protein binds directly to the capsid of the retrovirus, which accelerates uncoating and targets it for degradation (Kaiser et al., 2007, Stremlau et al., 2004). This involves degradation of CA domain of the Gag protein, which results in disassembly of the virion, reducing the availability of virions for subsequent stages of the viral life cycle (Sakuma et al., 2007, Takeuchi et al., 2013). This degradation appears to involve

interaction with the proteasome, as when the proteasome is inhibited, viruses which are otherwise blocked by TRIM5 α are able to reverse transcribe, although they remain uninfected (Towers, 2007). The exact mechanism of this interaction is not yet well understood. The interaction between CA and TRIM5 α also appears to trigger a signalling pathway which produces a generalised antiviral state in the host via activation of the innate immune system (Pertel et al., 2011).

Different primate species encode TRIM5 α proteins with different species specificities, for example human TRIM5 α will not restrict HIV-1 in rhesus macaques, but rhesus macaque TRIM5 α will (Kaiser et al., 2007). HIV-2 is moderately restricted by human TRIM5 α , which may explain its reduced infectivity in humans compared to HIV-1 (Takeuchi et al., 2013). This may be because HIV-2 was transmitted to humans from sooty mangabeys, a fairly distantly related primate species, while HIV-1 was transmitted from chimpanzees, which are much closer relatives of humans (Takeuchi et al., 2013). This means that HIV-1 may have evolved the ability to evade the relatively similar chimpanzee TRIM5 α in its previous host, while HIV-2 was only exposed to the more distantly related sooty mangabey TRIM5 α (Takeuchi et al., 2013).

1.2.3.2. Translation

The Schlafen genes are a type of interferon-stimulated early response gene which are induced by pathogens, either directly or via interferon (Li et al., 2012). At least one of these genes, Schlafen11, has been shown to restrict HIV-1 (Li et al., 2012). In the presence of Schlafen11 there is a reduction in the level of viral proteins expressed by cells but no generalised reduction in cellular protein expression (Li et al., 2012).

The restriction on HIV-1 replication by Schlafen11 appears not to affect entry or reverse transcription but to reduce the formation of viral particles via inhibition of the production of viral proteins (Li et al., 2012). The limitation of

this effect to viral proteins seems to be the result of exploitation of the codon usage bias in viral, compared to host proteins (Li et al., 2012). Normally, HIV-1 will alter the tRNA concentrations in cells to promote viral synthesis, however, Schlafen11 interacts with tRNAs and prevents this, via a currently unknown mechanism (Li et al., 2012).

1.2.3.3. Release

Tetherin (encoded by the bone marrow stromal antigen-2 gene) is a transmembrane glycoprotein which is induced by interferon (Le Tortorec et al., 2011). The protein is unusual in that it is anchored to the cell membrane at either end, with the anchored regions connected by an extracellular domain (Le Tortorec et al., 2011). When expressed on the cell surface, Tetherin greatly reduces replication of lentiviruses lacking the *vpu* accessory gene, a gene only seen in HIV-1, SIVcpz, SIV gorilla (SIVgor) and some old world monkey SIV strains. Replication of HIV-1 lacking *vpu* in cells where this protein is stably expressed can be 80 to 100 fold lower than in cells not expressing the protein (Neil et al., 2008). This effect appears not to be specific to retroviruses, as *vpu*, an antagonist of Tetherin, also increases the release of particles of other viruses, such as the Ebola virus (Neil et al., 2008).

In the presence of tetherin, particles of HIV-1 lacking *vpu* are synthesised as normal but then accumulate at the surface of host cells, due to the “tethering” activity of the protein (Neil et al., 2008). These virions are then internalised into endosome and degraded (Neil et al., 2008). The effect of Tetherin appears to be physical anchoring of the virion to the plasma membrane, with Tetherin forming cross-links between the virion and the membrane (Le Tortorec et al., 2011). This effect depends on the topology, rather than the amino acid sequence of the protein (Le Tortorec et al., 2011).

The ability of HIV-1 to replicate in human cells expressing Tetherin is the result of its *vpu* accessory gene (Le Tortorec et al., 2011). For other SIVs the *nef* and *env* genes are also involved (Le Tortorec et al., 2011). The interaction

between Vpu and Tetherin is not completely understood but appears to be a physical interaction between the transmembrane domains of the two proteins (Le Tortorec et al., 2011). Vpu expression also targets Tetherin for degradation using ubiquitination and prevents trafficking of Tetherin to the cell membrane (Le Tortorec et al., 2011). In HIV-2, the Env protein antagonises Tetherin in much the same way as HIV-1 Vpu (Le Tortorec et al., 2011). In SIVs lacking *vpu*, the *nef* protein seems to reduce the availability of Tetherin at the host cell surface (Le Tortorec et al., 2011).

1. 3. Endogenous Retroviruses

When, by chance, a retrovirus integrates into a germline, rather than a somatic cell, it has the potential to become an ERV (Bannert and Kurth, 2006). Any offspring from the cell will have the ERV as part of their genome. Once a retrovirus has endogenised it is subject to selection, mutation and genetic drift like any other genetic element and can spread through the host population to fixation, or be eliminated from the population entirely (Jern and Coffin, 2008). An insertion which is deleterious to the host is unlikely to persist, but a neutral or advantageous insertion can increase in frequency in the population over time, through natural selection and random genetic drift (Bannert and Kurth, 2006). ERVs also have the potential to proliferate within a genome, through reinfection, retrotransposition in *cis* and complementation in *trans*. This section will describe the life cycle of ERVs, how they are controlled by the host and examples of ERVs which are beneficial and detrimental to the host.

1.3.1. Life Cycle and Evolution

1.3.1.1. Integration

The distribution of integration sites of ERVs is different to that of XRVs, probably as a result of selection against ERVs which have a strongly detrimental effect on the host genome (Brady et al., 2009). XRVs have various integration patterns, but often have an increased frequency within transcriptional units, while ERVs tend to be located outside of transcriptional units (Nellaker et al., 2012, Brady et al., 2009). This is indicative of strong negative selection against insertions in transcriptional units preventing their spread through the population (Nellaker et al., 2012). Accordingly, Brady et al. (2009) found that HERV-Ks in the human genome are more likely to be found outside of genes, while a resurrected exogenous HERV-K was more likely to integrate into transcriptional units, gene dense regions and regions associated with gene activity (Brady et al., 2009). When ERVs are found within introns, there tends to be a bias towards those integrated antisense to the gene (Nellaker et al., 2012). In mice (and most likely in other mammals) this is less pronounced for more recent integrations (Nellaker et al., 2012). Together, these results show that retroviruses which become endogenous have the same integration site preferences as other XRVs but that integration into non-coding regions confers a selective advantage (Nellaker et al., 2012).

1.3.1.2. Proliferation

It is advantageous for an ERV to generate as many copies of itself in the germline as possible (Bannert and Kurth, 2006). There are three main routes through which copy number can increase.

Replication competent ERVs, retroviruses which have integrated into somatic cells and other XRVs can all produce active viral particles, leading to further integrations into the germline. This is known as reinfection. Reinfection

requires a fully functional virus, so there is purifying selection for all the genes to maintain their ORFs (Bannert and Kurth, 2006). Many families of ERV show evidence of consistent purifying selection on all three genes, suggesting that they have predominantly spread via reinfection (Bannert and Kurth, 2006). Mutations causing stop codons cannot be transmitted via this route (Belshaw et al., 2004). Belshaw et al. (2004) found few shared stop codons in members of the HERV-K human mouse mammary tumour virus like (HML) 2 family, which suggests that transmission of this group has been predominantly via reinfection. Purifying selection can also be detected by comparing the proportion of synonymous mutations (mutations not changing the amino acid sequence) and non-synonymous mutations (mutations changing the amino acid sequence) in a gene, as non-synonymous mutations are more likely to be selected against in a functioning gene (Belshaw et al., 2004). HERV-K (HML-2) elements have an excess of synonymous mutations in all of their genes, including *env*, which again suggests that they have predominantly proliferated via reinfection (Belshaw et al., 2004).

The second route through which copy number can increase is retrotransposition in *cis* (Belshaw et al., 2005b). Here, the virus uses its own *gag* and *pol* proteins to copy itself and invade new areas of the genome within the same cell (Belshaw et al., 2005b). These ERVs function in the same way as LTR retrotransposons. There is purifying selection on *gag* and *pol* but none on *env*, as this strategy does not require entry into new cells. Therefore the *env* gene can degenerate while *gag* and *pol* are constrained to remain intact (Bannert and Kurth, 2006, Belshaw et al., 2005b). Based on analysis of 31 families of HERVs, it has been demonstrated that elements using retrotransposition in *cis* can reach substantially higher copy numbers than those using reinfection (Belshaw et al., 2005b). All families of HERVs with copy numbers greater than 200 have ratios of synonymous to non-synonymous mutation which were inconsistent with reinfection (Belshaw et al., 2005b). One of these, the HERV-K (HML-3) group, showed a far higher proportion of non-synonymous mutations on its *env* gene than on *gag* and

pol, a pattern which suggests retrotransposition in *cis* (Belshaw et al., 2005b). Magiorkinis et al. (2012) looked at the relationship between the size of an ERV lineage (in terms of copy number) and its proliferation mechanism by detailed analysis of the intracisternal A-type particle (IAP) group of ERVs, found in basal primates, rodents and lagomorphs (rabbits and hares) (Magiorkinis et al., 2012). The extent to which an IAP lineage has proliferated showed a significant positive correlation with the extent of *env* degradation, with most of the largest expansions lacking an *env* gene (Magiorkinis et al., 2012). The extent of *env* degeneration was not related to the age of the lineage (Magiorkinis et al., 2012). Magiorkinis et al. (2012) also compared unusually abundant lineages of ERVs in each of 38 mammalian genomes, described as “megafamilies”, with smaller lineages in the same genome. These megafamilies were responsible for an average of 80% of insertions and all but one appeared to have proliferated via retrotransposition in *cis*, as their *env* genes were highly degraded (Magiorkinis et al., 2012). *Gag* genes in the megafamilies and *env* genes in smaller families were not similarly degraded (Magiorkinis et al., 2012). These results suggest that ERV families tend to initially spread via reinfection but the most successful families later adapt and become intracellular retrotransposons (Magiorkinis et al., 2012). It is not clear whether the increased success of retrotransposing lineages is due to retrotransposition being a more efficient way to proliferate than reinfection, if the lack of *env* on the cell surface prevents activation of the host immune system or if loss of *env*, preventing reinfection, decreases the fitness cost of ERV proliferation, either by direct harm caused by Env protein or through increased insertional mutagenesis in somatic cells with a reinfection strategy (Magiorkinis et al., 2012).

Finally, copy number can increase via complementation in *trans*, with viruses “hitchhiking” by using transcripts from intact proviruses or other transposons to provide the proteins they need to replicate (Magiorkinis et al., 2012, Belshaw et al., 2005b). This does not introduce purifying selection on any gene in the hitchhiking virus (Belshaw et al., 2005b). However, certain regulatory

motifs within the LTR are required for viral packaging (Belshaw et al., 2004). This route is less common than the other two proliferation mechanisms, possibly because it requires two related insertions to be expressed simultaneously in the same cell (Magiorkinis et al., 2012, Belshaw et al., 2005b). The most abundant subgroup of the HERV-H family of ERVs appears to have proliferated via complementation in *trans* (Belshaw et al., 2005b). Members of this subgroup share several large deletions and seem to have used proteins from related, intact HERV-H elements to proliferate through the genome (Belshaw et al., 2005b).

1.3.1.3. Degeneration

Unless there is purifying selection maintaining the function of a proviral gene, it will lose its ability to produce active proteins over time, accumulating mutations at the host mutation rate (Katzourakis et al., 2005). Almost every ERV in the human genome has inactivating mutations (Belshaw et al., 2004). Mutations are also acquired during the reverse transcription process in reinfection and retrotransposition (Katzourakis et al., 2005) and induced by host restriction factors such as APOBEC3G (section 1.2.3).

ERV coding sequences can also be lost by recombinational deletion (Belshaw et al., 2007). LTRs provide regions of high sequence similarity within the host genome, which can lead to recombination between these regions, removing the central coding region (Belshaw et al., 2007). The remaining sequence is known as a solo-LTR. Recombinational deletion of ERVs is common and solo-LTRs are far more numerous in host genomes than intact sequences (Belshaw et al., 2007).

1.3.2. Host Control of ERVs

1.3.2.1. Control of Transcription

Despite their degeneration, ERVs often continue to be transcribed by cellular factors in the host. As for XRVs, the transcription of ERVs is initially controlled by regulatory sequences in the U3 region of the 5' LTR (see section 1.1.2.3) (Schön et al., 2001). However, Nellaker et al. (2006) observed transcription of HERV-W elements with absent or truncated 5' LTRs in cell lines and in vivo, suggesting that promoters outside the 5' LTR must direct expression of these elements. The transcribed elements were more likely than other HERV-Ws to be found within intronic regions of host genes, so may be transcribed by host regulatory promoters (Nellaker et al., 2006). Horse ERVs have also been shown to be more likely to be transcribed if they fall within introns (Brown et al., 2012).

Over time, depending on the position of the ERV and its effect on host fitness, the host can prevent transcription of the ERV using epigenetic modifications, such as DNA methylation and histone acetylation, of the 5' LTR. A selectively neutral insertion is likely to assume into the epigenetic state of the surrounding sequence over time (Reiss and Mager, 2007). If the insertion has a positive or negative effect this changes, for example the LTR of syncytin-1, an ERV env gene which has taken on a role in host placental development (section 1.3.3.1), has lower levels of DNA methylation in placenta-derived cell lines (Reiss and Mager, 2007). Rebollo et al. (2012) found a higher incidence of hypermethylation in the 5' LTRs and hypomethylation in the 3' LTRs of mouse ERVs and Laska et al. (2013) found that the transcription of HERV-Fc1 was greatly increased on treatment with a demethylating agent and decreased by remethylation. Histone acetylation was also shown to be involved, with induced hyperacetylation causing increased HERV-Fc1 expression (Laska et al., 2013). As the host controls epigenetic modifications, changes in the cell can result in changes in these modifications which alter ERV transcription (Laska

et al., 2012). DNA from tumour cells often has altered DNA methylation (Ehrlich, 2002), which may partly explain the differences in ERV transcription profiles between normal and cancer cells (Stauffer et al., 2004).

ERVs transcription can also be affected by the activity of other transposable elements. For example, integrated HERV-W “pseudogenes” have been identified, which are structurally similar to retroviral mRNA, with a polyA tail, but lack normal retroviral structure and do not have intact LTRs (Pavlíček et al., 2002). As LTRs regenerate during normal retroviral reverse transcription (section 1.1.3.4) these are proposed to be generated by reverse transcriptase from LINE elements (Pavlíček et al., 2002). This hypothesis is strengthened by the presence of direct repeats of variable length around the insertions, which is characteristic of LINE activity (Pavlíček et al., 2002).

As well as internal factors, the transcription of ERVs is controlled to some extent by external stressors, such as infection. Some of the pathogenic effects of ERVs may result from activation of ERVs by microorganisms. Young et al. (2012) investigated the role of immunity in control of ERVs by comparing ERV expression in wild-type mice with B and T-cell deficient mice and found that increased bacterial colonisation in immunocompromised mice resulted in increased ERV transcription (Young et al., 2012). One ecotropic MLV locus, *Env2*, showed particularly elevated expression, shown to be the result of repair to a mutation in the *pol* region which inactivates this MLV locus in wild-type mice, by recombination with another ERV locus with a functional *pol* gene (Young et al., 2012). This suggests that antibodies are involved in preventing the emergence of infectious eMLV recombinants when ERVs are induced by microbial products (Young et al., 2012).

Nellaker et al. (2006) investigated the impact of viral infections, specifically herpes simplex virus I and influenza A, on expression of the HERV-W family. After infection with either virus there was a relative increase in *gag* and *env* related transcripts in some cell types (Nellaker et al., 2006). Herpesviridae are the viruses most commonly shown to be associated with ERV activation and

with multiple sclerosis (MS) (Perron et al., 2009). Therefore, it is possible that activation of ERVs by herpesviruses could be involved in MS pathogenesis (Perron et al., 2009).

Infection with exogenous retroviruses has also been shown to affect ERV transcription. For example, blood plasma from HIV-1 patients contains significantly more HERV-K transcripts, mainly from the HML-2 group (van der Kuyl, 2012, Contreras-Galindo et al., 2012). There is no significant difference in HERV-H transcription in plasma from infected and control patients, so the increase in transcription does not appear to be universal across all HERV groups (van der Kuyl, 2012). HML-2 transcripts from a few specific loci predominate (van der Kuyl, 2012, Contreras-Galindo et al., 2012). There are several hypotheses as to how the increase in transcription at these loci may occur (van der Kuyl, 2012). For example, HIV-1 accessory proteins or cellular proteins upregulated by HIV-1 infection may specifically activate some HERV loci (van der Kuyl, 2012). Opportunistic infections by other pathogens as a result of HIV-1 induced immunosuppression may also alter HERV transcription (van der Kuyl, 2012).

Parasites may also play a role in ERV activation. Frank et al. (2006) looked at the HERV expression profiles of cells infected with *Toxoplasma gondii*, the protozoan involved in toxoplasmosis. Transcriptional activation of members of various HERV families was observed (Frank et al., 2006). Again, this correlation suggests that HERVs are activated by *T. gondii* infection but could be the result of other factors, such as stress affecting the availability of transcription factors (Frank et al., 2006).

1.3.2.2. Innate Immunity

If ERVs retain their ability to produce active viral particles, the innate immune system of the host may be responsible for limiting their detrimental effect (Yu et al., 2012a). In mice, which have several families of active ERVs, three

nucleic acid recognising Toll-like receptors (TLRs) have been identified which are involved in ERV viraemia (Yu et al., 2012a). Loss of these genes led to high retroviral viraemia and the appearance of T-cell lymphoblastic leukaemia (Yu et al., 2012a). Experimental infection of mice with ERV-derived MLV leads to a TLR-dependent response (Yu et al., 2012a). These results suggest that innate immunity is involved in reducing the pathogenic effect of active ERVs in mice, a result which may also apply to other hosts with active ERV insertions.

1.3.2.3. Restriction Factors

Finally, the restriction factors which inhibit various stages of the exogenous retroviral life cycle (section 1.2.3), plus other specific restriction factors often show activity against ERVs and prevention of ERV activity may have played a part in their evolution (Stoye, 2012). It is not clear if restriction factors evolved predominantly to control ERVs or to control exogenous viruses which then became endogenous (Yap and Stoye, 2013) If restriction factors evolved as a consequence of ERV activity, the ERV should still be found in the host where the restriction factor is active, whereas if they evolved as a response to an exogenous virus, this virus may no longer be circulating, either in this host or at all (Yap and Stoye, 2013).

Several ERVs show evidence of inactivation by the APOBEC family of proteins. Esnault et al. (2008) found that human APOBEC3G results in a decrease in the number of transposed copies of IAP and MusD elements in a human cell line. This appears to be due to degradation of reverse transcripts, as seen for exogenous viruses in the presence of APOBEC3G. Integrated copies of IAP and MusD also contain evidence of G to A editing by APOBEC3G, leaving mutation insertions which are less likely to produce functional proteins (Esnault et al., 2008). Reconstituted HERV-K (HML-2) elements are sensitive to human APOBEC3F but resistant to human APOBEC3G and TRIM5 α (Lee and Bieniasz, 2007). Relics of APOBEC3 activity are also present in endogenous gammaretroviruses of chimpanzees and macaques and the mutation patterns

present are consistent with the types of APOBEC which are active in each host (Perez-Caballero et al., 2008). Many of the stop codons in these ERVs appear to be the result of APOBEC activity, suggesting that APOBECs were responsible for their inactivation (Perez-Caballero et al., 2008). Polytrophic and modified polytrophic endogenous MLVs in mice also have a high proportion of G to A mutations, which seem to have appeared between reverse transcription and integration (Jern et al., 2007). This suggests that APOBEC proteins played a part in the inactivation of these ERVs (Jern et al., 2007). These results together demonstrate that APOBEC proteins can be an effective defence against pathogenic activity of existing ERVs.

TRIM5 α may also play a role in defence against ERV activity. The European rabbit (*Oryctolagus cuniculus*) has several copies of the endogenous lentivirus rabbit endogenous lentivirus type K (RELK) in its genome, while its relative the pika (*Ochotona princeps*) does not (Yap and Stoye, 2013). Both species encode TRIM5 α proteins but the rabbit protein is considerably more active against RELK (Yap and Stoye, 2013). There is evidence of strong positive selection acting on the part of TRIM5 α which interacts with the lentiviral capsid (Lemos de Matos et al., 2011). It is possible that the presence of endogenous RELK in the rabbit genome provided the selection pressure to maintain or increase TRIM5 α activity against RELK (Yap and Stoye, 2013). However, the differences in activity may also indicate either that the pika was not exposed to exogenous RELK or that the pika successfully defended itself against ancient exogenous RELK then lost its restriction ability against this virus due to a lack of ongoing selection pressure (Yap and Stoye, 2013).

1.3.3. Benefits of ERVs

ERVs can have beneficial, detrimental or neutral effects on host fitness.

1.3.3.1. Capture of ERV Genes by the Host

Genes which entered the host genome within ERVs sometimes take on an essential role in the biology of their host. This is probably most well-documented in the case of “syncytin” proteins captured by placental mammals. Env proteins expressed on the surface of infected host cells can allow them to fuse with nearby cells with the right receptors and form large, multinucleated cells known as “syncytia” (Lavialle et al., 2013). This fusogenic function has allowed retroviral *env* genes to be co-opted on several independent occasions by mammalian hosts for a role in cell-to-cell fusion in formation of the syncytiotrophoblast, the multinucleated syncytial layer which separates maternal and foetal tissues in pregnancy (Palmarini et al., 2004, Lavialle et al., 2013).

It has been hypothesised that this capture of *env* genes was essential for the transition between egg-laying and placental mammals during vertebrate evolution (Lavialle et al., 2013). Placentas have convergently evolved frequently in different groups of organisms and appear to have originated more times than any other organ, with independently acquired ERVs selected for convergent roles in placental development (Palmarini et al., 2004, Dunlap et al., 2006). Env proteins are required for placental development in at least some members of the primates (Cáceres et al., 2006), rodents (Dupressoir et al., 2005), lagomorphs (Heidmann et al., 2009), carnivores (Cornelis et al., 2012) and ruminants (Cornelis et al., 2013) (see section 1.4.3.9 for discussion of origins of syncytins in different hosts). However, these syncytins appear to have originated considerably more recently than the appearance of placental mammals, approximately 170 million years ago (Lavialle et al., 2013). It has been proposed that placental mammals emerged as the result of capture of a syncytin from an ancient ERV but that the function of this syncytin has been repeatedly replaced throughout the mammalian lineage with new *envs* from more modern ERVs (Lavialle et al., 2013). If this is the case, syncytins should be present in all placental mammals and, accordingly, candidate syncytin

genes had been identified in at least all mammals screened prior to 2012 (data is unavailable after this date) (Cornelis et al., 2012). The hypothesis predicts that “lost syncytins” will be present in mammalian genomes, *env* genes which used to encode syncytins but were replaced (Lavialle et al., 2013). This hypothesis is consistent with the wide diversity of placental structures seen in placental mammals, as these could be the result of novel syncytin acquisitions (Lavialle et al., 2013).

1.3.3.2. Protection Against Other Retroviruses

It is also possible for hosts to protect themselves against retroviral attack using existing endogenous insertions (Jern and Coffin, 2008).

One mechanism through which this can occur is the expression of endogenous retroviral Env proteins on the cell surface, which block receptors which would otherwise allow related exogenous viruses to enter (Ikeda and Sugimura, 1989). This mechanism is seen with the murine *Fv4* gene, a truncated endogenous MLV which expresses Env, which binds to receptors and blocks ecotropic MLV infection (Ikeda and Sugimura, 1989). Two other *env* genes have been similarly adopted by mice, the *Rmcf* and *Rmcf2* Env proteins block entry by polytrophic MLVs (Wu et al., 2005). This mechanism is also seen in chickens, in which the Env protein expressed by endogenous ALV can block receptors and prevent binding of exogenous Rous sarcoma virus (Palmarini et al., 2004). In sheep, receptor interference is one route through which endogenous Jaagsiekte sheep retrovirus (JSRV) restricts exogenous JSRV (Palmarini et al., 2004).

Proteins derived from ERVs can also block invading retroviruses later in the retroviral life cycle as demonstrated by the *Fv1* gene of mice, which is derived from an ERV-L *gag* (Pincus et al., 1971, Best et al., 1996). This gene can restrict various exogenous retroviruses, including foamy viruses, EIAV and HIV-1 (Yap et al., 2014). Fv1 protein binds to CA of invading retroviruses after

entry and prevents nuclear entry (Yap et al., 2014). In sheep, enJSRV restricts exogenous JSRV after integration and transcription but before release of viral particles (Palmarini et al., 2004). The mechanisms through which these ancient ERV proteins block exogenous retrovirus infection are not well understood.

Finally, ERVs have been demonstrated to protect against exogenous retroviruses through disruption of the host immune system. The MMTV *sag* gene (section 1.1.4.1) provides an example of this mechanism. MMTV requires a pool of B-cells and of T-cells with an appropriate receptor to proliferate in the host (Holt et al., 2013a). Mice with an active endogenous MMTV SAg during embryonic development will delete these T-cells during their deletion of “self” antigens as the immune system develops (Holt et al., 2013a). This means the reactive pool of these T-cells required for successful exogenous MMTV infection is not present, so these mice are less vulnerable to exogenous MMTV (Holt et al., 2013a).

1.3.3.3. Gene Regulation

Retroviral promoters can have powerful effects on the expression levels of nearby genes. For example, the expression level of the agouti coat colour gene in mice is governed by the degree of methylation of the LTR of an upstream retroviral insertion (Reiss and Mager, 2007). Depending on the expression level of the insertion, mouse coat colour can vary from yellow through intermediate colours to wild-type (Reiss and Mager, 2007). In general, the regulatory effects of retroviruses seem to be detrimental to their hosts, as Rebollo et al. (2012) looked at the location of ERVs in relation to host genes and found evidence of negative selection pressure against ERVs close to the 5' and 3' ends of genes.

However, there are examples of regulation of host genes by ERV promoters having beneficial effects for the host. For example, ERV insertions provide the

initiation sites for transcription of salivary amylase genes in humans and other apes (Meisler and Ting, 1993, Stoye, 2012, Breslin 2013). Without these ERV insertions, it appears that these genes would only be expressed in the pancreas and not the saliva (Meisler and Ting, 1993, Stoye, 2012). Production of salivary amylase allows “predigestion” of starch in the diet and is advantageous for species with high starch diets (Breslin). This gene is present in a high copy number in humans and may have contributed to the move from a hunter-gatherer lifestyle to development of agriculture (Breslin).

1.3.4. ERVs and Disease

ERVs also have detrimental effects on their hosts and roles have been proposed for ERVs in many human and veterinary diseases.

1.3.4.1. Active ERVs

The most obvious mechanism through which an ERV can cause disease is through the production of active viral particles. However, any insertion which is detrimental to the host is subject to strong negative selection so there are few ERVs which consistently produce pathogenic viral particles.

The exceptions to this are very new insertions and recombinant viruses, which have not been subject to selection over an extended period. For example, in the AKR strain of mice, a high incidence of thymomas is linked to the presence of an endogenous ecotropic MLV (Stoye et al., 1991). The disease causing agents in these mice are viruses formed by recombination between these ecotropic MLVs and endogenous polytrophic MLVs (Stoye et al., 1991). The recently integrated koala retrovirus (KoRV) appears to be active and to be associated with neoplastic disease (Tarlinton et al., 2005).

1.3.4.2. Transcription and Expression

ERVs which are not actively producing new viral particles can still be involved in disease in their host. Many studies have demonstrated correlations between transcription and translation of HERV sequences and human disease, some of the most significant of these will be discussed here.

Stauffer et al. (2004) analysed the transcription patterns of intact ORFs in several families of HERV in normal and cancerous human tissue. HERVs were shown to be expressed in both tissue types, but with a significant difference in expression pattern (Stauffer et al., 2004). Many studies have found significant increases in the transcription of specific HERV groups in cancerous tissue compared to normal tissue. A small sample of these studies, covering 15 of the 20 most common types of cancer diagnosed in England in 2011 (Office for National Statistics) are listed in Table 4. Clearly, transcriptional differences in HERVs exist between cancerous and normal tissue and are widespread amongst different types of cancer and different HERVs. However, various factors interacting with HERV transcription (section 1.3.2.1) may be responsible for these differences, for example general hypomethylation of cancerous tissue or differences in exposure or vulnerability to other microorganisms. The transcriptome of cancerous tissue is also generally very different to that of normal tissue, so differences in HERV transcription are not unexpected (Rhodes and Chinnaiyan, 2005).

Table 4: Common types of cancer with a reported significant increase in transcription of at least one HERV family.

Cancer	HERV	References
Breast	HERV-K (HML-2)	Wang-Johanning et al., 2001, Contreras-Galindo et al., 2008
Ovarian	HERV-K (HML-2), HERV-E, ERV3	Wang-Johanning et al., 2007
Prostate	HERV-E, HERV-K (HML-2), HERV-H	Wang-Johanning et al., 2003, Goering et al., 2011, Stauffer et al., 2004
Colon	HERV-H	Liang et al., 2012, Wentzensen et al., 2007
Lung	HERV-R	Andersson et al., 1998
Melanoma	HERV-K (HML-2)	Schmitt et al., 2013, Singh et al., 2013
Lymphoma	HERV-K (HML-2)	Contreras-Galindo et al., 2008
Bladder	HERV-H	Wang et al., 2006, Stauffer et al., 2004
Kidney	HERV-E	Cherkasova et al., 2011
Brain	HERV-K, HERV-H, HERV-W	Balaj et al., 2011
Pancreas	HERV-H	Wentzensen et al., 2007
Leukaemia	HERV-K (HML-2)	Depil et al., 2002
Uterine	HERV-K(HML-2)	Wang-Johanning et al., 2007
Stomach	HERV-H	Stauffer et al., 2004, Wentzensen et al., 2007
Oral	HERV-K (HML-2)	Stauffer et al., 2004

The presence of ERV proteins can provide a more convincing link between ERVs and cancer. HERV proteins have been detected in tissue from germ cell tumours, melanoma, breast cancer, ovarian cancer, endometrial carcinoma and neuroblastoma, amongst others (Ruprecht et al., 2008). These cancers all also show significantly increased HERV transcription (Table 4).

Testicular germ cell tumours show increased HERV-K (HML-2) transcription, express HERV-K (HML-2) proteins and sometimes release defective viral particles (Moyes et al., 2007, Ruprecht et al., 2008). Patients with these tumours have a specific immune response against HERV-K (HML-2) proteins (Moyes et al., 2007, Ruprecht et al., 2008). HERV-K Rec can induce testicular carcinoma in mice and Np9 interacts with the Notch signalling pathway, which is involved in cancer (Kaufmann et al., 2010). Rec and Np9 proteins of HERV-K have been shown to bind to the promyelocytic zinc finger, a protein which acts as a transcriptional repressor in spermatogenesis (Moyes et al., 2007). This binding may impair the function of the protein and promote cell proliferation, forming tumours (Moyes et al., 2007, Kaufmann et al., 2010). Together, these results suggest that HERV-K may be directly involved in tumorigenesis for this particular type of cancer.

Syncytin-1, one of the ERV Env proteins involved in placental development in humans (section 1.3.3) is overexpressed in tissue from breast and endometrial cancer (Ruprecht et al., 2008). Fusion between cancer cells and endothelial cells is common and can alter the behaviour of tumours and the fusogenic activity of Syncytin may play a role in this (Bjerregaard et al., 2006). A proportion of breast cancer patients present the receptor required for Syncytin mediated fusion on cell surfaces and inhibition of Syncytin prevents cancer-endothelial cell fusions in breast cancer tissues (Bjerregaard et al., 2006). Knockout of Syncytin in endometrial cancer reduces proliferation of cells and cell to cell fusion (Strick et al., 2007).

There is also debate about the role of ERV transcription in autoimmune disease, particularly MS (Moyes et al., 2007). Several HERVs are upregulated

at the site of inflammation in MS, including HERV-W, HERV-K and HERV-H (Moyes et al., 2007, Garcia-Montojo et al., 2013). The *Env* protein of a particular HERV-W locus, multiple sclerosis associated retrovirus (MSRV) has known inflammatory properties (Moyes et al., 2007, Garcia-Montojo et al., 2013). Demyelinated lesions in MS patients show overexpression of MSRV *env* transcripts (Perron and Lang, 2010). Garcia-Montojo et al. (2013) found an elevated copy number of MSRV-like HERV-Ws in peripheral blood mononuclear cells in MS cases compared to controls. The proviral load was higher in more clinically severe cases (Garcia-Montojo et al., 2013). This suggests that MSRVs may continue to transcribe, integrate and retrotranspose in MS (Garcia-Montojo et al., 2013). Activation of MSRV *env* by herpesviruses may be involved in MS onset or progression (Perron et al., 2009) (see section 1.3.2.1).

Schizophrenia and bipolar disorder have also been hypothesised to involve ERVs. There have been several studies into the potential role of ERVs in linking environmental factors with onset of schizophrenia; environmental factors such as infection can induce ERV transcription (section 1.3.2.1), which could then activate further factors involved directly in the pathogenesis of schizophrenia. Perron et al. (2012) found significantly elevated HERV-W transcription in peripheral blood mononuclear cells from schizophrenia and bipolar disorder patients compared to controls. Similarly, Huang et al. (2011) found HERV-W *env* transcripts and high HERV-W reverse transcriptase activity in schizophrenia patients. The effect of HERV-W Env protein on the expression of three genes known to be associated with schizophrenia was also examined (Huang et al., 2011). All three of these genes produced higher mRNA levels in the presence of HERV-W Env. Therefore, it was proposed that external factors inducing HERV-W expression could activate these genes and allow them to express proteins involved in schizophrenia pathogenesis (Huang et al., 2011). Deb-Rinker et al. (1999) found an MSRV-like transcript in the placenta in the affected members of three pairs of monozygotic twins discordant for schizophrenia. Previous work suggested that schizophrenia may

be associated with disrupted foetal development, so the expression of this sequence in the placenta could potentially explain this disruption (Deb-Rinker et al., 1999). Also, MS, schizophrenia and bipolar disorder have all been shown to involve myelin impairment or inflammation, although in different areas of the brain, so, as all three disorders have elevated MSR *env* transcription it is possible these transcripts play a role in this degeneration (Perron et al., 2012).

1.3.4.3. Chromosome Disruption

Endogenous retroviruses can also affect the host genome by providing regions of similar or identical genetic sequence in different areas, allowing mispairing and unequal crossing over to occur (Deb-Rinker et al., 1999). Kamp et al. (2000) demonstrated that this phenomenon may cause a microdeletion, known as AZoopermia factor a, on the human Y chromosome which causes male infertility, as the deleted region is flanked by identical retroviral insertions.

1. 4. Endogenous retroviruses in vertebrate genomes.

1.4.1. Overview

Endogenisation is almost ubiquitous amongst the retroviruses, with endogenous examples found in six out of seven retroviral genera (only endogenous deltaretroviruses have not been identified). Endogenous retroviruses have been found in every vertebrate genome screened to date, including mammals, birds, reptiles, amphibians and fish. Figure 9 gives a broad overview of the diversity of ERVs in vertebrate genomes.

The retroviruses are traditionally divided into classes based on sequence similarity, with gamma- and epsilon- retroviruses as class I, alpha- and beta- retroviruses as class II and spumaviruses as class III. Spumaviruses are

generally considered to be found in all classes of vertebrates (Kambol and Tristem, 2005, Herniou et al., 1998, Chong et al., 2012, Bolisetty et al., 2012). For the class I retroviruses, gammaretroviruses are likely to be limited to mammals, reptiles and birds (2009, Niewiadomska and Gifford, 2013) and epsilonretroviruses are generally considered to be viruses of fish and amphibians (Herniou et al., 1998). Class II retroviruses are divided into the betaretroviruses, infecting mammals, and the alpharetroviruses, infecting birds (Bolisetty et al., 2012). Lentiviruses have, to date, only been identified in mammals (Gifford et al., 2008, Gilbert et al., 2009).

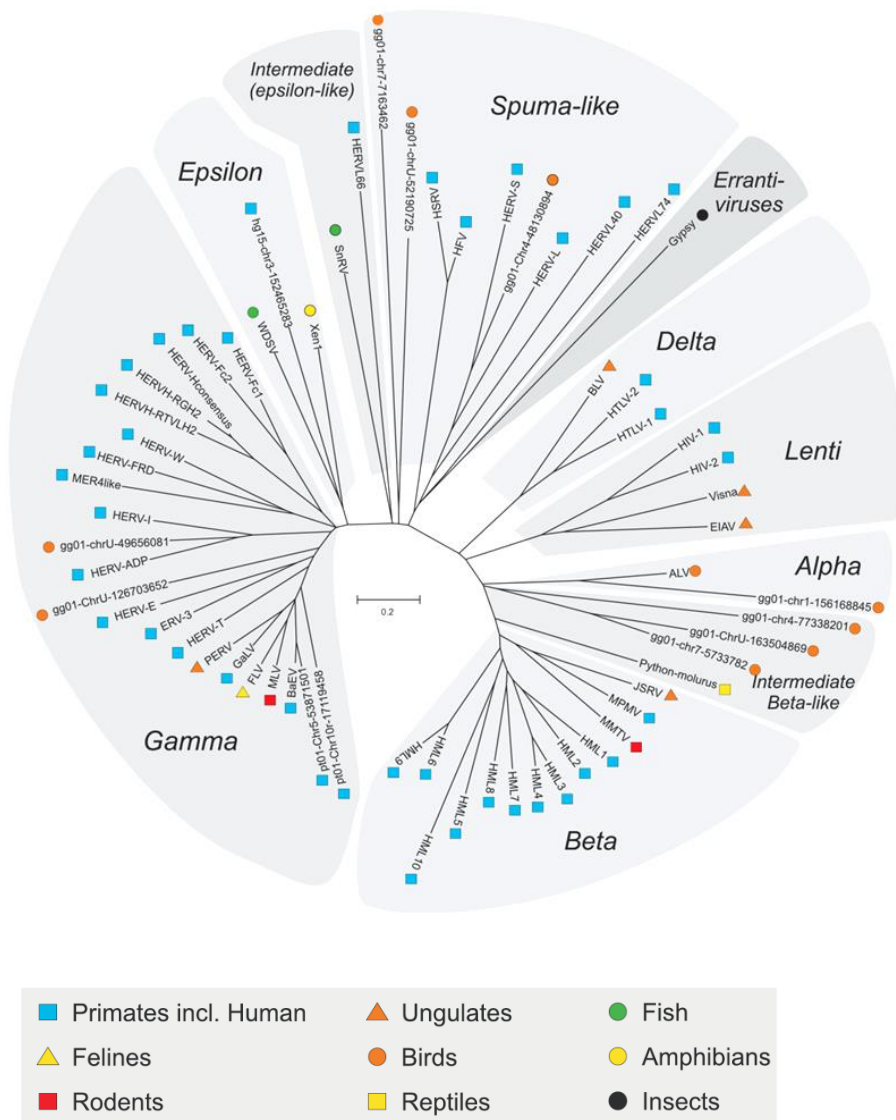


Figure 9: Pol gene phylogeny of the retroviruses showing the seven retroviral genera and their hosts.
From Jern et al. (2005).

1.4.2. Vertebrate Evolution

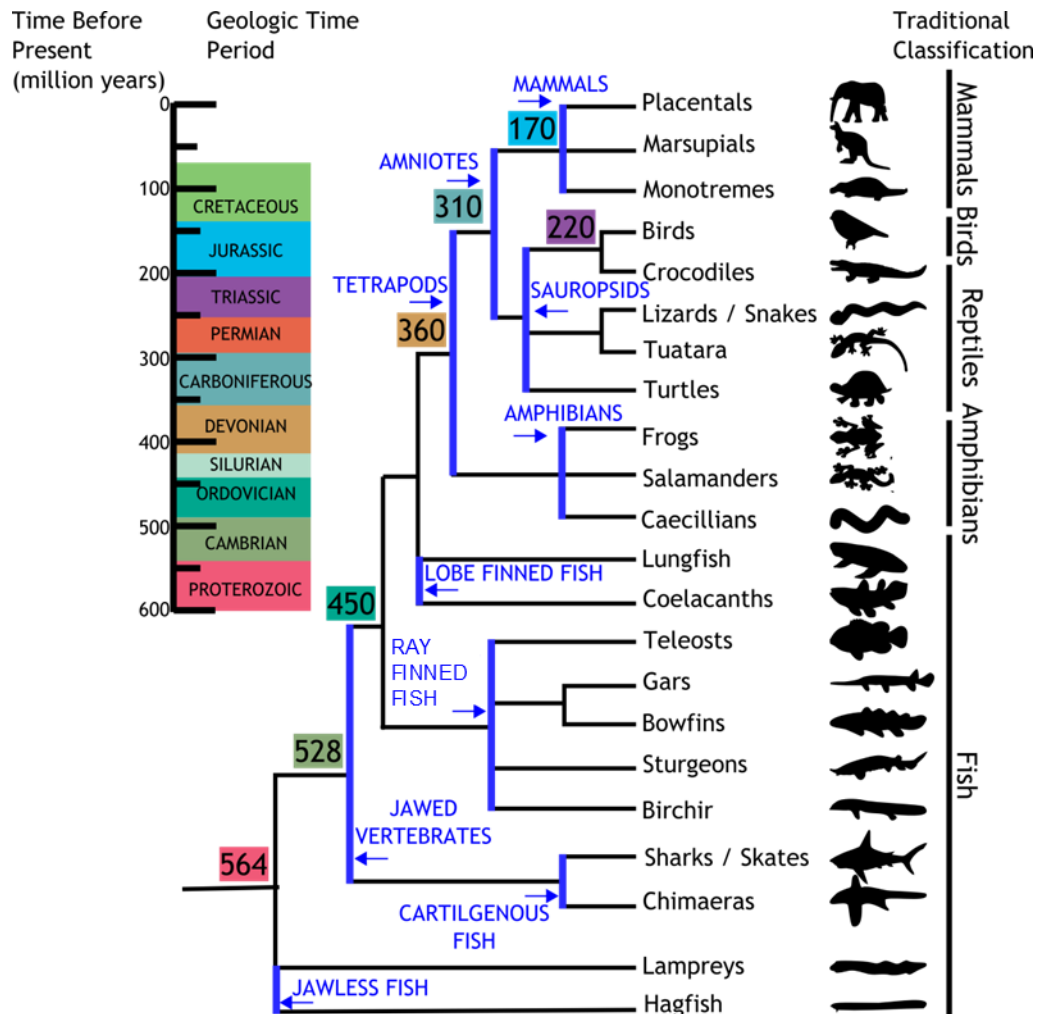


Figure 10: Overview of vertebrate phylogeny.

Tree adapted from Meyer and Zardoya (2003), node timepoints represent the approximate date of the split in millions of years since the present, from Kumar and Hedges (1998), geologic time scale is from Walker et al. (Walker et al., 2012) using the standard geologic time scale colour code (Walker et al., 2012). Notable monophyletic groups of vertebrates are marked and labelled in blue.

A brief overview of vertebrate evolutionary history is needed in order to understand retroviral evolution (Figure 10). Traditionally, vertebrates have been divided into mammals, birds, reptiles, amphibians and fish, based on similarities in morphology and lifestyle. However, this categorisation is not consistent with the accepted evolutionary history of these species, which is shown in Figure 10. The earliest major split in vertebrate evolution is believed

to be the division between the lineages leading to the jawless fish, hagfish and lampreys, and the jawed vertebrates, approximately 564 million years ago (Kumar and Hedges, 1998, Meyer and Zardoya, 2003) (Figure 10). The next split, divided the cartilaginous fish (sharks, skates and chimaeras) from other vertebrates (Kumar and Hedges, 1998) (Figure 10). This means the fish form a paraphyletic group, as there is no ancestor shared by all fish but no other class of vertebrates. The ray finned fish then split from the lobe-finned fish (coelacanth and lungfish) and tetrapods (amphibians, reptiles, mammals and birds) (Kumar and Hedges, 1998) (Figure 10). The earliest split within the tetrapods was between the amphibians and the amniotes, approximately 360 million years ago. Within the amniotes, the reptiles are again paraphyletic, as the most recent common ancestor of all reptiles is shared with birds.

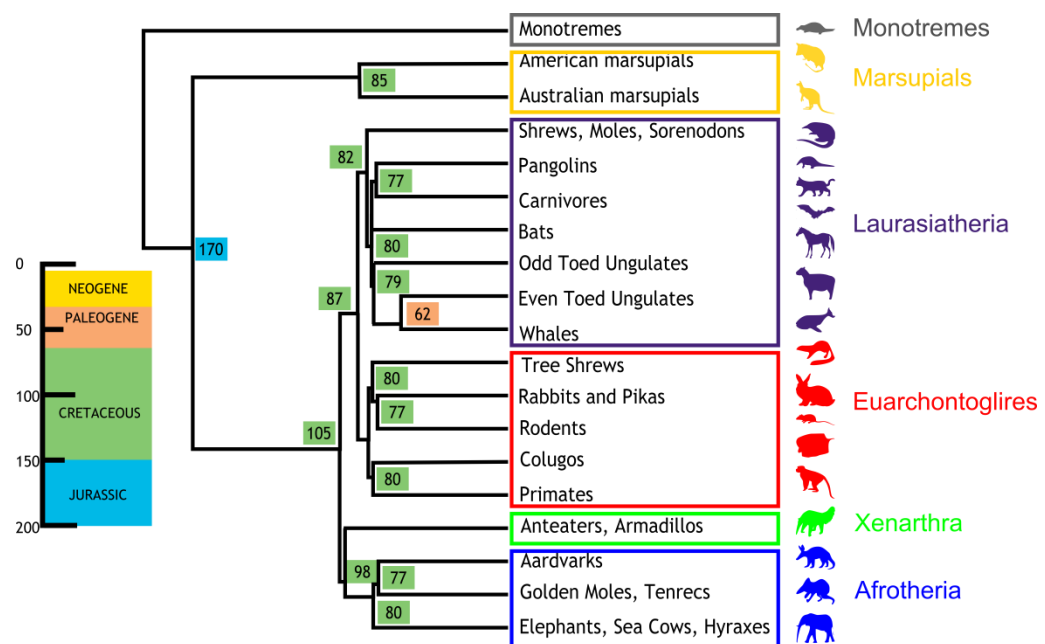


Figure 11: Overview of mammalian phylogeny.

Tree adapted from Meredith et al. (2011), node timepoints represent the approximate date of the split in millions of years since the present. Geologic time scale is from Walker et al. (Walker et al., 2012) using the standard geologic time scale colour code (Walker et al., 2012).

Figure 11 provides a similar overview of phylogeny within the mammals, divided into five major monophyletic groups. Mammals first divided into marsupials, monotremes and placental mammals, then, within the placental

mammals, into three groups, the Afrotheria, Xenarthra and Boreoeutheria (Figure 11) (Meredith et al., 2011). The majority of modern mammals are Boreoeutheria, which split into two groups approximately 87 million years ago, the Euarchontoglires and the Laurasiatheria (Figure 11) (Meredith et al., 2011).

This project concentrates on the Euarchontoglires: primates, rodents, lagomorphs (rabbits and hares) and tree shrews. Colugos (flying lemurs) are also in this superorder but no colugo has been sequenced to date. Figure 11 includes an overview of the phylogeny of this group, with two major clades, containing primates and the other lagomorphs and rodents (Meredith et al., 2011). The tree shrews are phylogenetically ambiguous and are sometimes considered to be closer to the rodents and sometimes to be closer to the primates (Martin, 2008).

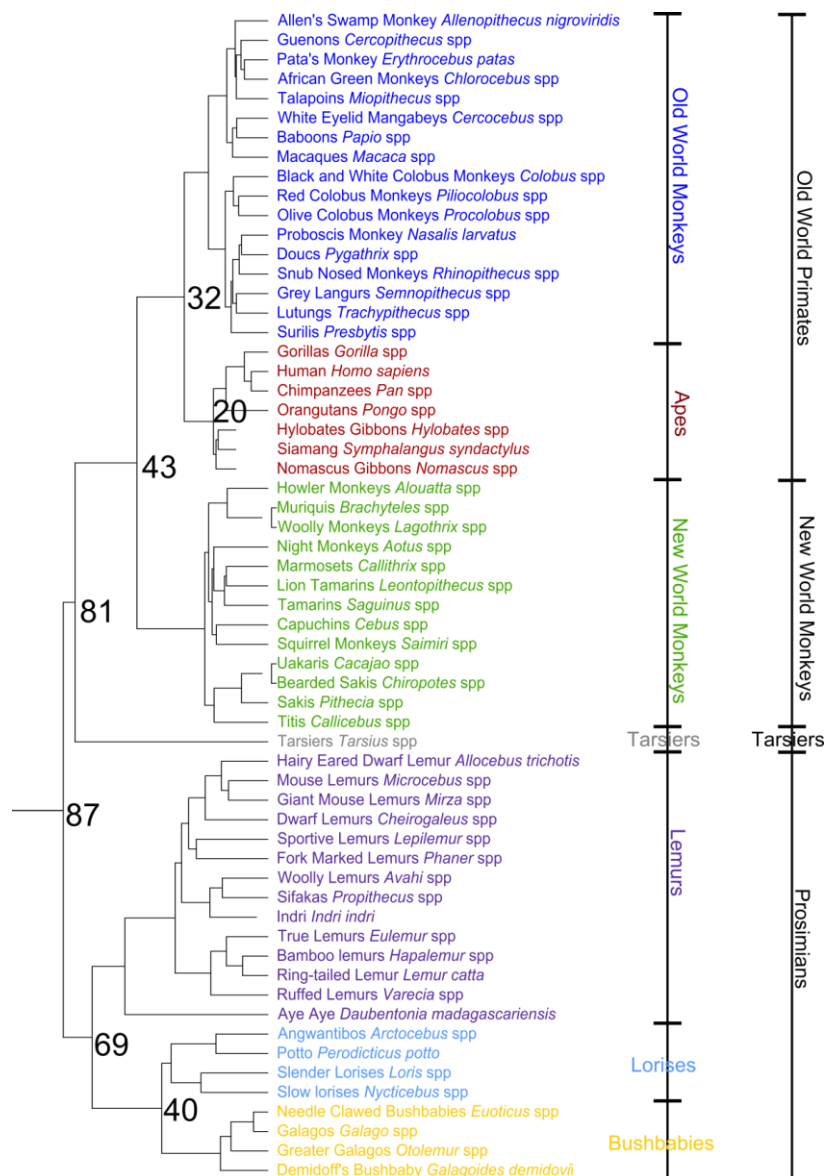


Figure 12: Phylogeny of the major groups of primates.

Nodes represent the age of the split in millions of years. Tree data from Arnold et al. (2010), dates from Perelman et al. (2011).

A primate phylogenetic tree is shown in Figure 12. The first division in this family was between the prosimian primates (lemurs, lorises, tarsiers and bushbabies) and the simian primates, 87 million years ago (Perelman et al., 2011). Lemurs, living on Madagascar, form a unique group, separate from the other prosimians, living in mainland Africa (Perelman et al., 2011). Within the simian primates, the major division is between the new world primates (native to the Americas) and old world primates (native to Africa and Asia), around 43 million years ago (Perelman et al., 2011). The old world primates are usually further divided into the apes, consisting of humans, chimpanzees, gorillas, orangutans and gibbons and the old world monkeys (Perelman et al., 2011).

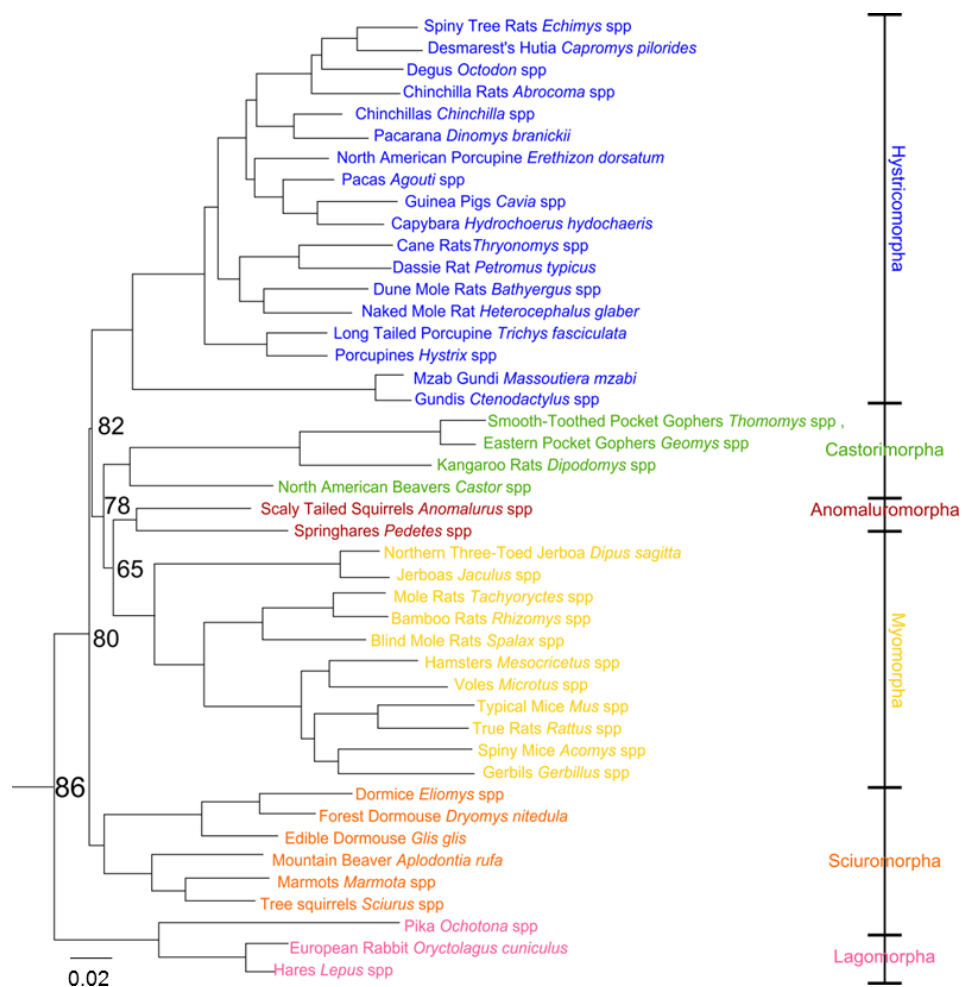


Figure 13: Phylogeny of the major groups of rodents and lagomorphs. Nodes represent the age of the split in millions of years. Tree adapted from Blanga-Kanfi (2009), dates from Hedges et al. (2006).

Figure 13 shows an overview of the phylogeny of rodents and lagomorphs. Rodents fall into five fairly clear, monophyletic suborders (Blanga-Kanfi et al., 2009) (Figure 13). The Sciuromorpha (squirrels and relatives) are considered to be the most basal group, however all five major groups seem to have appeared at approximately the same time (Blanga-Kanfi et al., 2009, Hedges et al., 2006). The Hystricomorpha can also be referred to as the “new world” rodents, as these species are native to south America. The lagomorphs (rabbits, hares, pikas) are a distinct order to the rodents (Blanga-Kanfi et al., 2009).

1.4.3. Gammaretroviruses

The gammaretroviruses have predominantly been identified in mammals, with a few exceptions in reptiles, amphibians, birds and fish.

Figure 14 shows some of the key groups of gammaretroviruses and how they relate to each other.

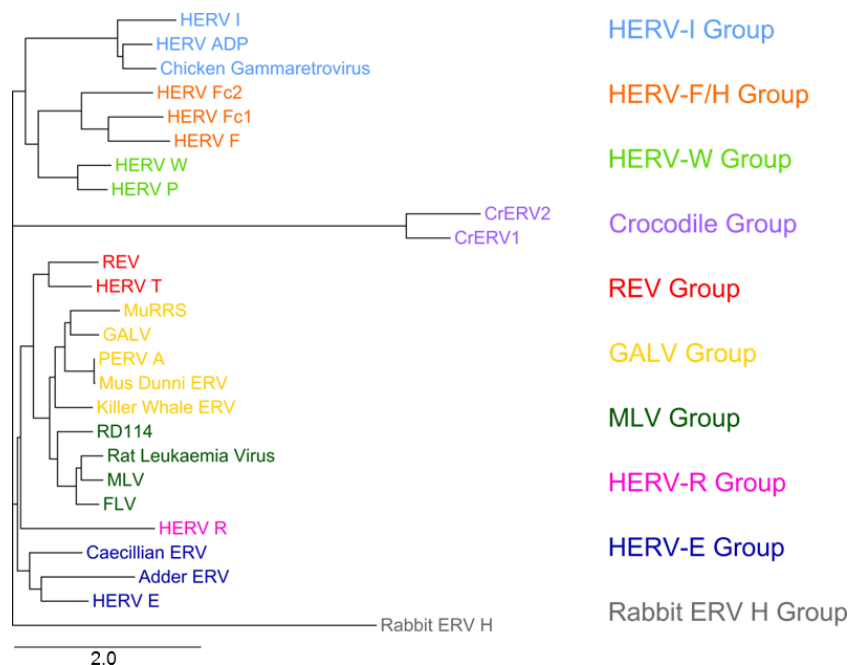


Figure 14: The relationship between the *pol* genes of the major groups of gammaretroviruses.

Sequences listed in Appendix B.1. Aligned using MAFFT localpair with 1000 iterations, tree built using PhyML under the GTR model.

1.4.3.1. HERV-I Group

HERV-I elements are found in old world monkeys, apes and possibly new world monkeys (Perelman et al., 2011, Lee and Kim, 2006, Seifarth et al., 2000). HERV-I-like elements appear to be unusually widely distributed amongst the vertebrates compared to other gammaretroviruses. Two lineages of HERV-I, one of which includes a full-length insertion, have been identified in the chicken genome (Borysenko et al., 2008, Niewiadomska and Gifford, 2013) and related fragments isolated from the budgerigar (*Melopsittacus undulatus*) and house sparrow (*Passer domesticus*) (Figure 14)(Martin et al., 1997, Niewiadomska and Gifford, 2013). The two chicken gammaretroviruses appear to have diverged in an ancestor to modern chickens, while the budgerigar and house sparrow ERVs are almost certainly the result of separate integration events (Figure 14). The only published ERV fragment from cartilaginous fish (from the lemon shark *Negaprion brevirostris*) is also HERV-I-like (Martin et al., 1997, Herniou et al., 1998) (Figure 14). Similarly, the only known gammaretrovirus of an anapsid (lizard, snake or tuatara) reptile is a HERV-I like insertion in the komodo dragon, *Varanus komodoensis* (Martin et al., 1997) (Figure 14). All current estimates (e.g. (Cui et al., 2012, Polavarapu et al., 2006b, Polavarapu et al., 2006a, Garcia-Etxebarria and Jugo, 2010, Lee and Kim, 2006) suggest that gammaretroviruses emerged well within the timescale of mammalian evolution and several hundred million years after the cartilaginous fish, reptiles and birds split from the lineage leading to humans. This means the HERV-I-like elements shown in Figure 14, with the exception of the two chicken insertions, are almost certainly each the result of independent transmission events from a group of retroviruses with a broad host range (Lee and Kim, 2006).

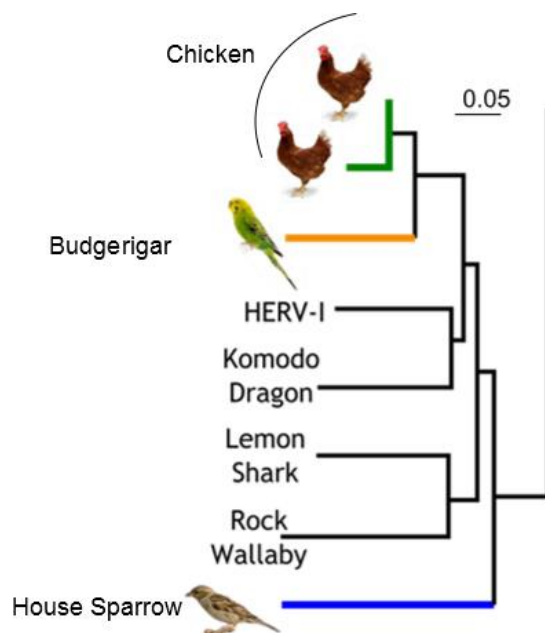


Figure 15: The phylogeny of HERV-I-like elements from birds, reptiles, cartilaginous fish and humans.

Adapted from Niewiadomska and Gifford (2013).

1.4.3.2. HERV-F/H Group and HERV-W

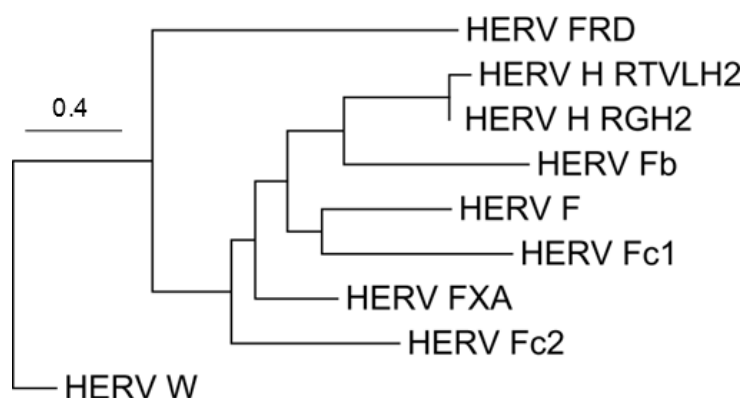


Figure 16: The relationship between the *pol* genes of HERV-H and HERV-F lineages of HERV.

Sequences listed in Appendix B.1. Aligned using MAFFT localpair with 1000 iterations, tree built using PhyML under the GTR model.

The largest group of HERVs is the gammaretroviral HERV-H / HERV-F superfamily (Bannert and Kurth, 2006) (Figure 16). Jern et al. (2004) found that HERV-H and HERV-F are not phylogenetically distinct from each other and identified 926 HERV-H and 198 HERV-F insertions in the human

genome. HERV-H elements are subdivided into two groups, RGH2-like and RTVLH2-like, while HERV-F elements are divided into HERV-F, HERV-FXA, HERV-Fb, HERV-Fc1, HERV-Fc2 and HERV-FRD groups (Bénit et al., 2003).

This group includes the provirus HERV-Fc1, which has highly divergent LTRs but intact ORFs for pro and env and only minor defects in the pol and gag ORFs (Jern et al., 2004). Contrary to the usual pattern, the RGH2-like HERV-H group appears to be older than the RTVLH2-like HERV-H group according to its LTR divergence, but has more intact ORFs (Jern et al., 2004). The “midwife” hypothesis is an attempt to explain this discrepancy. This hypothesis states that during divergence of the HERV-H group, intact elements (possibly HERV-Fc1), provided functions and proteins in trans, allowing older, less intact RTVLH2 like groups to reintegrate, renewing their LTRs and making them appear to be more modern (Jern et al., 2004). This model could apply to other primate ERVs, so it is useful to characterise the degree of gene sequence degradation alongside LTR divergence.

Elements related to the HERV-H/F group have been found in apes, old world monkeys, new world monkeys and prosimians (Bénit et al., 2003, Bannert and Kurth, 2006). This is generally assumed to mean that they entered the germline prior to the division of prosimians and simians, 87 million years ago (Bénit et al., 2003, Bannert and Kurth, 2006, Perelman et al., 2011). However, no analysis to date has confirmed shared integration sites between the different primate groups, so it is possible that the HERV-H/F retroviruses were widespread at some more recent point and integrated separately into various primate lineages.

The HERV-H/F group seems to be the origin of one of the syncytin genes co-opted by primates for an essential role in placental morphogenesis (Blaise et al., 2003) (section 1.3.3.1). Blaise et al. (2003) identified HERV-FRD Env as a fusogenic protein, expressed in the placenta, with an intact coding sequence and named it Syncytin-2. Closely related, intact *env* sequences were found to

be conserved in apes, old world monkeys and new world monkeys but not present in prosimians (Blaise et al., 2003).

The HERV-W lineage of ERVs is thought to be considerably more modern than the HERV F/W group, having entered the germline between the common ancestor of the old world monkeys and apes, 32 million years ago, and the ancestor of the new and old world monkeys, 43 million years ago (Bannert and Kurth, 2006, Perelman et al., 2011). Similarly to HERV-FRD, a HERV-W Env protein, Syncytin-1, has been co-opted by primate hosts to act in placental development (Cáceres et al., 2006). HERV-W like insertions are common to all old world monkeys and apes but Syncytin-1 in old world monkeys is inactive, with multiple mutations (Cáceres et al., 2006). Only apes have an intact *syncytin* gene capable of producing a functional protein (Cáceres et al., 2006). It is not clear if the gene acquired its function in placental development before the split between old world monkeys and apes and later lost this function in old world monkeys or if the function was acquired after the split and the old world monkey copies degraded due to lack of purifying selection (Cáceres et al., 2006).

1.4.3.3. Crocodile Group

Jaratlerdsiri et al. (2009) identified a novel lineage of ERVs, the crocodile ERVs (CrERVs). Divergent CrERV insertions are found within the same or closely related species (Jaratlerdsiri et al., 2009, 2012). This suggests that numerous retroviruses of this type were circulating amongst *Crocodylus* at some point and endogenisation was widespread.

1.4.3.4. REV Group

Birds have an unambiguous group of pathogenic exogenous gammaretroviruses: reticuloendotheliosis virus (REV), duck infectious anaemia virus (DIAV) and spleen necrosis virus (SNV) (Niewiadomska and

Gifford, 2013). These viruses are thought to be recombinant, as their *gag* and *pol* genes cluster with the gammaretroviruses but their *env* genes with the betaretroviruses (Niewiadomska and Gifford, 2013). The closest known endogenous relatives to REV are found in two species of Malagasy mongoose (*Galidiinae* spp.) and in the echidna *Tachyglossus aculeatus*, these are also recombinant viruses and all three genes cluster similarly to REV (Niewiadomska and Gifford, 2013). Closely related ERVs have not been found in birds (Niewiadomska and Gifford, 2013). The mongoose ERVs are orthologous, so they inserted prior to the divergence of these two species approximately eight million years ago (Niewiadomska and Gifford, 2013). They are not found in the fossa (*Cryptoprocta ferox*), a more distantly related species which diverged approximately 20 million years ago, placing the common ancestor of the Malagasy mongoose ERVs between these two dates (Niewiadomska and Gifford, 2013, Yoder et al., 2003). The echidna is a monotreme found in Australia and, as REV-related ERVs have not been found in other mammals, it is unlikely that their integration predates the monotreme-marsupial-placental mammal split, so instead REV-like viruses may have been widespread worldwide over a period encompassing the divergence of the two mongoose genera or may have entered the echidna after human colonisation of Australia (Niewiadomska and Gifford, 2013).

Due to the lack of diversity of the strains of REV, SNV and DIAV isolated from birds, Niewiadomska et al. concluded that these viruses entered birds very recently via a single founder and proposed the evolutionary history shown in Figure 17. The founder event is suggested to have been via stocks of contaminated *Plasmodium lophurae* parasites used to experimentally infect birds between the 1930s and 1980s, the period during which DIAV and SNV were circulating. These parasites were cultured from a single isolate from a pheasant in a zoo in New York, therefore it is possible that the original stock was contaminated with a retrovirus from an exotic animal. Two lineages of viruses are thought to have originated from this contaminant. The first diverged in culture into the similar SNV and DIAV strains and resulted in

outbreaks of disease through repeated experimental infection of ducks with *P. lophurae* (Niewiadomska and Gifford, 2013) (Figure 17). Unusually, the second lineage, leading to REV, appears to have emerged at around the same time but via integration into the genome of two larger viruses infecting birds, fowlpox virus (FWPV) and gallid herpesvirus 2 (GHV-2) (Marek's disease virus), both of which cause disease in chickens (Niewiadomska and Gifford, 2013) (Figure 17). The spread of REV through birds was partially due to use of attenuated GHV-2 with REV insertions in avian vaccines (Niewiadomska and Gifford, 2013). REV now circulates in birds using these FWPW and GHV-2 viruses as vectors (Niewiadomska and Gifford, 2013).

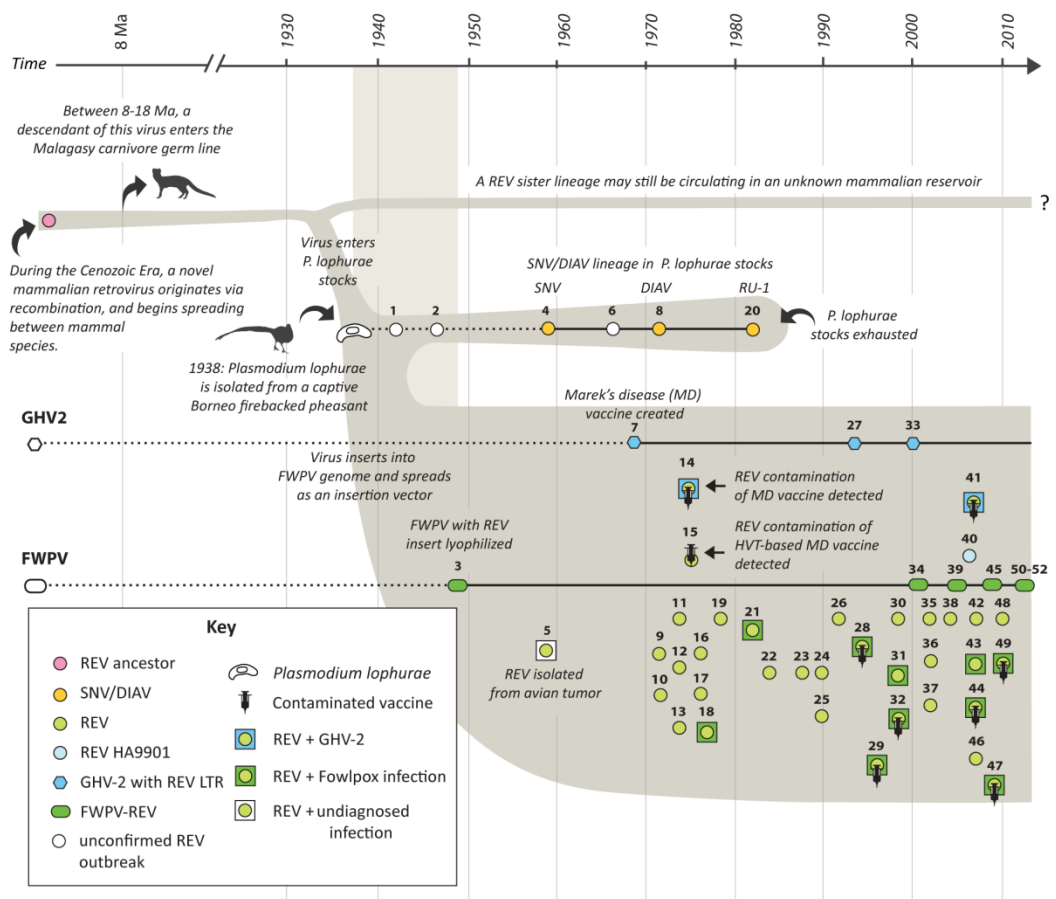


Figure 17: The proposed evolutionary history of SNV, DIAV and RSV.
From Niewiadomska et al. (2013). Numbers refer to individual cases.

The cetacea (whales in Figure 11) have a family of gammaretroviruses which are clearly distinct from REV but form a sister group to the REV lineage (Wang et al., 2013). This group is estimated to have integrated 10 to 19 million years ago and has been found in all cetacea tested to date (Wang et al., 2013).

1.4.3.5. GALV Group

The group of gammaretroviruses surrounding gibbon ape leukaemia virus (GALV) in the gammaretroviral phylogeny (Figure 14, Figure 18) again includes several examples of cross-species transmissions and retroviruses isolated from primates and rodents but also Australian marsupials, whales and even-toed ungulates. These mammals are found in diverse geographic locations and cover over 170 million years of mammalian evolution (Figure 11). Figure 18 shows an overview of the major groups of retroviruses within this group.

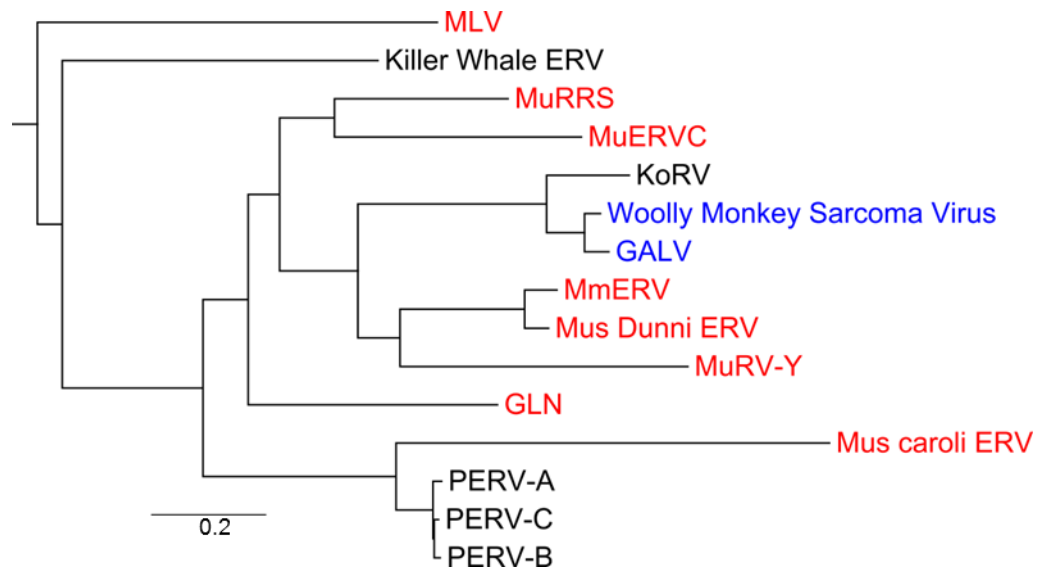


Figure 18: The relationship between the *pol* genes of members of the GALV group of gammaretroviruses.

Red: Rodents, Blue: Primates. Sequences listed in Appendix B.1. Aligned using MAFFT localpair with 1000 iterations, tree built using PhyML under the GTR model.

GALV is an exogenous pathogenic gammaretrovirus which was identified as causing haematopoietic neoplasias in captive white handed gibbons

(*Hylobates lar*) in the 1970s (Kawakami et al., 1972). A related retrovirus, woolly monkey sarcoma virus (WMSV) was identified at around the same time, causing sarcoma in a single woolly monkey (*Lagothrix* spp.) (Theilen et al., 1971). Over the next five years, related viruses were identified in several other gibbon colonies in the USA (Todaro et al., 1975), Thailand (Kawakami et al., 1975) and Bermuda (Krakower et al., 1978, Reitz et al., 1979). Later, GALV was found to be expressed in various cell lines (Okabe et al., 1976, Chan et al., 1976, Burtonboy et al., 1993). GALV is still considered to be a circulating pathogen of gibbons although no outbreaks have been described since the 1970s.

There is no evidence that GALV has ever become endogenous in gibbons and no close relatives of GALV have been identified in primates. The closest known relative of GALV is koala retrovirus (KoRV), an ERV of koalas (*Phascolarctos cinereus*) (Tarlinton et al., 2006) (Figure 18). Endogenous KoRV is polymorphic between koala populations and appears to be in transition between an exogenous and endogenous form, having entered the koala population within the last 200 years (Tarlinton et al., 2006).

Gibbons and koalas are distant from each other both evolutionarily and geographically, so it is generally considered that the virus was transmitted between the two groups via a vector species, most likely a rodent (Hayward et al., 2013a, Tarlinton et al., 2006, Tarlinton et al., 2008). However, this vector has not been identified to date. Several groups of endogenous murine retroviruses (Figure 18) form the closest phylogenetic group to GALV and KoRV. The most likely candidates as the origins of GALV and KoRV are thought to be retroviruses of Asian mice, related to a degenerate ERV in the Asian Ryukyu mouse (*Mus caroli*) and an intact virus in the Asian Earth coloured mouse (*Mus terricolor*, previously *Mus dunni*) (Lieber et al., 1975, Stocking and Kozak, 2008, Tarlinton et al., 2008). However, Hayward et al. (2013a) found endogenous insertions in the house mouse (*Mus musculus*) which were as close to GALV and KoRV as *Mus dunni* ERV (MDERV).

	MuRRS	MuRV-Y	GLN	MmERV
<i>Mus mus musculus domesticus</i>				
<i>castaneus</i>				
<i>musculus</i>				
<i>spicilegus</i>				
<i>spretus</i>				
<i>cervicolor</i>				
<i>cookii</i>				
<i>caroli</i>				
<i>Mus coelomys</i>				
<i>Mus nannomys</i>				
<i>Mus pyromys</i>				

Figure 19: The species of *Mus* in which MuRRS, MuRV-Y, GLN and MmERV have been detected.

Adapted from Stocking and Kozak (2008).

Several groups of murine ERVs surround GALV and KoRV in phylogenetic analysis, falling between *Mus caroli* ERV and GALV-KoRV (Stocking and Kozak, 2008) (Figure 18). Besides MDERV, these are the mouse retrovirus related sequence (MuRRS), murine ERV C (MuERV-C), *Mus musculus* ERV (MmERV), murine repeated virus on the Y-chromosome (MuRV-Y), and murine retrovirus using tRNA^{Gln} (GLN) groups. Figure 19 shows the species of mouse in which these groups have been detected (the remaining species have not been screened so may also contain these ERVs). The MuRRS group was one of the first groups of murine ERVs to be detected (Schmidt et al., 1985, Stocking and Kozak, 2008). This group is thought to have entered the mouse genome within the last 9 million years and, as many copies of MuRRS share the same *pol* gene defects, to have spread through the genome with the aid of an intact “helper” virus (Stocking and Kozak, 2008). There are 30-50 MuRRS copies in the mouse genome and these are highly degraded but have recognisable *gag*, *pol* and *env* genes. MuERV-C group is similarly degenerate and is found in 10-20 copies in the mouse genome (Stocking and Kozak,

2008). 10 copies are clustered on the X chromosome (Stocking and Kozak, 2008). MuRRS insertions have only been detected in the *Mus* subgenus of mice to date (Figure 19) and MuERV-C only in *Mus musculus* (Stocking and Kozak, 2008). The MmERV group contains close relatives of MDERV. Therefore, the MDERV was possibly part of a much wider circulating group found in mice or rodents worldwide (Stocking and Kozak, 2008). MuRV-Y is unusual in that almost 500 copies are found on the Y-chromosome of some *Mus* species (Stocking and Kozak, 2008) (Figure 19). Finally, the GLN group of murine ERVs has also been found all screened subgenera of *Mus* and is present in about 80 copies in the *Mus musculus* genome, including a fully replication competent copy (Ribet et al., 2008).

Unexpectedly, three groups of pig ERVs which produce replication competent particles, porcine ERV (PERV)-A, PERV-B and PERV-C, also fall into this group of gammaretroviruses, close to *Mus caroli* ERV (Figure 19). PERV-A and PERV-B are present in all pig breeds and have been shown to be capable of infecting human cells, while PERV-C is absent in some breeds and cannot infect human cells (Yu et al., 2012b). There is evidence of some past recombination between the three groups. None of the three groups are found in peccaries, the closest relatives of pigs, which separated from the pig lineage approximately 7.5 million years ago (Tönjes and Niebert, 2003) Therefore, these retroviruses must have circulated since this time and their phylogenetic relationships again suggest a rodent origin (Yu et al., 2012b) (Figure 19).

A final member of this group is found in the killer whale (*Orcinus orca*) a member of another mammalian order, the cetacea. Killer whale ERV (KWERV) was first isolated by LaMere et al. (2009) and falls between the MLV group and the GALV group in the gammaretroviral phylogeny (Figure 18). Related cetaceans and even-toed ungulates (hippopotamus, pig, beluga whale, fin whale) are negative for KWERV but it is found in at least partially in bottlenose dolphins, dwarf whales, pygmy sperm whales and harbour porpoises (LaMere et al., 2009). Therefore, KWERV-like insertions may be

ubiquitous in the dolphin family (Delphinidae) (which includes the killer whale) and possibly the porpoises (Phocoenidae) (LaMere et al., 2009). There are only two to four copies of the virus in the killer whale genome, so it may have become replication deficient fairly soon after integration (LaMere et al., 2009). The most intact insertions in the dolphin are considerably more degenerate than those in the killer whale, so while the initial integration of KWERV into the Delphinidae may predate their divergence 12 million years ago, some of the killer whale insertions appear to be considerably more recent than this (LaMere et al., 2009).

1.4.3.6. MLV Group

The MLV group of gammaretroviruses is one of the most widely studied groups. The term “MLV-like” is often used to refer to all gammaretroviruses [e.g. (Tristem et al., 1996)], however here the term refers only to the phylogenetic group clustering most closely around MLV (Figure 14).

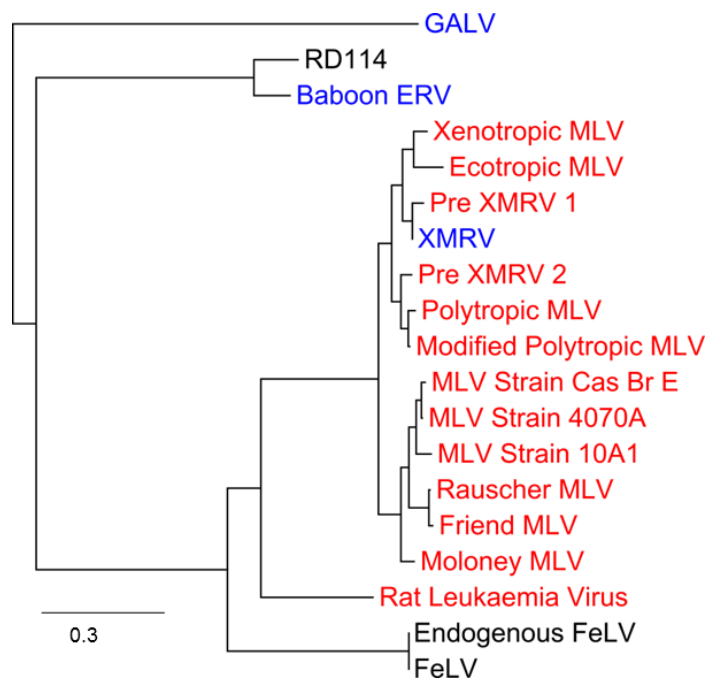


Figure 20: The relationship between the *pol* genes of members of the MLV group of gammaretroviruses.

Red: Rodents, Blue: Primates. Sequences listed in Appendix B.1. Aligned using MAFFT localpair with 1000 iterations, tree built using PhyML under the GTR model.

The MLV lineage is thought to have entered the murine genome recently (within the last 1.5 million years), is insertionally polymorphic between mouse strains and includes some replication competent members (Stocking and Kozak, 2008, Jern et al., 2007). Most MLVs fall into one of four groups, the ecotropic, xenotropic, polytropic and modified polytropic groups, all of which have endogenous members (Stoye and Coffin, 1987). The categorisation of MLVs into these groups depends on their ability to replicate in cells from different hosts: ecotropic MLVs can only replicate in murine cells, xenotropic MLVs only in non-murine cells (due to loss of the appropriate receptor after endogenisation) and polytropic and modified polytropic MLVs in murine and non-murine cells (the two polytropic groups differ in the structure of their *env* gene). Polytropic, modified polytropic and xenotropic MLVs are more closely related to each other in terms of their *env* gene sequence, which determines receptor specificity, than they are to ecotropic MLVs (Stoye and Coffin, 1987). Xenotropic MLVs are not monophyletic (Cingöz and Coffin, 2011). MLVs are only found in members of the *Mus* genus, with polytropic and modified polytropic MLVs identified in wild *Mus domesticus* and xenotropic MLVs in *Mus musculus* and *Mus castaneus* (Dudley et al., 2011). Inbred laboratory mice contain variable numbers of each group of MLVs (Dudley et al., 2011). Proliferative diseases in mice resulting from these ERVs are usually the result of recombination between different insertions or between ERVs and XRVs (Dudley et al., 2011). Rats do not have endogenous MLV but do have an ERV which falls somewhere between the FeLVs and MLVs phylogenetically, known as rat leukaemia virus (Lee et al., 1998).

A retrovirus closely related to the xenotropic MLVs, known as xenotropic murine leukaemia related virus (XMRV), also falls into this phylogenetic group. The virus was first identified as a novel gammaretrovirus in prostate tumour samples in 2006 (Urisman et al., 2006) and was later associated with, and proposed to be a causative agent for, chronic fatigue syndrome (CFS) (Lombardi et al., 2009, Mikovits et al., 2010) (Lo et al., 2010). The results of these studies had a large impact on both the research community and the

community of CFS patients. XMRV screening was made commercially available, anti-retroviral therapy was considered as a potential CFS treatment and blood donation by CFS patients was banned in many countries worldwide (Wainberg and Jeang, 2011). However, the results of many other studies contradicted any association between this virus and either prostate cancer or CFS [for example (Erlwein et al., 2010) (Hohn et al., 2009, van Kuppeveld et al., 2010, Switzer et al., 2010)]. A number of studies published in the same issue of *Retrovirology* demonstrated the high likelihood mouse contamination of laboratory samples being responsible for the XMRV findings, with mouse contamination identified in human tissue samples (Robinson et al., 2010, Oakes et al., 2010), laboratory reagents (Sato et al., 2010) and cell lines (Hue et al., 2010, Smith, 2010) in quantities sufficient for XMRV to be detected. Later studies confirmed the absence of XMRV in either CFS (Alter et al., 2012, Simmons et al., 2011) or prostate cancer (Lee et al., 2012) and on the basis of these studies the original papers describing the association were retracted. XMRV has since been shown to be a recombinant between xenotropic and polytropic MLVs found in laboratory mice, known as pre-XMRV-1 and pre-XMRV-2, produced by a common laboratory cell line (Cingöz and Coffin, 2011). Therefore, it appears that XMRV in human samples was the result of laboratory contamination resulting from this or a related cell line (Cingöz and Coffin, 2011).

Domestic cats and other members of the *Felis* genus have endogenous versions of FeLVs which do not code for infectious virus (Polani et al., 2010). Other members of the cat family (Felidae) appear not to have FeLVs (Polani et al., 2010). FeLVs appear to have invaded the germline more than once in the evolutionary history of the *Felis* lineage, as some lineages show evidence of integration before the lineage diverged while others appear to have proliferated in some species but not others (Polani et al., 2010). There are no FeLV ERVs in the domestic cat that are not also found in the wild cat (Polani et al., 2010).

Two recombinant ERVs also cluster within this group in analysis of the *pol* gene, baboon endogenous retrovirus (BaEV) and RD114 virus. These viruses have gammaretroviral *gag* and *pol* genes but betaretroviral *env* genes and are discussed in depth in section 1.4.7.2.

1.4.3.7. HERV-E Group

The HERV-E family of retroviruses initially entered primate genomes between the divergence between prosimians and simians, 87 million years ago, and the divergence between old and new world primates, 43 million years ago (Figure 12) (Yi and Kim, 2006). These ERVs are not replication competent. There are 35-50 HERV-E copies in the human genome and HERV-E LTRs are known regulatory elements for human genes (Yi and Kim, 2006). There is a large increase in HERV-E copy number in apes compared to old and new world monkeys and it is thought that a proliferation of this lineage occurred between 6 and 14 million years ago (Yi and Kim, 2006).

1.4.3.8. Rabbit ERV H Group

The first reported rabbit ERV lineage was identified by Griffiths et al. (2002). This lineage is unusual in that it was originally identified as a human XRV associated with disease (Griffiths et al., 2002, Griffiths et al., 1997). The human isolate, known as human retrovirus 5 (HRV-5), was thought to be associated with inflammatory disease and non-Hodgkins lymphoma (Griffiths et al., 1997, Kozireva et al., 2001). However, the retrovirus was later shown unambiguously to be a rabbit ERV and its appearance in human samples to be the result of either laboratory contamination or human interactions with rabbits (Griffiths et al., 2002).

1.4.3.9. Syncytins

Syncytins found in different orders of mammals are all derived from gammaretroviral *env* genes and have closely related functional properties but have distinct evolutionary origins (Cornelis et al., 2013) (Figure 21). This seems to be the result of convergent evolution leading to independent capture of separate gammaretrovirus *env* genes (Cornelis et al., 2013, Cornelis et al., 2012).

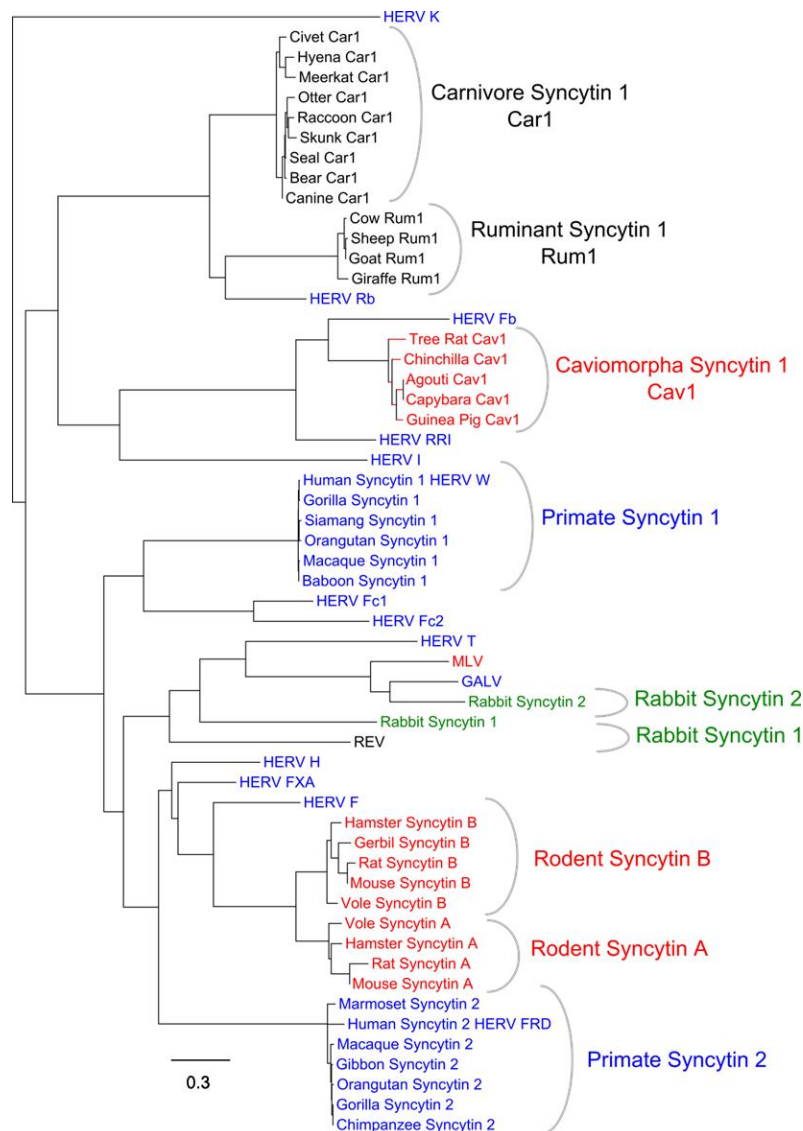


Figure 21: The relationship between syncytins and other gammaretroviral *env* genes.

Red: Rodents, Blue: Primates. Sequences listed in Appendix B.1. Aligned using MAFFT localpair with 1000 iterations, tree built using PhyML under the GTR model.

Primate Syncytin-1 and Syncytin-2 are unambiguously derived from the HERV-W and HERV-FRD lineages respectively (Dupressoir et al., 2005). Syncytin proteins have also been identified in rodents, lagomorphs, carnivores and ruminants and the phylogenetic relationships of these proteins are more ambiguous. The mouse family of rodents (Muridae) share two *syncytin* genes, coding for the Syncytin-A and Syncytin-B proteins (Dupressoir et al., 2005). Both seem to have circulated before the Muridae appeared approximately 20 million years ago (Dupressoir et al., 2005). Both of these genes code for functional proteins and show evidence of purifying selection (Dupressoir et al., 2005). Like the human syncytins, syncytins A and B have a fusogenic effect and are specifically expressed in the placenta (Dupressoir et al., 2005). Rodents within and outside of the Muridae have physiological differences in placental structure, possibly due to the presence and absence of these proteins (Dupressoir et al., 2005). Syncytins A and B fall either within the HERV F/H superfamily or closer to HERV-W in the gammaretroviral phylogeny (Redelsperger et al., 2014, Cornelis et al., 2013). Syncytin-A and Syncytin-B form distinct phylogenetic clusters (Dupressoir et al., 2005) (Figure 21). Within these clusters there is evidence of host tracking, with branching order of mouse, rat, hamster and gerbil and vole syncytins in both groups matching the branching order of the hosts (Figure 13, Figure 21).

A third lineage of rodent syncytins is found in the Caviomorpha, the South American parvorder of Hystricomorpha (Figure 13) (Vernochet et al., 2011). This lineage, known as Cav1, is very distinct to Syncytin-A and Syncytin-B and likely to have originated separately (Figure 21) (Vernochet et al., 2011). Like the other rodent syncytins, Cav1 shows evidence of host tracking (Figure 13, Figure 21), which suggests it integrated before the diversification of the Caviomorpha 30 million years ago (Vernochet et al., 2011). There is no evidence that Cav1 is fusogenic, but it is expressed specifically in the placenta and has intact ORFs in all hosts (Vernochet et al., 2011). Therefore, Cav1 is likely to have a physiological role in placentation (Vernochet et al., 2011).

A syncytin lineage has been very recently identified in the Marmotini, a group of rodents consisting of marmots and their relatives (Figure 21) (Redelsperger et al., 2014). This lineage is known as Mar1. Sequence information is not yet publicly available for this group, but it is phylogenetically similar to the Cav1 lineage (Redelsperger et al., 2014). Mar1 is at least 25 million years old and meets all the criteria for a genuine syncytin gene.

Rabbits also have a unique family of syncytins, the Ory1 family (Heidmann et al., 2009). This group is part of a family of rabbit ERVs related to GALV (Heidmann et al., 2009). The gene appears to have integrated between 12 and 30 million years ago and is found in all members of the Leporidae (rabbits and hares) but none of the Ochotonidae (pikas) (Heidmann et al., 2009). Again, Ory1 genes are fusogenic, placenta specific and intact (Heidmann et al., 2009).

Ruminants also have a unique syncytin gene, known as Rum1 and estimated to have integrated at least 30 million years ago (Cornelis et al., 2013). The closest known primate relative of Rum1 is HERV-Rb (Cornelis et al., 2013) (Figure 21). Rum1 is present in all higher ruminants (part of the “even-toed ungulates” group in Figure 11) and shows evidence of fusogenic activity and intact ORFs in all species tested to date (Cornelis et al., 2013). A close relative of Rum1, also similar to HERV-Rb, is found in all Carnivora and is known as Car1 (Cornelis et al., 2012) (Figure 21). Car1 seems to predate the diversification of the Carnivora, meaning it is at least 65 million years old, the oldest of the known syncytin genes (Cornelis et al., 2012).

Given the recent discovery of syncytins and their presence in diverse mammals, especially within the Laurasiatheria and Euarchontoglires, it is very likely that there are further members of this group which are yet to be discovered.

1.4.4. Epsilonretroviruses

1.4.4.1. Fish epsilonretroviruses

The epsilonretroviruses are traditionally considered to be viruses of teleost fish, as four proliferative diseases in these fish have been confirmed to be epsilonretroviral and to be transmissible via cell-free tumour filtrates. These are walleye dermal sarcoma, associated with WDSV; walleye epidermal hyperplasia, associated with walleye epidermal hyperplasia viruses (WEHVs) one and two; muskellunge and northern pike lymphosarcoma and chinook salmon plasmacytoid leukaemia (Coffee et al., 2013, Bowser and Casey, 1993) (Table 5). A fifth, white sucker epidermal papilloma, has been tentatively confirmed to be associated with a retrovirus, as transmission studies have shown conflicting results (Coffee et al., 2013) (Table 5). Tumours from seven other fish diseases, listed in Table 5, also contain retroviral type-C particles but the link between the retrovirus and the disease has not been confirmed (Coffee et al., 2013, Bowser and Casey, 1993). The fish affected by these diseases are distributed throughout the teleost fish phylogenetic tree in species with no apparent genetic or morphological connections. However, these fish species are all important in the food, recreational fishing or diving industries or as aquarium fish, with the exception of the hooknose, in which tumours were identified during a large-scale survey of a fish population (Bowser and Casey, 1993). Therefore, it is likely that with more widespread screening similar retroviral pathologies would be observed in other species less commonly encountered by humans.

Table 5: Diseases of fish either confirmed to be or provisionally associated with retroviruses.

Disease	Host (common name)	Host (scientific name)	Confirmed	Sequenced	Complex/Simple
Walleye Dermal Sarcoma	Walleye	<i>Sander vitreus</i>	y	NC_001867	Complex
Walleye Discrete Epidermal Hyperplasia 1	Walleye	<i>Sander vitreus</i>	y	AF133051	Complex
Walleye Discrete Epidermal Hyperplasia 2	Walleye	<i>Sander vitreus</i>	y	AF133052	Complex
Atlantic Salmon Swim Bladder Sarcoma Virus	Atlantic Salmon	<i>Salmo salar</i>	n	NC_007654	Simple
Yellow Perch Discrete Epidermal Hyperplasia 1	Yellow Perch	<i>Perca flavescens</i>	n	n	?
Yellow Perch Discrete Epidermal Hyperplasia 2	Yellow Perch	<i>Perca flavescens</i>	n	n	?
Chinook Salmon Plasmacytoid Leukemia	Chinook Salmon	<i>Oncorhynchus tshawytscha</i>	y	n	?
Muskellunge and Northern Pike Lymphosarcoma	Northern Pike, Muskellunge	<i>Esox lucius</i> , <i>Esox masquinongy</i>	y	n	?
Atlantic Salmon Epidermal Papillomatosis	Atlantic Salmon	<i>Salmo salar</i>	n	n	?
Muskellunge and Northern Pike Smooth Type Epidermal Hyperplasia	Northern Pike, Muskellunge	<i>Esox lucius</i> , <i>Esox masquinongy</i>	n	n	?
Hooknose Cutaneous Fibroma/Fibrosarcoma	Hooknose	<i>Agonus cataphractus</i>	n	n	?
White Sucker Epidermal Papilloma	White Sucker	<i>Catostomus commersoni</i>	p	n	?
Angelfish Lip Fibroma	Angelfish	<i>Pterophyllum scalare</i>	n	n	?
European Smelt Spawning Papillomatosis	European Smelt	<i>Osmerus eperlanus</i>	n	n	?
Bicolor Damselfish Neurofibromatosis	Bicolour Damselfish	<i>Stegastes partitus</i>	n	n	?

Another exogenous retrovirus, snakehead retrovirus (SNRV) has been confirmed to infect fish, but this virus has not been associated with disease (Hart et al., 1996). This virus is sometimes considered to be an intermediate between the epsilonretroviruses and the foamy viruses (Jern et al., 2005, Hart et al., 1996). However, more recent work, incorporating more fish viruses, places SDRV more firmly amongst the epsilonretroviruses (Basta et al., 2009).

There are also endogenous epsilonretroviruses in fish. Full genome sequences are available for eight species of fish, (Flicek et al., 2012). Of these, screening studies including epsilonretroviruses have been published for four species: zebrafish (*Danio rerio*), puffer fish (*Tetraodon nigroviridis*), medaka (*Oryzias latipes*) and stickleback (*Gadus aculeatus*). In the context of the wider teleost fish phylogeny these species are reasonably divergent, in particular *D. rerio* is very distinct and split from the other species approximately 265.5 million

years ago (Hedges et al., 2006). Compared to mammalian genomes, all four of the screened fish genomes contain relatively few ERVs with low copy numbers (Basta et al., 2009). The ERVs are also very recent, based on LTR divergence none entered the genome more than four million years ago (Basta et al., 2009). None of the fish ERVs are known to cause disease but zebrafish ERV (ZFERV) has been shown to be transcribed in the zebrafish thymus (Shen and Steiner, 2004). Fish epsilon-like ERVs do not cluster phylogenetically in any way which mirrors the host phylogeny (Basta et al., 2009). The viruses are also thought to have diverged considerably more recently than their hosts, therefore epsilon like ERVs must have invaded fish genomes multiple times. As with the exogenously infected species, given that the species containing these ERVs are not apparently correlated in terms of phylogeny, morphology or distribution, epsilonretroviruses are almost certainly present in many other teleost fish species.

1.4.4.2. Amphibian epsilonretroviruses

Exogenous epsilonretroviruses have not been confirmed in amphibians. However, epsilon ERVs have been confirmed in three families of frog - the tongueless frogs (Pipidae), poison dart frogs (Dendrobatidae) and true frogs (Ranidae) – and in the caecilian (*Epicrionops marmoratus*) and two salamanders (the palmate newt *Triturus helveticus* and tiger salamander *Ambystoma tigrinum*) (Herniou et al., 1998). These families make up reasonably divergent branches of the amphibian phylogeny. These amphibian viruses consistently fall into the epsilonretrovirus clade, but within this clade they are interspersed with several endogenous and exogenous fish and reptile retroviruses, as shown in Figure 22. Therefore, these lineages appear to be recent integrations and it is unlikely that they predate the divergence of the amphibians from the amniotes or the bony fish. However, some do show a degree of host tracking (Herniou et al., 1998) (Figure 22A), for example the three viruses found in the true frogs are more similar to each other than to any

not possible to say if epsilon ERVs are less common in reptiles. The higher density of gammaretroviruses in reptiles may also overshadow the presence of epsilonretroviruses. Epsilonretroviruses have been identified in the saltwater crocodile *Crocodylus porosus*, gharial (*Gavialis gangeticus*), tuatara (*Sphenodon* spp.), pit viper (*Bothrops jararaca*) and slider turtle (*Chrysemys scripta*), five species representing the four orders of extant non-avian reptiles (Herniou et al., 1998). These ERVs cluster together phylogenetically (Figure 23) but the branching pattern of the retroviruses is not consistent with the sauropsid phylogeny, so it is unlikely that these insertions predate the diversification of the reptiles (Herniou et al., 1998).

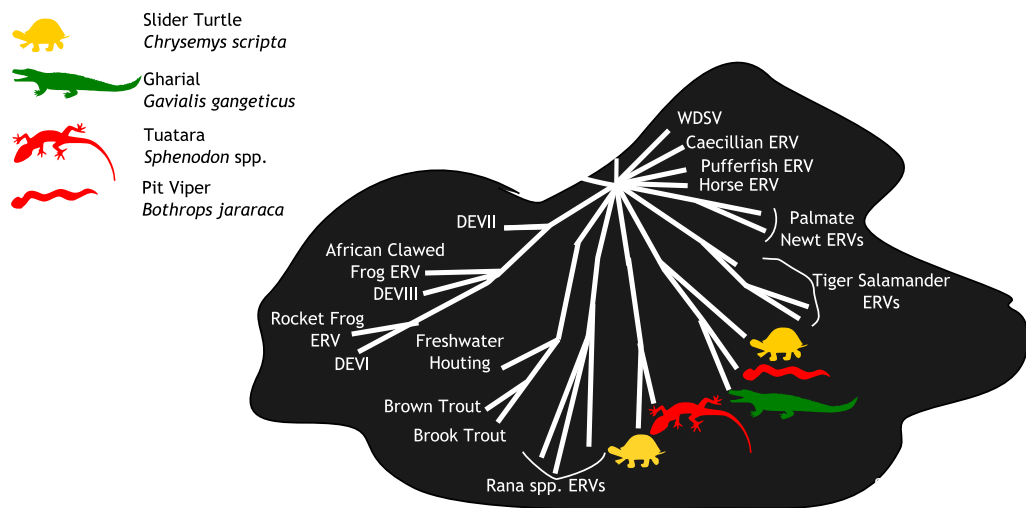


Figure 23: Phylogenetic tree showing reptile ERVs and their relationship with other epsilonretroviruses.

Host species are labelled according to their order (Testudines, Sphenodontia, Squamata and Crocodilia) and coloured by genus and species. Unlabelled symbols are ERVs which have not been assigned specific names. Adapted from Herniou et al. (1998).

Overall, the results of these studies suggest that epsilonretroviruses in fish and reptiles are not particularly ancient. Fish epsilonretroviruses seem to be especially modern, with none dated at more than four million years old (Basta et al., 2009). This is unexpected, since modern epsilonretroviruses are most commonly associated with fish. There is more evidence of an ancient origin in

the amphibian epsilonretroviruses, which show some evidence of host tracking. However, further work is needed to establish the diversity and age of epsilon ERVs in these hosts.

1.4.5. Spumaviruses

1.4.5.1. ERV-L Elements

Of all the retroviruses, the most ancient group is thought to have been the group that led to the ERV-L elements, part of the spumavirus genus. Some ERV-L elements are orthologous throughout the placental mammals, with insertions predating the divergence of this group and dated as 104 to 110 million years old (Lee et al., 2013). It is possible that ERV-L like elements entered genomes even earlier than this, as the integration of these viruses is described as nearing the “maximum achievable lookback time” with the limitations of sequence deterioration and the difficulty in identifying orthologous sites in distantly related genomes (Lee et al., 2013). Although it may not be possible to look for orthology between ERV-Ls in more distantly related species, it does appear that ERV-L like elements are ubiquitous in vertebrates, as, besides mammals, ERV-L like fragments have been described in bony and cartilaginous fish (Kambol and Tristem, 2005), amphibians (Herniou et al., 1998), reptiles (Chong et al., 2012) and birds (Bolisetty et al., 2012).

1.4.5.2. Foamy viruses

Foamy viruses infect various mammals exogenously but are only known to have become endogenous in three species: the coelacanth fish (*Latimeria chalumnae*), the sloth (*Choloepus hoffmanni*) and the aye-aye (*Daubentonia madagascariensis*) (Han and Worobey, 2012a). The phylogenetic tree for the known hosts of exogenous foamy viruses and the tree for the viruses

themselves share a very similar topology and the branch lengths of the virus tree are highly significantly correlated with the host divergence times (Han and Worobey, 2012a). This provides strong evidence for codivergence of the viruses and their hosts, suggesting that foamy viruses were circulating approximately 400 million years ago and diversified within different host species as they diverged from each other (Han and Worobey, 2012a). This suggests that at least the foamy viruses have been active throughout the evolution of the tetrapods and that the association between vertebrates and their retroviruses is extremely ancient.

1.4.6. Alpharetroviruses

The alpharetroviruses are a large, diverse, relatively well-studied group of avian retroviruses. A major group of avian pathogens, the avian leukosis viruses (ALVs) are oncogenic exogenous alpharetroviruses (Payne and Nair, 2012). These viruses are unusual in that, depending on their subgroup, they are transferred horizontally between animals, from mother to offspring via infection and genetically from mother to offspring in the form of ERVs (Payne and Nair, 2012). These diseases are commonly considered to be diseases of chickens, but can infect other species of fowl including turkeys and ducks (Payne and Nair, 2012).

There is some ambiguity as to where the alpharetroviruses end and the betaretroviruses begin. Bolisetty et al. (2012) concluded that the betaretroviruses are ancestral to the alpharetroviruses, betaretroviruses are widespread in birds and the “true” alpharetroviruses are only a subsection of this group, originating in Galliform birds (Bolisetty et al., 2012). However, in the majority of retroviral literature this system is not used, as class II retroviruses isolated from mammals and birds tend to fall into separate, unambiguous monophyletic clusters (Gifford et al., 2005). Therefore, all class II avian retroviruses are generally considered to be alpharetroviruses, and this is the classification which will be used here.

Gifford et al. (2005) identified alpharetroviruses in at least 15 orders of birds using degenerate polymerase chain reaction (PCR). There are very likely other alpharetroviruses in these hosts which were not identified (Gifford et al., 2005). The chicken genome has been analysed in more detail and approximately 2% of the genome appears to be derived from a diverse range of alpharetroviral lineages (Huda et al., 2008). There does not appear to be any correspondence between the phylogeny of the alpharetroviruses from diverse hosts and the host phylogeny (Gifford et al., 2005). Dimcheff et al. (2000) identified ALV-like *gag* genes in 26 species of Galliformes and significant host-tracking at a genus level within parts of this group, with congruence between the host and *gag* phylogenies in the *Gallus* (chicken) and *Perdix* (partridge). Given these two datasets, it seems like most alpharetroviruses invaded avian genomes less than 66 million years ago, when the majority of bird orders and superorders began to diverge but before the radiation of the avian families, genera and species (Brown et al., 2008). This fits well with the estimated age of chicken alpha ERVs based on LTR similarity calculated by Huda et al. (2008), which has a mean of 58 million years. These dates have a wide range from 1.5 million years ago to 140 million years ago, which would explain the inconsistency of host tracking in different clades of the alpharetrovirus tree (Dimcheff et al., 2000, Huda et al., 2008, Gifford et al., 2005).

1.4.7. Betaretroviruses

The betaretroviruses have only been identified in mammals. Figure 24 shows some of the key groups of gammaretroviruses in rodents and primates and how they relate to each other.

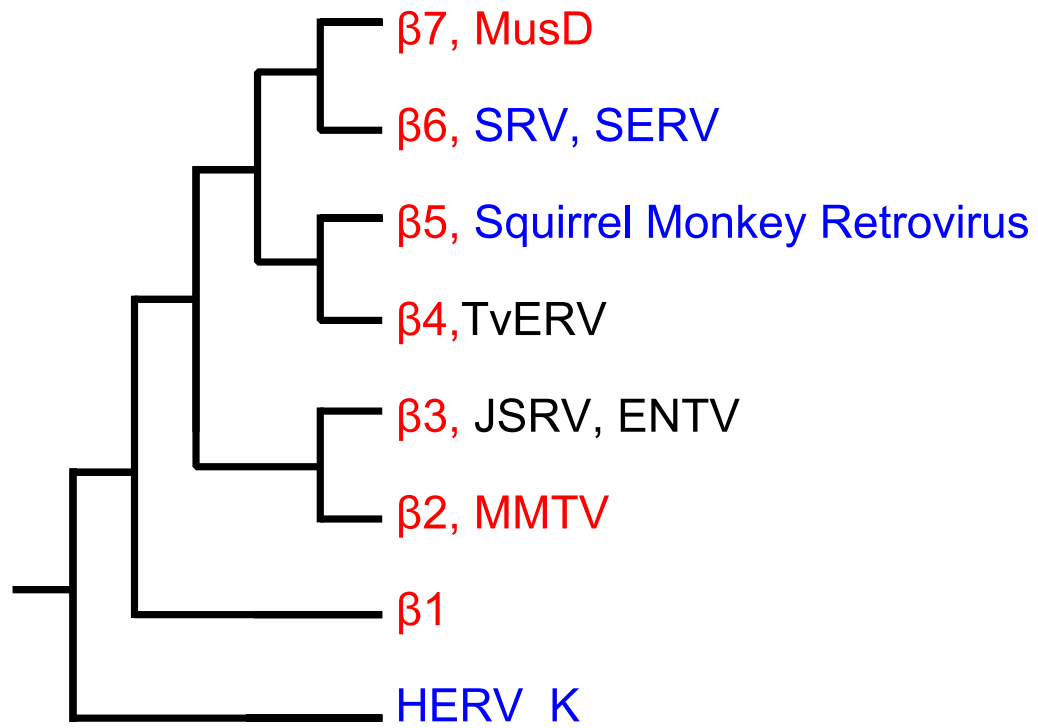


Figure 24: *Pol* gene phylogeny of primate and rodent betaretroviruses and their relatives. Adapted from Baillie et al. 2004.

Sequences from primates are shown in blue, sequences from rodents are shown in red.

1.4.7.1. HERV-K and $\beta 1$

Human endogenous betaretroviruses form the HERV-K group, which is probably the most widely studied group of ERVs. There are at least 550 HERV-K elements in the human genome (Bannert and Kurth, 2006). The group is subdivided into 10 families: human mouse mammary tumour virus like 1 to 8 (HML-1-8), HERV-K(C4) and HERV-K(14C) (Bannert and Kurth, 2006). The HERV-K group appears to be unique to primates, with no close relatives known in other taxa (Baillie et al., 2004).

The majority of known HERV-K like betaretroviruses integrated into the common ancestor of all old world primates at least 32 million years ago (Bannert and Kurth, 2006, Perelman et al., 2011, Greenwood et al., 2005). HERV-K HML-5 seems to be the oldest group, as HERV-K HML-5 proviruses are found at common sites in new world and old world monkeys, so integrated at least 43 million years ago (Bannert and Kurth, 2006, Perelman et al., 2011,

Greenwood et al., 2005). HML-5 proviruses in the human genome are highly degraded, which fits with their ancient origin (Lavie et al., 2004).

The HERV-K HML-2 group is unusual in that although it was circulating at least 32 million years ago, the most recent integrations appear to have become fixed after the human and chimpanzee lineages diverged approximately six million years ago, with at least 29 insertions unique to humans (Shin et al., 2013, Bannert and Kurth, 2006, Mayer et al., 1999, Reus et al., 2001). At least six HERV-K HML-2 loci are polymorphic within the modern human population, with the proviruses present in 15-50% of humans and an empty insertion site in others (Turner et al., 2001, Shin et al., 2013). Therefore, genetic differences both between humans and chimpanzees and within the human population may be partly due to HERV-K HML-2 ERVs. The HERV-K HML-2 group also includes at least 17 full-length betaretroviruses, of which three are known to contain full ORFs for *gag*, *pro*, *pol* and *env*: HERV-K113, HERV-K115 and HERV-K119 (Shin et al., 2013, Subramanian et al., 2011). HERV-K113 and HERV-K115 are truly polymorphic and HERV-K119 is present only as a solo-LTR in some individuals (Shin et al., 2013, Subramanian et al., 2011). 23 HERV-K loci, including HERV-K115 but not HERV-K113 or HERV-K119, have been shown to be transcriptionally active (Flockerzi et al., 2008). An increase in expression at one HERV-K (HML-2) locus was shown to occur in malignant testicular tissue (Flockerzi et al., 2008), however HERV-K expression has not been definitively shown to be associated with any disease. The sequence of the theoretical ancestral progenitor to HML-2 has been deduced *in silico* and constructed, producing an intact element, known as Phoenix, which can generate structurally intact viral particles (Dewannieux et al., 2006). These particles are able to infect human and other mammalian cells and have an integration pattern consistent with that of other HERV-Ks. Existing HERV-K (HML-2) ERVs in the human genome have the potential to recombine to produce these infectious viral particles, although they currently appear not to do so (Dewannieux et al., 2006).

The widespread presence of recently integrated HERV-K loci seems to be unique to humans, while in chimpanzees gammaretroviruses have undergone recent expansion (Jern et al., 2006). Jern et al. (2006) examined ERV loci with less than 2% LTR divergence (estimated to have integrated less than five million years ago) and found seven loci unique to the human genome, all HML-2, and 24 loci unique to the chimpanzee genome - one HERV-K and 23 gammaretroviruses. HERV-K appears to have proliferated in the human genome predominantly via reinfection rather than retrotransposition or complementation (Belshaw et al., 2004). These results together suggest that HML-2 viruses have been continuously active in the human lineage since they first entered the common ancestor of the old world primates and that they may still be circulating.

The $\beta 1$ group of ERVs found in mice and rats clusters close to HERV-K, although the two clusters are distinct (Baillie et al., 2004), $\beta 1$ ERVs are non-recombinant betaretroviruses.

1.4.7.2. **SERV, SRV, BaEV, RD114, MusD and TvERV**

Outside of humans, primate betaretroviruses have not been widely studied. With the exception of the newest HML-2 insertions, the betaretrovirus complement of the chimpanzee is very similar to that of humans and all known chimpanzee betaretroviruses have human orthologues (Polavarapu et al., 2006a). In other primates, most work has focussed on the phylogenetic group which contains the simian ERVs (SERVs), simian retroviruses (SRVs), baboon endogenous virus (BaEV) and feline RD-114 ERV. This group is more closely related to MMTV, MusD and other rodent retroviruses than the HERV-K group (Baillie et al., 2004).

SERVs are a group of ancient endogenous simian betaretroviruses found in the Papionini and Cercopithecini tribes of old world monkey, estimated to have entered the germline between 12 and 18 million years ago (van der Kuyl et al.,

1997, Perelman et al., 2011). SERVs appear to be responsible for an unusually high number of recombination events and cross-species transmissions.

SRVs are a group of exogenous retroviruses with betaretroviral *gag* and *pol* genes but gammaretroviral *env* genes which frequently cause disease in captive macaques (van der Kuyl et al., 1997). These retroviruses cause an AIDS-like immunodeficiency syndrome in several species of macaque and have often been used as a model of AIDS for this reason (van der Kuyl et al., 1997). SRVs are common in laboratory populations of rhesus macaques, with a prevalence exceeding 50% in some populations and are thought to be a potential confounding variable in any study involving these animals (Lerche and Osborn, 2003). These viruses have also been detected in humans working closely with non-human primates and in wild macaques, baboons and langurs (Lerche et al., 2001, Grant et al., 1995, Nandi et al., 2003, Sommerfelt et al., 2003). Exogenous SRVs are thought to be derived from recombination between the *gag-pol* region of an SERV and the *env* gene of a gammaretrovirus (Sonigo et al., 1986, van der Kuyl et al., 1997) (Figure 25). Therefore, SRVs present an atypical case, in that they appear to be the result of recombination of an ancient ERV with another retrovirus resulting in an active pathogen.

SERV is also thought to be the parent of another recombinant virus, BaEV, which has a gammaretroviral *gag-pol* region from *Papio cynocephalus* endogenous retrovirus (PcEV) and an *env* from an SERV (Mang et al., 1999) (Figure 25). BaEV-like sequences are only seen in members of the Papionini tribe and in *Chlorocebus* species (previously known as *Cercopithecus aethiops*) (van der Kuyl et al., 1995). Viruses from the Papionini group form a phylogenetic cluster, as do the *Chlorocebus* viruses (van der Kuyl et al., 1995). BaEV appears have entered the germline between 24,000 and 400,000 years ago, and is likely to have integrated separately into the *Chlorocebus* and Papionini germlines, rather than into a common ancestor, as the last common ancestor of these groups lived approximately 11.5 million years ago (Perelman et al., 2011, van der Kuyl et al., 1997). The recombination event which led to

this virus must have occurred in a species which harboured both SERV and PcEV (van der Kuyl et al., 1997, Mang et al., 1999).

RD-114 ERV has a betaretroviral *env* gene but gammaretroviral *gag* and *pol* genes and is found in the genome of all cats. It was recently identified in an infectious form in several live attenuated animal vaccines, probably due to culture of vaccines and seed viruses in feline cell lines (Miyazawa et al., 2010, Yoshikawa et al., 2010, Yoshikawa et al., 2011b, Yoshikawa et al., 2011a). Although RD-114 is not known to be associated with disease it has pathogenic potential and replicates in at least human, feline and canine cells (Yoshikawa et al., 2012). RD114 has a recombinant genome structure involving SERV – the *env* gene appears to be derived from BaEV, while the remainder of the genome is derived from the cat gammaretrovirus *Felis catus* endogenous retrovirus (FcEV) (van der Kuyl et al., 1999) (Figure 25). This suggests a cross species transmission of BaEV from primates to cats (van der Kuyl et al., 1999). At some point, an ancestor of the cat lineage must have harboured both BaEV and FcEV for a recombination event to have occurred (van der Kuyl et al., 1999).

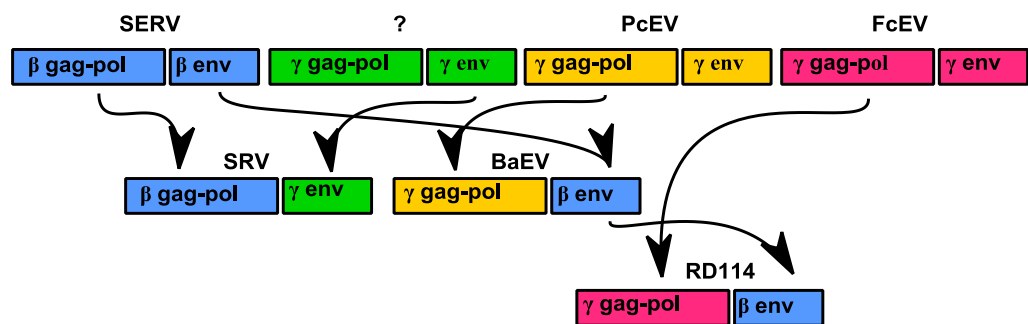


Figure 25: The relationship between the *gag*, *pol* and *env* genes of SERVs, SRVs, BaEV and RD114.

Colours show the transmission of genes between retroviruses.

Two other retroviruses form an unexpected clade with the SERVs: possum (*Trichosurus vulpecula*) ERV (TvERV) and squirrel monkey endogenous retrovirus (SMRV). These insertions are non-recombinant betaretroviruses but are unusual in that they are from geographically distant regions to the other viruses in the clade, which predominantly affect old world monkeys

living in Africa and Asia (Baillie and Wilkins, 2001). SMRV is one of very few known ERVs in new world primates, infecting a group of species found only in South America. TvERV infects possums, which are very distinct from primates evolutionarily, geographically and biologically (Baillie and Wilkins, 2001). These viruses not only reveal the potential for transmission of SRV-like retroviruses to very distinct species, but also raise questions as to how the viruses have been transmitted between such distant areas.

The *pol* gene of the $\beta 4$ group of rodent betaretroviruses clusters closely with TvERV, $\beta 5$ with SMRV and $\beta 6$ with SERV (Figure 24) (Baillie et al., 2004). However, like the SRVs, some of these insertions appear to be recombinant, as the $\beta 5$ and $\beta 6$ groups have gammaretrovirus like *env*-genes (Baillie et al., 2004). Within the $\beta 4$ group, some insertions with a $\beta 4$ like *pol* have a beta-like *env* while others have a gamma-like *env*. It appears that the tendency for this phylogenetic group to recombine extends into the rodents.

The $\beta 7$ group of mouse and rat retroviruses consists of the MusD elements (Baillie et al., 2004). These appear as an outgroup to the rest of the SERV-like group and usually lack *env* genes (Gifford et al., 2005, Baillie et al., 2004). MusD elements are numerous compared to other betaretroviruses, with around 100 copies in the mouse genome (Ribet et al., 2007). Only three of these copies are capable of retrotransposition (Ribet et al., 2007). The wide distribution of these elements seems to be the result of their intracellular lifestyle, as they lack part of the *gag* sequence which should target them to the plasma membrane (Ribet et al., 2007). Instead, the MusD elements can directly reintegrate within the same cell after reverse transcription, which is more efficient (Ribet et al., 2007).

1.4.7.3. Mouse Mammary Tumour Virus

MMTV is an exogenous pathogen causing mammary carcinomas in laboratory mice (Holt et al., 2013b). Two to eight endogenous copies are also found in a

typical inbred mouse genome (Holt et al., 2013b). Baillie et al. (2004) identified MMTV-like endogenous insertions in published mouse and rat genomes, designated as the $\beta 2$ group of rodent betaretroviruses, plus a similar insertion in the cow, and Garcia-Etxebarria and Jugo (2010) identified nine MMTV-like insertions in the cow genome. MMTV like insertions are also found in at least the horse (Brown et al., 2012, van der Kuyl, 2011) giraffe, bison and musk-ox [all Gifford et al. (2005)] but none in the human (Bannert et al., 2010) or chimpanzee (Polavarapu et al., 2006a) genomes. Therefore, it is possible that primates lack insertions in this group.

1.4.7.4. Jaagsiekte Sheep Retrovirus and Enzootic Nasal Tumour Virus

JSRV is a major betaretroviral pathogen of sheep, causing infectious pulmonary carcinoma (Palmarini et al., 2004). Enzootic nasal tumour virus (ENTV) is closely related to JSRV and has a similar pathogenic effect (Palmarini et al., 2004). There are around 20 endogenous copies of JSRV/ENTV viruses in the genomes of sheep and goats, the majority of which appear to have integrated after these species diverged four to 10 million years ago (Palmarini et al., 2004). Two of these insertions appear to predate the split between sheep and goats and could possibly be up to 18 million years old, which would also result in their presence in Bovidae (cows) and Cervidae (deer) (Palmarini et al., 2004). Accordingly, close relatives of JSRV/ENTV appear to be present in cows but not in pigs (Garcia-Etxebarria and Jugo, 2010) and in the caribou and white-tailed deer (Gifford et al., 2005), both Cervidae. Baillie et al. (2004) found the $\beta 3$ group of retroviruses in mice and rats, which is related to JSRV/ENTV. This relationship appears to be more distant than the relationship between the ovine, cervid and bovine JSRVs, however further analysis is needed to confirm this (Baillie et al., 2004).

1.4.8. Lentiviruses

1.4.8.1. RELIK

Until recently, it was thought that lentiviruses were unable to become endogenous, as no endogenous examples had been found (Katzourakis et al., 2007). This was hypothesised to be either because lentiviruses emerged too recently to have had time to become endogenous, or because of a biological barrier to germline invasion by lentiviruses (Katzourakis et al., 2007). However, Katzourakis et al. (2007) discovered the first example of an endogenous lentivirus in the European rabbit (*Oryctolagus cuniculus*). Around 25 full-length copies of RELIK and 150 solo LTRs were found in the rabbit genome (Katzourakis et al., 2007). RELIK clusters with EIAV in the lentivirus phylogenetic tree (Katzourakis et al., 2007). EIAV and RELIK also have similar genomic organisation, simpler than that of other lentiviruses, and both lack a *vif* gene, which suggests that these viruses are more similar to the unknown precursor of modern lentiviruses, thought to have a simpler genomic organisation (Katzourakis et al., 2007). Keckesova et al. (2009) found related insertions in the European brown hare (*Lepus europaeus*) which are orthologous to the rabbit insertions. The presence of these orthologues implies that RELIK entered the genome of a common ancestor of *L. europaeus* and *O. cuniculus*, providing a minimum integration date of 12 million years ago (Keckesova et al., 2009).

1.4.8.2. pSIVgml and pSIVfdl

Gifford et al. (2008) screened genome data from 21 primate species for further endogenous lentiviruses and found several regions with significant homology to lentiviruses in the genome of the grey mouse lemur (*Microcebus murinus*). Three regions contained the 5' and 3' LTR sequences and nearly complete *gag* and *pol* coding sequences of an endogenous lentivirus (Gifford et al., 2008). This virus was named grey mouse lemur prosimian immunodeficiency virus

(pSIVgml). The *M. murinus* genome was shown to contain at least 10 distinct pSIVgml insertions, including 2 full-length insertions and 8 solo-LTRs, so, as the published *M. murinus* genome has 30% coverage, the full genome is estimated to contain around six full-length insertions and 24 solo-LTRs (Gifford et al., 2008). The sequence was amplified and a consensus pSIVgml genome generated, containing four in-frame stop codons and one frameshifting indel (Gifford et al., 2008). Phylogenetically, pSIVgml falls between the primate and feline lentiviruses (Gifford et al., 2008). The structure of the pSIVgml genome is also transitional between these two groups (Gifford et al., 2008). The integration date of the insertion was estimated at 1.9 – 3.8 million years ago (Gifford et al., 2008).

Gilbert et al. (2009) characterised a second prosimian lentivirus, prosimian immunodeficiency virus fat-tailed dwarf lemur (pSIVfdl) in the genome of *Cheirogaelus medius*. This virus has 93 to 96% sequence similarity to pSIVgml (Gilbert et al., 2009). Southern hybridisation was used to look for proviruses related to pSIVgml in nine species of Malagasy lemur and was successful in *C. medius*, *M. murinus* and at a low copy number in *M. grisorufus* (Gilbert et al., 2009). *C. medius* was shown to contain at least four potentially full-length insertions (Gilbert et al., 2009). All pSIVgml and pSIVfdl sequences were used to generate a more intact consensus sequence, without stop codons or frameshifts (Gilbert et al., 2009). There were no shared orthologous insertions between *M. murinus* and *C. medius* and the total genetic distance between pSIVfdl and pSIVgml copies gave an estimated integration date of 3.75 – 18.75 million years ago (Gilbert et al., 2009). As pSIVgml and pSIVfdl are estimated to have integrated no more than 18.75 million years ago but are found in species which diverged 24 million years ago, and are found at different positions in the two hosts, it seems that the species were colonised separately by circulating pSIVs (Gilbert et al., 2009).

Exogenous SIVs have only been identified to date in mainland African primates, so if pSIV is the ancestor of modern SIVs the viruses must somehow

have been transferred between Madagascar and the mainland. Gifford et al. (2009) proposed two potential routes through which this could have occurred. First, the SIVs may have diverged into prosimian and simian lineages during the divergence of their hosts (Gifford et al., 2008). The most recent common ancestor of the simian and prosimian primates is estimated to have lived 87 million years ago (Perelman et al., 2011)(Figure 12). If SIVs and primates did codiverge, all simian primates, including new world primates, would either have SIVs or have independently lost their SIV lineage (Gifford et al., 2008). A second option is that species migrating between Madagascar and the mainland transmitted the virus. Lemur ancestors migrated to Madagascar on mats of vegetation from eastern Africa 50 to 54 million years ago and are not known to have been in contact with mainland African primates since this time. However, other species have crossed this divide considerably more recently and may have acted as vectors for the virus (Gifford et al., 2008). Finally, the virus may have been transferred even more recently via an aerial vector, such as a bird or bat (Gifford et al., 2008).

1.4.8.3. MELVs

Endogenous lentiviruses have also been identified in the genomes of members of the weasel family, the Mustelidae (Cui and Holmes, 2012, Han and Worobey, 2012b). These are known as the Mustelidae endogenous lentiviruses (MELVs). MELVs have been found in all the members of the Mustelinae and Lutrinae subfamilies tested to date but in none of the Martinae, giving an estimated integration date of nine to 11 million years ago (Han and Worobey, 2012b). This is consistent with the date estimated using LTR divergence (Han and Worobey, 2012b). The phylogenetic position of the MELVs is inconsistent, as they have been variously placed as an outgroup to the primate lentiviruses, the feline lentiviruses and to all the non-primate lentiviruses (Cui and Holmes, 2012). Weasels are a member of the Carnivora order of mammals (Figure 11)

so it is feasible that MELVs are an ancestor of modern FIVs, however more work is needed to show if this is the case.

1.4.9. Gypsy Elements and Errantiviruses

No retroviruses have been classified which predate the appearance of the first vertebrates. The genome of one species of amphioxus, the non-vertebrate species thought to be the most closely related to the vertebrates, has been sequenced and 0.5% of its genome were classified as ERV-like LTR retrotransposons, however no further detail is available about these insertions (Putnam et al., 2008).

Gypsy elements, a class of LTR retrotransposon, are found throughout the eukaryotes, including in invertebrates (Terzian et al., 2001). These elements resemble retroviruses in structure, as they have a *gag*-like region, encode integrase and reverse transcriptase enzymes and a proportion also encode an *env* like ORF (Terzian et al., 2001). The *gag* sequences of these elements do not share structural characteristics with retroviral *gag*. The RT protein of Gypsy elements is distinct from retroviral RT but the two are thought to share a common ancestor and the Gypsy RT is more similar to retroviral RT than to the RT of other retrotransposons (Xiong and Eickbush, 1990). Similarly, the IN protein of the Gypsy elements is distinct from, but shares a common ancestor with, the IN protein of retroviruses (Malik and Eickbush, 1999). The relationship between the Env protein of Gypsy elements and retroviruses is more ambiguous, however in some insects there is homology between the protein encoded by the gypsy *env* ORF and the envelope protein of the baculoviruses and structural similarities to retroviral Env proteins (Terzian et al., 2001, Malik et al., 2000). Gypsy elements in insects which encode Env have sometimes been described as insect ERVs or as the insect errantiviruses (Terzian et al., 2001). These have been identified in four species of *Drosophila*, plus *Trichoplusia ni* (the cabbage looper moth) and *Ceratitis capitata* (the Mediterranean fruit fly) (Terzian et al., 2001). These viruses form a single,

monophyletic group distinct from the vertebrate retroviruses regardless of which gene is used to build the phylogeny (Terzian et al., 2001). Although there is some shared ancestry, the insect errantiviruses are not thought to be direct precursors to the retroviruses (Terzian et al., 2001). Instead, it is proposed that both the insect errantiviruses and the vertebrate retroviruses are the result of separate acquisitions of *env* genes by retrotransposons (Kim et al., 2004). Malik et al. (2000) discussed two other *env* acquisition events by LTR retrotransposons. *Caenorhabditis elegans* have a group of LTR retrotransposons called “Cer” elements which have *env* genes (Malik et al., 2000). These are glycoproteins acquired from an ancestor resembling the modern Phleboviruses, most likely acquired during coinfection of a host cell. Similarly, *env* genes in *Tas* elements in another nematode, *Ascaris lumbricoides*, strongly resemble the gB glycoprotein of the Herpesviridae.

Chapter 2. Materials and Methods

This chapter will review methods previously used to identify and characterise ERVs and describe the pipeline developed here.

2. 1. Genome Screening for ERVs

The first step in this analysis was to identify candidate ERV-like regions in the genomes of interest. Section 2.1.1 briefly outlines the various genome screening techniques available. The technique selected for this project is described in more depth in sections 2.1.2 and 2.1.3. Details of how this technique was adapted and used are provided in sections 2.1.4 to 2.1.6.

2.1.1. Genome Screening: Techniques

Various computational techniques have been used to identify retroviral fragments in different mammalian genomes. ERVs can be difficult to detect as they are often very degenerate, so it is important to select a technique which takes this into account. Different techniques have different advantages and disadvantages and the optimum method depends on how the result will be used. A brief summary of the most commonly used bioinformatics-based methods for ERV detection is provided in this section.

2.1.1.1. BLAST and BLAT

The majority of ERV screening attempts to date have used BLAST (Altschul et al., 1990) or related algorithms [for example (Tristem, 2000, Villesen et al., 2004, Polavarapu et al., 2006a, Baillie et al., 2004, Pontius et al., 2007)]. In a BLAST-based genome screen, query sequences are compared to known retroviral sequences and the regions of the genome demonstrating the highest

similarity to these sequences are analysed (Altschul et al., 1990). The selection of appropriate input sequences is necessary for BLAST results to be comprehensive. TBLASTN, which compares translated nucleotide queries to a translated nucleotide database, is probably the most commonly used BLAST algorithm for locating ERVs. BLAT (Kent, 2002) is an algorithm related to BLAST, but is designed only to search target genomes for highly similar matches.

2.1.1.2. Retrorector

Recently, screening techniques based on conserved retroviral motifs have been developed and can be used as an alternative to sequence based screening.

Retrorector is a software tool developed to detect retroviral sequences based on conserved motifs, the distances between these motifs and codon usage (Sperber et al., 2009, Garcia-Etxebarria and Jugo, 2010). The detection of conserved motifs uses codon-by-codon comparison with a consensus amino acid sequence (Sperber et al., 2009). LTRs are detected, then the regions between LTRs are searched for other conserved retroviral motifs (Sperber et al., 2009). The program then attempts to reconstruct the *gag*, *pro*, *pol* and *env* proteins (Sperber et al., 2009). Retrorector favours longer sequences, as several retroviral fragments need to be present in the right order (Sperber et al., 2009). This increases accuracy but means shorter regions could be missed (Sperber et al., 2009). Retrorector is not useful for detecting solo-LTRs (Benachenhon et al., 2009). Many genomes have been screened for ERVs using Retrorector, including human (Sperber et al., 2007), chimpanzee (Sperber et al., 2007), rhesus macaque (Sperber et al., 2007), cow (Garcia-Etxebarria and Jugo, 2010), horse (Garcia-Etxebarria and Jugo, 2012), dog (Martínez Barrio et al., 2011), mouse (Sperber et al., 2007) and three avian genomes (chicken, turkey and zebra finch) (Bolisetty et al., 2012).

2.1.1.3. LTR_STRUC

Another widely used tool in searching for ERVs is LTR_STRUC. This program scans the genome to find pairs of regions which are similar to each other, within the length range of typical LTRs and within the length of a typical full-length ERV or LTR retrotransposon (Polavarapu et al., 2006a). These regions are then scanned for regions resembling primer binding sites, polypurine tracts and target site repeats (Polavarapu et al., 2006a). LTR_STRUC cannot locate solo-LTRs (Benachenhon et al., 2009). As LTR_STRUC only searches for LTRs, it may be able to detect elements with more variable genomic structures than other methods (Garcia-Etxebarria and Jugo, 2010). The mouse (McCarthy and McDonald, 2004), cow (Garcia-Etxebarria and Jugo, 2010), horse (Garcia-Etxebarria and Jugo, 2012) and chicken (Huda et al., 2008) genomes, amongst others, have been screened using this technique.

2.1.1.4. RepeatMasker and Repbase

The majority of published genomes are available pre-screened for repetitive elements, including ERVs, as it is often useful to mask these elements when analysing the genome, for example if only protein coding regions are of interest (Smit, 1996). The majority of repeat screening techniques identify multiple types of interspersed repeat, not just LTRs (Smit, 1996). This analysis is often performed using RepeatMasker (Smit, 1996). The program annotates the repeats and classifies them as LTR elements, short interspersed nuclear elements (SINEs), long interspersed nuclear elements (LINEs) and DNA elements (Smit, 1996). The LTR elements are subdivided into retroviruses, mammalian LTR retrotransposons and MER4 LTRs (Smit, 1996). This classification depends on their similarity to known elements in the RepeatMasker database, which is based on the Repbase dataset (Jurka et al., 2005). The algorithm used for this comparison can be selected by the user, commonly used algorithms include BLAST (Altschul et al., 1990) and cross_match (Green, 1996).

The success of RepeatMasker in detecting LTRs depends on its sequence libraries. The Repbase database aims to include “representative eukaryotic repetitive sequences and other biologically relevant information derived from printed journal articles, electronic journals, and public databases” (Jurka et al., 2005). Baillie et al. (2004) found that RepeatMasker was more successful in identifying groups of repeats within a species when there are large numbers of closely related members of that group in the host (Baillie et al., 2004). Low copy number repeats are less likely to be identified and ERVs can be mislabelled as other types of repeat (Benachenhou et al., 2009).

2.1.1.5. Comparison of Techniques

Direct comparison of genome screening techniques is difficult, because ERV content varies dramatically between genomes. The various techniques are also designed with different goals, for example to identify modern, intact ERVs, ancient degenerate insertions or solo-LTRs. Using the results of existing studies for comparative purposes is especially problematic, given the differences which exist between input sequence databases, genome builds and versions of the screening software.

Garcia-Etxebarria et al. (2010, 2012) attempted a direct comparison of three methods of ERV screening in the cow and horse genomes: LTR_STRUC, Retrorector and TBLASTN. In the cow genome, 4,487 elements were identified using LTR_STRUC, 9,698 using Retrorector and 928 using TBLASTN (Garcia-Etxebarria and Jugo, 2010) (Figure 26). In the horse genome, 291 elements were identified using LTR_STRUC, 1,615 using Retrorector and 378 using TBLASTN. Insertions identified using Retrorector were more likely to include all three major retroviral genes (*gag*, *pol* and *env*) and so were more intact (Garcia-Etxebarria and Jugo, 2010). As more sequences were identified using Retrorector and they were of higher quality, this appears to be the most efficient of these three ERV detection methods.

These comparisons of techniques have several limitations. First of all, all results are considered to be legitimate ERVs, when in reality all three techniques are likely to detect some false positives. As Figure 26 demonstrates, there was relatively little overlap between the regions identified in the cow genome using LTR_STRUC and the other two methods and in the horse genome only 31 ERVs were common to all three methods (Garcia-Etxebarria and Jugo, 2010, Garcia-Etxebarria and Jugo, 2012). The different techniques appear to be detecting different subsets of a larger overall ERV content, the size of which is unknown. BLAST results are also highly sensitive to the database of sequences used as an input and these studies both used a relatively small input of 12 sequences.

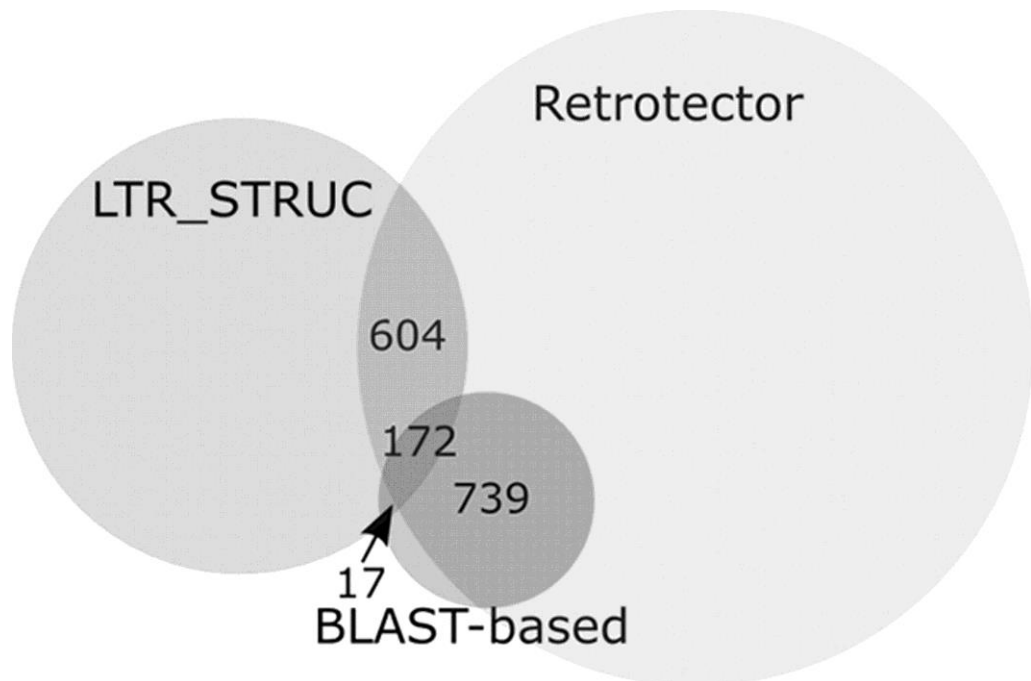


Figure 26: Comparison of the ERV sequences identified in the cow genome using LTR_STRUC, Retroector and BLAST.
From Garcia-Etxebarria et al. (2010).

A more robust comparison between genome screening algorithms may be to generate a pseudo-genome with a similar structure to the genome of interest and insert a known number of ERVs or ERV-like sequences with a range of properties at specific locations. This test genome could then be used to directly

compare the ability of the different genome screening algorithms to detect ERVs.

2.1.2. Exonerate

The sequence alignment algorithm selected for ERV screening in this study is part of the Exonerate package (Slater and Birney, 2005). This section describes how this algorithm works and how it was used here.

The Exonerate algorithm (Slater and Birney, 2005) was developed to allow “rapid approximation of exhaustive sequence alignment”. Exhaustive alignment models, such as the Smith-Waterman algorithm (Smith and Waterman, 1981) identify the optimum alignment of two sequences while heuristic strategies, such as BLAST (Altschul et al., 1990) generate valid alignments which are not necessarily the optimum. Heuristic strategies are significantly faster than exhaustive strategies (Slater and Birney, 2005). Exonerate is a heuristic method but approximates the alignment which would be produced using a user-specified exhaustive model (Slater and Birney, 2005).

The protein2genome model in Exonerate is a complex model built from components of other models (Slater and Birney, 2005). The model is used to map a protein onto genomic DNA (Slater and Birney, 2005). It begins with the Smith-Waterman algorithm, which compares two sequences by looking for the pair of segments from those sequences which have the highest degree of similarity, allowing for insertions and deletions (Smith and Waterman, 1981). The Smith-Waterman-Gotoh model builds upon this, reducing the amount of computational power required to run the model by using “affine gap penalties” (Gotoh, 1982). This means that while under the Smith-Waterman model, the costs of opening a gap and extending a gap are the same and gaps of all possible lengths need to be tested, under the Smith-Waterman-Gotoh model gap opening has a higher cost than gap extension, as fewer, longer gaps are

biologically more likely (Gotoh, 1982). The protein2genome model includes this change. Next, translation is incorporated into the protein2genome model, testing all possible translations and frameshifts (Slater and Birney, 2005). Finally, the model is altered so that introns can be present in the alignment (Slater and Birney, 2005). Using the Exonerate algorithm allows this complex model to be applied without using excessive computational resources.

2.1.3. Exonerate Pipeline

Exonerate has previously been incorporated into a pipeline to screen genomes for candidate ERV sequences. This pipeline has been used successfully to screen the horse (Appendix A.1) (Brown et al., 2012) and dog (Appendix A.2) (Tarlinton et al., 2012) genomes for ERVs. This pipeline is available at <https://github.com/ADAC-UoN/predict.genes.by.exonerate.pipeline>. This pipeline consists of five Perl scripts. The interactions between these scripts and their functions are summarised in Figure 27. The pipeline uses Exonerate to find candidate regions of similarity between a series of protein sequences and a genome sequence, then selects the highest quality result for each region and outputs these to the user in various formats suitable for further analysis. The input for the pipeline is a FASTA file containing protein sequences from known retroviruses and a genome sequence divided into FASTA files by chromosome (or sequences of similar length to chromosomes). It requires five user-specified options to run: the path to the file containing the protein sequences, the path to the directory where the chromosomes are stored, a prefix for the output sequence titles, the minimum allowable length of the hits (in amino acids) and number of introns to allow in a hit. These parameters are passed to a Perl script, “predict.genes.by.exonerate.pl”, a wrapper script which co-ordinates the other four scripts (Figure 27). The output is a table of non-overlapping candidate ERV regions which meet the requirements specified by the user and a FASTA file of these sequences. The intermediate output files produced by the pipeline are also stored.

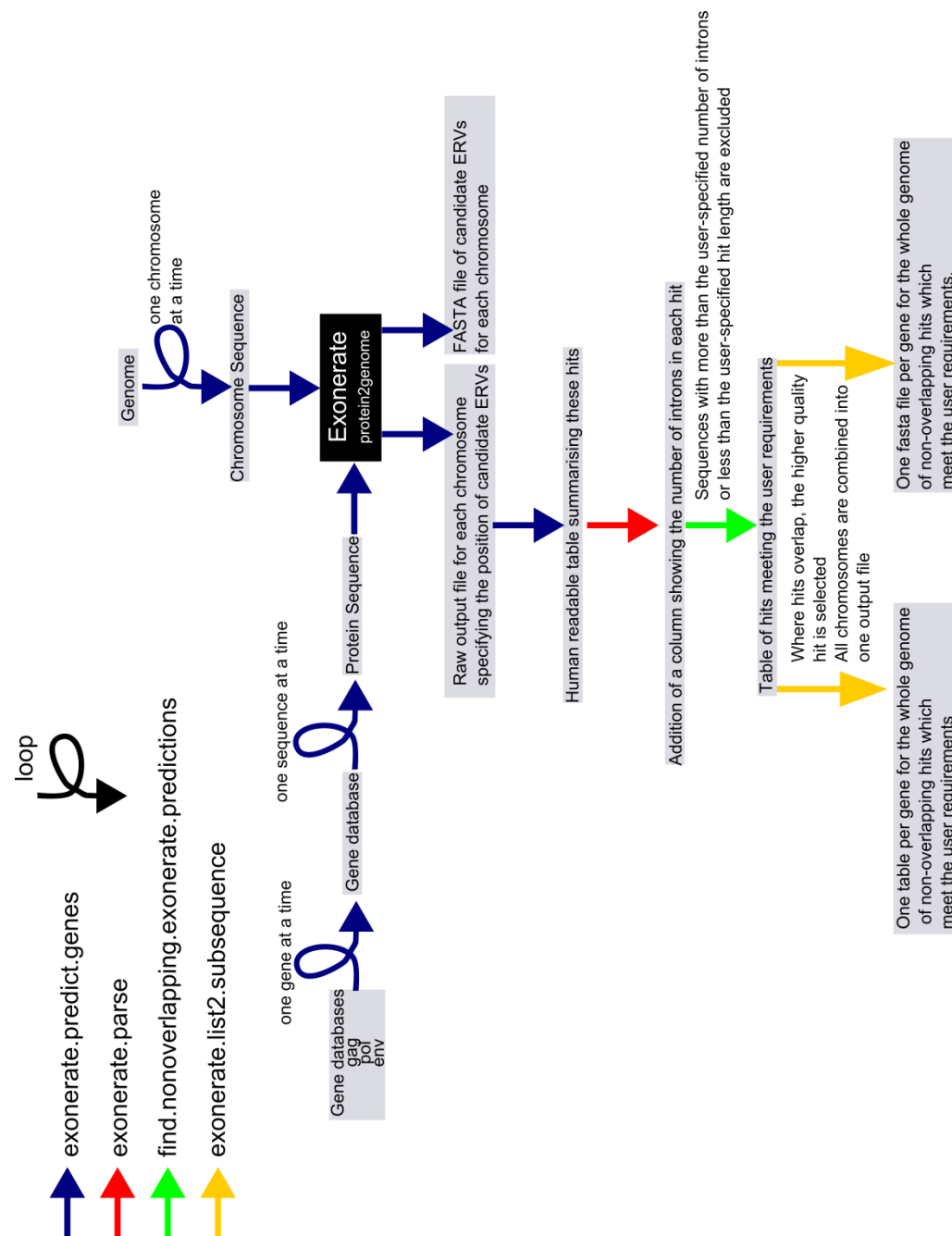


Figure 27: Flow chart representing the Exonerate pipeline used to detect candidate ERVs.

Arrows are colour-coded according to the Perl script which performs the task. The script “predict.genes.by.exonerate” co-ordinates this by activating the scripts one by one and passing the input and output files between them.

As Exonerate has previously been used to detect ERVs in the horse genome, it is possible to compare its performance in ERV detection to the other algorithms described in section 2.1.1.

Table 6: Comparison of the number of *pol* genes and of Class I *pol* genes identified in the horse genome with Exonerate, Retrotector, LTR_STRUC and TBLASTN.

Algorithm	N <i>pol</i> genes	N Class I	Reference
Retrotector	1,575	998	(Garcia-Etxebarria and Jugo, 2012)
LTR_STRUC	41	-	(Garcia-Etxebarria and Jugo, 2012)
TBLASTN	262	183	(Garcia-Etxebarria and Jugo, 2012)
Exonerate	813	768	(Brown et al., 2012)

Table 6 is a comparison between Exonerate, Retrotector, LTR_STRUC and TBLASTN in terms of how many ERV *pol* genes were detected in the horse genome. *Pol* genes were selected because these are the most useful for retroviral phylogeny and, as the longest gene, *pol* is easiest to detect. The horse study using Exonerate (Brown et al., 2012) was focussed on Class I sequences and used a single gammaretroviral query sequence, Moloney MLV. The number of Class I sequences identified was comparable with the number detected using Retrotector, which is usually found to identify the most ERVs (Garcia-Etxebarria and Jugo, 2012, Garcia-Etxebarria and Jugo, 2010). Despite the similarity between the algorithms, significantly more ERVs are detected using Exonerate protein2genome than TBLASTN, suggesting this algorithm is more appropriate for ERVs. This is especially apparent because Exonerate was used here with a single query sequence while Garcia-Etxebarria et al. used 12 sequences as a BLAST input. Increasing the size of the Exonerate input database may allow it to overtake other methods in terms of *pol* gene detection. However, this comparison of approaches is subject to the limitations discussed in section 2.1.1.5 and again may be improved by generation of an artificial dataset.

To verify that the insertions identified using Exonerate are genuine ERVs, a comparison between the proportion of insertions identified on each horse chromosome with each screening technique was performed. The results of this analysis are shown in Figure 28. There was a strong and highly significant ($p < 0.001$) positive correlation between the proportion of ERVs identified on each chromosome using Exonerate and each of the other three methods. The correlation with Retrorector was particularly high, which suggests that Retrorector and Exonerate detect a similar subset of ERV insertions.

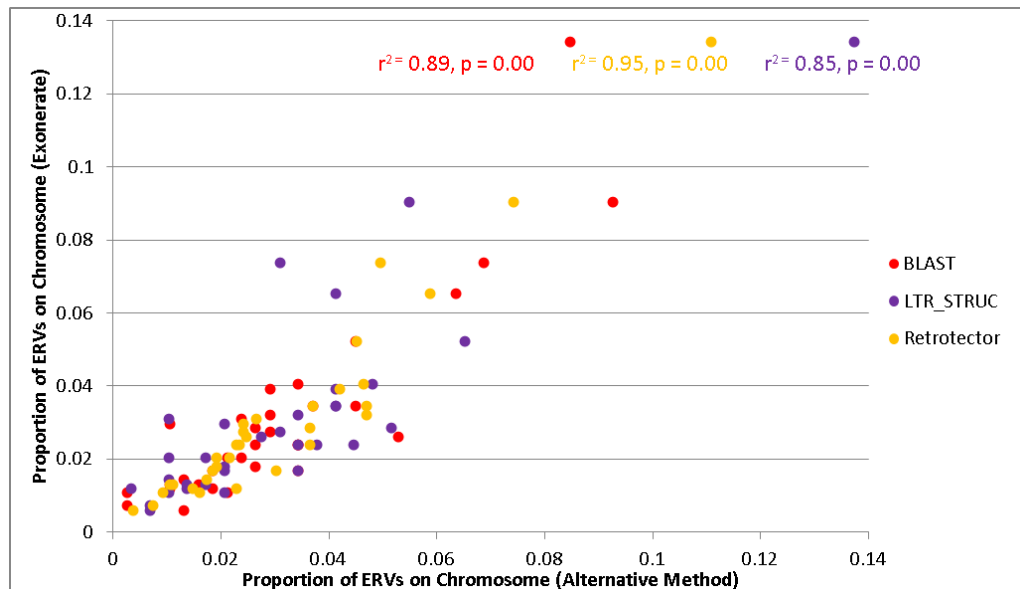


Figure 28: Scatter graph showing the proportion of ERVs identified on each chromosome of the horse genome using Exonerate compared to BLAST, LTR_STRUC and Exonerate and the Pearson's correlation coefficient for each comparison.

Finally, the horse ERVs identified using Exonerate were verified using BLAST. Each *pol* gene Exonerate hit from the horse genome was tested against the translated database of known retrovirus sequences described in section 2.1.4 using BLASTX. The output of this analysis was parsed to remove hits less than 75 nucleotides in length and with less than 30% identity to part of a known retrovirus. 791 out of 813 Exonerate sequences generated BLASTX hits which

exceeded these criteria. This shows that *pol* sequences identified by Exonerate are good candidates as ERVs.

2.1.4. Input Dataset

The size and quality of the input dataset used with Exonerate is expected to have a significant impact on the comprehensiveness of the output, as the algorithm can only identify regions matching sequences in the input dataset. A larger, more diverse input dataset should lead to a larger, more diverse output dataset. Phylogenetic analyses also require a comprehensive set of known sequences to provide reliable results. Therefore, an input dataset was built to encompass, as far as possible, the diversity of known exogenous and endogenous *gag*, *pol* and *env* sequences at the time of compilation.

106 journal articles providing or referring to sequences for endogenous or exogenous retroviruses were identified. Full references for these articles are listed in Appendix B.1. Sequences were either downloaded from Genbank or Repbase, copied directly from the manuscript or extracted from the appropriate region of the host chromosome between 02-Jun-2012 and 30-Jul-2012. Genome regions were downloaded from the UCSC genome browser using the genome build described in the original paper (Kent et al., 2002). Where sequences were copied directly from the manuscript, their genome position was identified using UCSC BLAT (Kent, 2002). The journal articles used included several genome-wide bioinformatics based ERV screens, in the human (Romano et al., 2006), baboon (Yohn et al., 2005), chimpanzee (Polavarapu et al., 2006a, Jern et al., 2006, Yohn et al., 2005), gorilla (Yohn et al., 2005), macaque (Yohn et al., 2005), mouse (Jern et al., 2007, McCarthy and McDonald, 2004), chicken (Huda et al., 2008), cow (Garcia-Etxebarria and Jugo, 2010), dog (Tarlinton et al., 2012) and pig (Yu et al., 2012b). Other groups of XRVs and ERVs described in the literature based on laboratory screening were also added. 651 sequences from different HIV strains were extracted from the National Institute of Health (NIH) HIV sequence database

(Antony et al., 2004). 96 human ERV strains were extracted from Repbase. Genbank was also extensively searched for sequences from other ERVs, including unpublished sequences.

The final nucleotide database is available as a MS Excel spreadsheet (Appendix B.1). The database has 4124 retroviral sequences from 321 host species, including mammals, birds, fish, amphibians and reptiles. 1119 sequences are full or partial *gag* genes, 1607 are *pol* genes and 1398 are *env* genes. Sequences have a mean length of 1310 bp. This dataset will be referred to from this point forward as the full previously known retrovirus (FULL_PREVKNOWN) dataset.

The FULL_PREVKNOWN dataset is of an unrealistic size to use as an Exonerate input and many of the sequences are redundant. The protein2genome model also requires amino acid, rather than nucleotide sequences as an input. Therefore representative nucleotide and amino acid datasets were created. Sequences in the FULL_PREVKNOWN dataset were placed into approximate groups by hand, according to their description (for example, HIV sequences were divided by group, sequences from whole genome screens were divided by host and viral genus). The longest sequence in each group was then selected as a potential “representative sequence”. Each sequence in the group was individually aligned to this representative using MUSCLE (Edgar, 2004) and its percentage identity to the representative sequence calculated. Sequences with less than 85% identity to the representative sequence were incorporated into the parsed dataset individually, otherwise only the representative sequence was added. This left 1590 sequences, described from here forward as the parsed untranslated previously known retrovirus (PARSED_UT_PREVKNOWN) dataset. This dataset is available as a FASTA file in AppendixC.2.

This dataset was then translated in all six reading frames using the seqinr package in R, in order to generate the amino acid sequences needed for the protein2genome Exonerate model (R Core Team, 2014) and the translation

with the least stop codons was added to an amino acid sequence dataset. 155 sequences were too degenerate to translate, with four or more stop codons in all reading frames, these were removed from the dataset. Sequences less than 200 nucleotides in length were also removed. A BLASTP search was performed against the Uniprot database for each of the remaining sequences and any sequence not generating a significant match (as defined by BLASTP) against a retroviral *gag*, *pol* or *env* gene was removed. The final Exonerate input dataset consists of 1361 amino acid sequences (367 Gag, 655 Pol and 339 Env) which represent 89.48% of the original nucleotide dataset. This dataset is available as a FASTA file in Appendix C.3 and will henceforth be referred to as the parsed translated previously known retrovirus (PARSED_T_PREVKNOWN) dataset. The mean similarity between each sequence in the FULL_PREVKNOWN dataset and the sequence by which it is represented in the PARSED_T_PREVKNOWN dataset is 96.3%.

To test the PARSED_T_PREVKNOWN dataset as an Exonerate input, three analyses were performed on the human genome. The first used only the *gag*, *pol* and *env* genes of Moloney MLV as an input. The protein2genome model was run on the GRCh37.p10 version of the human genome with 0 introns and a minimum overlap of 200 amino acids. This analysis identified 1439 ERV-like fragments (82 *gag*, 1327 *pol*, 30 *env*). The second analysis used the same settings with the RefSeq retroviral *gag*, *pol* and *env* genes as an input, as listed in Appendix B.3. RefSeq is a curated database of annotated reference sequences which included 33 members of the Retroviridae, amounting to 99 gene sequences, when this analysis was carried out (Pruitt et al., 2012a). This analysis identified 3602 ERV-like fragments (442 *gag*, 2792 *pol*, 368 *env*). The same analysis was then performed with the PARSED_T_PREVKNOWN input dataset. This resulted in 8945 ERV-like fragments (1709 *gag*, 6171 *pol*, 1065 *env*). The fragments from all three analyses were verified using BLASTX against PARSED_T_PREVKNOWN with a threshold of 40% identity over 100 bp. With MLV only as an input, 98.3% of hits pass this test, with the RefSeq sequences 98.1% pass and with the full input 93.3% pass. There is a 5%

decrease in quality in the hits with the full input dataset compared to the MLV-only input, however only 16% of the high quality hits identified with the full dataset are identified with only MLV as a query. Therefore, using the full input dataset leads to a considerably more comprehensive output.

2.1.4.1. Categorisation of Reference Sequences

The 1590 known retrovirus sequences in the untranslated PARSED_UT_PREVKNOWN were combined into 242 approximate groups of interest using a combination of prior knowledge about the retroviruses and sequence similarity. This grouped dataset was used later to provide closely related genetic groups for phylogenetic analyses (section 2.2.3). Of the 1590 sequences, 1205 were already members of well-characterised groups and had been subject to previous informative phylogenetic analyses. These were divided accordingly into 203 groups. The remainder were less well-characterised and so were compared sequence by sequence to each of the 1205 well-characterised sequences.

This comparison, and subsequent sequence-by-sequence comparisons used to identify ERVs, used the water function of EMBOSS, which is based upon the Smith-Waterman algorithm (Rice et al., 2000, Smith and Waterman, 1981). This function aligns two sequences, finds regions of local similarity and assigns a score based on the quality of the alignment of these regions (based on length and sequence similarity). The uncharacterised known sequences were assigned to groups based upon their highest score after comparison with each of the 1205 well-characterised sequences. New groups were created for sequences or groups of sequences which did not reach a threshold score (here a score of 300 was used after testing using sequences from well-characterised categories). This led to creation of 36 additional groups. The groups into which each previously known retrovirus sequence was placed are listed in Appendix B.4. The grouped dataset of known sequences from

PARSED_UT_PREVKNOWN is referred to as the grouped previously known retrovirus (GROUPED_PREVKNOWN) dataset.

2.1.5. Input Genomes

Genome sequences were downloaded for 33 species of vertebrate. Species details are listed in Table 7. 30 species are members of the Euarchontoglires, the taxonomic superorder which includes all known primates, rodents, lagomorphs, tree shrews and colugos (flying lemurs). 15 species of primate, 11 species of rodent, 2 species of lagomorph and 2 species of tree shrew had been sequenced on the date this analysis was performed (08-Mar-2013) (Table 7). Three species outside of the Euarchontoglires were also screened: the ferret genome was screened due to the presence of known endogenous lentiviruses in its genome and the chicken and turkey genomes to investigate the possibility of amniote-specific retroviruses.

All genomes were downloaded on 08-Mar-2013 from one of the following sources, in this order of preference: RefSeq release 57 (<http://www.ncbi.nlm.nih.gov/refseq/>) (Pruitt et al., 2012b), National Center for Biotechnology Information (NCBI) Genome (<http://www.ncbi.nlm.nih.gov/genome/>), Ensembl release 70 (<http://www.ensembl.org/>) (Flicek et al., 2012), NCBI Whole Genome Shotgun (WGS) (<http://www.ncbi.nlm.nih.gov/genbank/wgs>). Full details of the genome builds used are listed in Table 8. Genome quality was variable between hosts, with some genomes, for example human and mouse, sequenced to a high quality and others in an early draft stage of sequencing.

Table 7: Taxonomic details of the 33 vertebrate genomes screened using Exonerate.

Scientific Name	Common Name	Abbreviated Name	Prefix	Taxonomic Group
<i>Callithrix jacchus</i>	Common marmoset	Marmoset	Cjac	Primate
<i>Daubentonia madagascariensis</i>	Aye-aye	Aye-aye	Dmad	Primate
<i>Gorilla gorilla</i>	Western gorilla	Gorilla	Ggor	Primate
<i>Homo sapiens</i>	Human	Human	Hsap	Primate
<i>Macaca fascicularis</i>	Crab-eating macaque	Crab Eating Macaque	Mfas	Primate
<i>Macaca mulatta</i>	Rhesus macaque	Rhesus Macaque	Mmul	Primate
<i>Microcebus murinus</i>	Gray mouse lemur	Lemur	Mmur	Primate
<i>Nomascus leucogenys</i>	Northern white-cheeked gibbon	Gibbon	Nleu	Primate
<i>Otolemur garnettii</i>	Northern greater galago	Bushbaby	Ogar	Primate
<i>Pan paniscus</i>	Bonobo	Bonobo	Ppan	Primate
<i>Pan troglodytes</i>	Common chimpanzee	Chimpanzee	Ptro	Primate
<i>Papio anubis</i>	Olive baboon	Baboon	Panu	Primate
<i>Pongo abelii</i>	Sumatran orangutan	Orangutan	Pabe	Primate
<i>Saimiri boliviensis</i>	Black-capped squirrel monkey	Squirrel Monkey	Sbol	Primate
<i>Tarsius syrichta</i>	Tarsier	Tarsier	Tsyr	Primate
<i>Tupaia belangeri</i>	Northern Treeshrew	Northern Treeshrew	Tbel	Tree Shrew
<i>Tupaia chinensis</i>	Chinese Treeshrew	Chinese Treeshrew	Tchin	Tree Shrew
<i>Cavia porcellus</i>	Guinea Pig	Guinea Pig	Cpor	Rodent
<i>Chinchilla lanigera</i>	Long-tailed chinchilla	Chinchilla	Clan	Rodent
<i>Cricetulus griseus</i>	Chinese hamster	Hamster	Cgri	Rodent
<i>Dipodomys ordii</i>	Ord's kangaroo rat	Kangaroo Rat	Dord	Rodent
<i>Heterocephalus glaber</i>	Naked mole rat	Naked Mole Rat	Hgla	Rodent
<i>Ictidomys tridecemlineatus</i>	Thirteen-lined ground squirrel	Ground Squirrel	Itri	Rodent
<i>Jaculus jaculus</i>	Lesser Egyptian jerboa	Jerboa	Jjac	Rodent
<i>Microtus ochrogaster</i>	Prairie vole	Vole	Moch	Rodent
<i>Mus musculus</i>	House mouse	Mouse	Mmus	Rodent
<i>Octodon degus</i>	Degu	Degu	Odeg	Rodent
<i>Rattus norvegicus</i>	Brown rat	Rat	Rnor	Rodent
<i>Ochotona princeps</i>	American pika	Pika	Opri	Lagomorph
<i>Oryctolagus cuniculus</i>	European rabbit	Rabbit	Ocun	Lagomorph
<i>Gallus gallus</i>	Domestic chicken	Chicken	Ggal	Bird
<i>Meleagris gallopavo</i>	Domestic turkey	Turkey	Mgal	Bird
<i>Mustela putorius</i>	Domestic Ferret	Ferret	Mput	Carnivore

Table 8: Assemblies and source databases for the 33 vertebrate genomes screened using Exonerate.

Abbreviated Name	Assembly	Assembly ID	Source	Assembly Level
Marmoset	Callithrix jacchus-3.2	GCF_000004665.1	RefSeq	chromosome
Aye-aye	DauMad_1.0	GCA_000241425.1	WGS	contig
Gorilla	gorGor3.1	GCF_000151905.1	RefSeq	chromosome
Human	GRCh37.p10	GCF_000001405.22	RefSeq	chromosome
Crab Eating Macaque	MacFas_Jun2011	GCA_000222185.1	NCBI	chromosome
Rhesus Macaque	Mmul_051212	GCF_000002255.3	RefSeq	chromosome
Lemur	micMur1	micMur1	Ensembl	scaffold
Gibbon	Nleu_3.0	GCF_000146795.2	RefSeq	chromosome
Bushbaby	OtoGar3	GCF_000181295.1	NCBI	scaffold
Bonobo	panpan1	GCF_000258655.1	RefSeq	scaffold
Chimpanzee	Pan_troglodytes-2.1.4	GCA_000001515.4	RefSeq	chromosome
Baboon	Panu_2.0	GCF_000264685.1	RefSeq	chromosome
Orangutan	P_pygmaeus_2.0.2	GCF_000001545.4	RefSeq	chromosome
Squirrel Monkey	SaiBol1.0	GCF_000235385.1	RefSeq	scaffold
Tarsier	tarSyr1	GCA_000164805.1	Ensembl	scaffold
Northern Treeshrew	ASM18137v1	GCA_000181375.1	WGS	contig
Chinese Treeshrew	TupChi_1.0	GCA_000334495.1	NCBI	scaffold
Guinea Pig	Cavpor3.0	GCF_000151735.1	RefSeq	scaffold
Chinchilla	ChiLan1.0	GCA_000276665.1	NCBI	scaffold
Hamster	CriGri_1.0	GCF_000223135.1	RefSeq	scaffold
Kangaroo Rat	dipOrd1	dipOrd1	Ensembl	scaffold
Naked Mole Rat	HetGla_female_1.0	GCA_000247695.1	NCBI	scaffold
Ground Squirrel	SpeTri2.0	GCA_000236235.1	NCBI	scaffold
Jerboa	JacJac1.0	GCA_000280705.1	NCBI	scaffold
Vole	MicOch1.0	GCA_000317375.1	NCBI	chromosome
Mouse	GRCm38.p1	GCF_000001635.21	NCBI	chromosome
Degu	OctDeg1.0	GCA_000260255.1	NCBI	scaffold
Rat	Rnor_5.0	GCF_000001895.4	RefSeq	chromosome
Pika	OchPri3.0	GCA_000292845.1	NCBI	scaffold
Rabbit	OryCun2.0	GCF_000003625.2	RefSeq	chromosome
Chicken	WASHUC2	WASHUC2	Ensembl	chromosome
Turkey	Turkey_2.01	GCF_000146605.1	RefSeq	chromosome
Ferret	MusPutFur1.0	GCA_000215625.1	NCBI	scaffold

2.1.5.1. Pre-processing

Of the 33 genomes listed in Table 8, to date, 15 have been assembled into chromosomes. The others are at various stages of assembly into scaffolds and contigs. The length of these scaffolds and contigs affects the probability of detecting intact ERVs (ERVs with a recognisable LTR-*gag-pol-env*-LTR structure, typically 7,000 to 10,000 bp in length) and the length of the gene fragments identified. With decreasing contig length there is an increased probability that ERVs or ERV genes will be broken across more than one contig.

Table 9 shows the probability of a 10,000 base pair intact ERV and of a 1,100 base pair gene fragment (the mean fragment length identified in the human genome) being split across more than one contig for two measures of average contig length: mean contig length and N50. N50 (provided by the source genome browser) is identified by sorting contigs by length, starting with the longest, then counting the bases in each contig until half the total genome length is reached (Yandell and Ence, 2012). The length of the contig in which this number is reached is the N50 length (Yandell and Ence, 2012). N50 tends to overestimate contig length while mean length does not take into account that, by definition, more of the genome will be contained in the longer contigs (Yandell and Ence, 2012). Therefore the real “middle point” of contig length is likely to fall between these two measurements.

For some genomes – aye-aye, hamster, kangaroo rat, lemur, northern tree shrew and tarsier - regardless of the measure of mean contig length used, full-length ERVs are unlikely to be identified and gene fragments are likely to be shorter than for more complete assemblies. The Chinese tree shrew and ground squirrel genomes are likely to have a reduced number of detectable full-length ERVs but fragment length should be mostly unaffected. For the remaining nine species assembled at a scaffold level there should be little

change in the probability of finding full-length ERVs or the length of ERV fragments compared to the genomes assembled to chromosome level.

Table 9: The probability that an average ERV (10,000 bp) and an average Exonerate ERV fragment (1100 bp) would span more than one contig or scaffold in each screened genome based on mean segment (contig or scaffold) length and on N50. Green, probability less than or equal to 1%; yellow, probability less than or equal to 5%; red, probability greater than or equal to 5%. Scaffold or contig N50 are provided depending on the “Assembly Level” given in Table 8.

Abbreviated Name	Mean Segment Length			N50		
	Length	Expected % Overlap 10,000 bp	Expected % Overlap 1110 bp	Length	Expected % Overlap 10,000 bp	Expected % Overlap 1110 bp
Aye-aye	884	100.00%	100.00%	3653	100.00%	30.39%
Bonobo	264,002	3.79%	0.42%	10,124,892	0.10%	0.01%
Bushbaby	323,332	3.09%	0.34%	13,852,661	0.07%	0.01%
Chinchilla	842,443	1.19%	0.13%	21,893,125	0.05%	0.01%
Chinese Tree shrew	56,090	17.83%	1.98%	3,670,124	0.27%	0.03%
Degu	419,943	2.38%	0.26%	12,091,372	0.08%	0.01%
Ground Squirrel	198,542	5.04%	0.56%	8,192,786	0.12%	0.01%
Guinea Pig	866,164	1.15%	0.13%	27,942,054	0.04%	0.00%
Hamster	21,986	45.48%	5.05%	156,635	6.38%	0.71%
Jerboa	260,185	3.84%	0.43%	22,080,993	0.05%	0.01%
Kangaroo Rat	11,048	90.51%	10.05%	36,427	27.45%	3.02%
Lemur	16,828	59.42%	6.60%	140,884	7.10%	0.78%
Naked Mole Rat	619,250	1.61%	0.18%	20,532,749	0.05%	0.01%
Northern Tree shrew	2,541	100.00%	43.68%	2,974	100.00%	37.32%
Pika	213,995	4.67%	0.52%	26,863,993	0.04%	0.00%
Squirrel Monkey	971,180	1.03%	0.11%	18,744,880	0.05%	0.01%
Tarsier	5,049	100.00%	21.98%	12,214	81.87%	9.01%

The Exonerate pipeline runs more efficiently against fewer, longer sequences. For the less assembled genomes, sequences were therefore concatenated into “artificial chromosomes”: sequences approximately the same length as a typical mammalian chromosome consisting of contig or scaffold sequences joined end to end. This is achieved using the Python script `make_chromosomes.py`, available in Appendix D.1. Briefly, this script divides the length of the genome by a user-specified number of chromosomes to give an approximate “chromosome length”. Contig sequences are then concatenated into strings of approximately this length, which are stored as FASTA files, with the position of each contig in each “chromosome” recorded in a text file.

2.1.6. Screening

The Exonerate pipeline was run for the full input database described in section 2.1.4 against each of the genomes listed in Table 8 with the following settings: model `protein2genome`; number of introns `0`; minimum overlap `200` amino acids. The number of introns was set to `0` as ERV sequences should not contain true introns and the `protein2genome` model already allows for gaps in the region matching the query sequence. This analysis generates three FASTA files and three tables of candidate ERV sequences for each genome, one for each gene. These analyses were performed on the University of Nottingham high performance computing cluster. The set of putative ERV fragments produced by this analysis will be referred to from this point forward as the raw Exonerate output (`RAW_EXO_OUT`) dataset.

2. 2. Parsing Output

Once the raw Exonerate output files had been generated, the output was parsed in various ways. First, poor quality results were removed using the method described in section 2.2.1. The high copy number of some ERVs and

the abundance of ERVs in the genome means that it was also usually necessary to group elements into families and represent each family with a representative or consensus sequence, as this minimises redundancy in the dataset and reduces the amount of computational power needed for further analysis. Various techniques have previously been used to cluster ERV sequences and these are reviewed here (section 2.2.2.1), and the technique selected discussed (section 2.2.2.2). Finally, the method used to identify ERV fragments is described in section 2.3.3.3.

2.2.1. Quality Control

To ensure that only genuine ERV sequences were passed to the next stage, all fragments in the RAW_EXO_OUT dataset were verified using BLASTX against the PARSED_T_PREVKNOWN dataset. Only sequences sharing at least 40% sequence identity over 100 bp with a PARSED_T_PREVKNOWN sequence were kept in the dataset. This edited dataset is referred to as the parsed Exonerate output (PARSED_EXO_OUT) dataset. Each sequence in PARSED_EXO_OUT was also assigned approximately to a genus using this technique. Sequences were first assigned to the genus against which they had the highest number of significant BLAST hits. If there was the same number of BLAST hits against two genera, the genus with the highest scoring hit was used.

New, parsed FASTA files and tables were generated from PARSED_EXO_OUT for each combination of host, genus and gene.

2.2.2. Clustering

2.2.2.1. Clustering Sequences: Techniques

The technique used for clustering depends on how related the input sequences are to each other and how much detail is required in the final analysis.

The simplest technique is to eliminate sequences which are very similar to each other. Bailie et al. (2004) examined a small section of the *pol* gene of related betaretroviruses, and all sequences with less than 95% sequence identity were kept in the analysis. Similarly, Bénit et al. (2001) and Han et al. (2007) excluded sequences with greater than 90% sequence identity.

Gentles et al. (2007) used a slightly more complex approach with the opossum genome, grouping sequences with more than 75% similarity over at least 50% of their length. Consensus sequences were then generated for each of these groups (Gentles et al., 2007). For more divergent sequences, this was carried out using a variation of the SWAT algorithm (Green, 1996). Each sequence was taken as a seed to which all others were aligned, then a majority-rule consensus sequence was built for each alignment (Gentles et al., 2007). Each of these consensus sequences was aligned in turn to each transposable element using SWAT, and the sequence with the highest overall alignment score was selected (Gentles et al., 2007).

Branch support in phylogenetic trees is often used to determine which groups are most likely to be legitimate families and can be represented by single sequences. In Polavarapu et al.'s analysis of the chimpanzee genome a neighbour joining phylogenetic tree was generated and for groups with high bootstrap support, the most recent intact element was taken as a representative (Polavarapu et al., 2006a). This gave one representative sequence for each ERV family. Tristem (2000) used a similar approach with the human genome, generating a neighbour-joining tree and removing all but three sequences from clusters with bootstrapping values over 95%. Garcia-Etxebarria and Jugo (2010) determined which species to include in their phylogeny of cow ERVs by comparing their results from neighbour-joining, maximum likelihood and Bayesian analysis. Clusters were selected which were significant in at least two analyses, significant groups had bootstrap values over 70% in neighbour joining and maximum likelihood analysis and posterior

probabilities of at least 95% in Bayesian analysis (Garcia-Etxebarria and Jugo, 2010).

A combination of techniques may be the most appropriate clustering methodology for ERVs. Approximate phylogenies describing a group of sequences can be generated quickly using hierarchical clustering techniques (Corpet, 1988). Sequence similarity within clusters can then be determined and sequences showing insufficient similarity to the rest of the cluster excluded. Hierarchical clustering relies on a distance matrix, first taking each item in the matrix as a single cluster then continuing to join the most similar clusters until one cluster remains.

2.2.2.2. Clustering Analysis

Here, for each of the FASTA files representing a host, gene and genus combination in PARSED_EXO_OUT, a clustering process was carried out in order to form reliable groups of ERVs. This process is described in Figure 29.

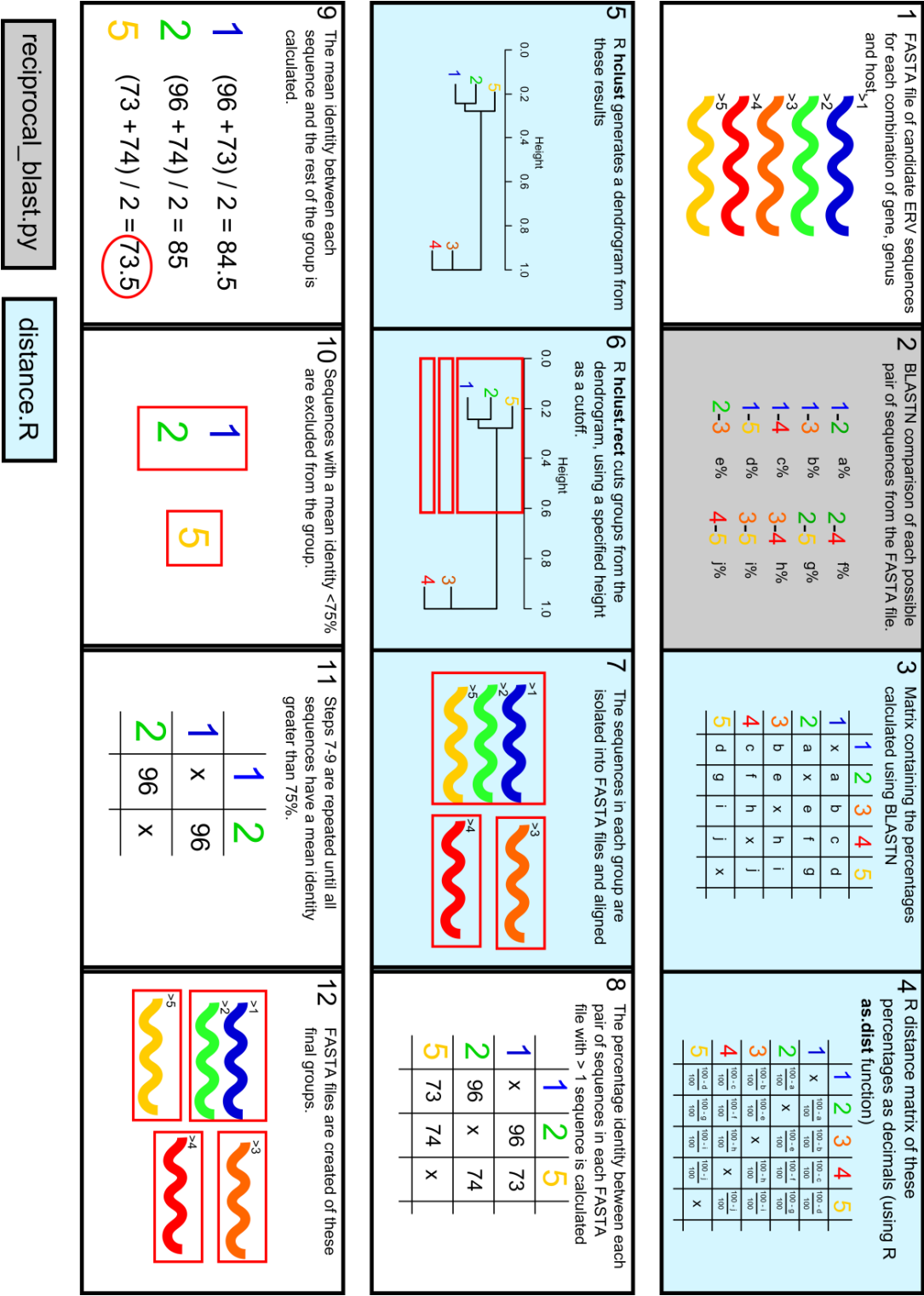


Figure 29: The process used to cluster candidate ERV sequences.
Grey and blue boxes represent tasks performed by the scripts `reciprocal_blast.py` (Appendix D.2) and `distance.R` (Appendix D.3) respectively. Numbers and sequence names are for illustration only.

The input files for the clustering process were the PARSED_EXO_OUT FASTA files, each of which represents a host, gene and genus combination. First, the similarity between sequence fragments was approximately established by using BLASTN searches to compare each pair of fragments in the file, via the Python script reciprocal_BLAST.py (Appendix D.2) (Figure 29.2). The results of these BLAST searches were combined into a matrix for each gene, genus and host combination showing the percentage identity of the most significant BLAST hit between each pair of sequences (Figure 29.3). Only pairs with over 100 identical bases were included, if there were no matches meeting this criterion a default value of 0% identity was assigned.

Based upon these matrices, the “hclust” function in R (R Core Team, 2014) was used to construct putative groups. This process is illustrated in Figure 29.4 and Figure 29.5 and the R script is available in Appendix D.3. This function produces a tree or dendrogram describing the clusters identified. A cut-off can then be specified below which clusters are isolated, using the R function rect.hclust (R Core Team, 2014) (Figure 29.6). A threshold of 35% similarity was selected, this is low but allows putative clusters to be identified which can then be checked by eye.

The FASTA files produced for each cluster by this process were aligned using MUSCLE (Edgar, 2004) under the default settings (Figure 29.7) and the percentage identity between each pair of sequences in the cluster (excluding gaps) calculated based on this alignment and combined into a matrix (Figure 29.8). Several rounds of checks were then carried out based on these matrices (Figure 29.8 - Figure 29.11). If all sequences in the matrix had a mean identity of at least 75% with the other sequences in the matrix the cluster was considered to be reliable. Otherwise, the alignments which formed the basis for each matrix were examined by eye. In some cases, there was a clear division within the alignment into two or three distinct groups, these groups were therefore split into subgroups accordingly. Single sequences which were noticeably poorly aligned were excluded. The new groups created at this stage

were realigned and new alignment matrices were generated. If all sequences in the matrix still did not have a mean identity of greater than 75% with all other sequences in the matrix the process was repeated. Groups which could not be resolved in this manner were split back into the original sequences.

A consensus sequence was built for each cluster using the Python script `make_cons.py`, available in Appendix D.4. This script uses a combination of the EMBOSS (Rice et al., 2000) functions `CONS` and `CONSAMBIG`, which generate standard and ambiguous consensus sequences respectively, to generate a final consensus without ambiguous characters which represents the input as closely as possible.

First, for each aligned FASTA file, both types of consensus sequence are created. For the standard consensus, the `CONS` plurality was set to 0.5, meaning a base present in more than half the sequences is used in the consensus. If no base is present in more than half the sequences, an “N” is added to the consensus at this position. For the ambiguous consensus, `CONSAMBIG` uses an IUPAC ambiguity code (Tipton, 1994) to describe sites where no base is present in more than half the sequences. If all bases are present at equal frequencies an “N” is added. Therefore, for each position where an “N” is present in the standard consensus, either an “N” or an IUPAC ambiguity code (Tipton, 1994) will be present in the ambiguous consensus.

The Python script `make_cons.py` (Appendix D.4) was written to represent these sites without the use of ambiguity codes. This script deletes sites with an “N” in both the ambiguous and standard consensus. Where an ambiguity code is present, one of the bases the code can represent is added to the final consensus, with all possible bases added at equal frequencies.

If the final consensus sequence produced by this process contained more than 10% “N”s, it was checked by eye and if necessary the alignment was split until a more reliable consensus could be produced. Each sequence in each cluster was then aligned to its consensus and their percentage identity calculated. The

minimum identity accepted was 70%, however mean identity between each sequence and its consensus was 96%.

The final parsed output dataset consists of these consensus sequences plus all sequences which were not sufficiently similar to another sequence to be incorporated into a consensus. This dataset is referred to as the clustered Exonerate output (CLU_EXO_OUT) dataset.

2.2.3. Identification of Output Sequences

Sequences in CLU_EXO_OUT were identified and categorised according to their similarity to previously known retroviruses.

First, newly identified sequences were divided by gene (*gag*, *pol* or *env*) and into three “classes”, gamma and epsilon-like retroviruses, alpha, beta, lenti and delta-like retroviruses and spuma-like retroviruses. This categorisation is based on the previously used classification system for HERVs into class I, class II and class III (Gifford and Tristem, 2003), which is consistent with the phylogeny of ERVs (Jern et al., 2005) but is broader than the genus categorisation. This was used to allow sequences originally classified as the wrong genus (in section 2.2.1) to move into the appropriate group but avoid time-consuming comparison of distant sequences.

Each sequence in CLU_EXO_OUT was aligned to each sequence in the untranslated parsed dataset of previously known ERVs (PARSED_UT_PREVKNOWN) of the same gene and type, using the Smith-Waterman algorithm (Smith and Waterman, 1981) via EMBOSS water (Rice et al., 2000) as described in section 2.1.4.1. The CLU_EXO_OUT sequences were grouped with the PARSED_UT_PREVKNOWN sequences with which they had the highest alignment score. These groups were then merged based on the categories in GROUPED_PREVKNOWN, described in 2.1.4.1. This generated 120 groups of sequences (as not all categories contained any newly identified sequences). These 120 groups are referred to as the GROUPED_EXO dataset.

2. 3. Phylogenetic Analysis

The next step was to generate representative phylogenetic trees for groups of interest. This section contains a review of alignment (section 2.3.1) and tree-building techniques (section 2.3.2) and a description of the techniques used here (section 2.3.3).

2.3.1. Aligning Retroviral Sequences: Techniques

As retroviral sequences are so degenerate and divergent, they can be difficult to align. Alignments are often performed using either Clustal (Larkin et al., 2007) or MUSCLE (Edgar, 2004) and adjusted by hand. It is unclear whether nucleotide or amino acid sequences are preferable for this purpose, as amino acids generate less ambiguous alignments but the degeneracy of ERV sequences means that amino acid sequences are often short and contain multiple stop codons (Polavarapu et al., 2006a). Additionally, as ERV sequences are not usually subject to selection to conserve protein function, amino acid alignments are not always meaningful. Baillie et al. (2004) used both techniques and minor differences can be seen between the two trees, however it is not clear which technique is more accurate. Clustal and MUSCLE are generally used interchangeably, but MUSCLE has been shown in some circumstances to run more quickly and have higher accuracy than Clustal (Edgar, 2004).

The MAFFT “L-INS-I” alignment technique has been shown to be more accurate than MUSCLE or Clustal in analysis of datasets of both similar and diverse sequences (Ahola et al., 2006, Katoh et al., 2002). This is a local, rather than global, alignment technique, so the first stage of the alignment process, pairwise comparison of the sequences, looks for regions of similarity in sequences rather than attempting to align every base in each sequence. As the Exonerate ERV fragments are often partial, a local alignment technique

may be preferable. The L-INS-I technique is iterative, meaning subsets of the initial alignment are realigned in an attempt to improve the alignment score. Alignment techniques using iteration have been shown to represent a significant improvement on non-iterative techniques, including for divergent and difficult datasets (Wallace et al., 2005). The L-INS-I technique is slow compared to other methods, so is more suitable for alignments with fewer than 200 sequences than for larger alignments (Katoh et al., 2002). The technique is also likely to be unnecessarily stringent for alignment of small groups of highly similar sequences, such as the alignments used for clustering in section 2.2.2.2. However, for the majority of ERV alignments, made up of partial, sometimes degraded and divergent sequences, this technique may be the most appropriate.

2.3.2. Building Trees: Techniques

Once sequences have been successfully aligned, they can be combined into phylogenetic trees. Various factors need to be considered in generating a robust phylogeny.

2.3.2.1. Choosing a Gene

Trees can be built based on LTRs, *gag*, *pol*, *env* or any combination of these. The type of element selected depends on the data available and the purpose of the tree. The *pol* gene has the longest retroviral ORF and its sequence is relatively conserved, so it is often used for phylogenetic analysis (Jern et al., 2005). Reverse transcriptase is present in a diverse range of non-retroviral elements, so it is useful in comparisons with these groups (Benit et al., 2001). The higher variability of the *env* gene means it is useful for distinguishing more closely related sequences, for example HIV sequences within an individual (Andréoletti et al., 2007). Multiple genes are often examined, this is particularly useful when looking for recombination events (Benit et al., 2001). In particular, *env* trees tend not to be identical to those generated using *gag*

and *pol* (Benit et al., 2001). LTR sequences are often used to differentiate between closely related sequences. For example Polvarapu et al. (2006) used this technique to show the relatedness of different members of the same family of retroviruses in chimpanzees. Belshaw et al. (2005) used LTR sequences to generate a phylogenetic tree of human HERV-K like insertions (Belshaw et al., 2005b).

2.3.2.2. Model Selection

In order to build a phylogenetic tree, a substitution model is usually needed, showing the probability of each possible mutation (Posada and Crandall, 2001). Neighbour-joining, maximum likelihood and Bayesian analyses all require a model to be selected before building a tree (Posada and Crandall, 2001). A maximum of ten parameters can be incorporated into a nucleotide substitution model – the frequency of each of the four nucleotides and the probability of each type of mutation (assuming mutation rates between two bases are equal in both directions) (Posada and Crandall, 2001). Parameter-rich models take longer to run and require more computational power (Posada and Crandall, 2001). Each parameter which is estimated introduces error, so it is ideal to use the model which incorporates as much complexity as needed but no more (Posada and Crandall, 2001).

The Jukes-Cantor (JC) model (Jukes and Cantor, 1969) is the simplest, assuming equal frequencies of all bases and equal probability of all mutations. This model is often used as it is less computationally demanding than more complex models and it is sufficient for many datasets. However, in general all bases and mutations are not equally probable. For example, for ERV sequences it is known that guanine to adenine mutations can be induced by the APOBEC family of restriction factors (section 1.2.3.1), so may be present in excess and Zsíros et al. (1999) found this was the case in HERV-K-like ERVs. Phylogenetic trees based on the chimpanzee genome screen were generated under the JC model (Polavarapu et al., 2006a).

The Kimura (K80) model, which incorporates differences in the rates of transitions and transversions, is also commonly used, for example by Andersson et al. (1998) in characterising ERV-3 and HERV-E like sequences. These simple models may be insufficient to represent complex retroviral datasets (Posada and Crandall, 2001). Posada and Crandall (2001) tested 56 models on various HIV-1 datasets, covering different genes and taxonomic levels. The JC and K80 models were never optimal and different models were optimal for different datasets (Posada and Crandall, 2001).

The RtRev amino acid substitution matrix was developed based on retroviral reverse transcriptase genes and has been shown to be the most appropriate for *pol* gene phylogenetic analysis of amino acid alignments (Dimmic et al., 2002). The JTT model (Jones et al., 1992) was shown in this study to be more appropriate for non-*pol* retroviral data (Dimmic et al., 2002). However, this result may not apply to all lineages or all regions of the *pol* gene and is not applicable to DNA alignments (Posada and Crandall, 2001).

Various statistical tests can be used to select an appropriate model for a particular dataset. One of the most common is the hierarchical likelihood ratio test (Posada and Crandall, 2001). Using this technique, a simple neighbour-joining Jukes-Cantor tree is constructed for the dataset and likelihood scores are calculated for each model, then each level is compared to the previous level of complexity using a likelihood ratio test (Posada and Crandall, 2001). The most complex test which is significantly more likely than the previous level of complexity is selected as the optimum (Posada and Crandall, 2001).

2.3.2.3. Tree Building Algorithm

Neighbour-joining techniques are most commonly used to build retroviral trees, as they are less computationally demanding and allow trees to be built quickly. Sequences are used to generate a distance matrix showing the estimated number of changes between each pair of sequences and a tree is

generated based on this matrix (Holder and Lewis, 2003). This is similar to the clustering technique used here to group related ERVs within a genome (section 2.2.2.2). When multiple substitutions occur at the same site, evolutionary distance will be underestimated, so a correction for multiple hits needs to be incorporated (Holder and Lewis, 2003). This technique is very fast and works well for simple datasets, but is less effective with complex relationships and older, more degraded sequences (Holder and Lewis, 2003). Therefore it is more appropriate for determining relationships between related exogenous retroviral sequences within genera and less appropriate for distant relationships and older, degraded endogenous sequences. Trees representing multiple genera or more than one type of repeat element, such as Han et al. (2007) and van der Kuyl et al. (2011), are often produced using this method because of the computational power required for more complex methods. Neighbour joining trees are also often used as a starting point for more complex methods (Holder and Lewis, 2003).

Maximum parsimony analysis generates trees based on the route through which the least mutations need to have occurred (Holder and Lewis, 2003). However, there is often more than one equally parsimonious route to the same tree (Holder and Lewis, 2003). Trees which can be generated through more pathways are more likely, but this is not recognised by the maximum parsimony technique (Holder and Lewis, 2003). Maximum parsimony analysis is also susceptible to “long branch attraction”, with unusually long branches tending to group together regardless of their relatedness (Holder and Lewis, 2003). Parsimony analysis is more useful when there has not been a large amount of convergent evolution (Holder and Lewis, 2003). Bénit et al. (2001) found very similar results for retroviral data using maximum parsimony and neighbour joining techniques. Maximum parsimony analysis was also used in identifying the feline ERV families located by Pontius et al. (2007).

Maximum likelihood analysis generates the tree with the highest probability of producing the observed data (Holder and Lewis, 2003). The substitution

model provides the probability of each sequence change and these are combined to generate the probability of a particular sequence (Holder and Lewis, 2003). Multiple substitutions at the same site are corrected for (Holder and Lewis, 2003). This technique is much slower than maximum parsimony or neighbour joining techniques but generally more accurate (Holder and Lewis, 2003). Maximum likelihood analysis is often carried out using PhyML (Guindon and Gascuel, 2003), for example this technique was used by Belshaw et al. (2005b) and Garcia-Etxebarria and Jugo (2010).

Neighbour joining, maximum parsimony and maximum likelihood trees all require bootstrapping to show the strength of the relationships within the tree (Holder and Lewis, 2003). This involves generating “pseudo-replicates” of the data based on sites from the original dataset and repeating the tree-building analysis on these (Holder and Lewis, 2003). The proportion of replicates which contain a particular grouping shows how likely this group would be to recur if more data was collected (Holder and Lewis, 2003). If an inappropriate method of data analysis has been used, the results of bootstrapping can be misleading, as repeating the analysis is likely to lead to the same result (Holder and Lewis, 2003). Bootstrapping is very time-consuming and computationally intensive, so various alternatives have been proposed. One of the most widely used of these is the approximate likelihood ratio test based method proposed by Anisimova et al. (2006). This technique is considerably faster than conventional bootstrapping and is thought to be similarly robust (Anisimova and Gascuel, 2006).

Bayesian analysis is closely related to maximum likelihood analysis, but produces a tree and a measure of the robustness of each group simultaneously (Holder and Lewis, 2003). Bayesian analysis relies on prior and posterior probabilities (Holder and Lewis, 2003). The prior probability of a dataset is the probability of an event before the dataset is taken into account, usually in Bayesian phylogenetics this means that every possible value of a parameter has the same prior probability (Holder and Lewis, 2003). The posterior probability

is proportional to this prior probability multiplied by the likelihood of each parameter value given the data and model (Holder and Lewis, 2003). The optimal tree has the highest posterior probability (Holder and Lewis, 2003). Maximum likelihood involves estimation of parameter values, while Bayesian analysis tests all possible parameter values (Holder and Lewis, 2003). The Markov chain Monte Carlo algorithm is used for this analysis (Altschul et al., 1990).

Although Bayesian analysis is usually more robust than maximum likelihood analysis, the degenerate nature of ERV sequences means that unrealistic levels of time and computational power are often required to resolve a phylogeny, especially for a large number of sequences. PhyML maximum likelihood analysis with aLRT-like branch support was therefore selected as the most appropriate technique which can be realistically used with the ERV datasets generated for this project.

2.3.3. Phylogenetic Analysis of Exonerate Output

2.3.3.1. Phylogenetic Test Datasets

For larger retroviral genera, it was not feasible to incorporate all of the known retroviral sequences of that gene and genus into every phylogenetic analysis. Therefore, basic test datasets were generated for these combinations of genes and genera, namely betaretrovirus *gag*, *pol* and *env*, gammaretrovirus *gag*, *pol* and *env* and spumavirus *pol*. Details of the sequences in these test datasets are provided in Appendix B.5 . These datasets were designed to be combined with more detailed datasets representing known members of the specific group being analysed. This strategy helps ensure that sequences have been assigned to the right group.

2.3.3.2. Model Selection

As a large majority of PARSED_EXO_OUT sequences fell into the groups represented by the test datasets, these datasets were used to select a substitution model for phylogenetic analysis. The datasets were aligned using MAFFT under the L-INS-I model with 1000 iterations. JModelTest version 2.1.6 (Darriba et al., 2012), which tests multiple nucleotide models and chooses the most appropriate for the data, was used on these alignments with a maximum likelihood starting tree, heuristic model filtering and model selection using the Akaike information criterion. Models allowing for a proportion of invariable sites were not tested, as ERV sequences are generally not subject to purifying selection to the same extent as gene sequences so there is no biological reason for sites to be invariable.

For five of the seven datasets, the general time reversible (GTR) model, which allows all ten parameters (the frequency of each nucleotide and the probability of each type of mutation) to vary, was the optimum (Figure 30.A). Figure 30.B and Figure 30.C demonstrate that all ten of these parameters do vary considerably for these datasets. For the remaining two datasets (betaretrovirus *env* and *gag*) the transversion model (TVM), which allows the four nucleotide frequencies and the probability of each transversion to vary but not the probability of each transition, was the optimum, as rates of A<>G and C<>T transitions were almost the same (Figure 30.C). However, the likelihood of the GTR model for these two datasets was almost identical to the likelihood of the TVM model (Figure 30.A) so this model is likely to be adequate to describe these datasets. Therefore, the GTR model was used for all subsequent phylogenetic analysis of all datasets.

Incorporating the gamma shape parameter into the model allows substitution rate to vary between sites. Guindon et al. (2010) propose that a shape parameter of less than 0.7 suggests high rate variation, 0.7 to 1.5 moderate variation and more than 1.5 low variation. Five of the seven datasets tested

here had moderate or high variation (Figure 30.A) and adding this parameter always improved the model, so an optimised shape parameter was incorporated into all subsequent phylogenetic analysis.

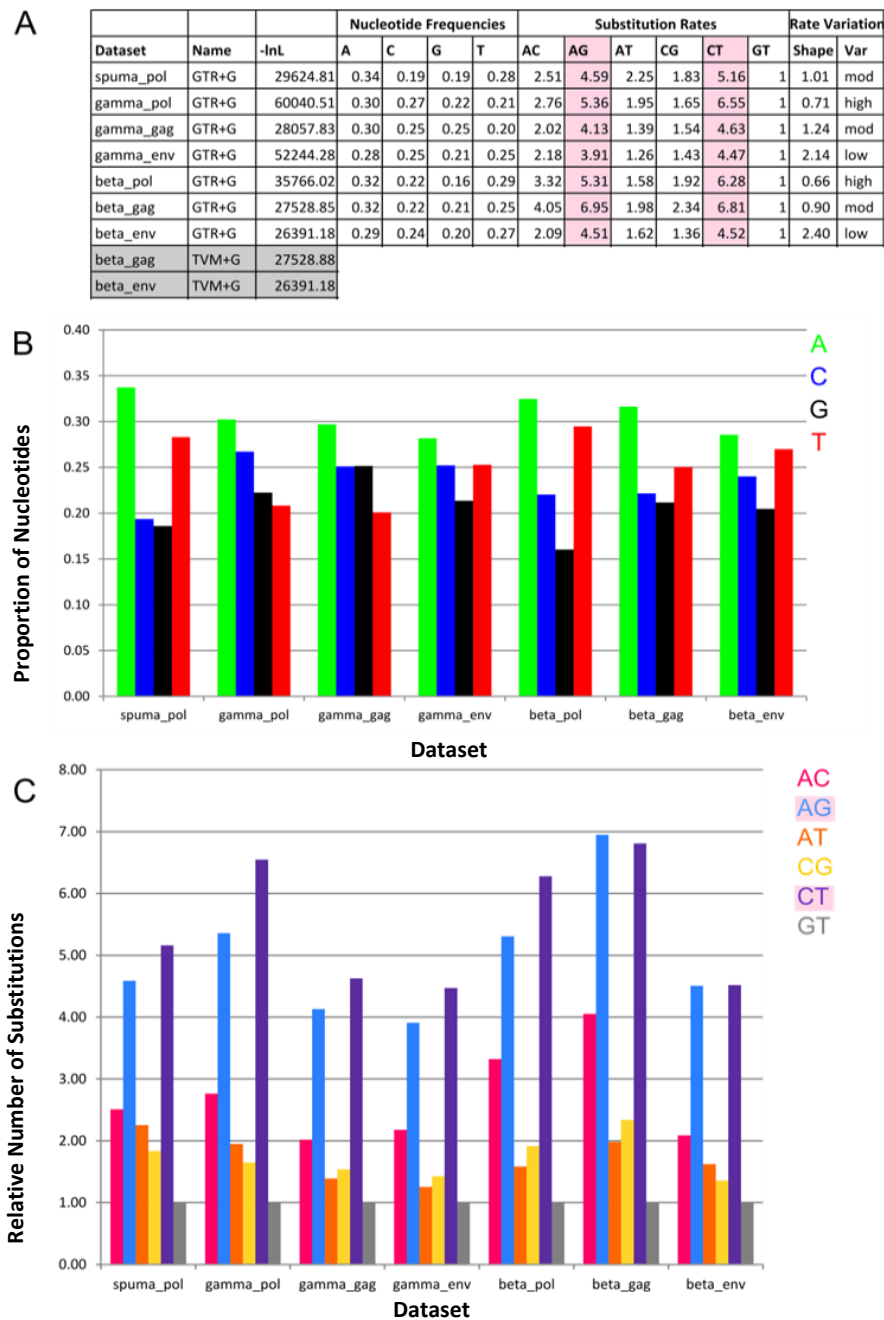


Figure 30: Results of model testing using the JModelTest software.

A) The negative log likelihood (-lnL) of the GTR+G model for each of the phylogenetic test datasets, the frequency of each nucleotide, the relative number of substitutions of each type, the calculated gamma shape parameter and the degree of variation this indicates according to Guindon et al. (2010). For the beta_gag and beta_env datasets the -lnL for the TVM+G model is also provided for comparison. B) Bar chart of relative nucleotide frequencies for each dataset. C) Bar chart of the frequency of each type of substitution. Transitions are highlighted in pink.

2.3.3.3. Alignment and Phylogenetic Analysis

Sequences were selected for phylogenetic analysis based on their categorisation in the GROUPED_EXO dataset.

Selected sequences were aligned with the appropriate test dataset (section 2.3.3.1) and sequences from the GROUPED_PREVKNOWN dataset corresponding with their group in the GROUPED_EXO dataset. Alignments were generated using the L-INS-I model in MAFFT, with 1000 iterations (unless otherwise specified).

All phylogenetic trees were constructed using PhyML (Guindon and Gascuel, 2003). Trees were built under the GTR model with no invariable sites and optimised across site variation. NNI tree-searching, optimised across site variation, a BioNJ starting tree and optimised tree topology were selected. Branch support was calculated using the aLRT method (Guindon and Gascuel, 2003).

2. 4. Characterisation of ERVs

The next stage in this analysis of the Exonerate output dataset was to characterise groups of ERVs of interest in more detail. To do this, several analyses were performed, as appropriate.

Phylogenetic analysis was used to determine if a particular lineage of ERVs is present or absent in a particular host (section 2.4.1) and how many copies are present (section 2.4.2). Another important feature of an ERV is whether it has retained the potential to produce functional viral particles. To produce these particles, the ERV needs a conserved LTR-*gag-pol-env*-LTR structure. The methodology used to identify regions with this structure is described in sections 2.4.3 and 2.4.4. Intact ORFs are needed to produce functional viral proteins, so these need to be detected in candidate regions, as described in section 2.4.5. Finally, the age of the ERV is important in determining its

evolutionary history. Several techniques were combined to approximate ERV age, these are described in section 2.4.6.

2.4.1. Determining Presence or Absence

In some cases, it was necessary to determine if a particular ERV was present or absent in each host. To do this, the EMBOSS water score for every candidate ERV sequence of a particular gene and genus against the ERV of interest was determined as described in section 2.1.4.1. The sequence with the highest score in each host was selected and these sequences for all hosts combined into a FASTA file. The ERV of interest and closely related retroviruses were appended to this file. These sequences were aligned and a PhyML phylogeny constructed. If the highest scoring sequence from a particular host clustered more closely with the ERV of interest than with any other retrovirus it was considered to be present in that host. If the highest scoring sequence clustered elsewhere in the phylogeny it was considered to be absent.

2.4.2. Determining Copy Number

Copy number was taken to be the number of Exonerate fragments in a particular host with a higher EMBOSS water alignment score against the ERV of interest than against any other retrovirus in the PARSED_UT_PREVKNOWN dataset for a particular gene and genus.

2.4.3. ERVs with Multiple Genes

To detect ERVs with more than one recognisable gene, sequences in the PARSED_EXO_OUT dataset were mapped back to their original chromosome positions. A map was constructed showing all the putative ERV fragments on each chromosome from each host, sorted by start position. The Exonerate pipeline occasionally detects two fragments of the same gene close together, these are likely to represent different parts of the same gene. Therefore, where

two sequences from the same gene were found separated by less than 1,500 bp they were merged to give a sequence encompassing both fragments.

The Python script “classify_sets.py” was used to identify and characterise regions of the genome containing fragments of more than one retroviral gene. This script is available in Appendix D.5. This program uses the chromosome maps described above to identify regions where the end of one ERV gene fragment falls within 5,000 bp of the beginning of another gene fragment. This distance was selected because the maximum length of a retroviral gene is approximately 3,000 bp and the threshold length for Exonerate hits used was 600 bp. Therefore, a 5,000 base pair gap should allow for hits from opposite ends of two neighbouring genes. A longer gap length increases the probability of ERV fragments resulting from different integration events being detected.

The output from this program was parsed to generate a table showing “ERV regions” in each chromosome - regions containing one or more ERV-like fragments – and the start and end positions of these regions.

2.4.4. LTRs

To characterise an intact retrovirus, LTR sequences also need to be identified. To isolate LTRs, 8,000 bp on either side of the identified ERV region were extracted from the original chromosome FASTA file. This distance was selected to encompass the maximum distance between the end of the retroviral gene detected and the end of the LTR.

Pairs from either side of each ERV region were aligned to each other using the Smith-Waterman algorithm (Smith and Waterman, 1981) via EMBOSS water (Rice et al., 2000). This software provides the co-ordinates of the section within the aligned sequences which aligns most closely, a FASTA file of these regions and an alignment quality score. The highest scoring sections of each pair of sequences were isolated. Sections which shared 75% sequence identity, were between 6,000 and 15,000 bp apart and between 300 and 1,500 bp in

length were considered to be candidate LTRs. These thresholds are based on the range of retroviral genome sizes and LTR lengths listed in Bannert and Kurth (2010). These candidate regions were classified using CENSOR and regions classified as ERV LTRs were considered to be LTR sequences (Jurka et al., 1996).

2.4.5. Identification of Open Reading Frames

To identify ORFs, the full span of the ERV was extracted from the original chromosome FASTA file. These regions were translated in all six reading frames using the seqinr package in R (R Core Team, 2014). The longest distance between two stop codons was recorded. Where this distance was long enough to potentially encode a full-length protein, the translated sequences were examined to identify further long ORFs which could represent the other genes. Each gene was confirmed using a BLASTP search against the UniProt database.

2.4.6. Determining Age

Several types of analysis can be performed to establish the approximate age of an ERV insertion.

2.4.6.1. LTRs and Degeneration

When an ERV has integrated into a host, assuming no selection, it will accumulate mutations at the host mutation rate. As the LTRs flanking an ERV are identical at integration, the number of differences which have been accumulated between these LTRs can be used to approximate an integration date, assuming the host mutation rate is known (Bannert and Kurth, 2006). This date is calculated using the equation $t = k/2N$, where t is time, k is divergence (number of sites at which the LTRs differ over LTR alignment length), and N is the neutral substitution rate of the host, assumed here to be

the human neutral substitution rate of 4.5×10^{-9} substitutions per site per year (Gifford et al., 2008). This is a common ERV dating technique [used for example in (Sinzelle et al., 2011), (Polavarapu et al., 2006a), (Gifford et al., 2008)].

The general degree of degeneration of an ERV also gives some clue as to its age. ERVs with identical LTRs, intact ORFs and with a recognisable structure, lacking large gaps between detectable gene fragments are likely to be more modern. However, as well as age, the degree to which an ERV is intact depends on selection acting to preserve or prevent gene function, which is variable between hosts and depends on the replication strategy of the ERV (Bannert and Kurth, 2006) (section 1.3.1.2).

2.4.6.2. Host Tracking and Locus-by-Locus Analysis

If an ERV entered the germline of the common ancestor of two hosts before the hosts diverged from each other, at this locus the divergence between the ERVs should parallel the divergence between the hosts, so the phylogeny of the hosts and the ERVs should be similar. This relationship becomes less clear if the ERV is present at multiple loci, as integration events at different sites may not have occurred at the same time.

In general, if an ERV entered the ancestor of two hosts and proliferated in this ancestor before host divergence, we would expect to see a similar copy number of the ERV in the two hosts and the evolutionary distance between the ERVs to be consistent with the evolutionary distance between the hosts. All or almost all integration sites will be shared between hosts. If the ERV entered the common ancestor of the hosts but also proliferated after their divergence later then there may be host tracking at some sites but not others and copy number will vary widely between hosts. Some integration sites will be shared and others will not. If the ERV entered both hosts after they diverged, the copy

number will be variable, host tracking is unlikely to be evident and no integration sites will be shared.

2.4.6.3. Locus-by-locus Analysis

It is not always feasible to trace if the locus at which an ERV appears is orthologous between hosts, as the data required is only available for some hosts, is variable in quality and the analysis requires a large amount of computational power. However, the Compara EPO six primate alignment (C6P) (Ensembl release 74), an alignment of the DNA sequence of human, chimpanzee, gorilla, orangutan, rhesus macaque and marmoset genomes, provides information which can be useful for these hosts.

Where this alignment was used, the positions of the ERVs were identified in the alignment and the corresponding positions from other genomes extracted. The positions in this alignment provide a fairly large orthologous region (usually approximately 50,000 bp) rather than an exact position. For detailed analyses (used for the epsilon-like ERVs in o) the exact position of the ERV was detected in the alignment and the sequence for each host extracted from this position. Sequences from this region were then aligned to the original ERV sequence and if there was at least 75% sequence identity between the ERV sequence and the sequence of any host within the ERV region, excluding gaps, the ERV was considered to be present in this host. Sequences from all hosts at each locus were aligned and PHYML phylogenetic trees were built for each locus.

For larger groups, such as HERV-K-like ERVs (section 4.6.1) all ERV-like sequences in PARSED_EXO_OUT from the larger region identified as orthologous to the region containing the ERV were extracted from all hosts. These sequences were aligned and phylogenetic trees built and any ERVs with a tree topology exactly matching that of the hosts were assumed to be the result of inheritance from a common ancestor.

2.4.7. Identifying Selection

A common technique to identify selection is to compare the number of synonymous substitutions (substitutions not affecting the amino acid sequence) (K_s) and non-synonymous substitutions (substitutions changing the amino acid sequence) (K_A) between two nucleotide sequences. An excess of synonymous mutations suggests that there is selection against change, or purifying selection, meaning the sequence is likely to be beneficial to the host. In this case the K_A/K_s ratio will be less than one. An excess of non-synonymous mutations suggests selection for change, or positive selection, which is often the result of an antagonistic interaction between a virus and its host (Sawyer et al., 2005). In this case the K_A/K_s ratio will be greater than one. This method has been used to detect selection in several previous studies into ERVs [e.g. (Dupressoir et al., 2005, Sawyer et al., 2005, Carre-Eusebe et al., 2009)]. Here, this analysis was performed using the software package DNASP version 5.10.01 (Rozas and Rozas, 1995). This type of analysis requires a very precise alignment with a reliable translation, so it was not appropriate for all ERVs.

2. 5. Host Phylogeny

A phylogenetic method which is appropriate for all Euarchontoglires and allows relationships to be established from species to order level was required to build a robust phylogenetic tree of the Euarchontoglires species screened here. This tree is needed for comparison of the phylogenetic relationships among the ERVs with the phylogenetic relationships among the hosts described in section 2.4.6.2.

Sixteen genes were selected for this purpose from Hovarth et al. (2008). These genes are widely sequenced and have been shown to be appropriate for family to species level primate phylogenetics (Horvath et al., 2008). Genes are listed in Table 10. Appropriate gene fragments were identified using sequences from Hovarth et al. (2008) as probes for a BLASTN search against the nr database,

limited to sequences from appropriate members of the Euarchontoglires. One sequence each was taken from each species with a significant hit covering the majority of the query site. Accession numbers for all hosts for the genes in Table 7 are available in Appendix B.6. Sequences for each gene were aligned using MAFFT under the L-INS-I model with 1000 iterations. Aligned sequences for all genes were combined into a single matrix, with missing segments replaced by gaps. A phylogenetic tree was built to represent this alignment using PhyML (Guindon and Gascuel, 2003), under the settings described above (section 2.3.3).

Table 10: Genes used for host phylogeny.

Gene Abbreviation	Gene Description
ABC1	ATP binding cassette protein 1
ADORA3	adenosine A3 receptor
AXIN1	axin 1
CFTR	cystic fibrosis transmembrane conductance regulator
ERC2	ELKS/RAB6-interacting/CAST family member 2
FRMD5	FERM domain containing 5
FGA	fibrinogen alpha chain
LLPPRC	leucine-rich pentatricopeptide repeat containing
LUC7L	LUC7-like
SLC11A1	solute carrier family 11
RAG1	recombination activating gene 1
RBP3	retinol binding protein 3
TTR	transthyretin
VWF	von Willebrand factor
ZNF202	zinc finger protein 202

Chapter 3. Overview of Results

This chapter provides an overview of the output from the methodology at each stage described in Chapter 2.

3. 1. Raw Output and Quality Control

Using the Exonerate pipeline, a total of 190,196 partial ERV gene sequences were identified across the 33 genomes screened. Of these, 40,627 were *gag*-like, 124,187 were *pol* like and 25,382 were *env* like. The mean length of these fragments was 1049.53 bp. This figure refers to individual gene fragments, so separate *gag*, *pol* and *env* fragments were often counted at single ERV loci. This is the RAW_EXO_OUT dataset (Figure 31). This dataset is available as a FASTA file in Appendix C.4 and details of each sequence are provided as an MS Excel spreadsheet in Appendix B.7.

A quality control step was carried out (as described in section 2.2.1), removing sequences which could not be verified by a highly significant match in a BLASTX search against the PARSED_UT_PREVKNOWN dataset. After this verification, 169,424 sequences remained. Of these, 35,223 were *gag* like, 111,711 were *pol* like and 22,490 were *env* like. This is the PARSED_EXO_OUT dataset, available as a FASTA file in Appendix C.5 and described in Appendix B.7 . Figure 31 shows the number of fragments which did not meet the quality control threshold for each gene. The proportion of sequences which were parsed out at this stage was approximately the same for each gene.

In 21,929 cases, it appeared that two partial gene sequences identified were likely to be from different regions of the same integrated gene, as two fragments of the same gene were identified less than 5,000 base pairs apart. These fragments are marked in Appendix B.7. The vast majority of these cases

(20,532) were *pol* genes, as these are the longest ERV genes and many sequences representing different segments of the *pol* gene were included in the Exonerate input (Figure 31). Taking into account these fragments, 147,496 may be a more accurate assessment of the total number of ERV genes (not ERV loci) detected across all genomes. Of these potential genes, 34,629 were *gag*, 91,179 *pol* and 20,532 *env* (Figure 31). The number of unique, BLAST-verified *gag*, *pol* and *env* genes identified is compared to the number of fragments in the raw Exonerate output in Figure 31.

Count data from this point forward refers to the number of gene fragments identified and, for the *pol* gene, is an average of 13% higher than the number of unique genes identified. Fragments of the same gene were merged when looking for intact loci.

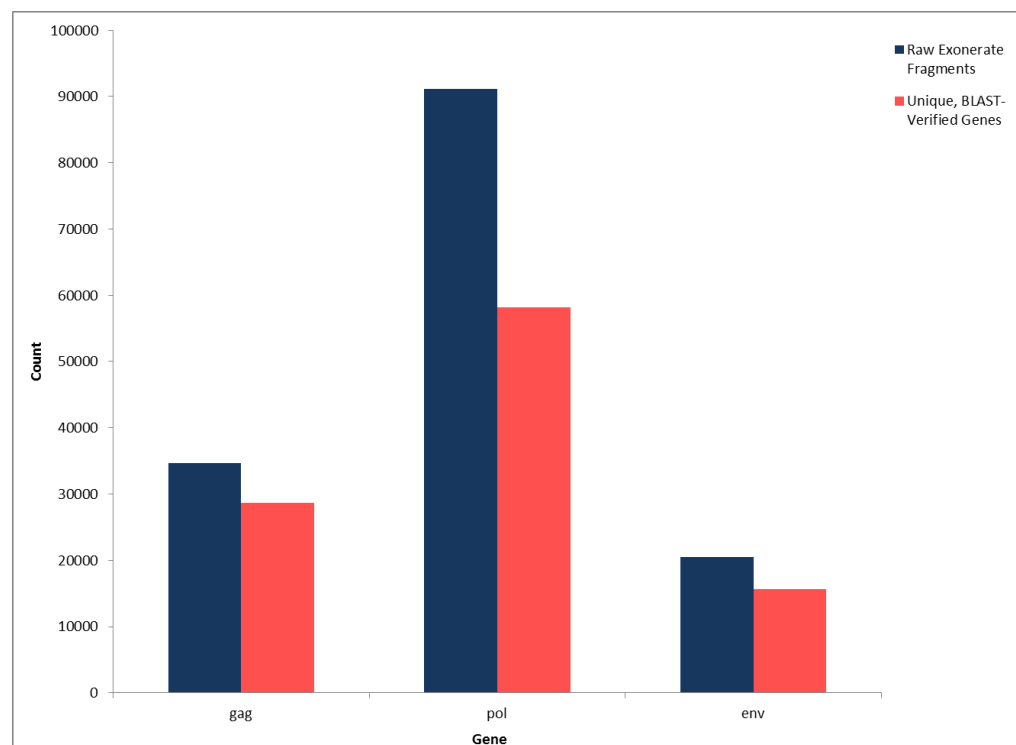


Figure 31: Graph showing the total number of candidate *gag*, *pol* and *env* ERV fragments identified and the proportions of these fragments which were both verified by BLAST and appeared to represent unique genes.

The number of fragments of each gene in the raw Exonerate output are shown in blue and the number of these fragments which passed the BLAST quality control step and were not duplicates representing different parts of the same gene are shown in pink.

The ERV fragments identified are not evenly distributed between the different host genomes (Figure 32). Noticeably more fragments were identified in several hosts, particularly the mouse, guinea pig and tarsier. These species have no particular phylogenetic or geographical connection. Fewer ERV fragments were detected in avian genomes than in those of the Euarchontoglires. The number of fragments meeting the quality control threshold ranged from 71.7% (vole) to 95.1% (mouse). Primates tended to have more fragments in this category than rodents.

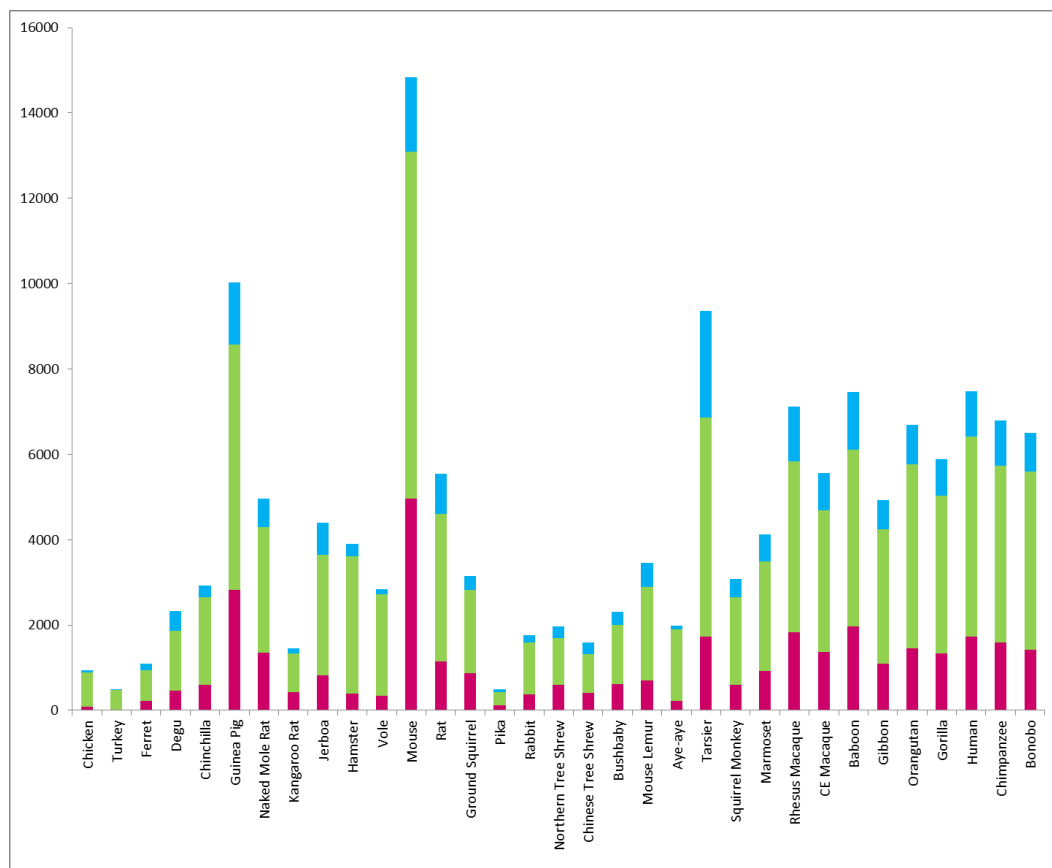


Figure 32: Graph showing the number of unique, BLAST-verified *gag*, *pol* and *env* genes identified using Exonerate in each host.
gag is represented in red, *pol* in green and *env* in blue.

Various metrics about each host genome (total length, total assembly gap length, number of scaffolds, scaffold N50, number of contigs, contig N50) were downloaded from NCBI Assembly (<http://www.ncbi.nlm.nih.gov/assembly>). These were then compared to the number of fragments identified using

Exonerate and the mean length of these fragments for each host to establish the extent to which genome size and quality affect the output of the Exonerate pipeline (Table 11). Pearson's correlation coefficient was calculated for each genome statistic compared to number of fragments in PARSED_EXO_OUT and mean length of fragments in PARSED_EXO_OUT for each Euarchontoglires host. Only one relationship showed a statistically significant ($p < 0.05$) correlation, a moderate positive correlation ($r = 0.444$) between number of scaffolds and number of fragments identified. This relationship is shown in Figure 33 and suggests that more fragments were identified in the species with less assembled genomes but that they were not significantly shorter than the fragments identified in other genomes. However, this relationship is completely dependent upon the tarsier genome, which has a very large number of scaffolds and contained a very large number of fragments. Excluding this unusual value, there is no significant correlation between these variables and the slope of the graph is very different ($r = -0.30$, $p = 0.137$) (Figure 33).

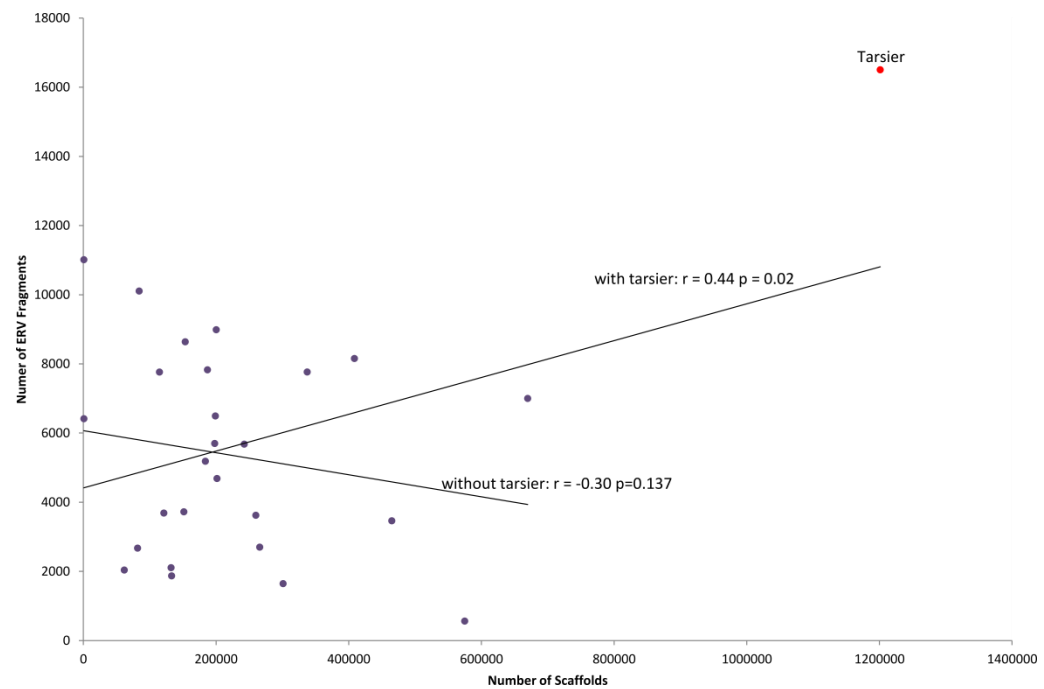


Figure 33: The relationship between number of scaffolds and number of ERV fragments identified using Exonerate, including and excluding results from the tarsier genome.

r is Pearson's correlation coefficient.

Table 11: Statistical comparison between the number of fragments and mean fragment length for each host in PARSED_EXO_OUT and various genome metrics. Pearson refers to Pearson's correlation coefficient (r).

Genome Statistic	Number of Fragments		Mean Fragment Length	
	Pearson	P-value	Pearson	P-value
Total sequence length	0.335	0.071	-0.024	0.905
Total assembly gap length	0.154	0.435	-0.026	0.897
Number of scaffolds	0.444	0.020	-0.141	0.482
Scaffold N50	0.133	0.507	0.254	0.201
Number of contigs	0.059	0.756	-0.085	0.675
Contig N50	0.264	0.159	0.002	0.993

3. 2. Clustering

The PARSED_EXO_OUT dataset was clustered into groups and consensus sequences built to represent these groups (section 2.2.2). In total, across all genomes, 17,185 clusters were identified, representing 138,713 sequences. The remaining 30,711 sequences were left in the dataset individually, so the total size of the CLU_EXO_OUT dataset is 47,896 sequences, 28.3% of the size of PARSED_EXO_OUT. Groups ranged in size from two to 1,374 sequences but large groups were relatively uncommon, with a mean group size of 8.07 sequences and a median group size of four sequences. Consensus sequences were named as “*prefix_gene_genus_nseqs_seqs*” where *nseqs* is the number of sequences represented by the consensus. The CLU_EXO_OUT dataset is available as a FASTA file in Appendix C.6 and the sequences which make up each consensus are listed in Appendix B.7.

The majority of fragments were represented by a consensus which is a good reflection of their original sequence. 6,018 of the sequences in PARSED_EXO_OUT were identical to the consensus sequence representing their group. The remaining sequences had a mean identity to their representative consensus of 94.7%, with the distribution shown in Figure 34.

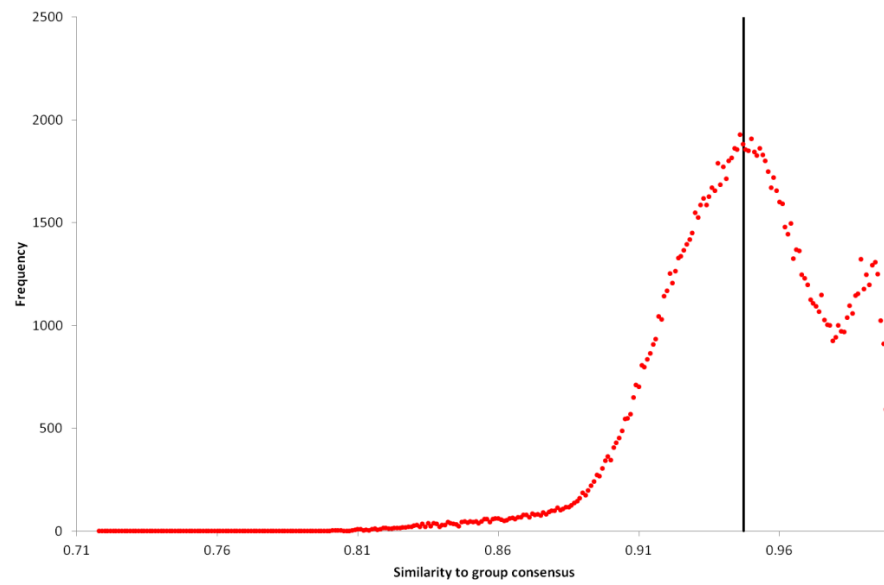


Figure 34: The distribution of the similarity of sequences in PARSED_EXO_OUT to the group consensus sequence by which they were represented in CLU_EXO_OUT, in cases where the group consensus and the original sequence are not identical.

3. 3. Intact ERVs

To identify potentially intact ERVs, regions resembling more than one ERV gene less than 5,000 bp apart were isolated. Across the 33 genomes, one or more genes were identified in 113,395 regions. Of these, 6,348 contained all three genes (*gag*, *pol* and *env*), 24,842 contained two of these genes and the remaining 82,205 contained only *gag*, *pol* or *env*. Details of these regions are provided in Appendix B.8. There are four potential locus types for regions with more than one gene, *gag-pol*, *pol-env*, *gag-env* and *gag-pol-env*. The ratio of these types across all genomes is shown in Figure 35 and the number of *gag-pol-env* insertions in each genome is shown in

Figure 36. *Gag-pol* insertions were the most common and *gag-env* insertions relatively rare (Figure 35). In general, genomes assembled into chromosomes or with fewer contigs or scaffolds contained a higher number of more intact insertions. It was rare to find more than 200 *gag-pol-env* insertions in a genome not assembled at the chromosome level, with the exception of the bonobo, guinea pig and naked mole rat. The majority of scaffold-assembled genomes had 10-200 *gag-pol-env* insertions. Only the aye-aye, which has a very large number of contigs, had less than 10 *gag-pol-env* insertions.

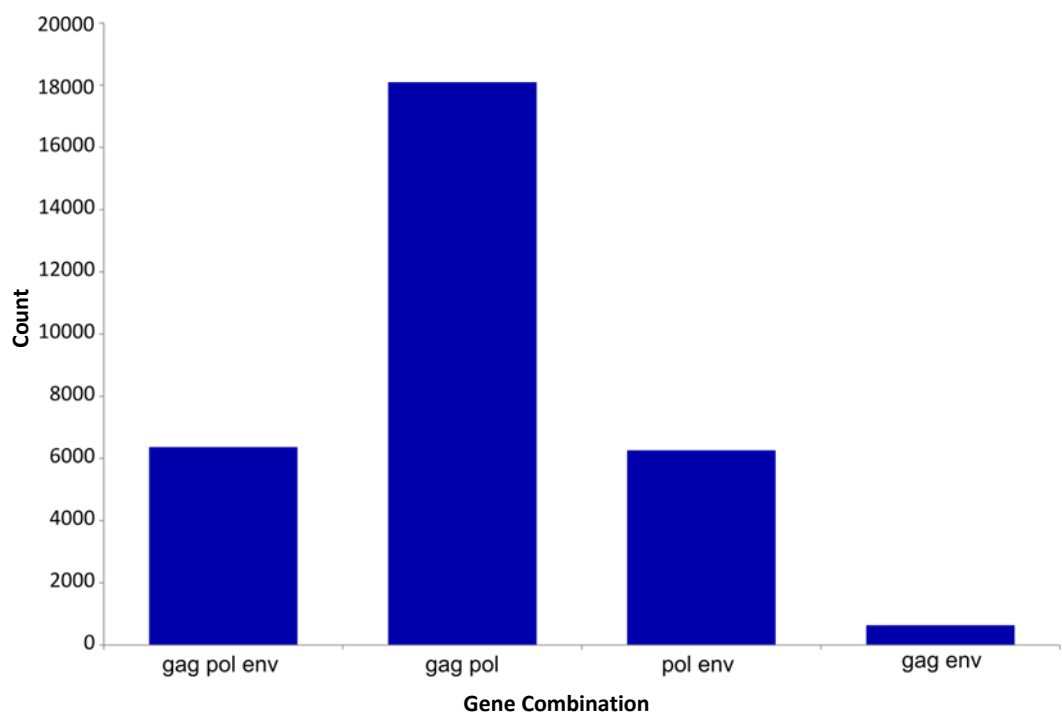


Figure 35: The number of regions with each possible combination of multiple ERV gene fragments across all hosts.

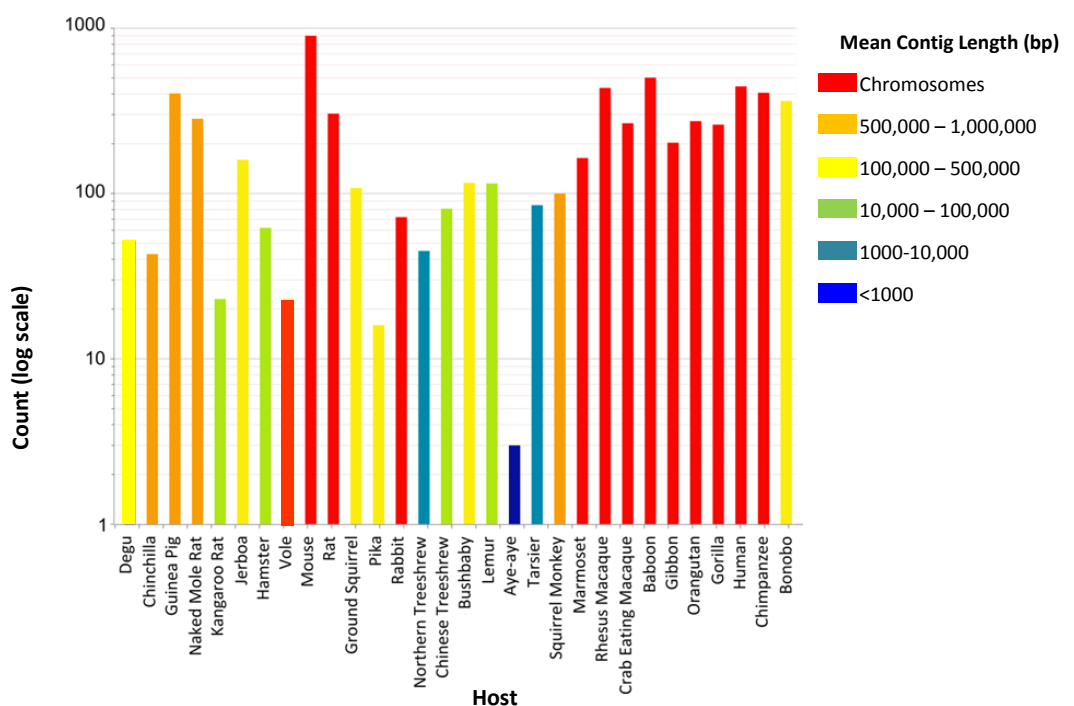


Figure 36: The number of *gag-pol-env* regions identified in each host.

Bars are coloured according to the number of contigs in the genome, genomes with the most contigs are blue and genomes assembled into chromosomes are red.

3. 4. Host Phylogeny

A host phylogeny for the sequenced genomes was generated using the methodology outlined in section 2. 5 for comparison with the phylogeny of the various groups of retroviruses in these genomes and is shown in Figure 37. The relationships in this tree are consistent with the literature (Perelman et al., 2011, Arnold et al., 2010, Blanga-Kanfi et al., 2009, Meredith et al., 2011).

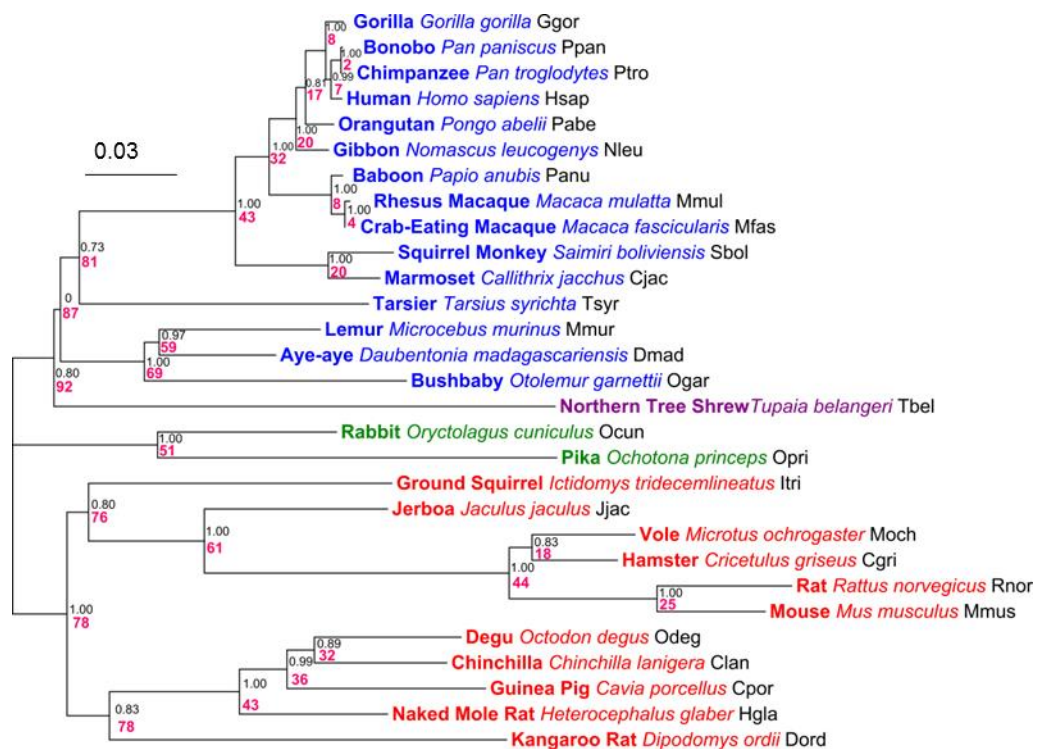


Figure 37: Phylogenetic tree based on 15 nuclear genes showing the relationships between the sequences of primates, rodents, lagomorphs and tree shrews. Rodents are shown in red, primates in blue, tree shrews in purple and lagomorphs in green. Black text tip labels represent abbreviated host names. Black node labels denote branch support, pink node labels approximate number of years since divergence [based on Perelman et al. (2011) and Hedges et al. (2006)].

Chapter 4. Genus-by-genus Analysis

This chapter provides an in-depth analysis of the ERVs identified in each genus.

4. 1. Overview

Sequences were assigned to a genus according to the sequence in PARSED_UT_PREVKNOWN with which they formed the highest scoring alignment. The distribution of sequences between the six genera (no deltaretroviruses were identified) for each gene is shown in Figure 38. Gammaretroviruses were always the most common, followed by betaretroviruses. A significant number of spuma-like *gag* and *pol* regions were identified but very few spuma-like *envs*. This is likely to be because ERV-L elements, the most common endogenous spuma-like insertions, lack *env* (Benit et al., 1999). 0.74% of *pol* gene insertions identified were epsilonretrovirus-like but no *gag* or *env* genes of this type were identified. Alpharetroviruses are generally considered to be avian retroviruses and, as only two of the host species screened were birds, a low overall proportion of alpharetroviruses is to be expected. Endogenous lentiviruses are known to be rare, however a few representatives of each gene were identified.

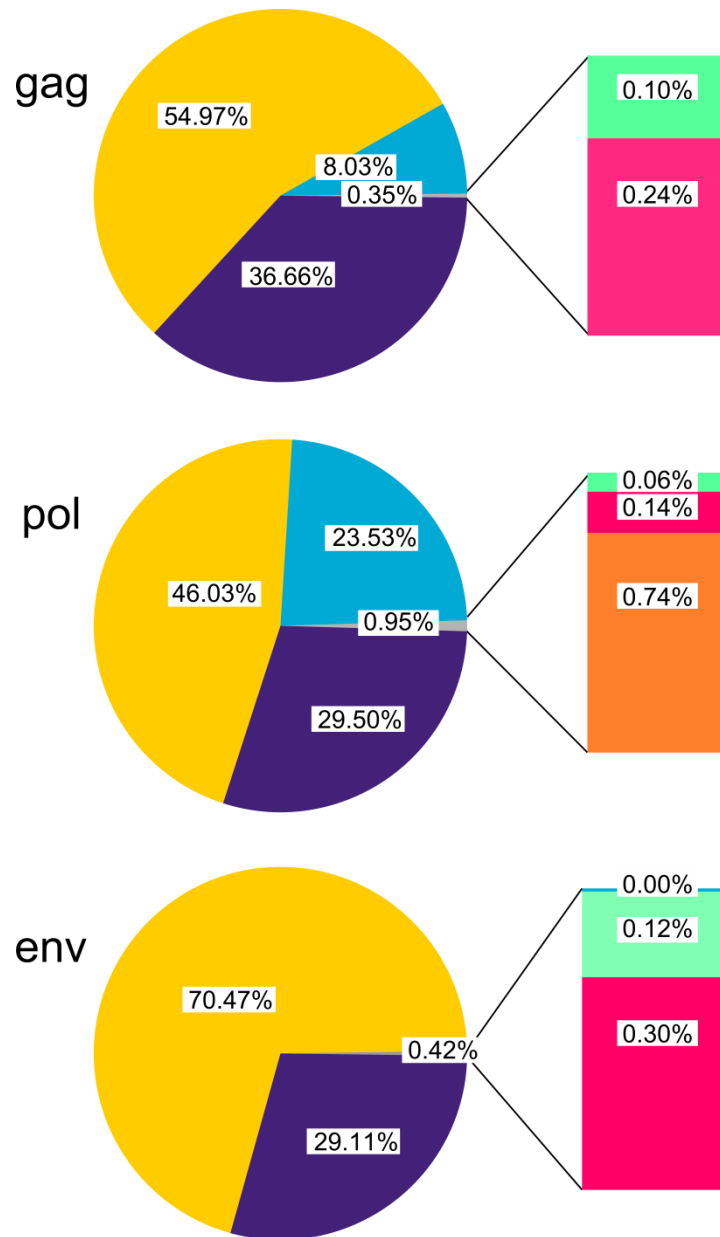


Figure 38: The distribution of *gag*, *pol* and *env* insertions identified in each host between genera.

Gammaretroviruses are shown in yellow, betaretroviruses in purple, spumaviruses in blue, alpharetroviruses in pink, lentiviruses in green and epsilonretroviruses in orange. Rectangular charts are expansions of the smallest segments of the pie charts.

4. 2. Gammaretroviruses

86,628 (19,363 *gag*, 51,417 *pol*, 15,848 *env*) gammaretrovirus like ERV fragments were identified across the 33 hosts screened. These were distributed between hosts as shown in Figure 39.

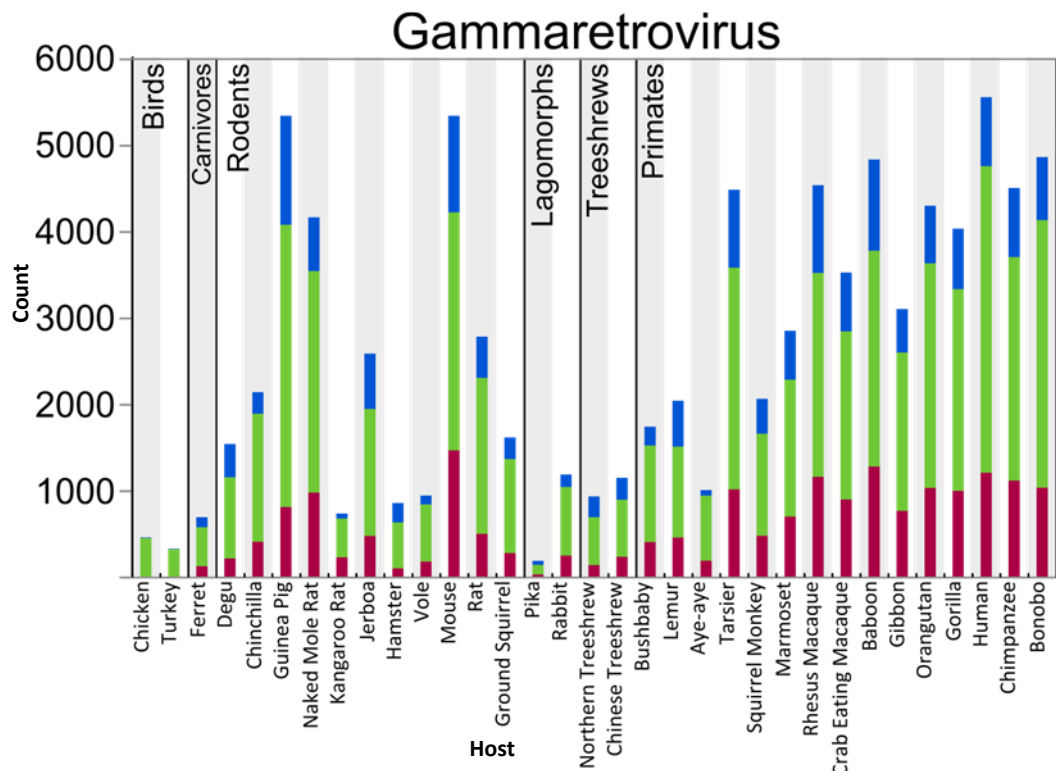


Figure 39: The distribution of ERV fragments between genomes for the gammaretrovirus genus.

gag fragments are represented in red, *pol* in green and *env* in blue.

Hayward et al. (2013a) carried out a similar analysis on 22 of these 33 genomes using Retroector. As demonstrated by Figure 40, the proportion of gammaretrovirus-like ERVs (referred to as Class I ERVs by Hayward et al. but not incorporating the epsilonretroviruses) identified in each genome using these two types of analysis was similar. There were three noticeable exceptions to this. Hayward et al. identified a very large number of insertions in the human and mouse genomes compared to related hosts, while our analysis did

not. Conversely, our pipeline identified a large number of insertions in the tarsier and lemur compared to other hosts, while Hayward et al.'s did not. This corresponds to genome quality, as, based on the genome metrics discussed in section 3.1 the human and mouse genomes are the highest quality sequences, especially in terms of contig N50, for which mouse and human have scores of over 30 million, compared to approximately 80,000 for the next highest scoring genome (NCBI Assembly, 2014). The tarsier and lemur genomes are of the lowest quality based on this metric, with scores of less than 4,000 (NCBI Assembly, 2014). Given this data, the Exonerate pipeline used here seems to be more appropriate for identification of gammaretroviruses in poor quality genomes. The algorithm used by Retrotector is more dependent on an intact ERV structure than the Exonerate algorithm (as discussed in section 2.1.1.2), so this result is not unexpected. Based on our results, the excess of gammaretroviruses in humans compared to other apes and in mice compared to other old world rodents identified by Hayward et al. (2013a) appears to be an artefact of their analysis method, rather than reflecting the true ERV complement of these species.

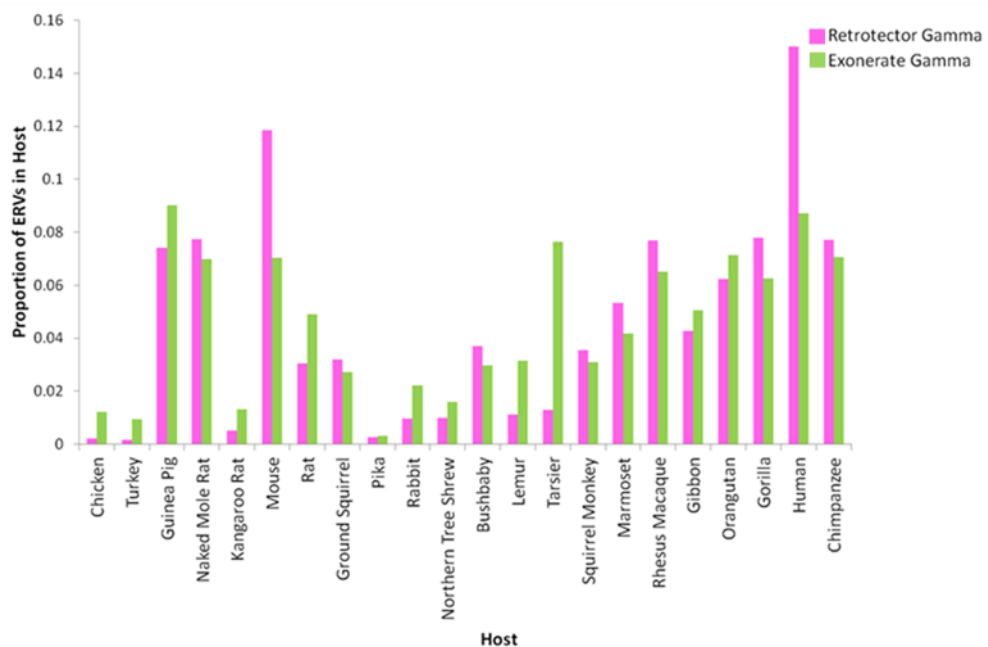


Figure 40: The proportion of gammaretroviruses identified in each host here (green) and by Hayward et al. (2013a) (pink).

Hayward et al. (2013a) divide the gammaretroviruses into five groups based on their phylogenetic analysis. These groups are described in Table 12. In their analysis, a subset of previously known retroviruses fell into each of these groups, along with a large number of newly identified sequences. Primate ERVs were especially abundant in groups II and III and rodent ERVs in group V.

Table 12: The gammaretrovirus groups identified by Hayward et al. (2013a) . Examples are provided of well-known previously classified sequences which were assigned to each group. Counts are the number of ERV regions detected in each type of host in each group.

	ERV Group				
	I	II	III	IV	V
Example Known Sequences	HERV-I HERV-ADP	HERV-W HERV-FRD HERV-Rb	HERV-F HERV-Fc HERV-H	HERV-R HERV-E	MLV FeLV GALV
Host Group	Count				
Bird	28	0	0	0	0
Rodent	2	68	530	137	1931
Lagomorph	2	19	22	1	53
Tree Shrew	0	1	0	26	51
Primate	389	1888	1309	776	681
TOTAL	421	1976	1861	940	2716

We propose a slightly different classification scheme, into the six groups outlined in Figure 41. 99.98% of the gammaretroviral *pol* gene fragments identified here fell into 22 groups in the GROUPED_EXO dataset. A single previously known sequence from each of these groups was analysed phylogenetically and six clusters were identified (Figure 41). A second phylogenetic analysis was performed to clarify relationships in Hayward et al.'s groups II and III (Figure 42). There are several differences between our classification scheme and that of Hayward et al. First, we did not find that

HERV-R and HERV-E form a monophyletic group, so propose splitting this category into HERV-R-like and HERV-E-like groups. Secondly, there was no distinction in our analysis between the ERVs classified as group II and group III by Hayward et al. (Figure 42), instead we propose combining these two groups. Finally, the REV-like group was distinct from the MLV-like group phylogenetically and in terms of host distribution, so division of these two groups may be more representative. The six clusters identified here will provisionally be referred to according to a well-characterised previously known sequence within the cluster: HERV-I like (equivalent to group I), HERV-F-like (groups II and III) HERV-R-like (group IV), HERV-E-like (group IV), REV-like and MLV-like (both parts of group V) (Figure 41).

The majority of groups of gammaretroviral *pol* gene sequences (319/398) identified in the previous screening projects listed in section 2.1.4.1 fell into one of these six groups when characterised by sequence similarity (section 2.4.5). Almost all of the sequences falling outside of these groups were from amphibians or reptiles. Therefore, these six groups represent a large majority of the gammaretroviral diversity in the Euarchontoglires and most likely the mammals. This grouping is much more representative of the gammaretroviruses than the subset of these viruses which is often used in phylogenetic analysis, which tends to be biased towards MLV-like insertions [for example (Elleder et al., 2012, Cui et al., 2012)] and may limit the ability to fully characterise the evolutionary relationships of newly discovered gammaretroviruses.

We have identified novel endogenous gammaretroviruses in each of the six groups outlined in Figure 41, each group will be discussed below.

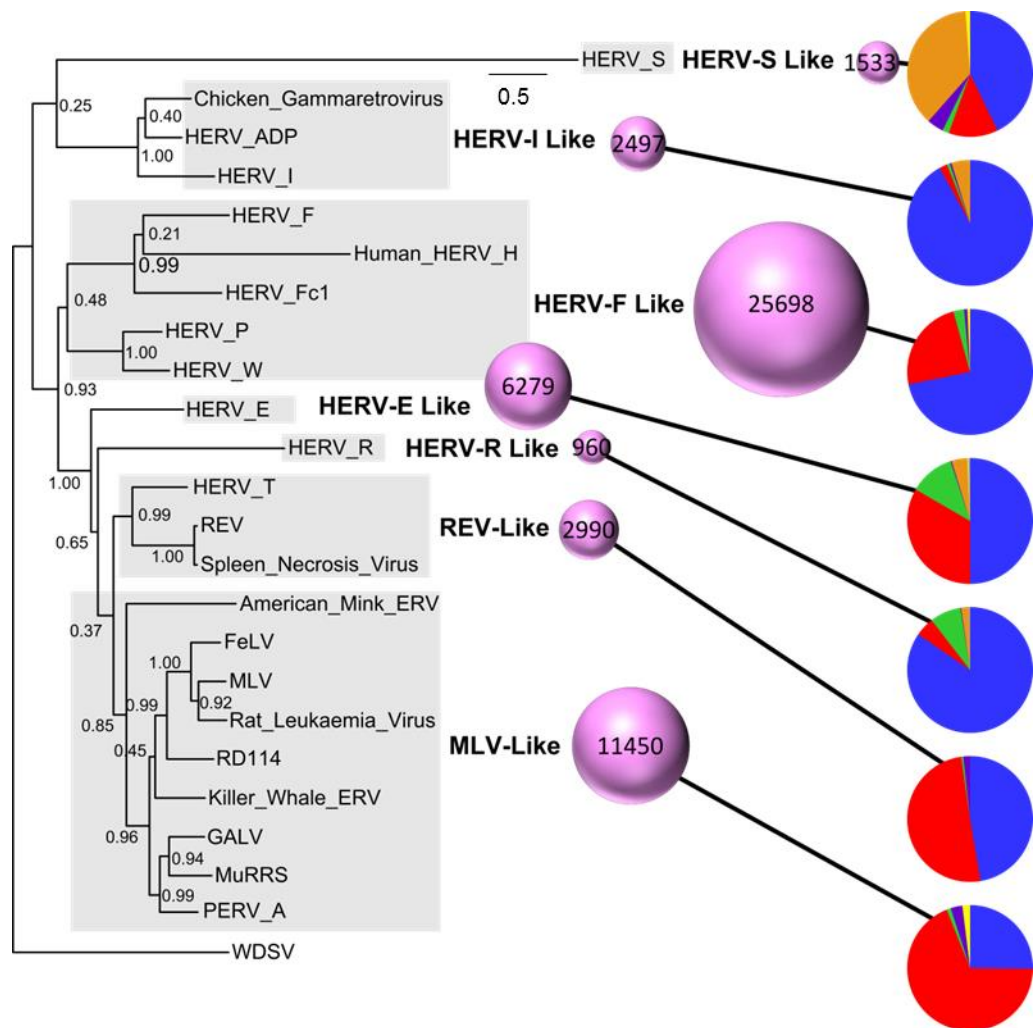


Figure 41: *Pol* gene phylogeny showing the seven proposed groups of gammaretroviruses in the Euarchontoglires.

Groups in the phylogeny are highlighted in grey. Pink circles are sized according to and labelled with the total number of new *pol* gene fragments identified in this group. Pie charts show the proportion of these fragments found in primates (blue), rodents (red), lagomorphs (green), tree shrews (purple), birds (orange) and ferrets (yellow). Details of previously known sequences are provided in Appendix B.2.

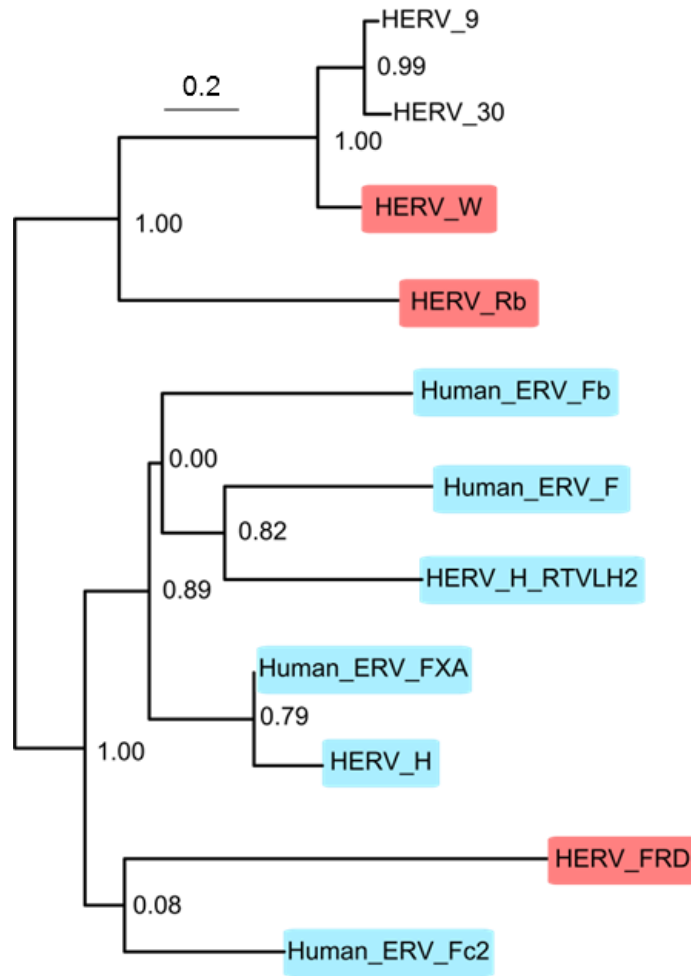


Figure 42: *Pol* gene phylogeny showing the HERV sequences classified as class ii (red) and class iii (blue) by Hayward et al. (2013a).
Details of previously known sequences are provided in Appendix B.2.

4.2.1. HERV-I Group

The HERV-I like group of gammaretroviruses is known to include three groups of HERVs: HERV-I (Maeda and Kim, 1990), HERV-ADP (Lyn et al., 1993) and HERV-IP10 (Seifarth et al., 2000). A chicken gammaretrovirus, ChiRV1 (Borysenko et al., 2008), several other avian ERVs (Niewiadomska and Gifford, 2013) and ERVs from the lemon shark, komodo dragon and wallaby (Martin et al., 1997) are also known to cluster close to HERV-I.

Several other previously characterised sequences fell into this group in the analysis in section 2.4.5. These were a subset of avian gammaretroviruses, characterised by Martin et al. (1999), chimpanzee endogenous gammaretrovirus groups 20 to 28 from Polavarapu et al. (Polavarapu et al., 2006a) and bovine ERV 5 from Garcia-Etxebarria et al. (2010).

Using Exonerate, 4,454 HERV-I like fragments were identified across the 33 genomes screened. The majority of these fragments were identified in primates (4,199 fragments) and birds (140 fragments). Details of the fragments are shown in Table 13.

Table 13: The number of HERV-I-like ERV fragments identified in each host type.
All: all 33 genomes. Blue rows labelled “T” represent total counts, yellow rows labelled “PG” represent the mean number of fragments per genome of this type.

		All	Gag	Pol	Env
All	T	4454	1200	2497	757
	PG	135	36	76	23
Primates	T	4199	1158	2296	745
	PG	280	77	153	50
Rodents	T	72	15	51	6
	PG	7	1	5	1
Lagomorphs	T	25	4	19	2
	PG	13	2	10	1
Tree Shrews	T	17	4	11	2
	PG	9	2	6	1
Birds	T	140	19	119	2
	PG	70	10	60	1
Ferret	T	1	0	1	0
	PG	1	0	1	0

There has previously been some ambiguity about the integration date of the HERV-I lineage into primate genomes. All previous studies have identified HERV-I like insertions in apes and old world monkeys, but Seifarth et al. (2000) and Greenwood et al. (2005) found evidence of HERV-I like insertions in new world monkeys while Lee et al. (2006) did not. Hayward et al. (2013a) did identify insertions in this group in new world monkeys, plus 13 insertions in prosimians.

Here, all apes and old world monkeys had approximately 200 HERV-I like *pol* gene fragments. New world monkeys had somewhat less, an average of 119 per genome. 190 chromosome regions were identified with consecutive HERV-I like *gag*, *pol* and *env* fragments. These were found in all apes, old world monkeys and new world monkeys screened and in no other hosts. Potential LTRs were identified flanking 111 of these regions, again distributed across all

simian hosts. Using the LTR dating approach described in section 2.4.6.1, these LTRs gave a range of potential of integration dates for the HERV-I like group of between three and 61 million years. Only seven loci gave a date greater than 43 million years, the estimated date of the divergence between old and new world primates (Figure 37). This may indicate that HERV-I entered the genomes of new and old world primates before this date but diversified separately in the two host groups.

HERV-I like regions with a *gag-pol-env* structure were screened for ORFs as described in section 2.4.5 to establish if any loci have the potential to produce functional viruses. No loci were found with *gag* or *pol* ORFs long enough to produce functional viral proteins. However, a conserved 662 to 675 amino acid *env* ORF was identified in nine simian primates: human, chimpanzee, bonobo, gorilla, orangutan, crab-eating macaque, rhesus macaque, baboon and marmoset. The positions of these fragments in each genome are listed in Table 14. These ORFs represent the ERV-Pb *env* gene first described by Aagaard et al. (2005), who identified this ORF in all old world primates. Aagaard et al. also found a fragment of this gene in the owl monkey (a new world monkey) but were unable to assign this to a locus or identify the rest of the gene. The position of this locus in the chimpanzee genome was identified in the Compara six primate alignment as described in section 2.4.6.3, which gave co-ordinates for the orthologous position in five other primates: human, gorilla, orang-utan, rhesus macaque and marmoset. These positions corresponded to the positions containing the *env* ORF in all six genomes. Therefore, we have found an unambiguous, full-length copy of this gene in the marmoset, a new world primate, at an orthologous position to the old world primate gene. The nine *env* ORF sequences from this locus were analysed phylogenetically with other known gammaretroviral *env* genes and formed a monophyletic group with a branching pattern identical to the host phylogeny (Figure 37). DNA sequences for this ORF from these nine hosts were aligned and the K_a/K_s ratio calculated to look for evidence of selection (as described in section 2.4.7). The mean value for this ratio was 0.68, with a range from 0.06 to 0.93, which is consistent

with purifying selection. These results combined suggest that this ORF has been conserved and capable of producing active protein for at least 43 million years and that it has been subject to selection in the host to confirm its function. This gene is phylogenetically distinct from all known syncytins (Figure 43) and has been shown to be poorly expressed in all human tissues tested to date (21 tissues) and not overexpressed in the placenta (Blaise et al., 2005). However, the protein has been shown to be fusogenic (Blaise et al., 2005). Further work is needed to establish how this protein benefits its primate hosts.

Table 14: The position of the HERV-I *env* ORF in various simian hosts.

Host	Chromosome	Start Position	End Position	Strand
Baboon	7	148757927	148759949	-
Bonobo	scaffold 1120388623549	9008405	9006426	-
Chimpanzee	14	92184238	92186257	-
Crab-eating Macaque	7	157604564	157606587	-
Gorilla	14	74409054	74411075	-
Human	14	93089235	93091254	-
Marmoset	10	118228238	118230221	-
Orangutan	14	93952896	93954917	-
Rhesus Macaque	7	156359342	156361366	-

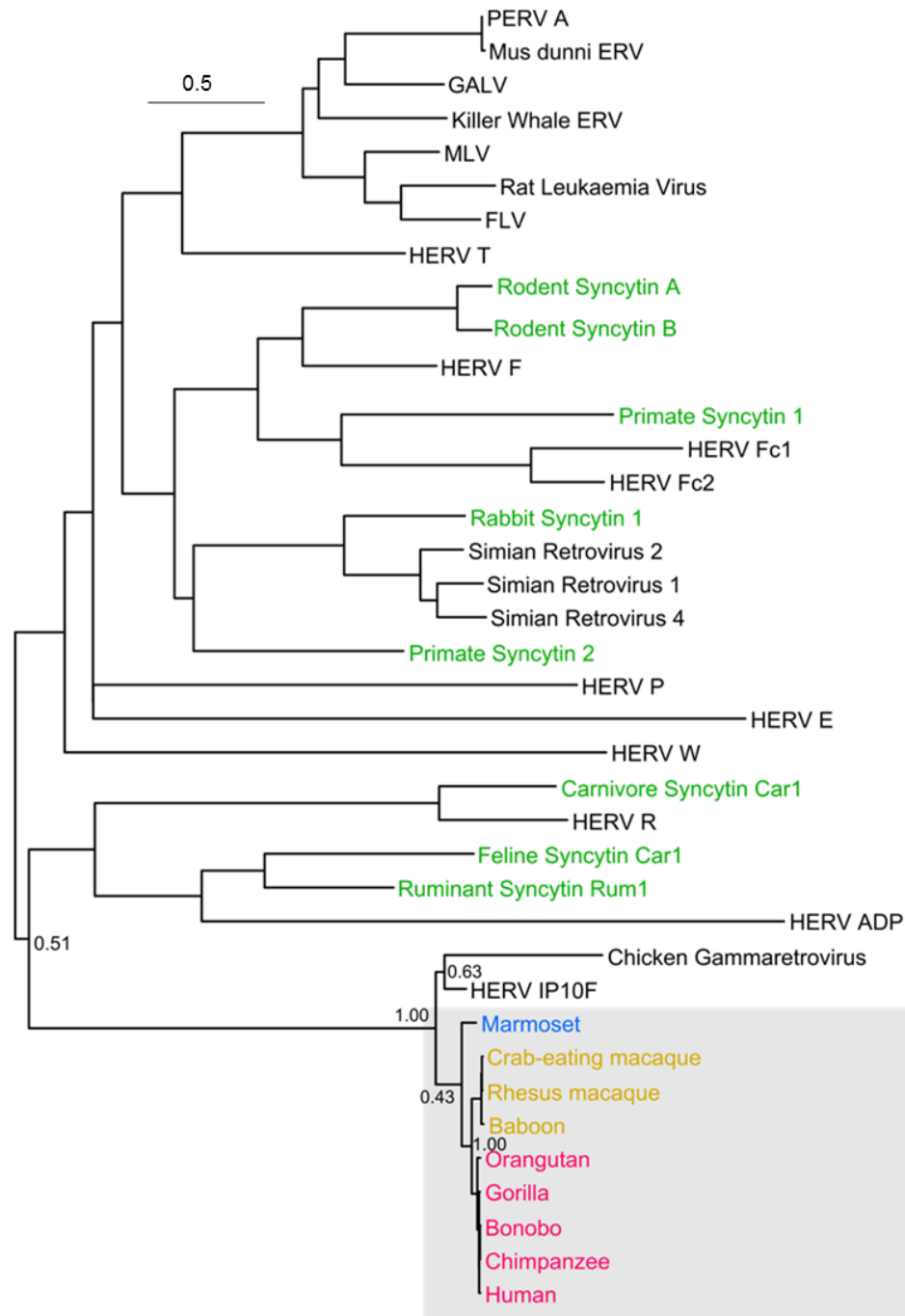


Figure 43: Phylogenetic comparison between the nucleotide sequences of the *env* ORFs identified in HERV-I related insertions and of other known gammaretroviral syncytin proteins and *env* genes.

Newly identified sequences are in the shaded region with sequences from apes in pink, old world monkeys in yellow and new world monkeys in blue. Known syncytins are shown in green. Branches leading to primate syncytins one and two and rodent syncytins A and B have been collapsed. Details of previously known sequences are provided in Appendix B.2.

A few scattered HERV-I like insertions were identified in lemurs, aye-ayes and rodents. These were analysed phylogenetically and the result of this analysis is shown in Figure 44. In order to represent the simian HERV-I insertions discussed above in this tree, a phylogeny was built for representative *pol* genes (genes from the most intact loci were used) and one sequence from each cluster selected, these are sequences g_1 to g_10 in this phylogeny. Two well supported clusters of prosimian and rodent HERV-I like insertions were found, the first consisting of insertions from the two species of lemur (aye-aye and mouse lemur) and from chinchilla, naked mole rat and hamster. The lemur insertions are more closely related to each other than to insertions from any other host, which may be indicative of a small group of HERV-I like insertions limited to Malagasy lemurs, however further work would be needed to confirm this. The second group consists of only sequences from rodents (naked mole rat, ground squirrel, jerboa and chinchilla) however again these are not closely related species. The rodent insertions are very scattered and show no particular correspondence to the phylogenetic or geographical relationships between the hosts.

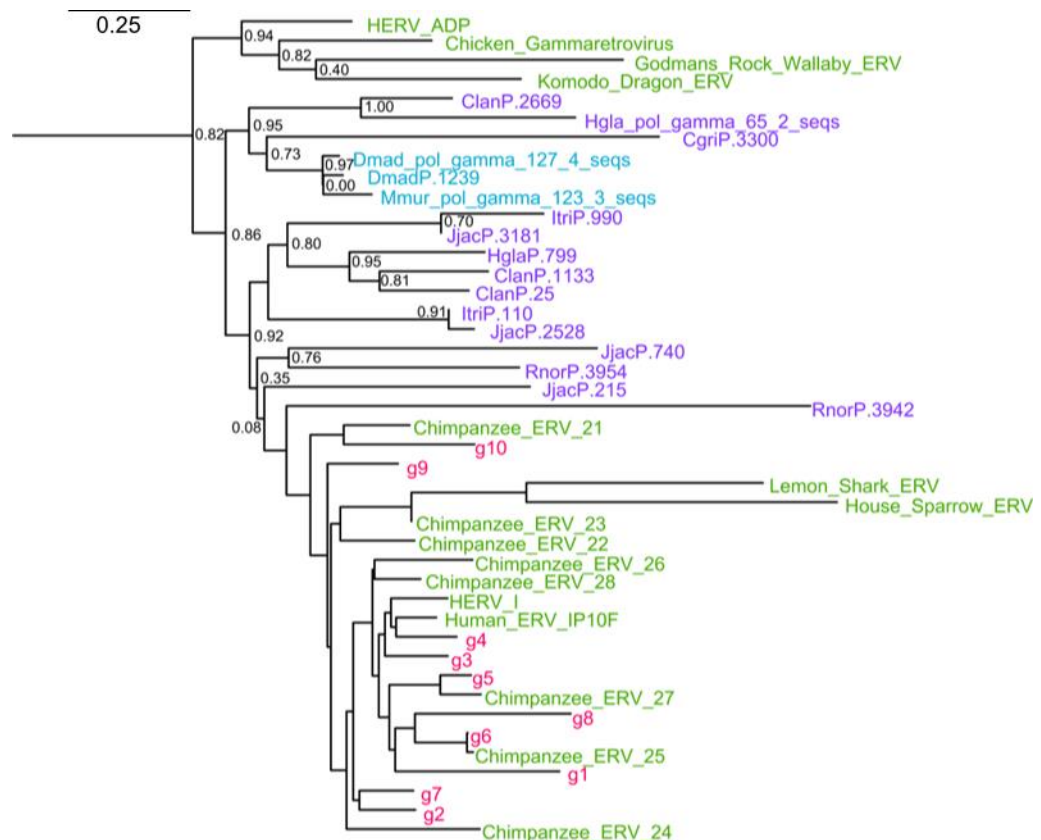


Figure 44: *Pol* gene phylogeny showing the relationship between the HERV-I like insertions identified in prosimians and rodents with those found in mammals. Prosimians are shown in blue, rodents in purple, representative sequences for the simian HERV-I clusters in pink and previously described HERV-I like proviruses in green. This tree is rooted on the basic gammaretrovirus dataset, however this group has been cropped for better visualisation. Details of previously known sequences are provided in Appendix B.2.

All HERV-I like *pol* gene sequences identified in birds were also combined with the gammaretrovirus test dataset and the previously defined HERV-I like sequences in the analysis shown in Figure 45. Bird HERV-I like sequences clustered robustly with marsupial and non-mammalian insertions, separately to the HERV-I like insertions identified in placental mammals. There are three strongly supported groups of avian HERV-I like insertions on this tree. Two contained only avian sequences, however the third contains gammaretroviruses from very diverse hosts (Figure 45). Sequences in this group have previously been identified in cartilaginous fish (Martin et al., 1997, Herniou et al., 1998), birds (Borysenko et al., 2008, Niewiadomska and Gifford, 2013), lizards (Martin et al., 1997), marsupials (Martin et al., 1997) and monotremes (Martin et al., 1997). HERV-ADP, first described by Lyn et al. (1993), inconsistently falls into this group. HERV-ADP was not identified using our pipeline, probably due to its degeneration through integration of other retroelements, as described in the original paper. No insertions clustering in this group were identified in mammals. This suggests some barrier to their integration, such as a restriction factor. The lineage appears to have proliferated considerably more successfully in birds (Figure 45), with diverse members of this group found in the chicken and turkey genomes. This result points to birds as a potential vector and reservoir host for these viruses, which may then have been transmitted to other hosts. This would explain the wide geographical distribution of their non-avian host species.

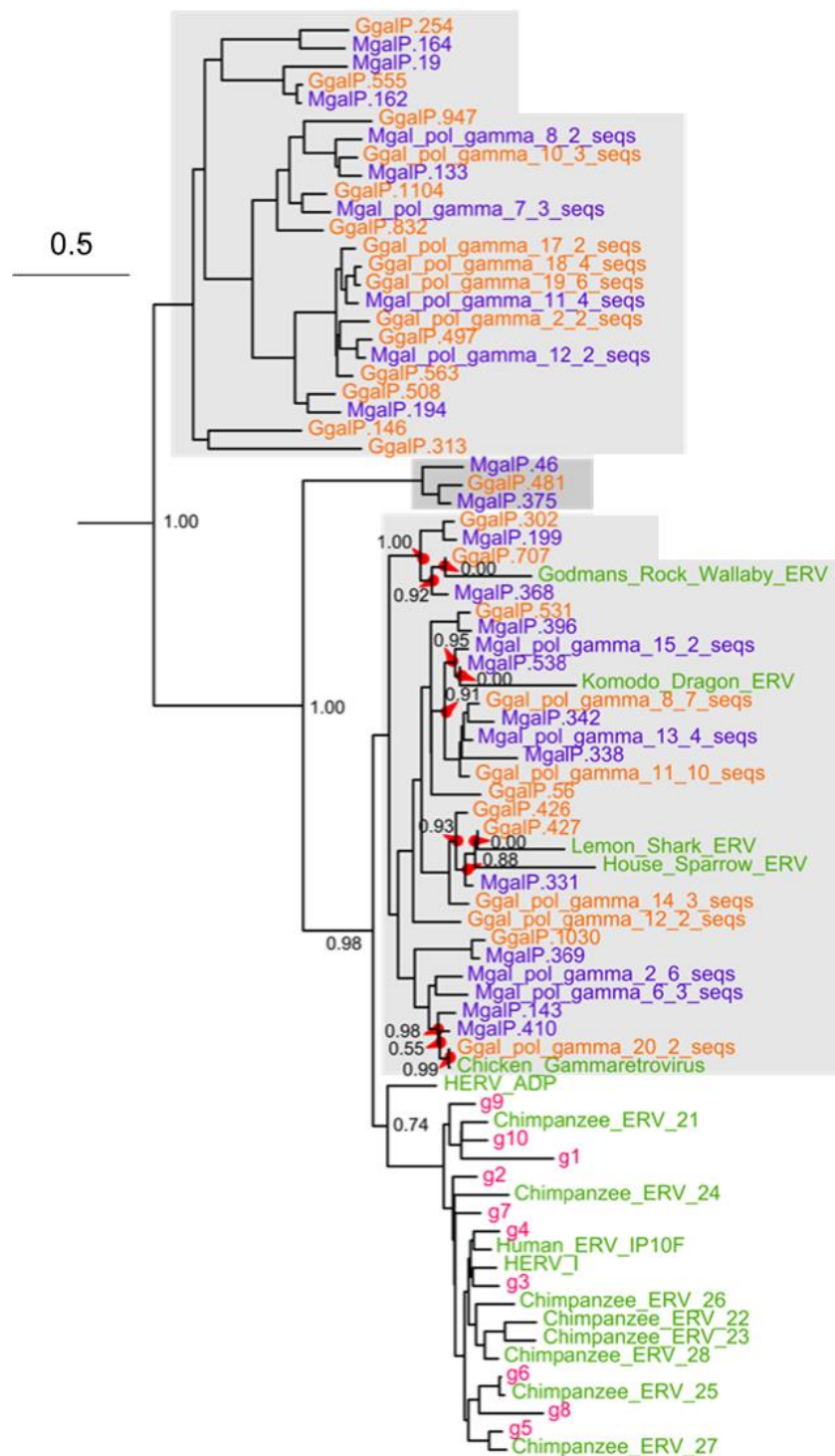


Figure 45: *Pol* gene phylogeny showing the relationship between the HERV-I like insertions identified in birds and those found in mammals.

Chicken insertions are shown in orange, turkey in purple, the sequences representing simian HERV-I in pink and previously described HERV-I like proviruses in green. This tree is rooted on the basic gammaretrovirus dataset, however this group has been cropped for better visualisation. In dense regions red arrows connect the node being described with its branch support value. Details of previously known sequences are provided in Appendix B.2.

4.2.2. HERV-F Group

There are 10 subgroups in the HERV-F group: HERV-F, HERV-FRD, HERV-FXA, HERV-Fc1, HERV-Fc2, HERV-Fb, HERV-W, HERV-P, HERV-H RGH2 and HERV-H RTVLH2. These have been characterised in some detail in old world primates and apes [for example (Bénit et al., 2003, Kim et al., 1999, Seifarth et al., 1995, Jern et al., 2004)] and reviewed by Bannert and Kurth (2006). This group includes many of the syncytin proteins discussed in section 1.4.3.9: primate syncytin 1 is related to HERV-W Env (Cáceres et al., 2006), primate syncytin 2 to HERV-FRD (Blaise et al., 2003), rodent syncytins A and B to HERV-F (Dupressoir et al., 2005) and the new world rodent syncytins to HERV-Fb (Vernochet et al., 2011).

A number of sequences from previous genome screens were found to be members of this group. These were Polavarapu et al.'s (2006a) chimpanzee gammaretroviruses 9 to 18 and 29, ERV-9 (La Mantia et al., 1991), HERV-4 (Taruscio and Mantovani, 1996), Garcia-Etxebarria et al.'s bovine ERVs 1 to 4 and 6 to 10 (2010), Benit et al.'s (2003) HERV-F-like elements from apes, old world monkeys, new world monkeys and lemurs, Huda et al.'s (2008) chicken ERV 21 and McCarthy et al.'s (2004) murine ERV-7 and murine ERV-9.

Here, 37,311 HERV-F like fragments were identified. Details of these fragments are provided in Table 15. These insertions were especially abundant in primates but were also identified in all other host groups.

Table 15: The number of HERV-F-like ERV fragments identified in each host type.
All: all 33 genomes. Blue rows labelled “T” represent total counts, yellow rows labelled “PG” represent the mean number of fragments per genome of this type.

		All	Gag	Pol	Env
All	T	37311	5193	25698	6420
	PG	1131	157	779	195
Primates	T	27391	3594	18829	4968
	PG	1826	240	1255	331
Rodents	T	8062	1220	5821	1021
	PG	733	111	529	93
Lagomorphs	T	983	217	674	92
	PG	492	109	337	46
Tree Shrews	T	598	110	193	295
	PG	299	55	97	148
Birds	T	37	0	37	0
	PG	19	0	19	0
Ferret	T	240	52	144	44
	PG	240	52	144	44

The presence or absence of each of the 10 HERV-F-like lineages and its approximate copy number was established for each host (as described in sections 2.4.1 and 2.4.2). The results of these analyses are shown in Figure 46 and Figure 47.

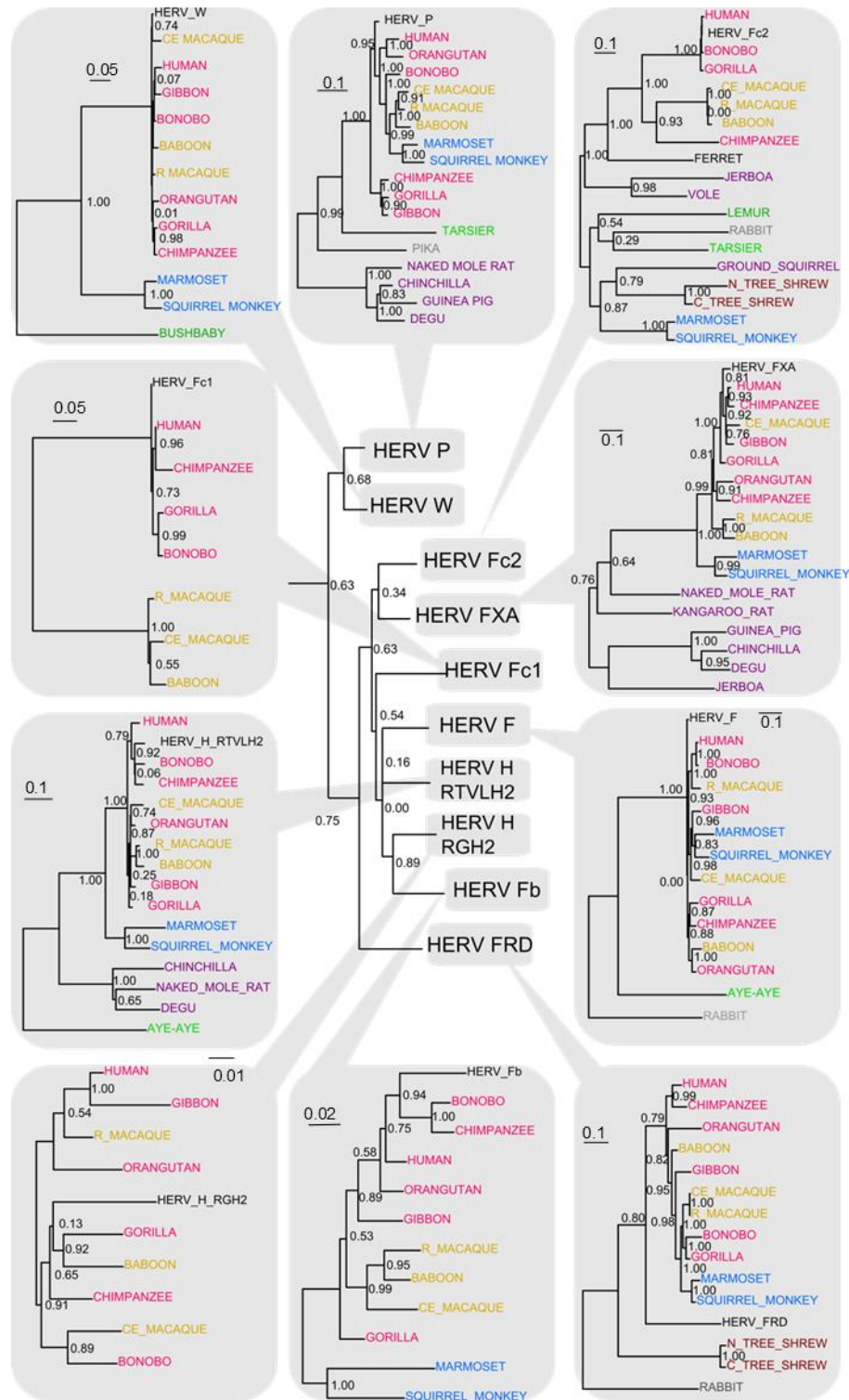


Figure 46: The *pol* gene phylogenetic relationships between each lineage of HERV-H, HERV-F, HERV-W and HERV-P and the closest matching sequence from each host species, where a sequence was present in the host.

Sequences from apes are shown in pink, old world monkeys in yellow, new world monkeys in blue, prosimians in green, tree shrews in brown, lagomorphs in grey, rodents in purple. Reference sequences are shown in black. Details of previously known sequences are provided in Appendix B.2.

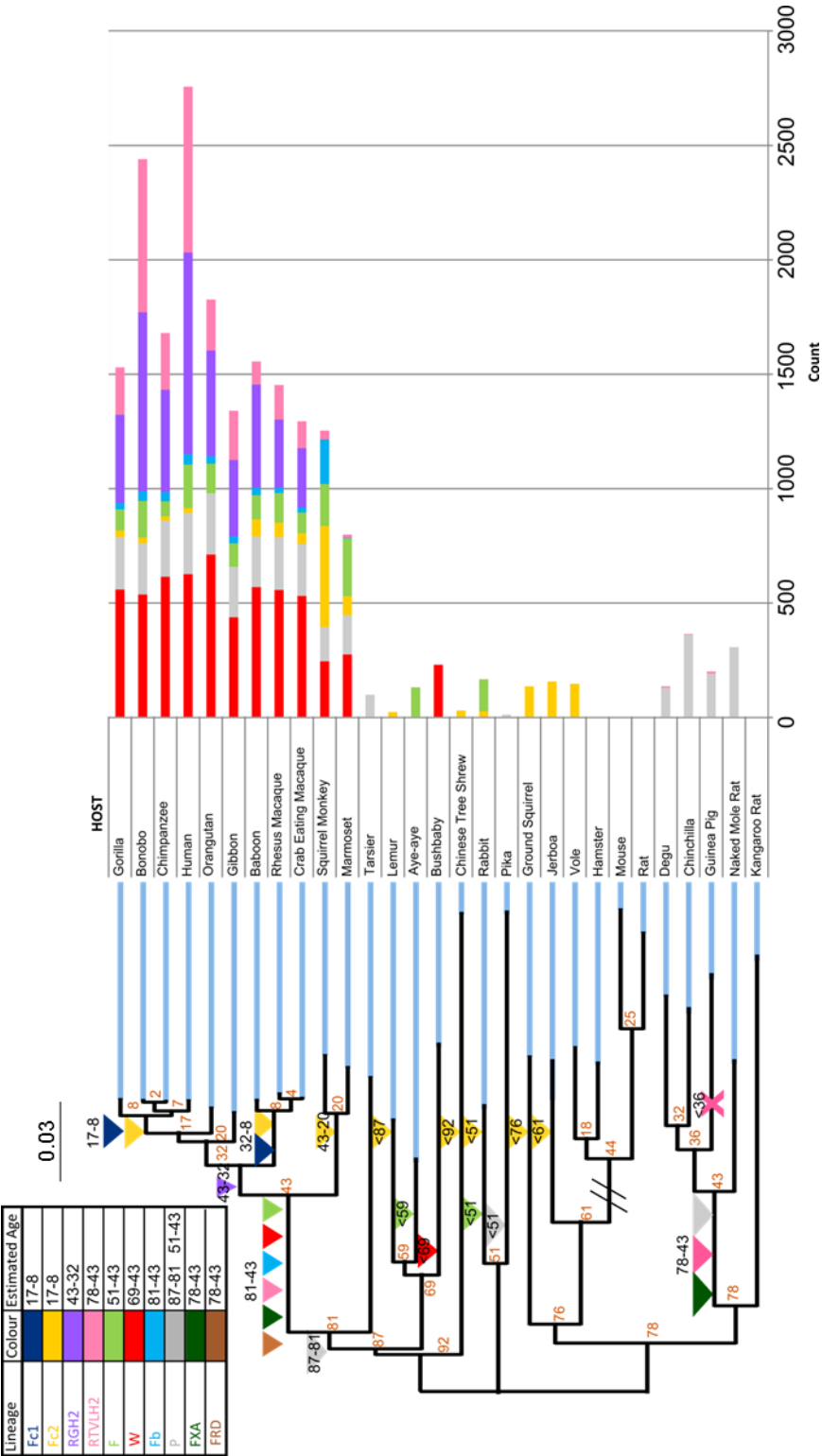


Figure 47: The copy number and estimated integration dates of the HERV-F/H/W family of ERVs.

Arrows represent estimated integration dates (million years ago). X's represent loss of an ERV from a lineage. Arrows and bars are colour coded according to the key and represent different ERV lineages. Orange numbers are estimated node ages in millions of years.

Our results suggest that the HERV-P group is more ancient than the other ERVs in this group, appearing for the first time at least 81 million years ago in an ancestor of tarsiers and simian primates. HERV-P-like *pol* gene fragments were identified at a moderate copy number (approximately 200 to 300 copies) in all new and old world primates and in tarsiers, the prosimian which shared the most recent common ancestor with the simians. The consistent copy number of this group and the phylogenetic relationships of shown in Figure 47 fit with the hypothesis that HERV-P entered this ancestor. Previously, Yi et al. (2007) identified HERV-P like *pol* gene fragments in a lemur species (*Lemur catta*) while our screen of the grey mouse lemur (*Microcebus murinus*) was negative. Hayward et al. (2013a) did not include HERV-P in their analysis.

HERV-P-like insertions were also identified in the new world rodents (chinchilla, guinea pig, naked mole rat and degu). These fragments cluster together within the HERV-P family, have a fairly consistent copy number and the naked mole rat insertion is more distantly related than those in the other hosts, which suggests that the integration event was prior to the divergence of these species (43 million years ago) but after their divergence from the kangaroo rat (78 million years ago). New world rodents arrived in south America from Africa approximately 41 million years ago and primates approximately 26 million years ago (Antoine et al., 2012). Therefore, the cross-species transmission leading to colonisation of these two groups by HERV-P-like ERVs must have occurred in Africa more than 41 million years ago.

As new world rodent HERV-P-like insertions have not been characterised previously, regions flanking the *pol* genes in this group were screened for *gag* and *env* genes. 11 loci were identified with all three genes: six in chinchilla, one in guinea pig, two in naked mole rat and two in degu. Separate *gag* and *env* phylogenies were generated for these sequences. The phylogeny for *gag* was consistent with Figure 47 (data not shown), however the *env* gene phylogeny showed an unexpected relationship, clustering with a carnivore syncytin gene in the group described by Cornelis et al. (2012). A more detailed *env* gene

phylogeny was therefore created including all known sequences in this group and the related ruminant syncytin group identified by Cornelis et al. (2013) and is shown in Figure 48. *Env* genes from the primate HERV-P-like group usually had clustering patterns matching their *gag* and *pol* phylogenies (data not shown). There were five exceptions to this. A single *env* gene neighbouring a HERV-P like *pol* but clustering with the carnivore syncytins was identified in each the two genomes of *Pan* species: chimpanzee and bonobo. Another insertion of this type was identified in the aye-aye. A single *env* gene neighbouring a HERV-P like *pol* but clustering with the HERV-R *env* was identified in each new world monkey genome: marmoset and squirrel monkey. These relationships are shown in Figure 48.

The newly identified new world rodent *env* sequences form a monophyletic, well supported group which is close to but distinct from the carnivore syncytins. This group is not closely related to the known new world rodent syncytins described by Vernochet et al. (2011). The aye-aye and *Pan env* genes are even more similar to the carnivore syncytins. The *env* regions of members of this group were screened for ORFs, however, no sequence longer than 384 nucleotides was identified, so these loci are not capable of generating a functional syncytin protein. The evolutionary distance between aye-ayes and *Pan* species, between these primates and the new world rodents and between the Euarchontoglires hosts of these newly identified *env* genes and the carnivores which harbour the Car1 syncytin gene are strongly indicative of cross-species transmission events. Transmission via bites from rodents or primates to carnivores while these recombinant viruses were active is a feasible transmission route. The gene may then have been co-opted for placental development in carnivores but allowed to degrade in the new world rodents and primates.

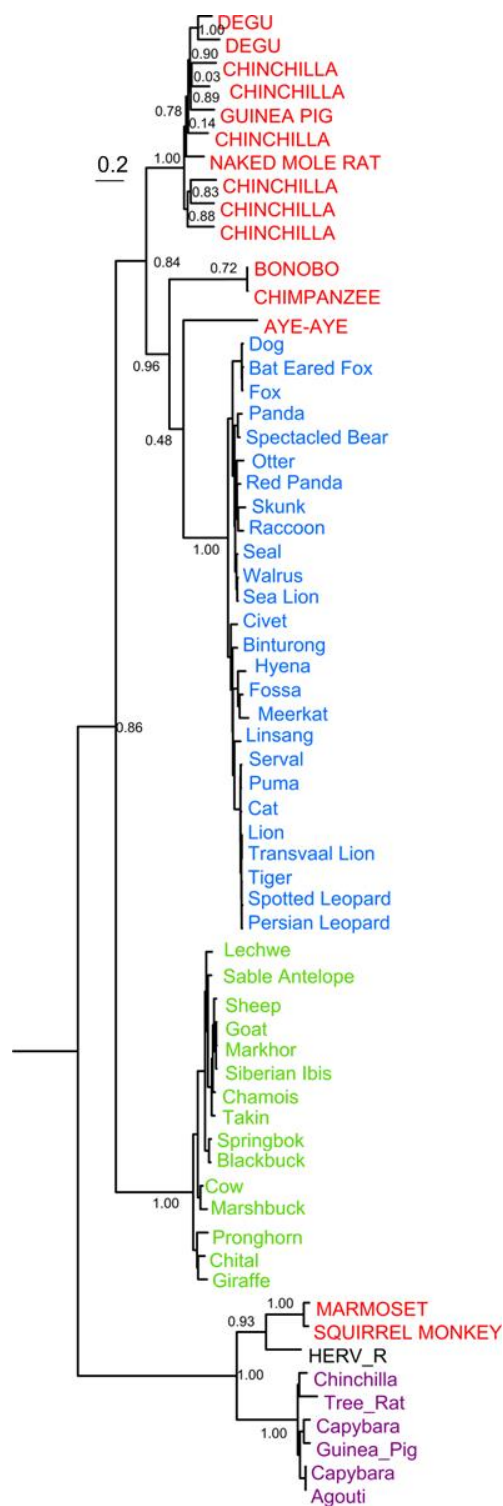


Figure 48: *Env* gene phylogeny showing the relationship between known carnivore, new world rodent and ruminant syncytins and the HERV-P *env* genes identified here.

Newly identified sequences are shown in red capitals, known carnivore syncytins in blue, known ruminant syncytins in green and known new world rodent syncytins in purple. Details of previously known sequences are provided in Appendix B.2.

Pol genes belonging to the HERV-W lineage, which is closely related to the HERV-P lineage, have only been described in detail in old world monkeys and apes to date, however HERV-W like LTRs have previously been detected in new world monkeys (Kim et al., 1999) and phylogenetic analysis by Hayward et al. (2013a) shows a very similar pattern to our Figure 46. We have identified insertions clustering with HERV-W in high copy numbers in all simian genomes screened, including new world monkeys. However, the intact *env* ORF encoding syncytin was only found in apes, as described by Caceres et al. (2006). As in Hayward et al.'s (2013a) phylogeny, the new world monkey HERV-W like *pol* fragments were relatively distant from those found in old world monkeys and apes and this, combined with the general lack of comprehensive previous screening attempts in new world monkeys, may explain why these have not generally been detected previously. This genetic distance, combined with the copy number increase in old world monkeys and apes compared to new world primates and the high similarity between old world monkey and ape HERV-W like insertions, suggests that HERV-W integrated into a common ancestor of new and old world primates but that it has also been more recently active in primates.

A number of HERV-W-like insertions were also identified in bushbabies but not in other prosimians, again this is consistent with Hayward et al.'s (2013a) phylogeny. The relationship between these insertions and the HERV-W like insertions of other primates is shown in Figure 49. Unless these insertions have been deleted in all other prosimian primates, these are likely to have an origin in the last 69 million years in the lineage leading to the bushbabies (Figure 47). Therefore, it is likely that a bushbaby ancestor was infected horizontally by another primate at some point in their evolutionary history. Bushbabies are widespread in the savannahs of southern Africa, a habitat shared with several other primate species, so a cross-species transmission would be possible.

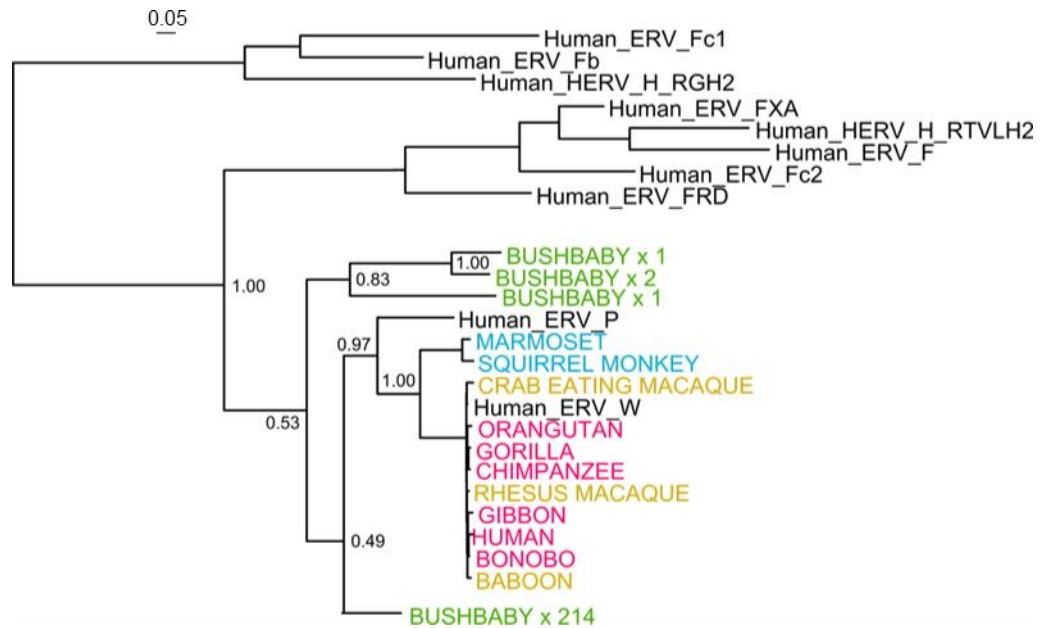


Figure 49: *Pol* gene phylogeny showing the relationship between the bushbaby HERV-W-like insertions and those of other primates. Details of previously known sequences are provided in Appendix B.2.

HERV-F, HERV-Fb and HERV-H RTVLH2 were detected in all simian primates but not tarsiers. Therefore, they are likely to have integrated between the tarsier-simian (81 million years ago) and new world primate – old world primate (43 million years ago) divergence events (Figure 37). These dates are consistent with those reviewed by Bannert and Kurth (2006) and the phylogenies for these sequences shown in Figure 46 are consistent with those in Hayward et al.'s (2013a) phylogenetic analysis. HERV-H-RGH2 appears to be more modern, as insertions were only identified in apes and old world monkeys, which points to an origin of this group 43 to 32 million years ago. This contradicts the results of Jern et al. (2004) who considered HERV-H RTVLH2 to be more modern than HERV-H RGH2.

We identified a particularly high copy number of both HERV-H lineages (RGH2 and RTLH2) in humans and bonobos compared to other apes. The most recent common ancestor of humans and bonobos is shared with the chimpanzee but humans and bonobos have a two-fold increase in RGH2 and a three-fold increase in copy number of RTLH2 compared to chimpanzees. There is no clear phylogenetic distinction between the human, bonobo and chimpanzee insertions, as Figure 50 demonstrates, so the reason for the

success of these retroviruses in humans and bonobos but not chimpanzees is unclear. Jern et al. (2006) also noted the recent expansion of HERV-H insertions in the human but not the chimpanzee genome and proposed the “midwife” hypothesis to explain this, with HERV-H co-opting proteins from another more intact provirus to allow reintegration (section 1.4.3.2). This would be consistent with our results and a similar phenomenon may have occurred in the bonobo, which has not previously been extensively screened for ERVs.

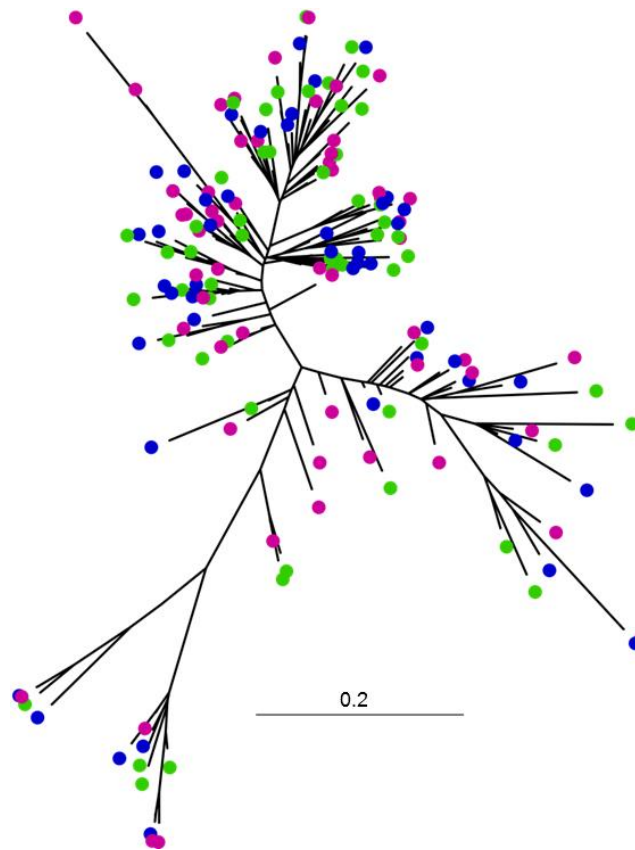


Figure 50: *Pol* gene phylogeny showing the relationship between HERV-H like clusters identified in humans (green), bonobos (pink) and chimpanzees (blue).

Outside of the primates, HERV-F like insertions were identified in the aye-aye and rabbit and RTVLH2-like insertions in some new world rodents: naked mole rat, chinchilla and degu. Hayward et al. (2013a) identified the same

naked mole rat and rabbit insertions. Four of the HERV-F like insertions in the rabbit and two of the insertions in the aye-aye had a *gag-pol-env* structure, so an *env* gene phylogeny was constructed to look for recombination (data not shown). Both *env* genes were somewhat similar to the syncytin 1 group in primates but contained many stop codons. Although RTVLH2 was present in new world rodents it was absent in guinea pig, which limits its integration date to less than 36 million years ago, which is more recent than its estimated integration date in primates. However, copy number in rodents is very low (less than 10 copies per genome) and insertions in the other new world rodents were very degraded, so it is possible that guinea pig insertions exist but were not detected. Alternatively, RTVLH2 may have circulated twice.

Rodent syncytins A and B are similar to HERV-F *env*. Outside of the *env* ORF, these insertions are known to be very degenerate (Dupressoir et al., 2005). Accordingly, the *env* genes resembling syncytins A and B identified here were found in isolation, not close to any other ERV fragments. We detected syncytin A ORFs in mouse, rat and vole and syncytin B ORFs in mouse, rat, vole and hamster. This is consistent with the literature (Dupressoir et al., 2005). No non-rodent hosts had similar ORFs.

Our results for HERV-Fc1 were consistent with those of Benit et al. (2003) in the identification of low copy numbers of HERV-Fc1 in the human, chimpanzee, gorilla and baboon genomes but not in the orangutan or gibbon genomes. However, Benit et al. did not identify HERV-Fc1 in the macaque, whereas here it was found in both macaque species screened. As the copy number of HERV-Fc1 is low, it may have been absent in the sequence data available for rhesus macaque when the Benit et al. paper was published but it is identifiable in the 2010 RheMac3 genome build used here. Like the previously identified human and baboon HERV-Fc *env* genes, one rhesus macaque ERV-Fc locus encoded a full length *env* ORF. The sequence of this ORF was very close to that of the baboon *env* ORF, as shown in Figure 51. Chimpanzees, bonobos and gorillas had degraded copies of the same gene

while no HERV-Fc-like *env* fragments were detected in the crab-eating macaque. The absence of HERV-Fc1 in orangutan and gibbon suggests that it was circulating less than 17 million years ago, when orangutans diverged from the other great apes and more than 8 million years ago, when gorillas diverged from the ancestor of humans, chimpanzees and bonobos. The *env* gene of this lineage appears to have been subject to selection to remain intact in humans, baboon and macaques but this is not apparent in chimpanzees, gorillas or bonobos. This suggests that this gene has been co-opted for a functional role more than once or that its function has been replaced by another factor in some hosts.

HERV-Fc2 also had a low copy number in primates, with less than 100 copies per genome, with the exception of the squirrel monkey, where it is very abundant with more than 400 copies. A previously undescribed HERV-Fc2 *env* ORF was identified in the bonobo (Figure 51). The highly similar human and bonobo sequences were aligned and a Ka/Ks ratio of 0.767 was calculated using the methodology described in section 2.4.7. This value indicates weak purifying selection. The presence of an Fc2 *env* ORF in bonobo and human suggests that the gene was once present in chimpanzees, however, it was not detected here and Benit et al. (2003) only found a very degraded copy. Gorillas contained a recognisable Fc2 but with multiple stop codons. This is similar to the pattern detected for HERV-Fc1 and again suggest that either this protein has been co-opted for a functional role twice, once in bonobos and once in humans, or that it has lost its function in some hosts.

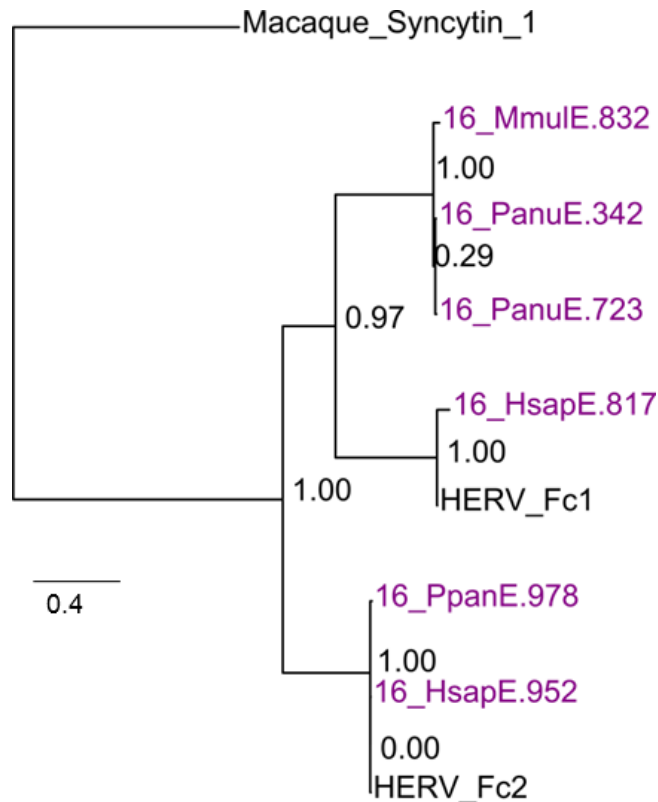


Figure 51: *Env* gene phylogeny showing the regions identified with ORFs corresponding to HERV Fc1 and HERV Fc2.

Newly identified sequences are shown in purple.

Fragments resembling HERV-Fc2 were also identified in the lemur, Chinese tree shrew, rabbit, ground squirrel, jerboa and vole (Figure 47). None of these species share a common ancestor which is not also shared with other hosts lacking these fragments (Figure 37). Insertions from these hosts form a separate, monophyletic cluster to those identified in simian primates and are likely to represent a distinct lineage of ERVs to the HERV-Fc2 lineage.

A consistent but low number of copies of HERV-FXA, ranging from 3 to 24, was found in all simians and all new world rodents. No HERV-FRD like *pol* genes were detected here. This is not unexpected, as only a single copy of HERV-FRD is found in the human genome and the *pol* gene is known to be degenerate. The HERV-FRD *env* ORF gene which has been co-opted to as primate syncytin 2 was detected in all simian primates, as described by Blaise et al. (2003). Hayward et al. (2013a) did identify HERV-FRD-like *pol* gene

fragments in all simian primates, however as sequence details are not provided in this paper it is not possible to compare these to our results.

4.2.3. HERV-E Group and HERV-R Group

The HERV-E group was built around one known HERV: HERV-E (Taruscio and Manuelidis, 1991). Several previously known retroviruses clustered with HERV-E in the analysis discussed in section 2.1.4.1. These were Polavarapu et al.'s (2006a) chimpanzee ERVs four to eight, HERV-1 and HERV-33 (both uncharacterised to our knowledge but available via Repbase), Tristem et al.'s (1996) American mink, grey seal, Mexican bat, cow and sheep ERVs and Garcia-Etxebarria et al.'s (2010) bovine ERVs 14, 15, 17 and 18.

Using Exonerate, we have identified 8,349 fragments in this group. Primate and tree shrew genomes had a higher HERV-E content than other host groups. Details of the fragments identified are provided in Table 16.

Table 16: The number of HERV-E-like ERV fragments identified in each host type.
All: all 33 genomes. Blue rows labelled "T" represent total counts, yellow rows labelled "PG" represent the mean number of fragments per genome of this type.

		All	Gag	Pol	Env
All	T	8349	2070	6279	0
	PG	253	63	190	0
Primates	T	5818	1628	4190	0
	PG	388	109	279	0
Rodents	T	1721	263	1458	0
	PG	156	24	133	0
Lagomorphs	T	41	7	34	0
	PG	21	4	17	0
Tree Shrews	T	660	152	508	0
	PG	330	76	254	0
Birds	T	33	0	33	0
	PG	17	0	17	0
Ferret	T	76	20	56	0
	PG	76	20	56	0

Similarly, the HERV-R group was built around one known HERV: HERV-R (Andersson et al., 1998). No other sequences in the PARSED_UT_PREVKNOWN dataset were assigned to this group.

3,528 fragments were identified in this group. The majority of these were found in primate hosts. Details of these fragments are provided in Table 17.

Table 17: The number of HERV-R-like ERV fragments identified in each host type.
All: all 33 genomes. Blue rows labelled “T” represent total counts, yellow rows labelled “PG” represent the mean number of fragments per genome of this type.

		All	Gag	Pol	Env
All	T	3528	176	960	2392
	PG	107	5	29	72
Primates	T	2975	160	791	2024
	PG	198	11	53	135
Rodents	T	438	2	116	320
	PG	40	0	11	29
Lagomorphs	T	12	0	11	1
	PG	6	0	6	1
Tree Shrews	T	47	13	32	2
	PG	24	7	16	1
Birds	T	8	0	3	5
	PG	4	0	2	3
Ferret	T	48	1	7	40
	PG	48	1	7	40

The HERV-E and HERV-R groups were not selected for detailed analysis, as preliminary testing did not identify any important differences from the literature. However, a presence / absence analysis was performed as described in section 2.4.1 to establish if sequences clustering closer to HERV-E/R than to any of the other known retrovirus sequences assigned to these groups were

present or absent in each host. The results of this analysis are shown in Figure 52.

HERV-E was found in all simian primates plus bushbabies. Yi et al. (2006) also found this ERV in all simians but found it to be absent in two prosimians, the bushbaby *Otolemur crassicaudatus* and the lemur *Lemur catta*. We also found this ERV group to be absent in lemur, aye-aye and tarsier, therefore these results combine to suggest that the *O. garnettii* insertions are the result of a late cross-species transmission into this host. Hayward et al. (2013a) found a very similar phylogeny to for HERV-E to Figure 52, including these insertions in *O. garnettii*. HERV-R-like fragments were identified in all simian primates, tarsier, all new world rodents, rabbit and jerboa. The old world monkey and ape HERV-R *pol* fragments were very similar to each other and probably represent the group described by Kim et al. (2006). The remaining HERV-R-like fragments are much more distinct, which may be why they were not detected in this earlier study. All except the jerboa insertions were detected and characterised phylogenetically by Hayward et al. (Hayward et al., 2013a) (jerboa was not screened in this study) and showed relationships very similar to those shown in Figure 52. An in depth analysis of these groups would be a worthwhile extension of this work.

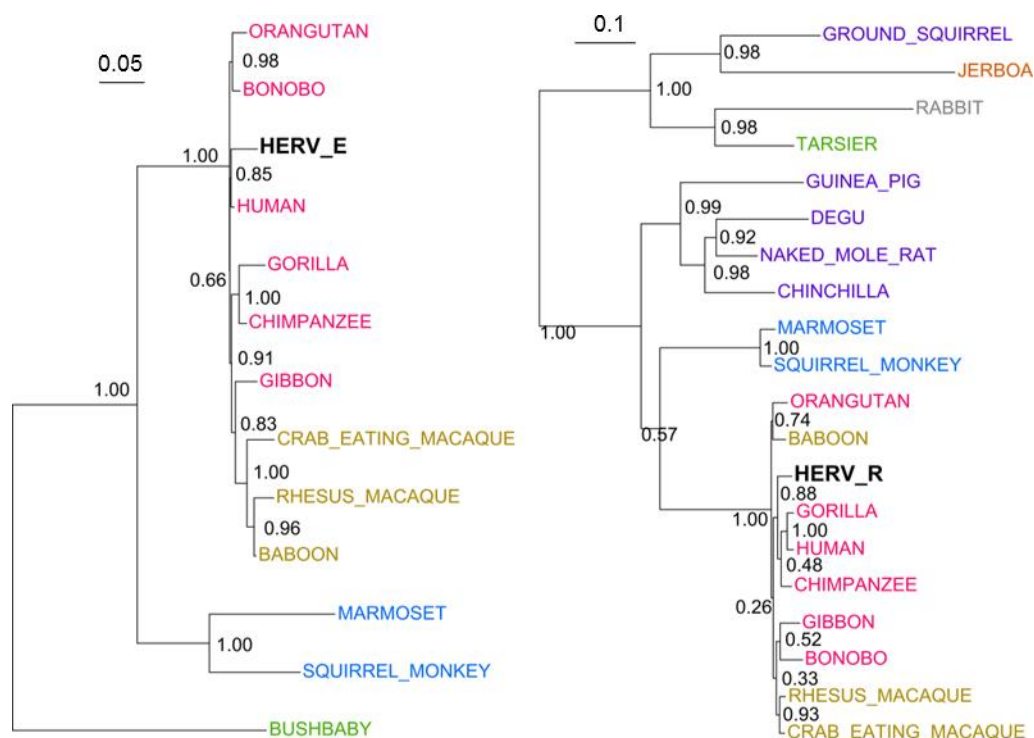


Figure 52: *Pol* gene phylogeny for the closest sequence identified in each host to HERV E (left) and HERV R (right).

Sequences which clustered more closely to another retrovirus in the group are not shown. Details of previously known sequences are provided in Appendix B.2.

4.2.4. REV-Like Group

The REV group was built based upon the exogenous REV viruses which are pathogenic in birds (REV, DIAV and SNV), their endogenous relatives in mongoose and echidna (Niewiadomska and Gifford, 2013) and the closely related HERV-T lineage found in all simian primates (Yi and Kim, 2007). Only one further previously known virus was added to this group, Polavarapu et al.'s (2006a) chimpanzee ERV 3.

5,409 REV-like fragments were identified in total, with approximately the same number in primates and rodents. Details of these fragments are provided in Table 18.

Table 18: The number of REV-like ERV fragments identified in each host type.
All: all 33 genomes. Blue rows labelled “T” represent total counts, yellow rows labelled “PG” represent the mean number of fragments per genome of this type.

		All	Gag	Pol	Env
All	T	5409	1516	2990	903
	PG	164	46	91	27
Primates	T	2690	675	1415	600
	PG	179	45	94	40
Rodents	T	2575	798	1507	270
	PG	234	73	137	25
Lagomorphs	T	32	8	16	8
	PG	16	4	8	4
Tree Shrews	T	105	35	52	18
	PG	53	18	26	9
Birds	T	0	0	0	0
	PG	0	0	0	0
Ferret	T	7	0	0	7
	PG	5409	1516	2990	903

These results are consistent with the literature in that no ERVs were found clustering closely with REV across all three genes. Instead, a series of ERVs were identified clustering with HERV-T and between REV and HERV-T, including several which have not previously been described in detail.

Considerably more insertions in this group were identified in the guinea pig and tarsier than in any other host, as shown in Figure 53.

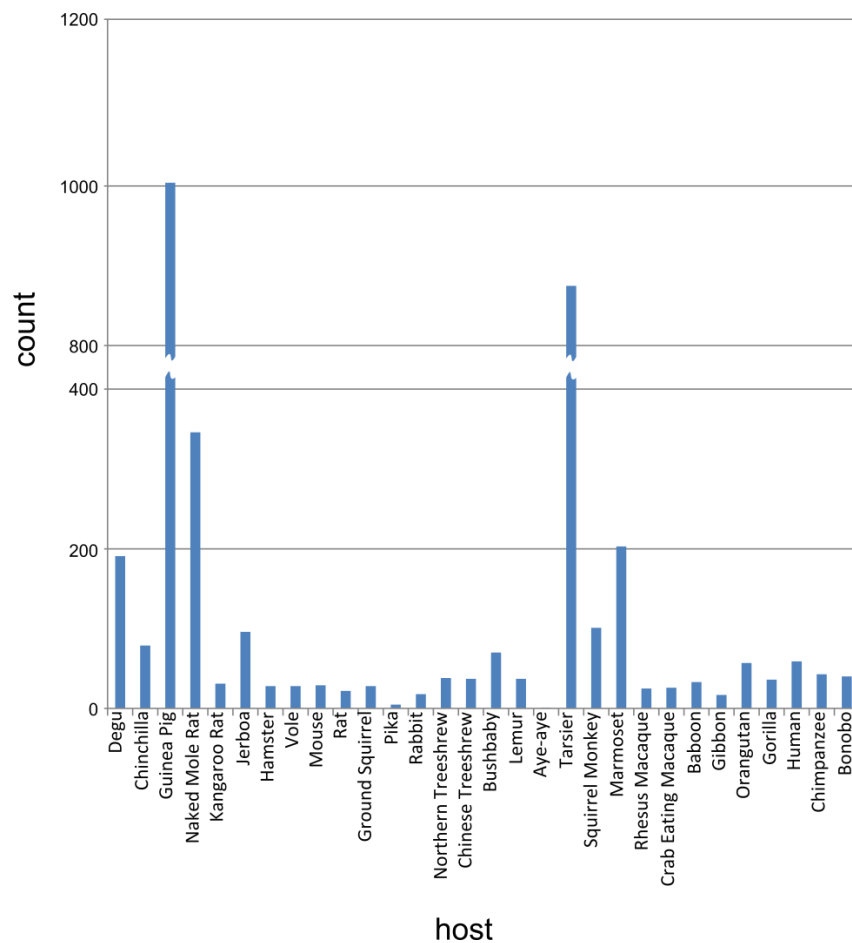


Figure 53: The number of REV/HERV-T like *pol* gene insertions identified in each host.

As REV is a recombinant virus, endogenous REV insertions would be expected to have *gag* and *pol* genes clustering with the gammaretroviruses but *env* genes clustering with the betaretroviruses. HERV-T like ERVs have gammaretroviral *gag*, *pol* and *env* genes. Therefore, all *pol* gene sequences were screened for neighbouring *env* genes similar to HERV-T *env* and neighbouring *env* genes in the betaretrovirus dataset.

400 non-recombinant full-length loci (loci with HERV-T like *gag*, *pol* and *env*) were identified. These were much more numerous in guinea pigs but were widespread amongst different hosts, with *gag-pol-env* regions identified in all species of ape, old world monkey, new world monkey and new world rodent, plus jerboa and bushbaby. *Pol* genes were clustered using a phylogenetic tree for each host and a single representative sequence selected for each cluster. A phylogenetic analysis was then performed on these representative sequences,

shown in Figure 54. ERVs closely related to HERV-T were found in all apes screened here. A second group of closely related viruses spanned old world monkeys and apes and a third group was found only in new world monkeys. This is consistent with the results of Yi et al. (2007), who suggested that the lineage leading to HERV-T integrated before the divergence of new and old world monkeys then proliferated 56, 47 and 31 million years ago.

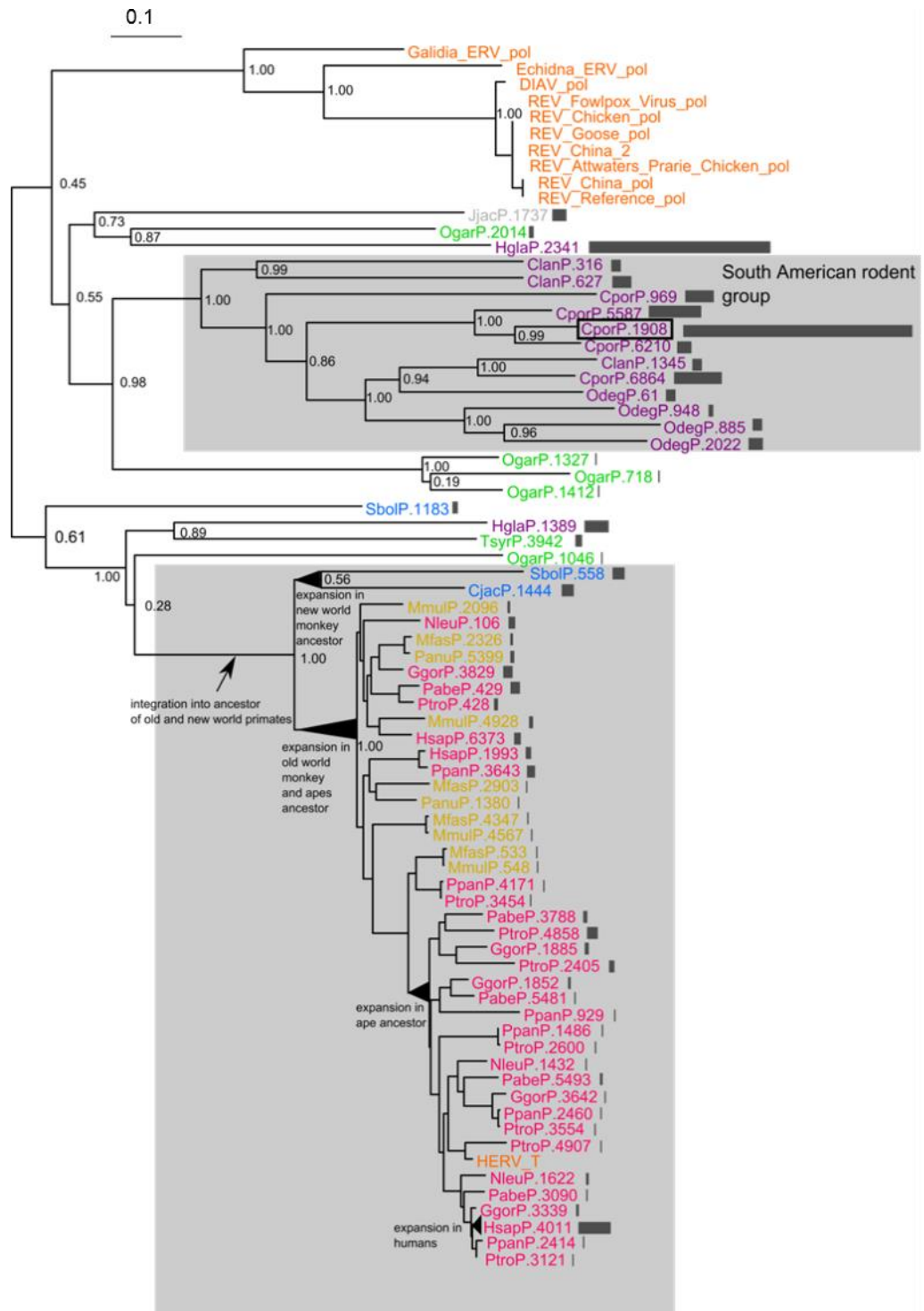


Figure 54: Phylogeny showing the relationships between the *pol* gene at selected REV-like non-recombinant loci and known gammaretroviruses.

Sequences from apes are shown in pink, old world monkeys yellow, new world monkeys blue, prosimians green, Hystricomorpha rodents purple, Myomorpha rodents grey, known gammaretroviruses orange. Bars represent the relative size of each group. Triangles on branches represent copy number increases. Black border indicates the potentially intact cluster described below. Details of previously known sequences are provided in Appendix B.2.

Non-recombinant HERV-T like loci were screened for ORFs (as described in section 2.4.5) and two guinea pig loci, one on scaffold_9 and one on scaffold_13, were potentially intact (scaffold names from the Cavpor3.0 genome build on the UCSC genome browser), with full-length ORFs for *gag*, *pol* and *env*. A full length *pol* ORF was identified at six other guinea pig loci, however these all had stop codons or frameshifts in either *gag* or *env*. The positions of ORFs within scaffold_11 and scaffold_9 the loci are shown in Table 19, along with the positions of putative conserved domains (identified using the NCBI conserved domain search at <http://www.ncbi.nlm.nih.gov/Structure/cdd/wrpsb.cgi>). The layouts of the two loci are depicted in Figure 55. The scaffold_9 locus has sufficiently long *gag*, *pol* and *env* ORFs to potentially code for functional proteins and contains the majority of conserved domains expected in a retrovirus. The LTRs of this locus differ by 4 bp out of 328, giving an estimated age of 678,000 years. However, in gammaretroviruses *gag* and *pro* are usually separated by a stop codon and at this locus they are not (Goff, 2007). There is also a gap between the stop codon at the end of *pol* and the beginning of *env*, which is not typical of an active retrovirus. This appears to be the result of a truncated *pol* protein. The scaffold_13 locus has a more typical structure, with *gag* and *pro* separated by a stop codon and *pol* and *env* separated by a stop codon and frameshift. The LTRs differ by four bases out of 1044, giving an estimated age of 213,000 years. These dates are very recent compared to most ERVs and these insertions could have the potential to produce functional viral particles and to propagate within the genome.

As six loci in guinea pig had the same *pol* gene ORF, the K_a/K_s ratio was calculated for these sequences across this gene, to establish if it may have been co-opted for a function in the host. The mean K_a/K_s over all six loci was 0.543, which suggests purifying selection preventing changes to this gene over time. K_a/K_s was lowest for the intact scaffold_13 locus. This suggests selection to maintain viral function and may mean that the divergence between these loci

occurred while the virus was circulating exogenously and that there has been little time for the host to counteract the viral activity.

The *pol* genes from both these loci cluster within the south American rodent group in Figure 54, close to each other and to CporP.1908. Guinea pigs have a considerably higher copy number of ERVs in this phylogenetic group than other rodents. Where LTRs could be identified flanking these insertions and those of other new world rodents in this phylogenetic group, these were used to date the insertions, as described in section 2.4.6.1. 17 guinea pig loci had estimated integration dates of less than one million years, which was not the case for any members of this group from other new world rodents. Several guinea pig loci had identical LTRs. This evidence combines to suggest that the chinchilla and degu insertions are degraded to some extent but that this virus remains active or was active very recently in guinea pigs. No endogenous guinea pig gammaretrovirus has been described in depth, however there are occasional references in the literature to “guinea pig leukaemia virus” and “guinea pig type C oncovirus”, which is also classified as a gammaretrovirus by the ICTV but does not appear to have been sequenced (Nadel et al., 1967, Opler, 1967, Davis and Nayak, 1977, International Committee on Taxonomy of Viruses, 2002). An updated analysis of this retrovirus is likely to be worthwhile, especially given the widespread use of guinea pigs and guinea pig cells in laboratory work, which could be contaminated by the release of active viral particles. Guinea pigs are also important as a food source and as pets and are often kept alongside other animals, so an understanding of their retroviral activity would be beneficial to assess cross species transmission risks as well as the impacts of any exogenous viruses on the species.

Table 19: Table showing the details of the potentially intact loci in the guinea pig genome.

	Scaffold_9					Scaffold_13				
Name	Start Position	End Position	Relative Start Position	Relative End Position	Length	Start Position	End Position	Relative Start Position	Relative End Position	Length
Total	29570779	29561579	1	9201	9200	13958978	13968902	1	9925	9924
5' LTR	29570779	29570451	1	329	328	13958978	13960022	1	1045	1044
gag-pol	29569402	29564805	1378	5975	4597	13960633	13962134	1656	3157	1501
MA	29569347	29568991	1433	1789	356	13960688	13961044	1710	2066	356
CA	29568759	29568130	2021	2650	629	13961276	13961905	2298	2927	629
pol						13962133	13965788	3156	6811	3655
PR	29567871	29567596	2909	3184	275	13962170	13962415	3192	3437	245
RT	29567325	29566687	3455	4093	638	13962839	13963351	3861	4373	512
RNAseH	29565933	29565499	4847	5281	434	13964117	13964533	5139	5555	416
IN	29565135	29564863	5645	5917	272	13964900	13965253	5922	6275	353
env	29564237	29562253	6543	8527	1984	13965788	13967792	6811	8815	2004
TM	29562736	29562503	8044	8277	233	13967313	13967540	8335	8562	227
3' LTR	29561907	29561579	8873	9201	328	13967858	13968902	8881	9925	1044

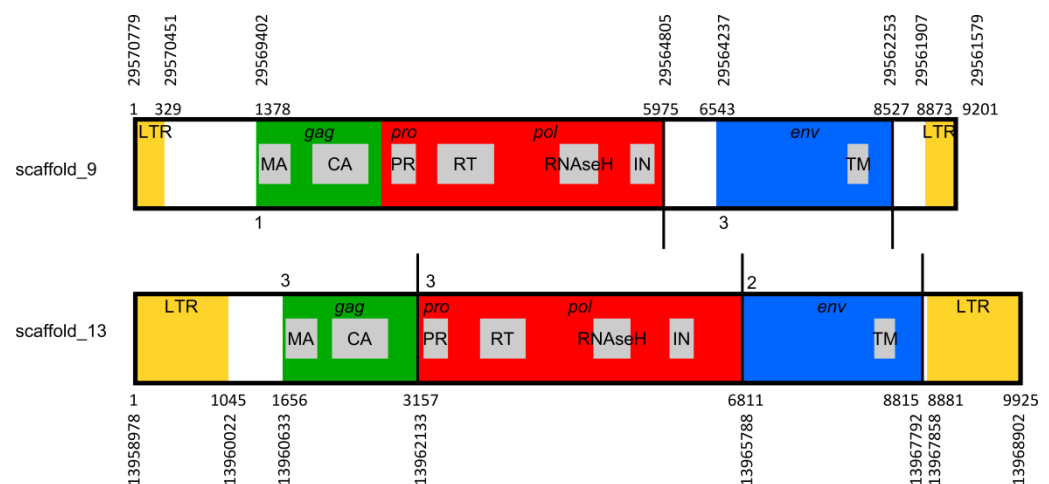


Figure 55: The structure of the two potentially intact gammaretroviral loci identified in the guinea pig genome.

LTRs are shown in yellow, *gag* in green, *pol* in red and *env* in blue. Grey boxes represent conserved domains. Vertical lines represent stop codons. Actual positions in the scaffold are shown vertically and relative positions within the locus are shown horizontally.

17 recombinant loci were identified, with *gag* and *pol* clustering with REV and HERV-T and *env* clustering with the betaretroviruses. Sixteen of these loci

were identified in guinea pigs and one in chinchillas. Separate phylogenetic analyses were performed for the *gag*, *pol* and *env* genes of these 17 loci (Figure 56). In the *gag* and *pol* gene analyses, the guinea pig and chinchilla sequences formed a clear, robustly supported phylogenetic cluster, with HERV-T as the most closely related known virus. The chinchilla sequences clustered outside of the guinea pig group. Guinea pig sequences fell into two well supported groups, marked as group 1 and group 2 in Figure 56. The *env* gene analysis also contained a cluster of guinea pig and chinchilla sequences with the chinchilla sequence as an outgroup, but one guinea pig sequence fell elsewhere in the tree. Guinea pig group one was still apparent but group two was not.

Potential LTRs were identified flanking all 17 of these recombinant loci. These were used to approximately date the insertions and gave dates ranging from four to 15 million years ago. These loci were screened for ORFs and the longest was a 442 amino acid *gag* gene fragment found at one locus in guinea pig. There is no evidence that these recombinant ERVs could produce functional viral particles. The estimate of four to 15 million years ago overlaps with the period during which the guinea pig genus (*Cavia*) diverged into the modern guinea pig species, which occurred between 6.2 and 0.4 million years ago (Dunnum and Salazar-Bravo, 2010). Therefore, these insertions may be present in all species of *Cavia* or in a subset of these species, depending on the exact integration dates.

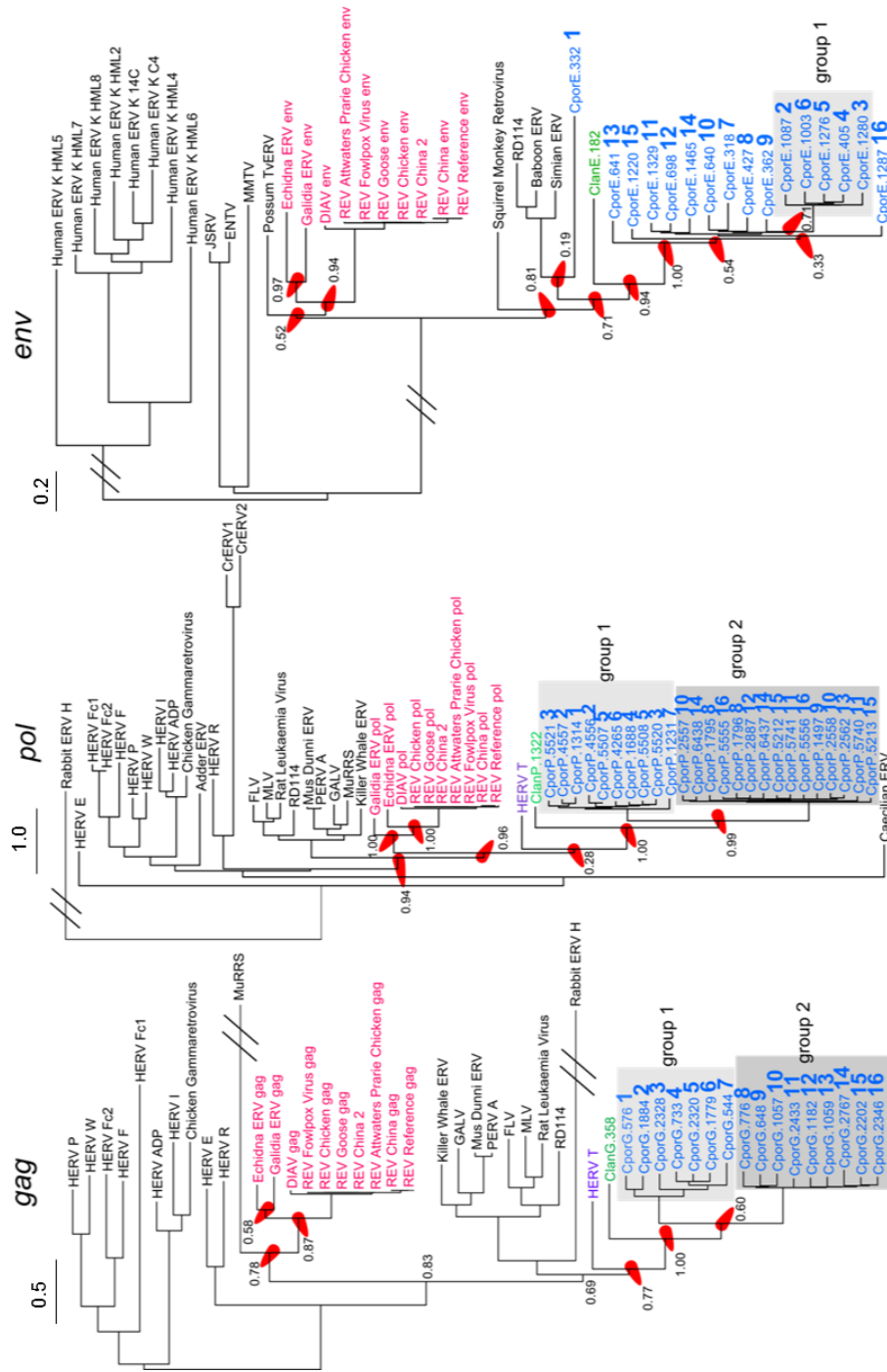


Figure 56: *Gag*, *pol* and *env* gene phylogenies of the REV/HERV-T like full length recombinant insertions identified in the guinea pig and chinchilla genomes. REV is shown in pink, HERV-T in purple, guinea pig in blue and chinchilla in green. In dense regions red arrows connect the node being described with its branch support value. Groups are shaded in grey. Details of previously known sequences are provided in Appendix B.2.

Although tarsier insertions were abundant in this group, they were not analysed in depth due to the short contig length of the currently available tarsier genome build, which rarely allows more than one ERV gene to be identified at a locus. As recombination is fundamental to this group, it is not possible to fully characterise the tarsier insertions at this point.

4.2.5. MLV-Like Group

The insertions identified using Exonerate in the MLV-like group are described in Table 20. This group is discussed in depth in Chapter 6.

Table 20: The number of MLV-like ERV fragments identified in each host type.
All: all 33 genomes. Blue rows labelled “T” represent total counts, yellow rows labelled “PG” represent the mean number of fragments per genome of this type.

		All	Gag	Pol	Env
All	T	24177	9207	11450	3520
	PG	733	279	347	107
Primates	T	9029	5611	2885	533
	PG	602	374	192	36
Rodents	T	14115	3409	7867	2839
	PG	1283	310	715	258
Lagomorphs	T	237	54	126	57
	PG	119	27	63	29
Tree Shrews	T	492	73	345	74
	PG	246	37	173	37
Birds	T	4	1	2	1
	PG	2	1	1	1
Ferret	T	300	59	225	16
	PG	300	59	225	16

4. 3. Epsilonretroviruses

Unexpectedly, epsilon-like *pol* genes were relatively abundant in primates, with 821 insertions identified, distributed across all primate and tree shrew genomes screened (Figure 57). Nine insertions were identified outside the primates: one in the ground squirrel and eight in the ferret. The ground squirrel insertion clustered with the gammaretroviruses on further analysis. Ferret insertions were not analysed in detail as they are outside the focus of this project. The 821 epsilon-like *pol* gene fragments are discussed in depth in o.

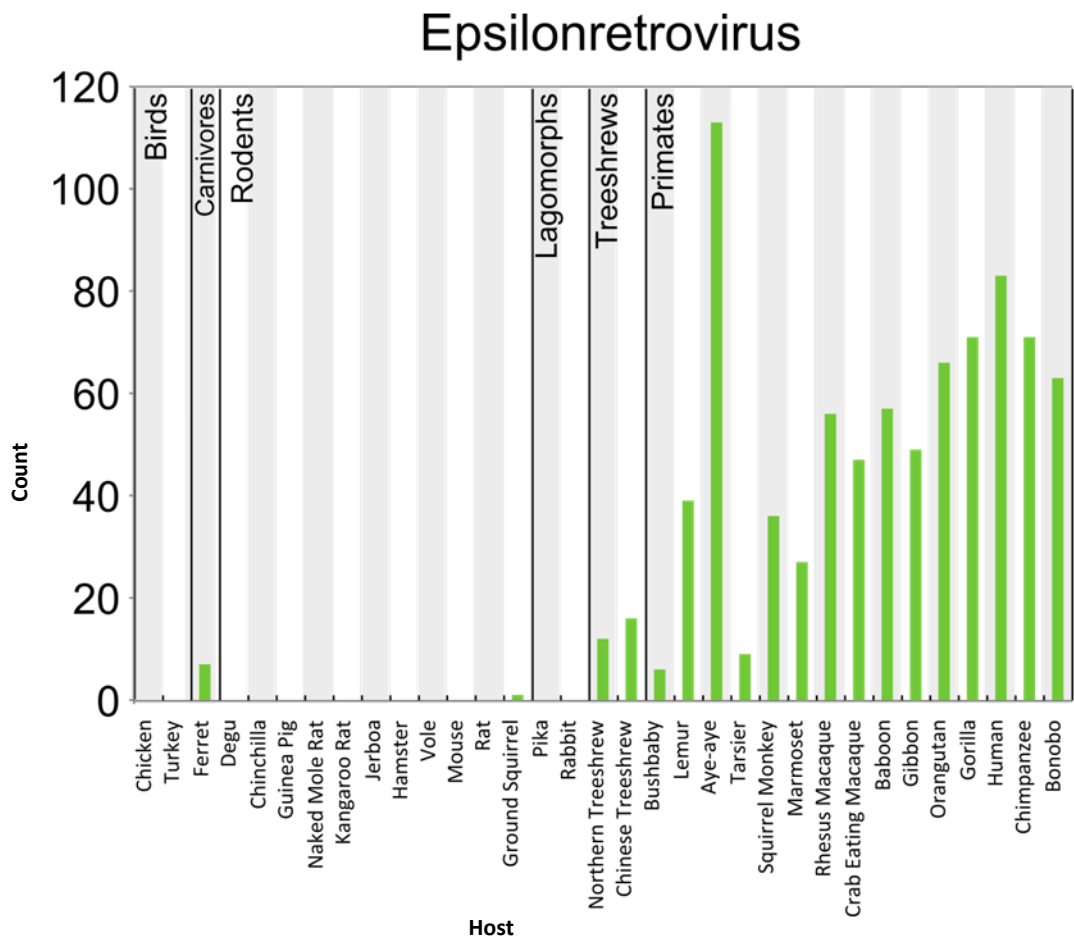


Figure 57 The distribution of ERV fragments between genomes for the epsilonretrovirus genus. *gag* fragments are represented in red, *pol* in green and *env* in blue.

4. 4. Spumaviruses

Spumavirus-like *gag* and *pol* genes were found in all hosts. The vast majority of these (29096 out of 29109) were ERV-L like elements, which are known to lack *env* (Benit et al., 1999). ERV-L insertions are known to be degenerate and inactive, so were not characterised in detail. However, several details about this group were noteworthy. First, there was a noticeable lack of ERV-L like elements in some species, particularly the kangaroo rat (3 elements) and pika (13 elements) (Figure 58). Secondly, the number of ERV-L elements in new world monkeys, old world monkeys and apes was much higher and more consistent than suggested in Benit et al. (1999).

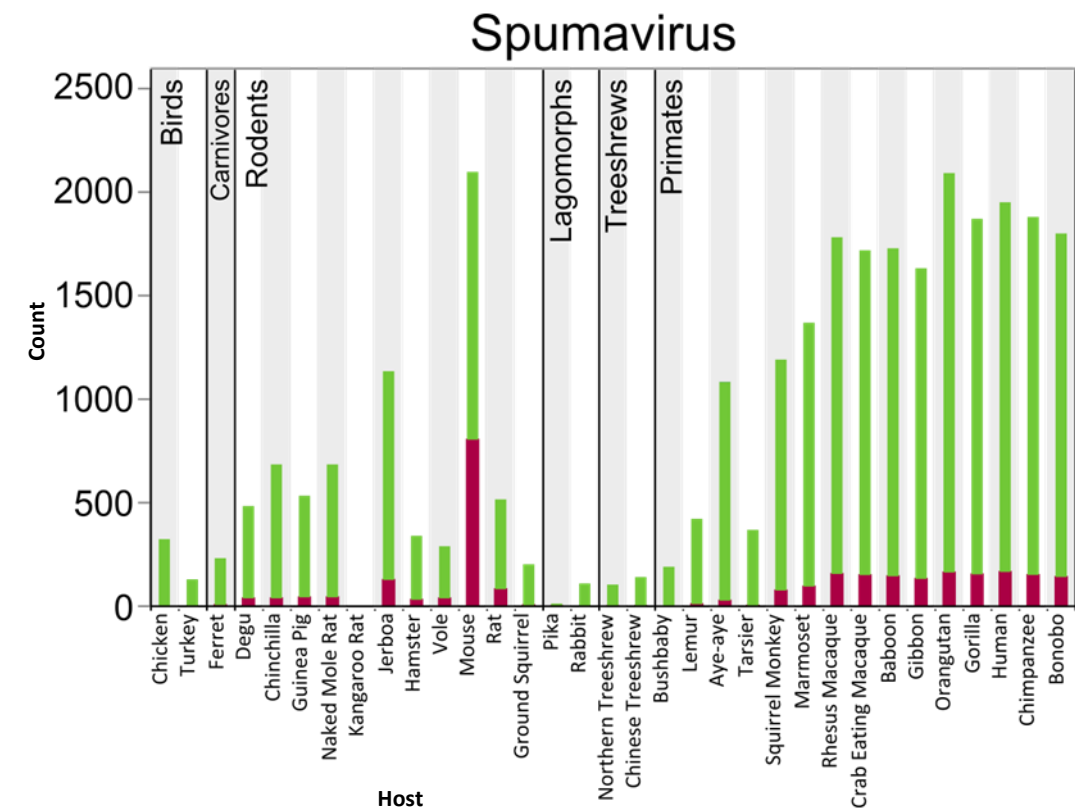


Figure 58: The distribution of ERV fragments between genomes for the spumavirus genus.
gag fragments are represented in red and *pol* in green.

4. 5. Alpharetroviruses

308 alpharetrovirus-like fragments were identified in preliminary analysis, 287 of which were in birds (Figure 59). The non-bird fragments clustered with the betaretroviruses in more detailed analyses (data not shown) so were not considered in depth. Avian alpharetroviruses have been well characterised by other groups [e.g. (Bolisetty et al., 2012)] so will not be considered here.

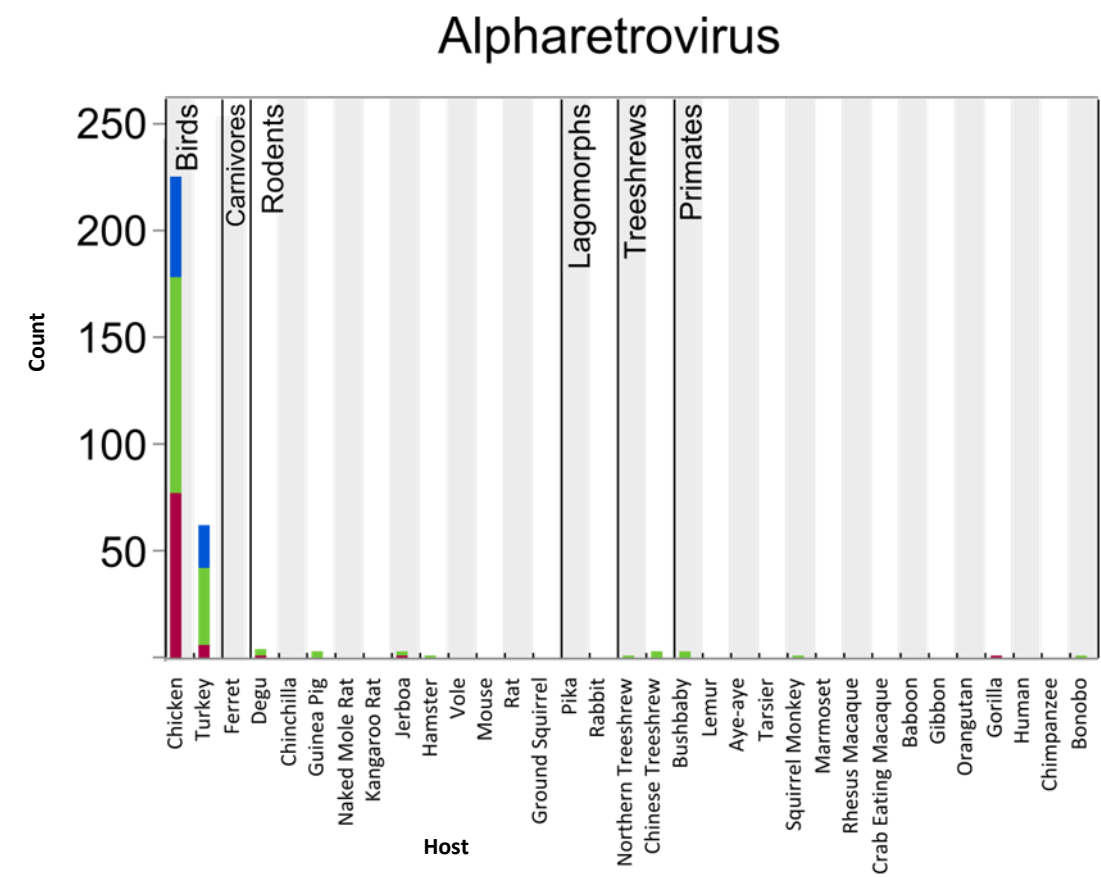


Figure 59: The distribution of ERV fragments between genomes for the alpharetrovirus genus.
gag fragments are represented in red, *pol* in green and *env* in blue.

4. 6. Betaretroviruses

In total, 52,415 betaretrovirus-like fragments were identified across the 33 genomes, 12,911 *gag*, 32,957 *pol* and 6,547 *env*. Figure 60 shows the distribution of these fragments between hosts. Primate betaretroviruses were somewhat less abundant than gammaretroviruses (except in the tarsier) and 38% of the betaretrovirus insertions were in primates, compared to 62% of gammaretroviruses.

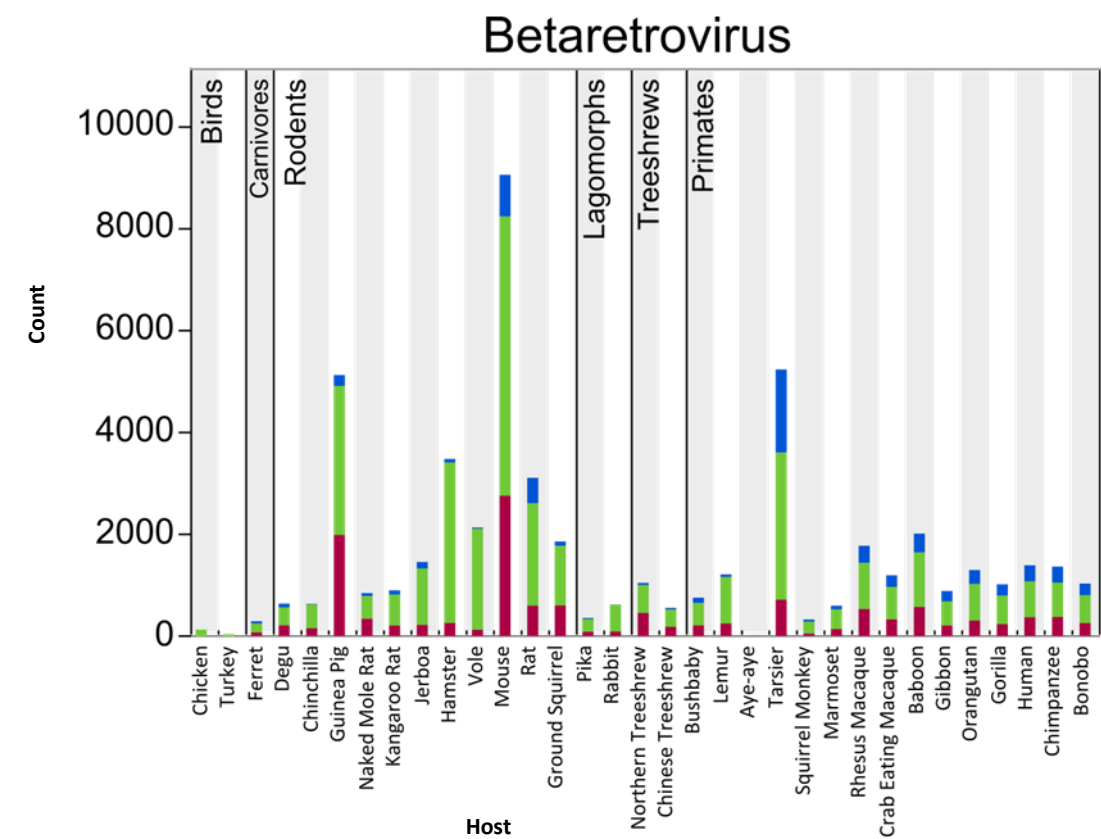


Figure 60: The distribution of ERV fragments between genomes for the betaretrovirus genus.
gag fragments are represented in red, *pol* in green and *env* in blue.

To our knowledge, no comprehensive genome scale analysis of mammalian betaretroviruses has been performed to date. 99.99% of betaretroviral *pol* genes identified using Exonerate fell into seven groups in the GROUPED_EXO

dataset: simian retrovirus (SRV)-like, simian endogenous retrovirus (SERV)-like, MusD-like, HERV-K-like, JSRV-like, hedgehog ERV-like and IAP-like. These betaretroviruses are described in section 1.4.7. One sequence from each of these groups was combined into the phylogeny shown in Figure 61. The sequences can be divided into four groups, outlined in Figure 61. Of the 194 previously known betaretroviral *pol* gene sequences in the FULL_PREVKNOWN dataset, 179 fell into one of these four groups. Eleven of the remaining 15 sequences were from marsupials, which may therefore harbour a fifth group of endogenous betaretroviruses, however this falls outside of the scope of this thesis. These four groups provide a good representation of the diversity of the betaretroviruses in the placental mammals.

We therefore propose classifying the betaretroviruses of these hosts into these groups, provisionally named according to a well-characterised sequence within the group: HERV-K like, SERV-like, JSRV-like and IAP-like. New sequences have been identified in each of these groups and each group will be discussed below.

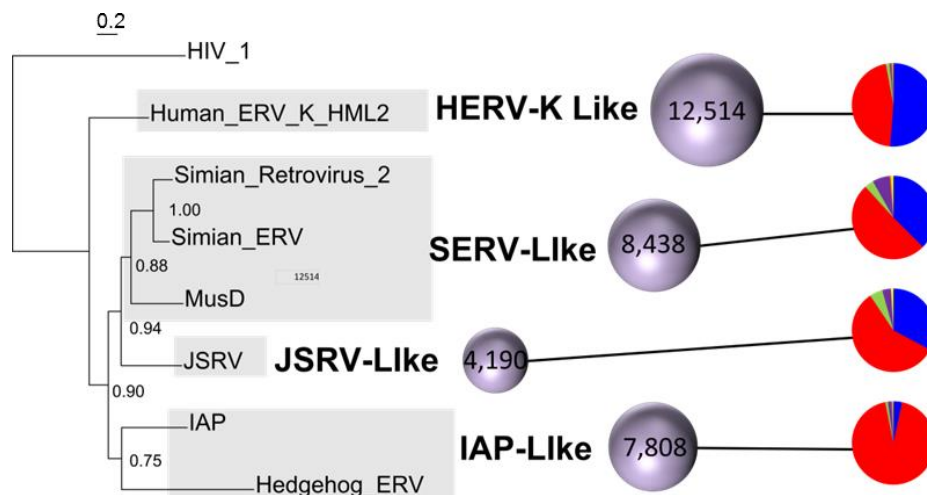


Figure 61: *Pol* gene phylogeny showing the four proposed groups of betaretroviruses in the Euarchontoglires.

Groups in the phylogeny are highlighted in grey. Purple circles are sized according to, and labelled with, the total number of new *pol* gene fragments identified in this group. Pie charts show the proportion of these fragments found in primates (blue), rodents (red), lagomorphs (green), tree shrews (purple), birds (orange) and ferrets (yellow).

4.6.1. HERV-K-Like Group

The HERV-K like group consists of the subgroups of known HERV-K insertions in the human genome, of which eight have previously been described in detail and Baillie et al.'s (2004) β 1 group of mouse and rat betaretroviruses. HERV-K like ERVs have previously been characterised extensively in the human genome but have often not been examined in detail in other hosts.

Besides the known human ERVs, several other previously described ERVs fell into this group in the analysis described in section 2.1.4.1. These were Polavarapu et al.'s (2006a) chimpanzee ERVs 30 to 39, all of Romano et al.'s (2006) chimpanzee HERV-Ks, Garcia-Etxebarria et al.'s (2010) bovine ERVs 21 to 24, Gifford et al.'s (2005) rice rat and shrew mouse ERVs, McCarthy et al.'s (2004) murine ERVs 13, 15, 17 and 18 and Wang et al.'s (2010) rat ERV K.

Using our pipeline, 22,506 HERV-K like ERV fragments were identified and are detailed in Table 21. 82% of primate insertions could be unambiguously assigned to one of the eight HERV-K groups. Rodent sequences formed a separate cluster, distinct from the primate HERV-Ks.

Table 21: The number of HERV-K-Like ERV fragments identified in each host type.

All: all 33 genomes. Blue rows labelled “T” represent total counts, yellow rows labelled “PG” represent the mean number of fragments per genome of this type.

		All	Gag	Pol	Env
All	T	22506	5821	12514	4171
	PG	1500	388	834	278
Primates	T	12050	3112	6418	2520
	PG	803	207	428	168
Rodents	T	9846	2546	5725	1575
	PG	895	231	520	143
Lagomorphs	T	239	43	178	18
	PG	120	22	89	9
Tree Shrews	T	192	88	81	23
	PG	96	44	41	12
Birds	T	76	1	75	0
	PG	38	1	38	0
Ferret	T	103	31	37	35
	PG	103	31	37	35

For primates, a presence/absence analysis (section 2.4.1) was performed for each of the eight lineages of HERV-K: HML-1 to HML-8 (section 1.4.7.1). The results of this analysis are summarised in Table 22 and the phylogenies resulting from the analysis are shown in Figure 62. Figure 62 provides *pol* gene phylogenies for the most similar sequence found in each host to each HERV-K reference sequence, where a related sequence was present, generated using the technique described in section 2.4.1. Figure 63 shows how the HERV-K lineages were distributed amongst primate hosts, their estimated integration dates based on this distribution and their approximate copy number in each primate.

Table 22: The presence or absence of each HERV-K lineage (HML-1 to HML-8) in each primate group.

Fully shaded cells signify that the lineage is present in all hosts in the group, half shaded cells signify that the lineage is present in some hosts in the group and absent in others.

	1	2	3	4	5	6	7	8
Apes								
Old World Monkeys								
New World Monkeys								
Prosimians								

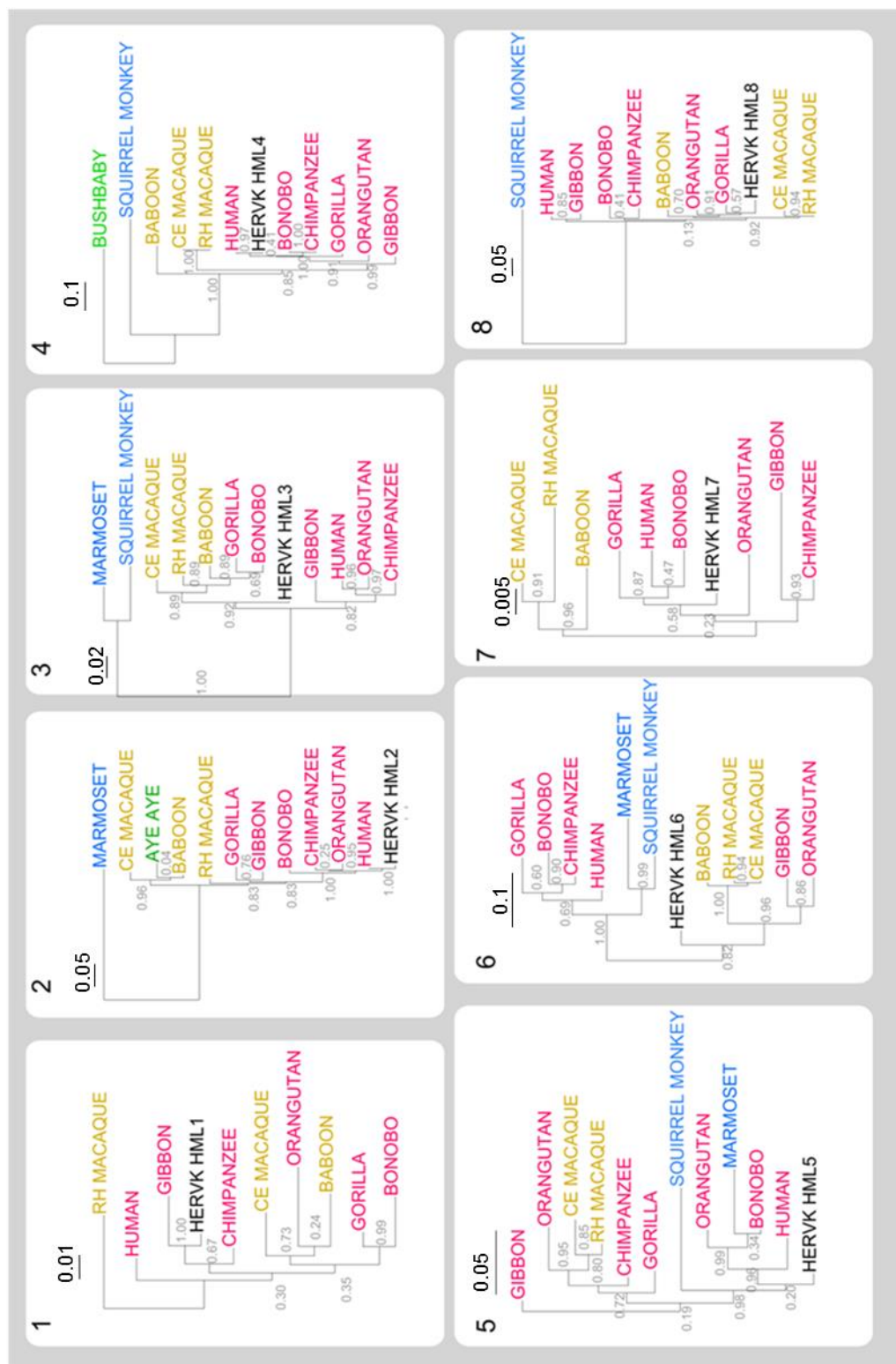


Figure 62: *Pol* gene phylogenies of the most similar insertion to each HML type sequence identified in each host where an insertion clustering with the type sequence was identified.

Apes are shown in pink, old world monkeys in yellow, new world monkeys in blue, prosimians in green.

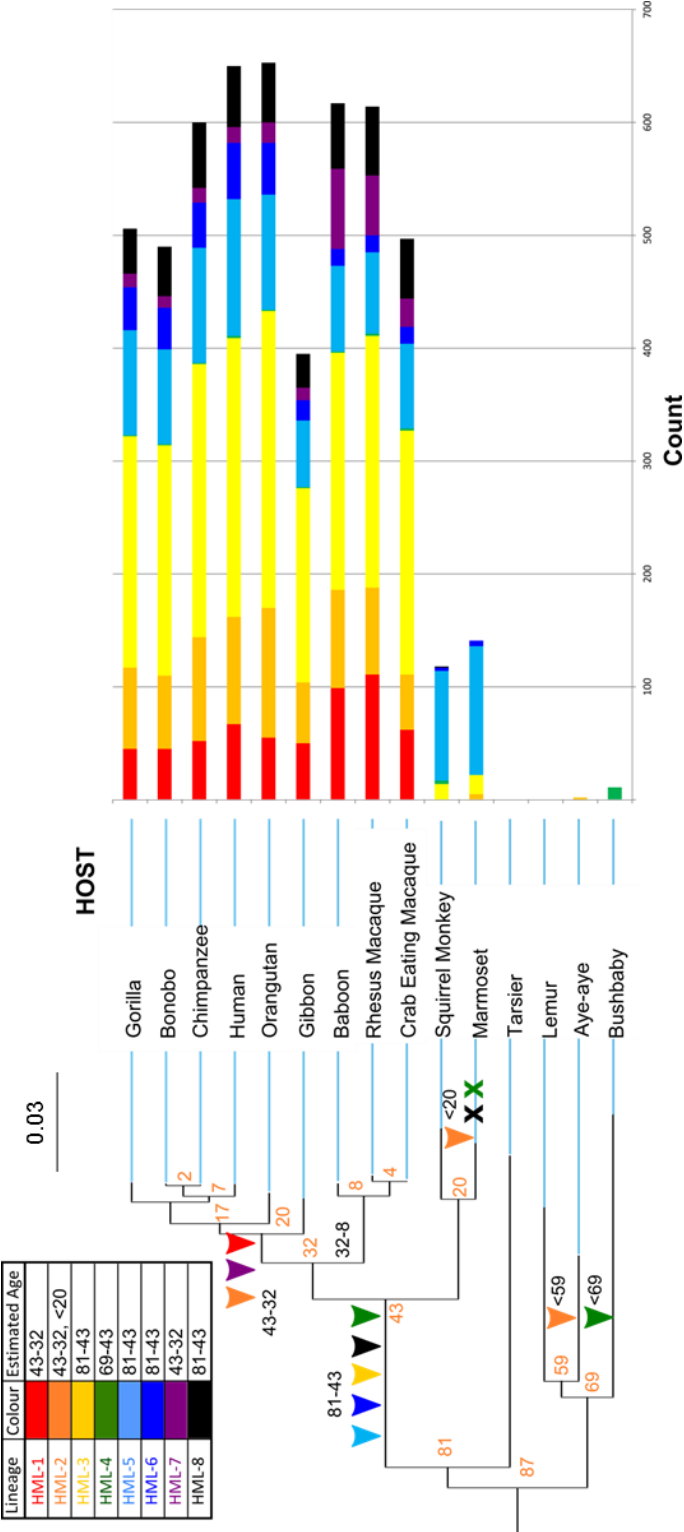


Figure 63: The copy number and estimated integration dates of the HERV-K family of ERVs in primates.

Arrows represent estimated integration dates (million years ago). X's represent loss of an ERV from a lineage. Arrows and bars are colour coded according to the key and represent different lineages in this family. Orange numbers are estimated node ages in millions of years.

The only HERV lineage which has previously been identified in new world monkeys is the HML-5 group. (Greenwood et al., 2005). However, here representatives of HML-2, HML-3, HML-4, HML-5, HML-6 and HML-8 were identified in new world monkeys.

In order to clarify the evolutionary history of these ERVs, the Compara six primate alignment was used to trace the orthologous positions of insertions in marmoset (a new world primate) with those in old world primates (rhesus macaque, orangutan, gorilla, chimpanzee and human) using the less computationally intensive locus-by-locus technique described in section 2.4.6.3. Sixteen of the marmoset insertions (out of 179 candidates) were found in regions covered by the Compara alignment and with an ERV in at least one other host. For each of these regions, the ERV sequences identified were aligned to the marmoset sequence from that locus and to the type sequences for HML-1 to HML-8 and phylogenetic trees were generated. At two loci, both HML-5, there was evidence that the insertion was orthologous in marmoset and in old world primates, as the phylogeny was consistent with that of the host species and all ERV sequences fell within a single, strongly supported phylogenetic cluster. Phylogenies for these loci are shown in Figure 64. This result suggests that at least some HML-5 loci appeared before the divergence of new and old world primates. No similar relationships were found for HERV-K lineages other than HML-5.

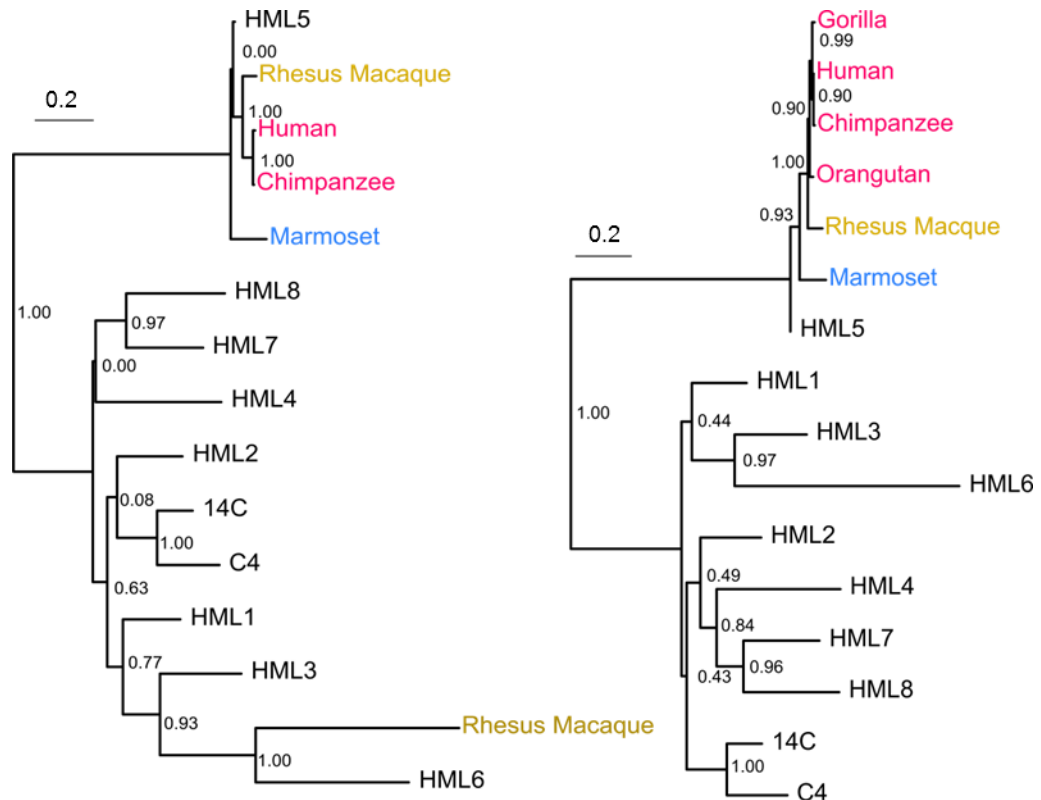


Figure 64: *Pol* gene phylogenetic trees for the two HML-5 loci which appear to predate the divergence of new and old world primates.

Apes are shown in pink, old world monkeys in yellow, new world monkeys in blue, type sequences for HERV-K lineages in black. Details of previously known sequences are provided in Appendix B.2.

No orthologous insertion sites were apparent for HML-2, HML-3, HML-4, HML-6 or HML-8. However, as Figure 63 demonstrates, there appears to have been a very significant copy number increase in all of these lineages in the old world primates compared to the new world primates, so a small number of shared loci could be present which have not been detected, especially given that the Compara alignment does not cover all of the marmoset genome. Conversely, HML-5 has a large number of insertions in the new world primates.

HML-3 and HML-6 were both detected in all simian primates (Table 22) at a low copy number. However, their phylogeny shows inconsistent host tracking, which suggests they are more modern than the common ancestor of the simian primates (estimated at 43 million years ago). Mayer and Meese characterised

HML-3 in humans and found that the virus appears to have been active approximately 36 million years ago, which is close to the date of the divergence of new and old world primates. Combining this results with our phylogenetic analysis, a small number of insertions may predate the old world monkey – new world monkey divergence but there has also been substantial activity since then in the old world primates. HML-6 shows very similar patterns. HML-6 insertions in old world primates have been shown to be at least 30 million years old (Medstrand et al., 1997) but our phylogeny suggests that members of this lineage have been active more recently.

HML-2, HML-4 and HML-8 were each identified in one of the two species of new world monkey screened here but not the other. As these two species share a common ancestor with each other more recently than with any of the other screened species, this suggests that either these integration events occurred after squirrel monkey and marmoset diverged approximately 20 million years ago or that the lineage has become unrecognisable in one of its new world monkey hosts.

HML-2 is considered to be ancient, having entered primate genomes approximately 35 million years ago, before the divergence of old world monkeys and apes (Bannert and Kurth, 2006). However, the structure of the HML-2 tree in Figure 62 suggests these viruses have also been active more recently, as no host tracking is evident. These viruses are known to have been active recently in the human genome (Shin et al., 2013). We have identified novel HML-2 like insertions in the marmoset and the aye-aye. The absence of these insertions in squirrel monkey, tarsier, lemur and bushbaby means the most parsimonious explanation is integration into the marmoset and aye-aye genomes after their divergence from ancestors shared with the other hosts screened here. Marmoset HML-2 like insertions were very distinct from those in old world monkeys and apes and are much less numerous. One marmoset locus had recognisable flanking LTRs, which gave an approximate integration date of 26 million years ago, consistent with a separate integration event to the

HML-2 loci shared between the old world primates, although possibly within a similar time period. The aye-aye insertion is seemingly modern and is very similar to those found in apes. None of the aye-aye loci had recognisable LTRs, so their age could not be estimated using this method.

Three copies of HML-4 were identified in the squirrel monkey but none in marmoset. The phylogeny of HML-4 shows strong evidence of host tracking, with relationships identical to those seen in the host, including for the bushbabies insertions, suggesting an ancient origin. However, the most recent common ancestor of bushbabies and the other hosts with HML-4 is shared with the other prosimians, which lack HML-4. The most parsimonious explanation is that HML-4 circulated 43 to 69 million years ago and entered the common ancestor of the old and new world primates and the ancestor of bushbabies and the low copy number insertions in marmoset are no longer recognisable. However, Seifarth et al. (1998) used Southern blotting to identify HML-4 in apes and old world monkeys but not new world monkeys. Only one species was screened, an *Aotes* night monkey. Night monkeys are closely related to marmosets (Figure 12), which also lacked HML-4 in our analysis. Therefore, this lineage could have been lost in the ancestor of the night monkey/marmoset/tamarin clade in the primate phylogeny but maintained in the lineage leading to squirrel monkeys and capuchins.

HML-8 has not been discussed in detail previously, however it is generally cited as having appeared in old world monkeys and apes after their divergence from new world monkeys (Bannert and Kurth, 2006). Here, one HML-8 like insertion was identified in squirrel monkeys and none in marmosets. The phylogeny of the HML-8 like insertions identified here is very similar to that of HML-2, suggesting an ancient integration into an ancestor of the new world and old world primates but also more recent activity in old world monkeys and apes. The single HML-8 insertion identified in the squirrel monkey is very degraded, with *pol* and *env* fragments but no recognisable LTRs or ORFs, suggesting an ancient origin. As only one insertion was identified and it was so

degenerate, it is very possible that the corresponding insertion in marmoset has deteriorated and that this integration event predates the common ancestor of the old and new world monkeys. However, this lineage also appears to have circulated more recently in old world monkeys and apes.

For HML-1 and HML-7, our results were consistent with the literature in that insertions were identified in old world monkeys and apes but not new world monkeys. The results in Figure 62 provide evidence that HML-7 circulated once, before old world monkeys and apes diverged, integrated into an ancestor of these species and then accumulated mutations at the host mutation rate. HML-1 appears to have circulated more recently, given its random distribution amongst hosts (Figure 62).

Clearly, rodents also have a large group of HERV-K like ERVs (Table 21). However, none of these ERVs clustered closely with any of the eight HERV-K HML lineages. Instead, there appears to be a cluster of ERVs in rodents which is not found in primates. This lineage is particularly abundant in hamsters and in voles, with 2,384 and 1,545 of the 5,725 rodent HERV-K like insertions identified in these hosts respectively. These hosts share a common ancestor approximately 18 million years ago (Figure 37) which is not shared with any of the other hosts screened here, so the expansion is likely to have occurred in this ancestor.

This rodent lineage is an interesting candidate for further analysis, as it has not been characterised in detail before. However, given that no insertions of this type were identified in primates, it is outside the scope of this thesis.

4.6.2. SERV-Like Group

The SERV-like group of ERVs is characterised by recombination between genes. Three “types” of retrovirus are present in this group. SERV, SMRV and TvERV are true endogenous betaretroviruses, with betaretrovirus-like *gag*, *pol* and *env* genes. BaEV and RD114 have gammaretrovirus-like *gag* and *pol* genes and betaretrovirus-like *env* genes. SRVs have betaretrovirus-like *gag* and *pol* genes and gammaretrovirus-like *env* genes. A diverse range of mouse and rat non-recombinant betaretroviruses are also present in this group, namely MMTV and the β 4, β 5, β 6 and β 7 groups characterised by Baillie et al. (2004).

Several further previously known retrovirus sequences were assigned to this group in the analysis described in section 2.1.4.1. These were Gifford et al.’s (2005) small mongoose ERVs, bison ERV, giraffe ERV, slow loris ERV, colobus ERV and musk ox ERV, Garcia-Etxebarria et al.’s (2010) bovine ERV 8 and bovine ERV 20 and McCarthy et al.’s (2004) murine ERV 12.

Table 23 describes the 12,836 SERV-like fragments identified here. For the purposes of this table, only betaretrovirus-like fragments are counted.

Table 23: The number of SERV-like ERV fragments identified in each host type.
All: all 33 genomes. Blue rows labelled “T” represent total counts, yellow rows labelled “PG” represent the mean number of fragments per genome of this type.

		All	Gag	Pol	Env
All	T	12836	3323	8438	1075
	PG	856	222	563	72
Primates	T	4118	627	3158	333
	PG	275	42	211	22
Rodents	T	7055	2032	4285	738
	PG	641	185	390	67
Lagomorphs	T	419	117	301	1
	PG	210	59	151	1
Tree Shrews	T	1075	507	567	1
	PG	538	254	284	1
Birds	T	52	0	52	0
	PG	26	0	26	0
Ferret	T	117	40	75	2
	PG	117	40	75	2

The regions surrounding the ERV fragments listed in Table 23 were screened for any gene of any genus using the methodology described in section 2.4.3.

Non-recombinant SERV-like betaretroviruses were detected in all old world monkeys and in no other hosts, which is consistent with the literature (van der Kuyl et al., 1997). A strongly supported cluster of sequences was identified in these hosts with *gag* and *pol* genes related to the exogenous simian retroviruses and *env* genes clustering with SERV, possum TvERV and SMRV (Figure 65). The sequences were indistinguishable from the SERV reference sequence. Many of the other hosts analysed here have not previously been screened for these insertions but our results confirm that they are likely to be unique to old world monkeys.

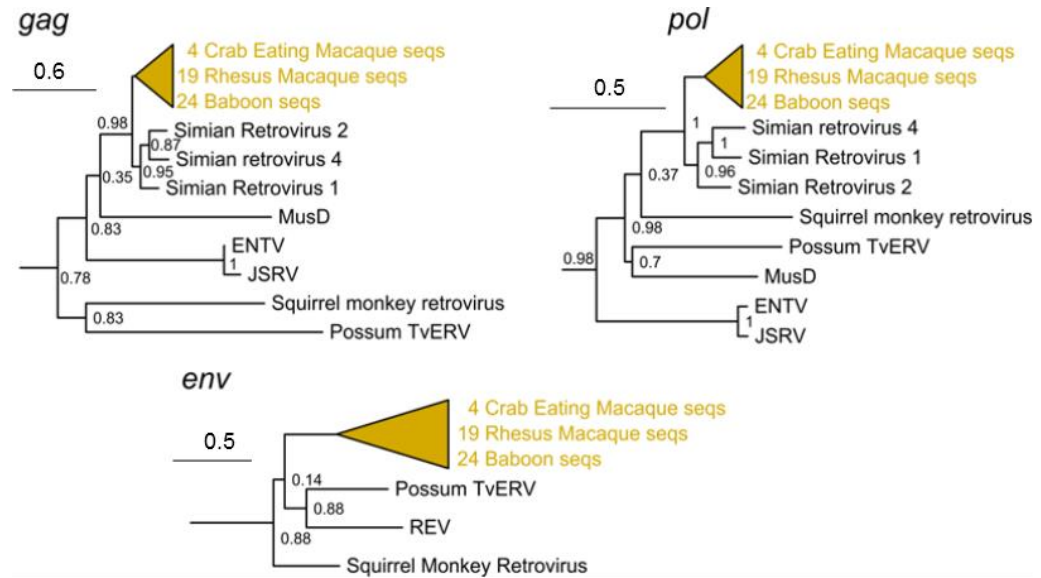


Figure 65: *Gag*, *pol* and *env* gene phylogenies of the SERV-like betaretroviruses identified in old world monkeys and reference betaretroviruses.

Old world monkey sequences are shown in yellow and have been collapsed. These trees are rooted on the appropriate betaretroviral test datasets, which has been cropped for better visualisation. Details of previously known sequences are provided in Appendix B.2.

SRV-like regions with a betaretroviral *gag-pol* and gammaretroviral *env* were also identified in several hosts. *Gag*, *pol* and *env* gene phylogenies for these regions are shown in Figure 66.

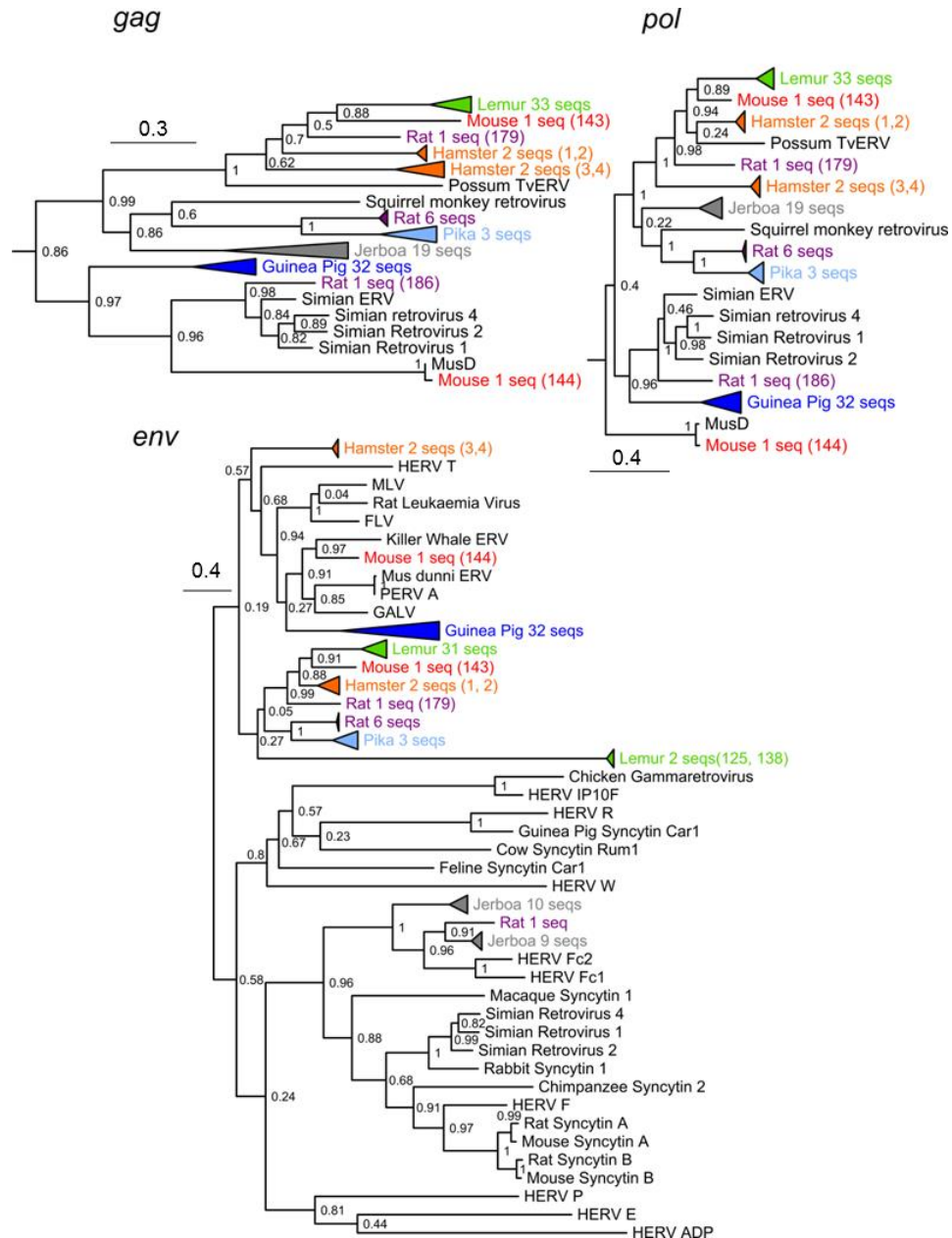


Figure 66: *Gag*, *pol* and *env* gene phylogenies of the SRV-like recombinants with betaretrovirus-like *gag* and *pol* genes and gammaretrovirus-like *env* genes with reference gamma- and betaretroviruses.

Each host is shown in a different colour. Clusters with more than one sequence from the same host have been collapsed. *Gag* and *pol* trees are rooted on the basic betaretrovirus datasets and the *env* tree on the basic gammaretrovirus dataset, these have been cropped for better visualisation. Details of previously known sequences are provided in Appendix B.2.

The betaretroviral *gag* and *pol* genes for these sequences cluster either with TvERV/SMRV, SRV/ SERV or MusD. The gammaretroviral *env* genes cluster

with MLV or SRV. For clarity, these sequences were divided into the five subcategories described in Table 24, referred to here as type I to type V.

Table 24: The four subcategories of SERV-like insertions with betaretroviral *gag-pol* regions and gammaretroviral *env* regions.

Type	Gag-pol (gamma)	Env (beta)
I	TvERV / SMRV	MLV
II	TvERV / SMRV	SRV
III	SRV / SERV	MLV
IV	SRV / SERV	SRV
V	MusD	MLV

Table 25 lists the number of regions of each type identified in each host and any previous description of these regions.

All of these insertions were screened for flanking LTRs and LTR pairs were identified around 38 of the 100 loci, encompassing all types except type IV. Using the methodology described in section 2.4.6.1, these LTRs gave the range of potential integration dates listed in Table 25. The majority of insertions were estimated at six to eight million years old. The rat type I insertions were somewhat more modern, approximately 3 million years old, and the pika type I insertions more ancient, approximately 15 million years old. The pika date is based on a single LTR pair so may not be reliable. All of these integration dates are after the host species diverged from each other, so integration into each host is likely to have been independent. Accordingly, the presence and absence of these insertions is not consistent with the phylogeny of their hosts, for example insertions were found in rat, mouse, hamster and jerboa but not vole, which contradicts the phylogeny shown Figure 37.

All 100 insertions were also screened for ORFs and potentially intact ORFs were identified in lemur type I, rat type I, guinea pig type III and mouse type IV insertions (Table 25). Figure 67 shows the structure of the most intact of

each of these insertion types and Table 26 provides details of their location. All except the mouse type IV insertion are sufficiently intact that they may have the potential to encode active viruses. In particular, the guinea pig and rat insertions have few defects, although the gaps between *pro* and *pol* may be the result of deterioration.

Table 25: The number of SRV-like insertions of each type identified in each host with *gag* and *pol* genes clustering with the betaretroviruses and *env* genes with the gammaretroviruses, any previous references to these insertions, the estimated age of the insertions and the length of the longest ORF.

Host	Type	Previous Description	Count	LTRs	Min Age	Max Age	Mean Age	Longest ORF
lemur	I	M. murinus_ERV-β4 (Baillie et al., 2004)	32	7	768,049	16,697,025	6,687,968	878
mouse	I	MmERV β4, MmERV β5 (Baillie et al., 2004)	1	0				682
hamster	I		4	3	3,035,823	8,318,479	6,104,953	567
rat	I	RnERV β4, RnERV β5 (Baillie et al., 2004)	7	6	214,087	9,841,270	2,950,730	936
pika	I		3	1	15,772,871	15,772,871	15,772,871	338
jerboa	II		19	1	6,880,734	6,880,734	6,880,734	456
guinea pig	III		32	19	500,501	15,799,016	6,657,738	875
rat	IV	RnERV β6	1	0				302
mouse	V	MmERV β7	1	1	7,335,491	7,335,491	7,335,491	868

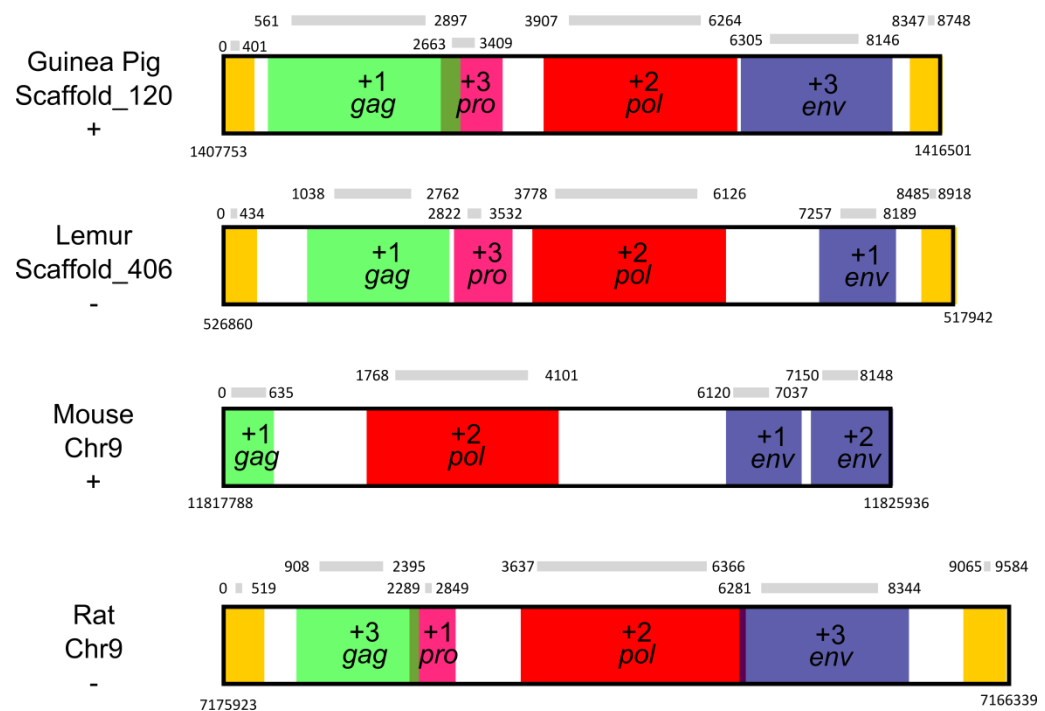


Figure 67: The structure of the most intact SRV-like insertions identified with betaretrovirus-like *gag* and *pol* genes and gammaretrovirus-like *env* genes. Numbers above diagrams represent the relative position of each part of the genome and numbers below the absolute position in the genome. Numbers on the diagrams represent reading frames, if the first base of the 5' LTR were in the +1 frame.

Table 26: The position of the most intact SRV-like insertions with betaretroviral *gag* and *pol* genes and gammaretroviral *env* genes in their host genomes. Scaffold numbers are from the genome builds listed in Table 8.

Host	Region	Chromosome / Scaffold	Strand	Absolute Start	Absolute End	Relative Start	Relative End	Length
Guinea Pig	All	scaffold_120	+	1407753	1416501			8748
	LTR			1407753	1408154	0	401	401
	<i>gag</i>			1408314	1410650	561	2897	2336
	<i>pro</i>			1410596	1411342	2663	3409	746
	<i>pol</i>			1411660	1414017	3907	6264	2357
	<i>env</i>			1414058	1415899	6305	8146	1841
	LTR			1416100	1416501	8347	8748	401
Lemur	All	scaffold_406	-	517942	526860			8918
	LTR			526860	526426	0	434	434
	<i>gag</i>			525822	524098	1038	2762	1724
	<i>pro</i>			523567	522857	2822	3532	710
	<i>pol</i>			523082	520734	3778	6126	2348
	<i>env</i>			519603	518671	7257	8189	932
	LTR			518375	517942	8485	8918	433
Mouse	All	chr_9	+	11817788	11825936			8148
	<i>gag</i>			11817788	11818423	0	635	635
	<i>pol</i>			11819556	11821889	1768	4101	2333
	<i>env</i>			11823908	11824825	6120	7037	917
	<i>env</i>			11824938	11825936	7150	8148	998
Rat		chr_9	-	7166339	7175923			9584
	LTR			7175923	7175404	0	519	519
	<i>gag</i>			7175015	7173528	908	2395	1487
	<i>pro</i>			7173392	7172832	2289	2849	560
	<i>pol</i>			7172286	7169557	3637	6366	2729
	<i>env</i>			7169642	7167579	6281	8344	2063
	LTR			7166858	7166339	9065	9584	519

The type I lemur group was previously identified by Baillie et al. (2004) as a single insertion in *Microcebus murinus*. Our results confirm this and show that at least 30 copies of this ERV are present in this host. These insertions are not shared by the closest host to the lemur screened here, the aye-aye, so are unlikely to be more than 59 million years old. Judging by the LTR divergence

and the degree to which the ORFs are intact in the lemur group, they may be considerably more modern than this, having appeared in the last 7 million years. However, only one insertion, the scaffold_406 insertion described in Figure 67, showed any potential to produce active virus. Screening of further species of lemur would clarify the integration date of this ERV, as an insertion occurring 15 to two million years ago would be shared amongst *Microcebus* lemurs while a more modern insertion would be unique to *M. murinus* (Perelman et al., 2011). The date at which non-native rodents reached Madagascar, the only geographical region where *M. murinus* is present, is currently unclear (Samonds et al., 2013). A more precise integration date of this rodent-like virus in a Malagasy lemur may help clarify this. The type I hamster and pika ERVs have not been characterised previously to our knowledge, while the mouse and rat insertions were also listed by Baillie et al. (2004).

The type III jerboa insertions identified here have not been characterised previously and appear to be unique in the relationship of their *gag*, *pol* and *env* genes, with *gag* and *pol* clustering with SMRV/TvERV and *env* with SRV. Therefore, a member of the gammaretrovirus group which recombined with SERV to produce the *env* genes of SRV has also recombined with an SMRV like betaretrovirus to produce these insertions.

Only guinea pig ERVs were identified as type III with SRV/SERV like *gag* and *pol* genes but *env* genes clustering closely with the MLV clade. As discussed in section 4.2.4, guinea pigs are known to produce defective viral particles, and these relatively intact insertions could again be responsible. As guinea pigs appear to contain diverse recombinant ERVs, it is also possible that these particles are produced by more than one type of ERV.

The rat type IV and mouse type V insertions have previously been described in detail (Baillie et al., 2004) so will not be characterised further here.

The diversity of these SRV like insertions in rodents in particular and the relatively modern origin of these insertions suggests that rodents may have been involved in the appearance of exogenous SRV and of endogenous SERV. SERV is not considered to be particularly ancient, estimated at 12 to 18 million years old (van der Kuyl et al., 1997), which overlaps with the range of estimated integration dates of the rodent insertions shown in Table 25. Therefore, these viruses may have been circulating simultaneously. These rodent SRV-like ERVs are found in hosts which are geographically widespread and often coexist with other species and, as the lemur insertions demonstrate, are able to survive and proliferate in primates, so rodent SRVs as intermediates in the SRV-SERV-BaEV evolutionary history are likely.

BaEV/RD114 like insertions, with betaretroviral SERV-like *env* genes but gammaretroviral MLV-like *gag* and *pol* genes were less widespread. Sequences clustering closely with BaEV in all three genes were only identified in baboons (Figure 68). This is as expected, as BaEV has only previously been detected in baboons, geladas, mangabeys, mandrills and African green monkeys (van der Kuyl et al., 1995), and baboons are the only species in this group for which a full-genome sequence is currently available. This result confirms the absence of BaEV in their closest sequenced relatives, the macaques, as discussed by van der Kuyl et al. (1995).

Recombinant viruses with SMRV-like *env* genes were also detected in guinea pigs (Figure 68). These insertions had MLV-like *gag* and *pol* genes clustering separately to BaEV. Preliminary analysis also showed candidate recombinant sequences in the marmoset and bushbaby. However, the *env* gene of the bushbaby sequence did not consistently cluster as betaretroviral (Figure 68). Similarly, the *gag* and *pol* genes of the marmoset insertion could not be robustly classified as gammaretrovirus-like. These two insertions were therefore excluded from further analysis.

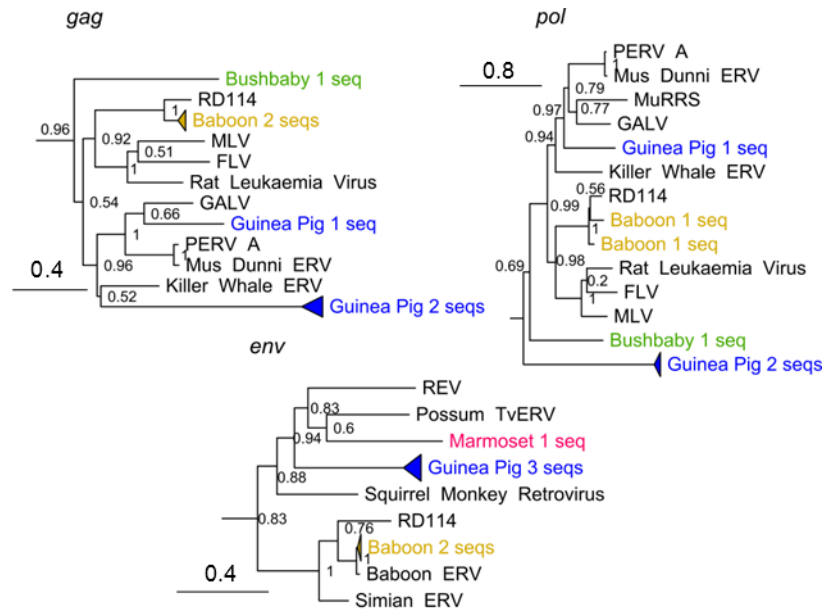


Figure 68: *Gag*, *pol* and *env* gene phylogenies of the BaEV-like recombinants with gammaretrovirus-like *gag* and *pol* genes and betaretrovirus-like *env* genes with reference gamma- and betaretroviruses.

Each host is shown in a different colour. Clusters with more than one sequence from the same host have been collapsed. *Gag* and *pol* trees are rooted on the basic gammaretrovirus datasets and the *env* tree on the basic betaretrovirus dataset, these have been cropped for better visualisation. Details of previously known sequences are provided in Appendix B.2.

These baboon insertions are already well known but the guinea pig ERVs are not, so they were again screened for LTRs and ORFs. No recognisable LTR sequences were identified flanking these sequences and the longest ORF was 424 amino acids in length. Therefore, these insertions are likely to be ancient.

4.6.3. IAP-Like Group

The IAP like group is based upon the murine IAP elements, a large group of murine endogenous betaretroviruses which lack *env*. Sequences clustering with Gifford et al.'s (2005) hedgehog ERV also fell within this group. No further previously known sequences in our input dataset were assigned to the IAP-like group.

Table 27 provides the details of the 10,408 IAP-like elements identified here.

Table 27: The number of IAP-like ERV fragments identified in each host type.

All: all 33 genomes. Blue rows labelled "T" represent total counts, yellow rows labelled "PG" represent the mean number of fragments per genome of this type.

		All	Gag	Pol	Env
All	T	10408	2600	7808	0
	PG	315	79	237	0
Primates	T	290	29	261	0
	PG	19	2	17	0
Rodents	T	9868	2569	7299	0
	PG	897	234	664	0
Lagomorphs	T	81	2	79	0
	PG	41	1	40	0
Tree Shrews	T	108	0	108	0
	PG	54	0	54	0
Birds	T	37	0	37	0
	PG	19	0	19	0
Ferret	T	24	0	24	0
	PG	24	0	24	0

As expected, a large majority of the insertions in this group were identified in rodents. All sequenced rodents are known to have IAPs (Magiorkinis et al., 2012), which was consistent with our results. 85% all IAP like *pol* genes were identified in the mouse, rat and guinea pig genomes and all primate insertions were in tarsier and lemur, again this is consistent with Magiorkinis et al. (2012).

4.6.4. JSRV-Like Group

The JSRV group is built around the exogenous and endogenous JSRV and ENTV retroviruses affecting ruminants (Palmarini et al., 2004) and the related $\beta 3$ group of ERVs found in rodents (Baillie et al., 2004).

In the analysis described in section 2.1.4.1 several other previously known sequences were assigned to this group: Garcia-Etxebarria et al.'s (2010) bovine ERVs 19, 20 and 23, Baba et al.'s (2011) bovine ERVs K1 and K2, Klymiuk et al.'s (2003) sheep ERVs one to three and Gifford et al.'s (2005) Risso's dolphin ERV, white-fronted deer ERV and caribou ERVs.

The 7,593 ERV fragments identified in this group using our Exonerate pipeline are described in Table 28.

Table 28: The number of JSRV-like ERV fragments identified in each host type.
All: all 33 genomes. Blue rows labelled "T" represent total counts, yellow rows labelled "PG" represent the mean number of fragments per genome of this type.

		All	Gag	Pol	Env
All	T	7593	1167	4190	2236
	PG	506	78	279	149
Primates	T	3866	782	1373	1711
	PG	258	52	92	114
Rodents	T	3211	323	2422	466
	PG	292	29	220	42
Lagomorphs	T	237	17	213	7
	PG	119	9	107	4
Tree Shrews	T	228	41	142	45
	PG	114	21	71	23
Birds	T	0	0	0	0
	PG	0	0	0	0
Ferret	T	51	4	40	7
	PG	51	4	40	7

Relatively few JSRV-like insertions were identified in primates, with the exception of the marmoset and the tarsier, in both of which this group appears to have undergone a recent expansion. A monophyletic group of JSRV-like insertions was identified in marmoset (Figure 69), clustering close to the JSRV group but distinct enough from this group that it is unlikely that this represents a cross-species transmission. Instead, these ERVs are likely to be part of a larger group of JSRV-like retroviruses. Two groups of tarsier JSRV-like ERVs were identified, one clustering similarly to the marmoset group and the other clustering with the Risso's dolphin JSRV-like ERV. Tarsiers and marmosets do not share a common ancestor which is not also shared with all other simian primates, so these ERVs are likely to have circulated since these hosts diverged from their closest screened ancestors.

A moderate number of rodent ERVs was also identified in this group. The copy number of these ERVs was fairly consistent across all rodent hosts, with 95 to 268 *pol* gene fragments identified, with the exception of the ground squirrel, which had a fourfold increase in copy number. These insertions represent the previously characterised $\beta 3$ group of rodent ERVs (Baillie et al., 2004), which we can confirm is ubiquitous in the rodents. Baillie et al. found LTR similarity of 84.3% for a single rat insertion and 98% for a single mouse insertion. The higher degree of divergence gives an estimated integration date of 17 million years (using the methodology described in section 2.4.6.1). This date is later than any pair of rodent hosts screened here diverged, so even if this group of ERVs entered the common ancestor of the rodents it has also been active much more recently. The increased copy number in ground squirrels also suggests recent activity.

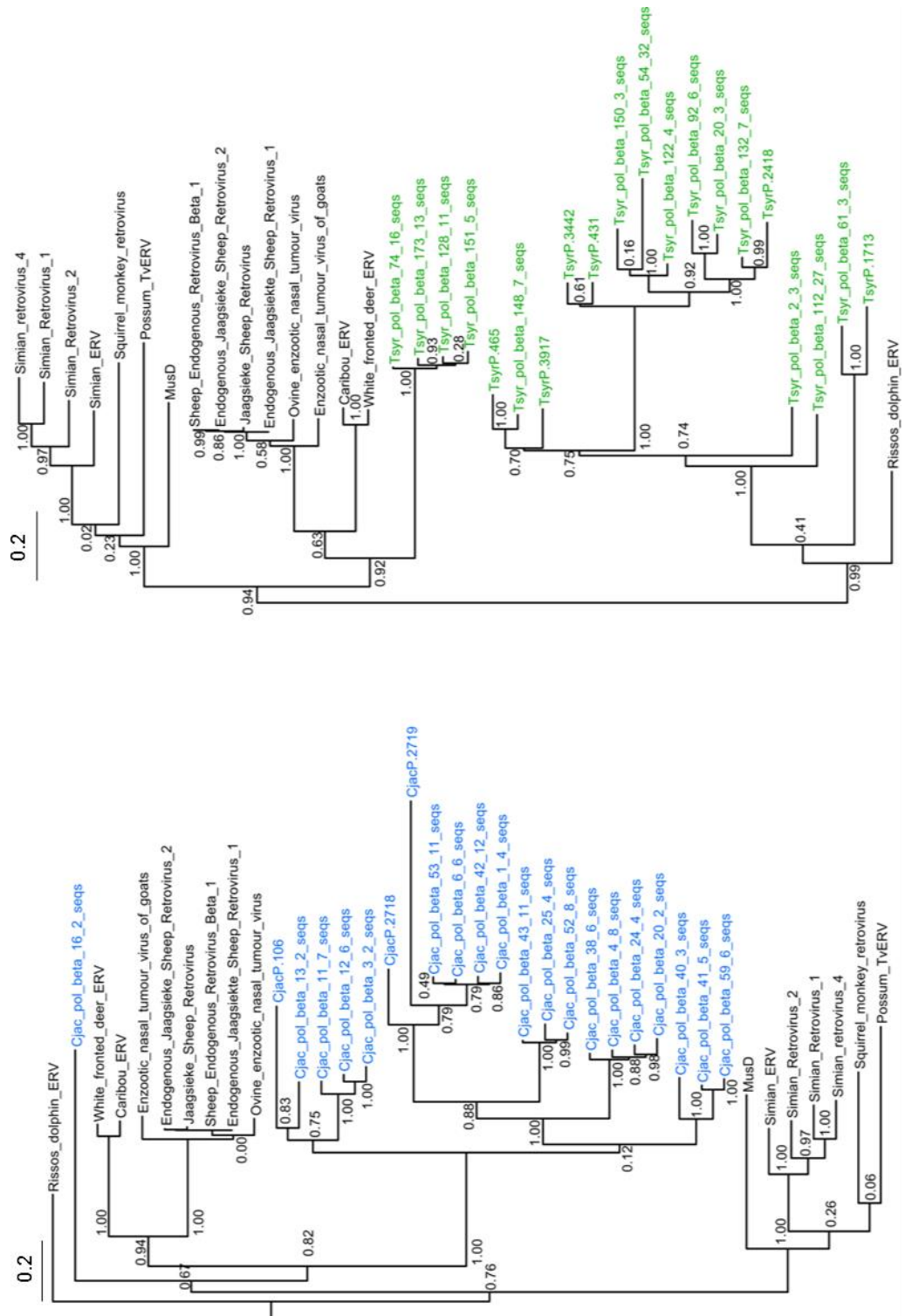


Figure 69: *Pol* gene phylogenies showing the relationships between the JSRV-like ERVs in the marmoset (blue, left) and the tarsier (green, right) and known members of this group.

Trees are rooted on the basic betaretrovirus dataset but this group has been cropped for better visualisation. Details of previously known sequences are provided in Appendix B.2.

4. 7. Lentiviruses

Lentiviruses were identified in the hosts in which endogenous lentiviruses have been previously described – the ferret (Han and Worobey, 2012b), the lemur (Gifford et al., 2008) and the rabbit (Katzourakis et al., 2007) (Figure 70). Copy number was noticeably higher in rabbit than in the other two hosts. A few scattered insertions were identified in other hosts in preliminary analyses, however in phylogenetic analyses these clustered with the betaretroviruses (data not shown).

A laboratory based technique was also used to identify endogenous lentiviruses, the results of this analysis are discussed in Chapter 5.

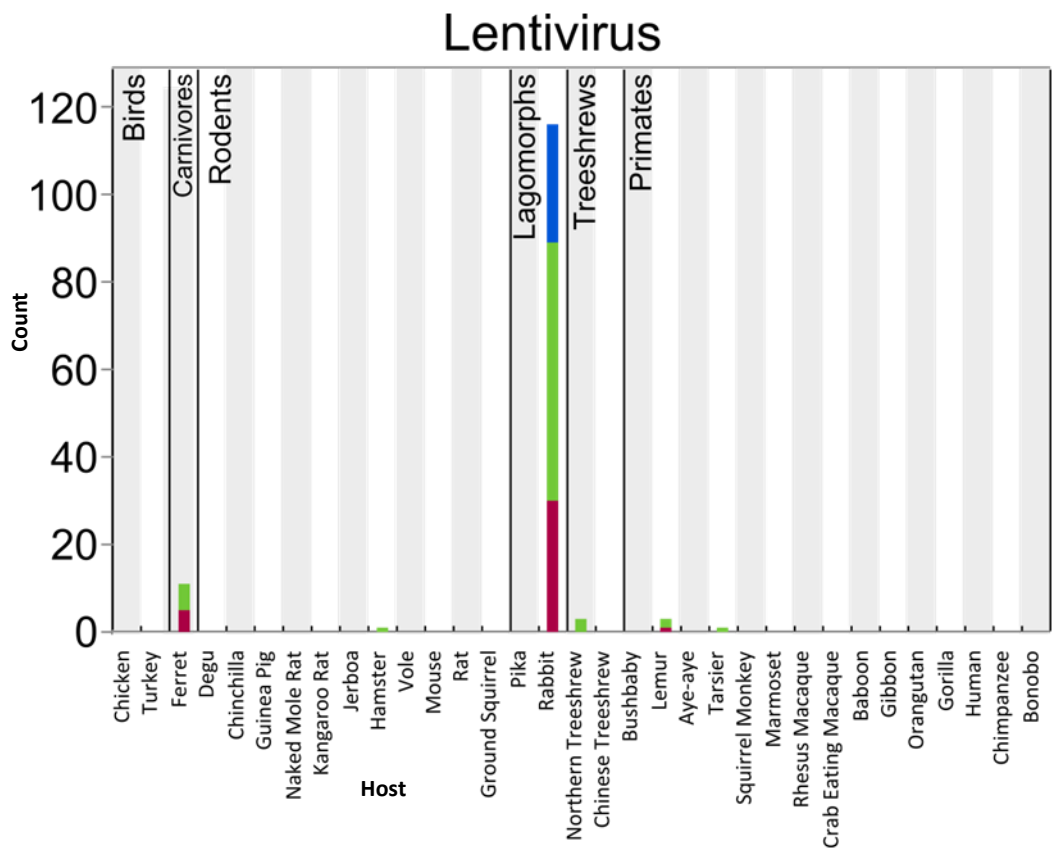


Figure 70: The distribution of ERV fragments between genomes for the lentivirus genus. gag fragments are represented in red, pol in green and env in blue.

Chapter 5. Endogenous lentiviruses in mainland African bushbabies provide insight into the origin of SIV.

Abstract

Simian immunodeficiency viruses are widespread in mainland African primates and cross-species transmission of one of these lentiviruses to humans led to the HIV-1 pandemic. The origin of SIVs in primates is not well understood. Endogenous lentiviral ancestors of SIVs have previously been found in two species of lemur living in Madagascar, raising questions about how these viruses passed from Madagascar to the mainland. We have identified and characterised the first endogenous lentivirus in a mainland African primate, the Mohol bushbaby, which appears to be another ancestor of modern SIVs. We therefore propose that SIVs in old world monkeys are the result of a direct transmission from other mainland African primates. The Mohol bushbaby lentivirus is extremely similar to the lemur lentiviruses, so we also propose routes through which lentiviruses circulating in mainland primates may have reached Madagascar.

5. 1. Introduction

HIVs are known to be the result of cross-species transmissions from a group of viruses which are widespread in monkeys and apes in mainland Africa, the SIVs. The origin of SIVs and their relationship to other retroviruses is ambiguous. Two recently identified ancient retroviruses of Malagasy lemurs seem to be predecessors to modern SIVs, giving some insights into the origin of SIVs, but leading to questions about how these viruses jumped between primates on Madagascar and those on mainland Africa (Gifford et al., 2008, Gilbert et al., 2009). We have identified a third ancient retrovirus in a mainland African primate, the Mohol bushbaby (*Galago moholi*), adding another piece to this puzzle and providing new information about the

evolutionary history of retroviruses and the potential transmission patterns of modern pathogens such as HIV.

The ancient lentiviruses identified in prosimians are ERVs. Until recently, it was thought that lentiviruses were unable to become endogenous, however, Katzourakis et al. (2007) discovered the first endogenous lentivirus in the European rabbit (*Oryctolagus cuniculus*). The two lemur endogenous lentiviruses were identified soon afterwards (Gifford et al., 2008, Gilbert et al., 2009). Lemurs are members of the “prosimian” group of primates, consisting of lemurs, lorises and bushbabies, so their retroviruses are known as prosimian immunodeficiency viruses (pSIVs). These viruses were found in the gray mouse lemur, *Microcebus murinus* (pSIVgml) (Gifford et al., 2008) and the fat tailed dwarf lemur, *Cheirogaleus medius* (pSIVfdl) (Gilbert et al., 2009). Endogenous lentiviruses have also been identified in the European brown hare (*Lepus europaeus*) and in several members of the weasel (Mustelidae) family (Cui and Holmes, 2012, Han and Worobey, 2012b). The known endogenous lentiviruses are discussed in detail in section 1.4.8.

pSIVfdl and pSIVgml entered the genomes of these lemurs three to five million years ago (Gifford et al., 2008, Gilbert et al., 2009). However, lemur ancestors migrated to Madagascar on mats of vegetation from eastern Africa 50 to 54 million years ago and are not known to have been in contact with mainland African primates since this time (Ali and Huber, 2010, Samonds et al., 2013). Therefore it is not clear how and when lentiviruses were transferred between Madagascar and the mainland. By screening prosimian samples for endogenous lentiviruses and identifying an insertion in a mainland species we have identified a possible route through which this could have occurred.

5. 2. Materials and Methods

15 species of prosimian primate were screened for endogenous lentiviruses. This project was approved by the University of Nottingham School of

Veterinary Medicine and Science Non-ASPA (animals scientific procedures act) Ethics Committee. DNA derived from archived fecal samples was kindly provided by Dr Christian Roos (Deutsches Primatenzentrum) for Mohol bushbaby (*Galago moholi*), potto (*Perodicticus potto*), red slender loris (*Loris tardigradus*), fat tailed dwarf lemur (*Cheirogaleus medius*), ring tailed lemur (*Lemur catta*), Verreaux's sifaka (*Propithecus verreauxi*), aye-aye (*Daubentonia madagascariensis*) and gray mouse lemur (*Microcebus murinus*). Archived tissue samples were kindly provided by the Zoological Society of London (ZSL) for pygmy slow loris (*Nycticebus pygmaeus*), greater galago (*Otolemur crassicaudatus*), gray slender loris (*Loris lydekkerianus*), common brown lemur (*Eulemur fulvus*), red bellied lemur (*Eulemur rubriventer*), red ruffed lemur (*Varecia rubra*) and black and white ruffed lemur (*Varecia variegata*). Surplus blood samples from veterinary procedures were kindly provided by Copenhagen Zoo for two further *G. moholi* individuals.

DNA was extracted from blood and tissue samples using the appropriate Nucleospin extraction kits (Machery-Nagel). Sample quality and species of origin were confirmed using PCR with cytochrome oxidase I primers COIbF and COIbR (Bitanyi et al., 2011). Three species could not be amplified using these primers (*Daubentonia madagascariensis*, *Varecia variegata*, *Microcebus murinus*).

Samples were screened for lentiviruses using the primer pairs FR (5' CCAAGAGTTAAAACAGTGGCC 3') (Gelman et al., 1992)– RR (5' ATGGTATGGTAAAATAAGCATC 3') (Gelman et al., 1992) and FG (5' GGGCAAGAACTTGGTATATCG 3') (Gifford et al., 2008) - Pol R2 (5' CCAAAACCACTTTGTTGGCT 3') with 2 minutes at 94°C, 40 cycles of 30s 94°C, 20s 58°C, 60s 72°C , 2 minutes at 72°C in 25ul reactions with 2X MgCl₂ free buffer (NEB), 2µM (FR-RR) or 3µM (FG – Pol R2) MgCl₂ (NEB), 10pmol each primer (FR-RR Invitrogen, FG – Pol R2 Sigma-Aldrich), 200µM each dNTP (NEB) and 2.5 units Taq DNA polymerase (NEB). All successfully

amplified PCR products were Sanger sequenced by Source Bioscience and aligned using CAP3 (Huang and Madan, 1999).

Assembled PCR fragments were aligned with sequences from representative known lentiviruses using the localpair setting of MAFFT (Katoh et al. 2002) with 1000 iterations. Lentiviruses included in this alignment are listed in Appendix B.9. A PhyML phylogenetic tree was built based on this alignment under the GTR model with aLRT branch support, no invariable sites, optimized across site rate variation and optimized tree topology. The host phylogenetic tree is from the 10K trees project (Arnold et al., 2010) incorporating all available genes.

Integration dates were estimated using the equation $t = k/2N$, where t is time, k is divergence (number of sites at which the sequences differ over alignment length), and N is the neutral substitution rate of the host, assumed to be between 4.5×10^{-9} substitutions per site per year (Gifford et al., 2008) and the mean prosimian rate of 7.17×10^{-10} substitutions per site per year (Perelman et al., 2011).

33 genomes were screened for pSIVs using the Exonerate algorithm (Slater and Birney, 2005) as described in section 2. 1. Query sequences were the two pSIVmb fragments identified here and pSIVgml and pSIVfdl consensus *pol* gene sequences (Gifford et al., 2008, Gilbert et al., 2009). Each candidate sequence identified was aligned to each of these using the Smith-Waterman algorithm via EMBOSS water (Smith and Waterman, 1981, Rice et al., 2000). High scoring sequences were verified using BLASTX against the NCBI nr database.

5. 3. Results

We screened samples from 15 species of prosimian primate (Figure 71) for strains of pSIV. DNA barcoding primers were used to confirm sample quality

and species of origin for the primate samples (these sequences were deposited in Genbank under accessions KJ543729 to KJ543742). Samples were then screened by PCR using two primer pairs designed against the pSIV *pol* gene. Only the bushbaby *G. moholi* and the two lemur species with known pSIV insertions (*M. murinus* and *C. medius*) gave positive results (Figure 71, Figure 72). Sequencing these fragments from all three hosts confirmed that they originated from the pSIV *pol* gene, with a total of 1190 bp identified in *G. moholi*. *G. moholi* pSIV will be provisionally referred to as pSIV Mohol bushbaby (pSIVmb). The pSIVmb fragments were deposited to Genbank (accessions KJ563276 and KJ563277). A second sample of *G. moholi* from a different individual also contained these two pSIV fragments. The two pSIVmb fragments fall at positions 2203 to 2503 and 3514 to 4403 of the pSIV consensus sequence (Gilbert et al., 2009). These fragments represent part of reverse transcriptase, RNaseH, dUTPase and part of integrase. The presence of dUTPase is consistent with pSIVgml and pSIVfdl (Gifford et al., 2008, Gilbert et al., 2009) and distinguishes pSIVs from SIVs. Attempts to amplify further regions of the pSIVmb genome have so far been unsuccessful. Southern blotting was not attempted due to the low volume and relatively poor quality of the primate DNA samples.

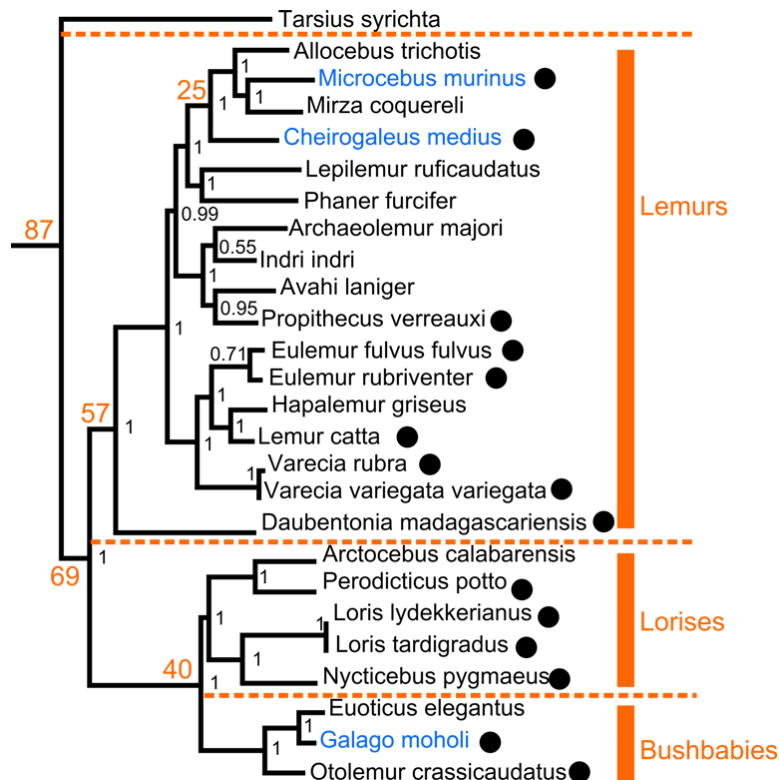


Figure 71: The phylogenetic relationships between the prosimian primates. Species marked with circles were tested for prosimian immunodeficiency virus (pSIV). Blue text indicates species which tested positive for pSIV. Black node labels indicate branch support, orange node labels indicate approximate divergence dates in millions of years. Tree data from the 10k trees project (Arnold et al. 2010), dates from Perelman et al. 2011.

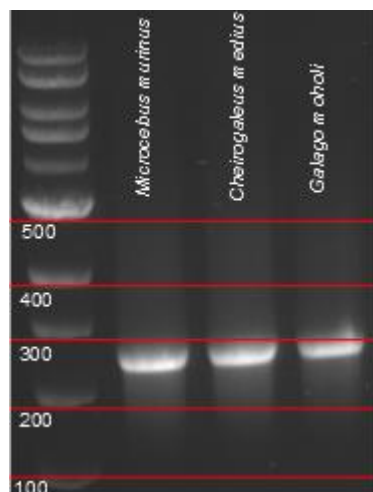


Figure 72: Gel electrophoresis photograph showing the 300bp band identified using the FR-RR primer pair in *M. murinus*, *C. medius* and *G. moholi*.

In silico screening for endogenous lentiviruses in 32 mammalian genomes, including three prosimian primates, identified all known endogenous lentiviruses but no novel endogenous lentiviruses in any host.

A phylogeny comparing pSIVmb with representative lentiviruses [selected based on (Gifford et al., 2008) and (Gilbert et al., 2009)] is shown in Figure 2. pSIVmb is clearly a member of the pSIV family. Over the two fragments the percentage identity between the sequences is as follows: pSIVmb–pSIVgml 98.15%, pSIVmb–pSIVfdl 95.46%, pSIVgml–pSIVfdl 95.63%. The divergence between pSIVmb and pSIVgml suggest that these viruses diverged between 2.05 and 12.95 million years ago, very similar to the dates estimated for the lemur pSIVs (Gifford et al., 2008, Gilbert et al., 2009). The divergence between pSIVgml and pSIVmb gives an earlier estimate of 5.04 to 31.79 million years ago.

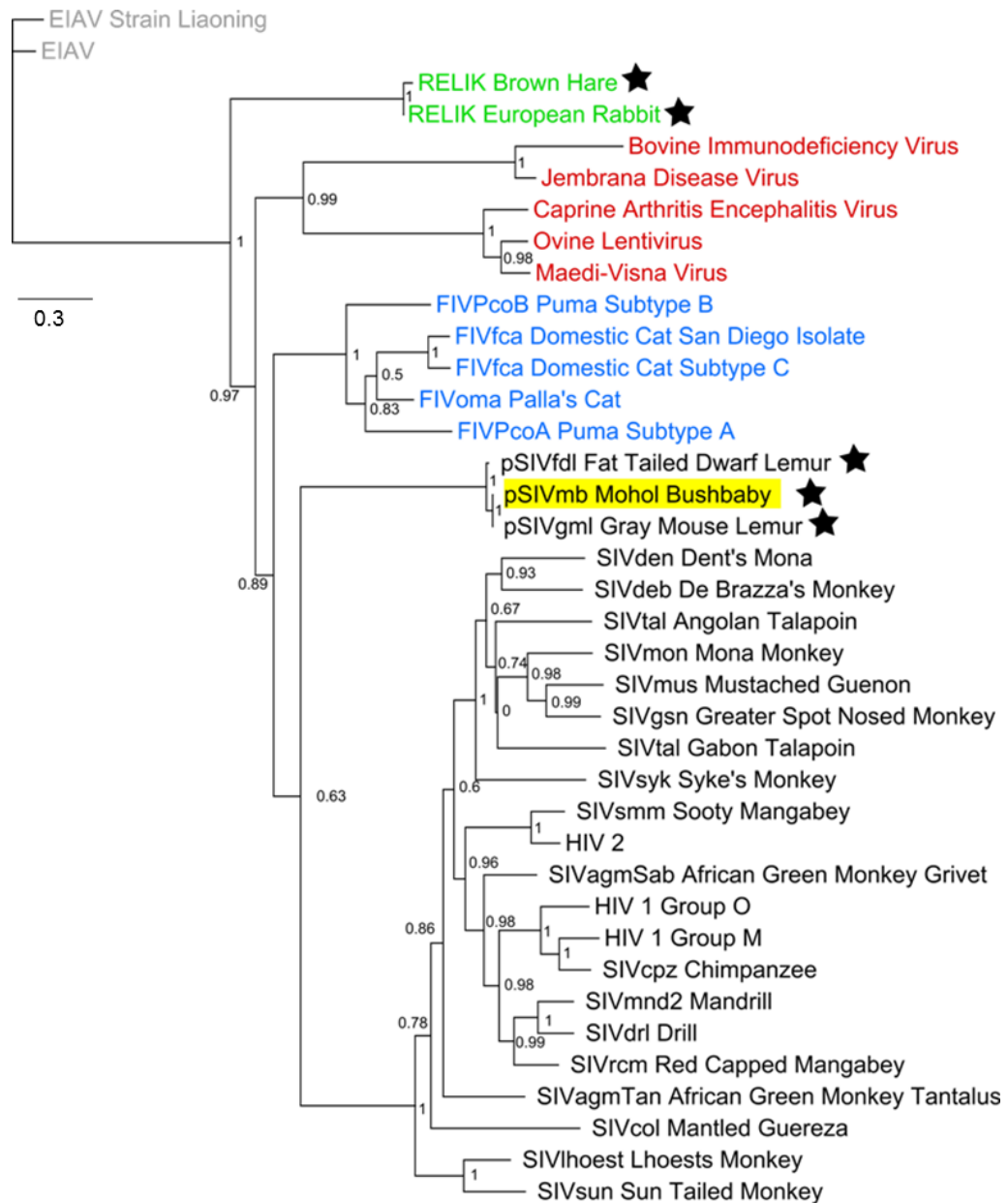


Figure 73: Maximum likelihood phylogenetic tree showing the phylogenetic relationship between pSIVmb (marked in yellow) and other lentiviruses.

Stars indicate endogenous lentiviruses, all others are exogenous. Host taxonomic groups are indicated as follows: black, primates; blue, carnivores; green, lagomorphs; red, bovids; grey, equids. Node labels indicate branch support. Abbreviations: equine infectious anemia virus, EIAV; feline immunodeficiency virus, FIV; simian immunodeficiency virus, SIV; prosimian immunodeficiency virus, pSIV; Mohol bushbaby, mb; gray mouse lemur, gml; fat-tailed dwarf lemur, fdl

5. 4. Discussion

These results confirm the presence of a pSIV strain in the Mohol bushbaby. Endogenous lentiviruses have now been identified in four families of mammals: bushbabies (Galagidae), mouse/dwarf lemurs (Cheirogaleidae), rabbits/hares (Leporidae) and weasels (Mustelidae). pSIV appears to be absent in all other lemurs, bushbabies and lorises screened here.

We propose that SIVs in old world monkeys result from a direct transmission from mainland African prosimians. *G. moholi* is widespread in the savannahs of southern Africa (Figure 74) a habitat shared with several SIV hosts. The phylogenetic analysis of pSIV here and previous phylogenetic analyses (Gifford et al., 2008, Gilbert et al., 2009) show pSIV to be closest to SIVcol, found in the mantled guereza (*Colobus guereza*), SIVsun, found in the sun-tailed monkey (*Cercopithecus solatus*) and SIVlhoest, found in the L'hoest's monkey (*Cercopithecus lhoesti*), compared to other SIVs. The geographical ranges of these three species overlap with that of *G. moholi*, so cross-species transmissions are feasible. There may also be unknown intermediate pSIVs or SIVs in other primates.

Our results are consistent with *G. moholi*, *M. murinus* and *C. medius* having contracted the same strain of pSIV at approximately the same time, two to five million years ago. *M. murinus* and *C. medius* are relatively close genetic relatives which diverged from a common ancestor approximately 25 million years ago, while lemurs diverged from bushbabies and lorises approximately 69 million years ago (Figure 71) (Perelman et al., 2011). *M. murinus* and *C. medius* also share a habitat (Figure 74) so the presence of pSIV in both of these species is less surprising than in *G. moholi*, which was not in contact with Malagasy primates during this period. If pSIVs had been circulating exogenously for 25 million years, they would be considerably more divergent from each other than they are. Therefore, assuming no contact between lemurs

and bushbabies since the migration of lemurs from the mainland, a vector species must have transmitted the virus between hosts.

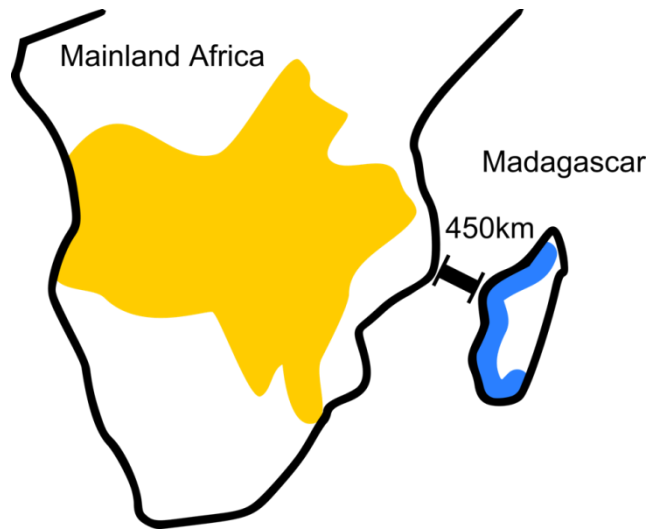


Figure 74: The geographical distribution of *Galago moholi* (yellow) and *Microcebus murinus* / *Cheirogaleus medius* (blue).
All data from IUCN (2013).

Only three groups of mammals migrated to Madagascar between 13 million years ago (the earliest possible divergence date calculated for pSIVmb and pSIVfdl) and human colonization: bats, hippopotamuses and possibly rodents (Samonds et al., 2013), all of which provide potential transmission routes. Bats are particularly effective at hosting and transferring viral pathogens and many bats colonized Madagascar in the last five million years ago (Calisher et al., 2006, Samonds et al., 2013). Native Malagasy rodents have lived on Madagascar for 20 to 24 my, but it is not clear when non-native rats and shrews reached the island (Samonds et al., 2013). Rodents seem to be particularly susceptible to retroviruses (Baillie et al., 2004, McCarthy and McDonald, 2004, Stocking and Kozak, 2008) and their wide distribution in different habitats makes them another attractive candidate as a vector species. Hippopotamuses have not been screened for retroviruses but do share habitats with bushbabies and potentially lemurs (the distribution of extinct Malagasy hippopotamuses is not known) so may have potential as a vector.

Another possibility is that the vector species here was an insect, rather than a mammal. Lentiviruses have diverse transmission routes and EIAV is usually spread mechanically via biting insects (Issel et al., 1988). Insect vectors also have the capability to mechanically transmit Jembrana disease virus (Soeharsono et al., 1995), bovine immunodeficiency virus (St.Cyr Coats et al., 1994) and small ruminant lentiviruses (Murphy et al., 1999), although these are generally transmitted via other routes. Malagasy invertebrates are not well studied, but there is evidence of recent dispersal across the ocean between mainland Africa and Madagascar of several insect families, including the small mayflies (Baetidae) (Monaghan et al., 2005), *Braunsapis* bees (Fuller et al., 2005) and *Papiliodemoleus* butterflies (Zakharov et al., 2004). Therefore, it is feasible that a vector insect crossed this divide approximately five million years ago.

The presence of a pSIV strain, circulating within the last five million years, in mainland African primates is an important step in establishing the evolutionary history of the lentiviruses and suggests that several potential cross-species transmission events between diverse hosts have occurred. If a mammalian or insect vector species was able to transfer these ancient lentiviruses between primate hosts in different geographical regions, this needs to be considered when modelling the transmission of modern lentiviral pathogens, including HIV.

Chapter 6. The origin and proliferation of gibbon ape leukaemia virus

Abstract

Gibbon ape leukaemia virus (GALV) is an exogenous gammaretrovirus causing haemopoietic neoplasias in gibbons. Several strains of this virus were identified in gibbons in southeast Asia, the USA and Bermuda during the 1970's and the virus is still widely considered to be an active pathogen with a high prevalence in gibbons. Here, through screening of tissue samples, genome screening, analysis of veterinary records and a review of documentation concerning this outbreak, we conclude that GALV is unlikely to be a currently circulating pathogen in gibbons. We have also identified additional relatives of GALV in the sequenced mouse, rat and hamster genomes, strengthening the hypothesis of the rodent origin of this group of viruses. Finally, we propose a route through which all known GALV outbreaks may be linked and therefore present the hypothesis that a single spill-over event from a rodent in south east Asia in the late 1960s may have been the origin of all known GALV isolates.

6. 1. Introduction

6.1.1. History of GALV

The first published report of haemopoietic neoplasia in a gibbon was malignant lymphoma in a male *Hylobates* gibbon in 1960 (Newberne and Robinson, 1960). This was followed by a case of malignant lymphoma, with an appearance similar to Burkitt's lymphoma, identified in 1966 in a white-handed gibbon (*Hylobates lar*). This gibbon was imported from South-East Asia to the University of Chicago in 1964. The causes of these two cases are not known.

The first outbreak of disease which is attributed to GALV was described by Johnsen et al. (1969). This paper describes four white-handed gibbons (*Hylobates lar*) in the Southeast Asia Treaty Organization (SEATO) medical research laboratory colony in Thailand which died of generalised malignant lymphoma between 1966 and 1968. At this time, no infectious agent could be isolated from these gibbons (Johnsen et al., 1969). However, in 1971, five further gibbons in this colony were identified with granulocytic leukaemia and a type-C (gammaretrovirus-like) retrovirus was identified in one of these cases as a possible causative agent (De Paoli et al., 1971). This virus was identified at the University of California at Davis (UC Davis) School of Veterinary Medicine Comparative Oncology Laboratory (COL), which also had a gibbon colony (De Paoli et al., 1971).

Soon afterwards, in 1971, a type C (gammaretrovirus like) exogenous retrovirus was identified in a woolly monkey (*Lagothrix lagotricha*) which was diagnosed with fibrosarcoma (Eiden and Taliaferro, 2011, Theilen et al., 1971). This monkey was kept as a pet in an apartment in San Francisco, alongside a lar gibbon (*Hylobates lar*) (Eiden and Taliaferro, 2011). Less than a year later, this gibbon was diagnosed with lymphosarcoma and another strain of the same retrovirus was identified as the cause (Kawakami et al., 1972). These strains were named WMSV and GALV San Francisco (GALV-SF) respectively. Both were diagnosed at the San Francisco Medical Center (SFMC) and identified at the UC Davis School of Veterinary Medicine.

GALV was soon reported in several other locations. Kawakami et al. (1973) looked for antibodies to GALV in sera of gibbons in various locations and found a high prevalence of these antibodies in the SEATO gibbons, SFMC gibbons and COL gibbons but none in gibbons from other US locations. The virus was then identified in frozen brain samples from gibbons imported from south east Asia in 1968 and stored in Louisiana at the Gulf South Primate Center. This strain is known as GALV brain (GALV-Br). In the late 1970's, a strain of GALV was also identified in a gibbon colony on Hall's island in

Bermuda, this is the GALV-Hall's island (GALV-H) strain (Krakower et al., 1978, Reitz et al., 1979). GALV was thought to be a widespread veterinary pathogen during this period and was described as infecting 11% of captive gibbons (Kawakami et al., 1975).

During this period, GALV also appeared as a cell line contaminant on several occasions. Okabe et al. (1976) and Chan et al. (1976) reported the strain GALV-X, similar to WMSV, in cells cultured from a single patient with acute myelogenous leukaemia. These studies were performed at the National Cancer Institute (NCI) in Maryland, USA. Later, in the 1990's, this strain was identified in HUT78 cells infected with HIV-1 strain ARV-2 at the University of Louvain in Belgium (Burtonboy et al., 1993, Parent et al., 1998).

No haemopoietic neoplasias in gibbons have been attributed to GALV since the 1970's. However, GALV is still often cited as a pathogen of gibbons, a risk to humans handling primates and a potential confounding factor in primate based research (Voevodin and Marx, 2009, Lerche and Osborn, 2003, Murphy et al., 2006, Fowler and Miller, 2008).

6.1.2. GALV Phylogeny

GALV falls into the MLV-like clade of retroviruses, as discussed in section 1.4.3.6. The closest known relatives of GALV are predominantly gammaretroviruses identified in species of mouse (Figure 75). The exception to this is KoRV, the most similar known retrovirus to GALV (Figure 75). KoRV is an active pathogen in koalas and entered the koala population within the last 100 years (Tarlinton et al., 2006). As gibbons and koalas are distant both evolutionarily and geographically, it is generally considered that the virus originated separately in both groups via another host, most likely a rodent (Eiden and Taliaferro, 2011). Asian mice share a habitat with gibbons and harbor GALV-like gammaretroviruses. In particular, the Asian mouse

retroviruses *Mus caroli* ERV (McERV) and *Mus dunni* ERV (MDERV) are known to be close relatives of GALV and KoRV (Lieber et al., 1975, Wolgamot et al., 1998, Martin et al., 1999) (Figure 75). Hayward et al. (2013) also identified GALV-like insertions in the house mouse (Hayward et al., 2013a). Simmons et al. (2014) screened bats and rodents which are either found in both Australia and southeast Asia, transit between these regions or which may otherwise be in contact with both gibbons and koalas for GALV-like ERVs. One of the species screened, the Grassland mosaic-tailed rat *Melomys burtoni*, contained *pol* and *env* gene fragments clustering between GALV and KoRV (Simmons et al., 2014). *M. burtoni* is an Australian native rodent also found in Papua New Guinea so provides an attractive possible vector species for GALV/KoRV. This ERV, known as *Melomys burtoni* ERV (MbERV) is more similar to GALV and KoRV than any other known rodent ERV. However, this species is not present on mainland southeast Asia and does not share a geographical range with gibbons, so a further vector species must also have been involved in this transmission (Simmons et al., 2014).

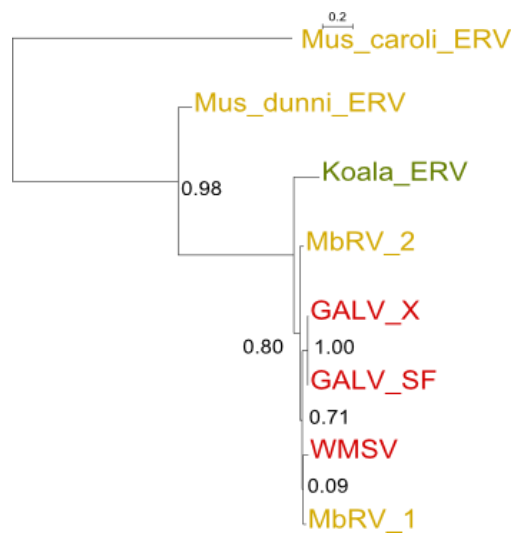


Figure 75: The relationship between the *pol* genes of GALV and the sequences described as its closest genetic relatives.
GALV sequences are shown in red, rodent sequences in yellow and KoRV in green.

6. 2. Materials and Methods

DNA was extracted from 23 blood samples from the following seven species: white-handed gibbon (*Hylobates lar*, 6 samples), siamang (*Symphalangus syndactylus*, 3 samples), red-cheeked gibbon (*Nomascus gabriellae*, 2 samples), southern white-cheeked gibbon (*Nomascus siki*, 2 samples), northern white-cheeked gibbon (*Nomascus leucogenys*, 6 samples), Bornean gibbon (*Hylobates muelleri*, 1 sample), agile gibbon (*Hylobates agilis*, 3 samples). All samples were sourced from the Deutsches Primatenzentrum, Goettingen, Germany. These samples are from primates confiscated for various reasons throughout the EU and so are of diverse origins.

All samples were screened via PCR using four primer pairs designed against the *pol* genes of GALV and KoRV. These analyses were performed prior to the commencement of this PhD project by Dr. Rachael Tarlinton using the primer pairs KoRV-pol-F to KoRV-pol-R and KoRVmgbf to KoRVmgbf designed against the KoRV *pol* gene, sequences and conditions described in Tarlinton et al. (2006), the primer pair ERV1 (5' TGG GCC GAG AAG GCA CCT AT 3') to ERVR1 (5' CCA TTC AAA CGC GAA CAA TG 3') designed against MLV, under the same conditions and the primer pair GALV-pol-F (5' AGA TCG ACC CGG CGT GTA CT 3') to GALV-pol-R (5' CCA TTC AAA CGC GAA CAA TG 3') designed against the GALV *pol* gene, again under the same conditions. During this PhD project, one sample from each species was also screened using the degenerate gammaretrovirus primer pair PRO (5' GTK TTI KTI GAY ACI GGI KC 3') to CT (5' AGI AGG TCR TCI ACR TAS TG 3') [from Martin et al. (1999)], which were designed to amplify MLV-related ERVs. These analyses were performed in 50µl reactions with 2x MgCl₂-free buffer, 20 pmol each primer, 3µM MgCl₂, 100µM each dNTP, 2.5 Units Taq polymerase with 2 minutes at 95°C, 40 cycles of 94°C for 20 seconds, 44.8°C for 30 seconds, 72°C for 60 seconds, followed by 2 minutes at 72°C.

DNA quality was confirmed for all samples using primers based on the β -actin gene as described in Tarlinton et al. (2006).

For *in silico* analysis, all available sequenced primate, lagomorph and rodent genomes were screened for candidate ERVs as described in section 2.1.6. Each candidate ERV identified was then aligned to the appropriate gene from the GALV reference sequence (Genbank NC_001885) using EMBOSS water. A BLAT search of the remaining sequenced mammalian genomes available via the UCSC genome browser on 18/04/14 was also performed under the default settings with the GALV reference sequence as a query. The highest scoring sequence from each genome was then aligned to the basic gammaretrovirus *pol* gene, with an epsilonretrovirus, WDSV, as an outgroup. A phylogenetic tree was built of this alignment as described in section 2.3.3.3.

Zoological records dating from 1964 to 2008 were obtained from the animal record keeping software (ARKS) records held at Twycross Zoo, Warwickshire, UK. 48 gibbons and 20 woolly monkeys died during this period. These animals included 11 white-handed gibbons, 19 siamangs, 10 pileated gibbons (*Hylobates pileatus*), one black-crested gibbon (*Hylobates concolor*), three agile gibbons (*Hylobates agilis*), one white-cheeked gibbon (*Nomascus leucogenys*) and one dwarf siamang (*Hylobates klossii*). All woolly monkeys were brown woolly monkeys (*Lagothrix lagotricha*).

The following documents were reviewed to trace the origin and epidemiology of GALV: scientific publications concerning GALV, 1963 to 1983 SEATO Medical Research Laboratory Annual Progress reports (<http://www.afirms.org/weblib/apr/aprF.shtml>), archived documentation from the US Department of State concerning gibbon transportation (documents 1974BANGKO17800, 1974STATE260768_b, 1974BANGKO17734_b, 1974STATE260770_b, 1974TAIPEI06749_b, 1974STATE244644_b, 1974BANGKO19028_b, 1975BANGKO15111_b, available in Appendix E) and scanned documents from this period available

via archived International Primate Protection League newsletters (<http://www.ippl.org/gibbon/current-news/newsletters-1970s>).

6. 3. Results

Despite being subjected to multiple PCR screens, none of the contemporary gibbon samples tested positive for GALV or KoRV in any analysis. These samples were all of sufficient quality for PCR amplification, as the control PCR using β -actin primers demonstrated. These primers and conditions have previously been used successfully to amplify these and other gammaretroviruses. Therefore it appears that GALV and KoRV are absent in the blood of these gibbons. The samples were from unlinked gibbons either from zoological collections or confiscated at various timepoints throughout Europe.

Veterinary records showed no confirmed cases of GALV in 48 captive UK gibbons and 20 captive UK woolly monkeys over a 44 year period. However, one pileated gibbon had clinical signs consistent with GALV at post-mortem, listed as “mediastinal and intestinal lymphadenopathy, lymphatic enlargement of spleen, liver and kidney and gastric ulceration”.

The phylogenetic tree in Figure 76 shows the closest relatives to GALV identified through in silico screening of 31 species of primate, rodent, lagomorph and tree shrew. No sequences clustering close to GALV were present in any primate genomes, including that of the gibbon *Nomascus leucogenys*. No new sequences were identified in any host falling within the GALV-KoRV-MbRV clade. The closest group to this clade was unique to old world rodents, specifically the Myomorpha. MDERV and the mouse MuRV-Y group (Stocking and Kozak, 2008) fell within this group. More distant relatives of GALV and KoRV are found in rodents, bats and pigs. The closest endogenous sequence found in a primate is a lemur ERV clustering with one of the bat ERVs identified by Cui et al. (2012) and red squirrel ERV 1A1,

discussed in section 1.4.3.5, however these are relatively distinct from GALV.

The closest gibbon ERV is very distant from GALV and resembles HERV-T.

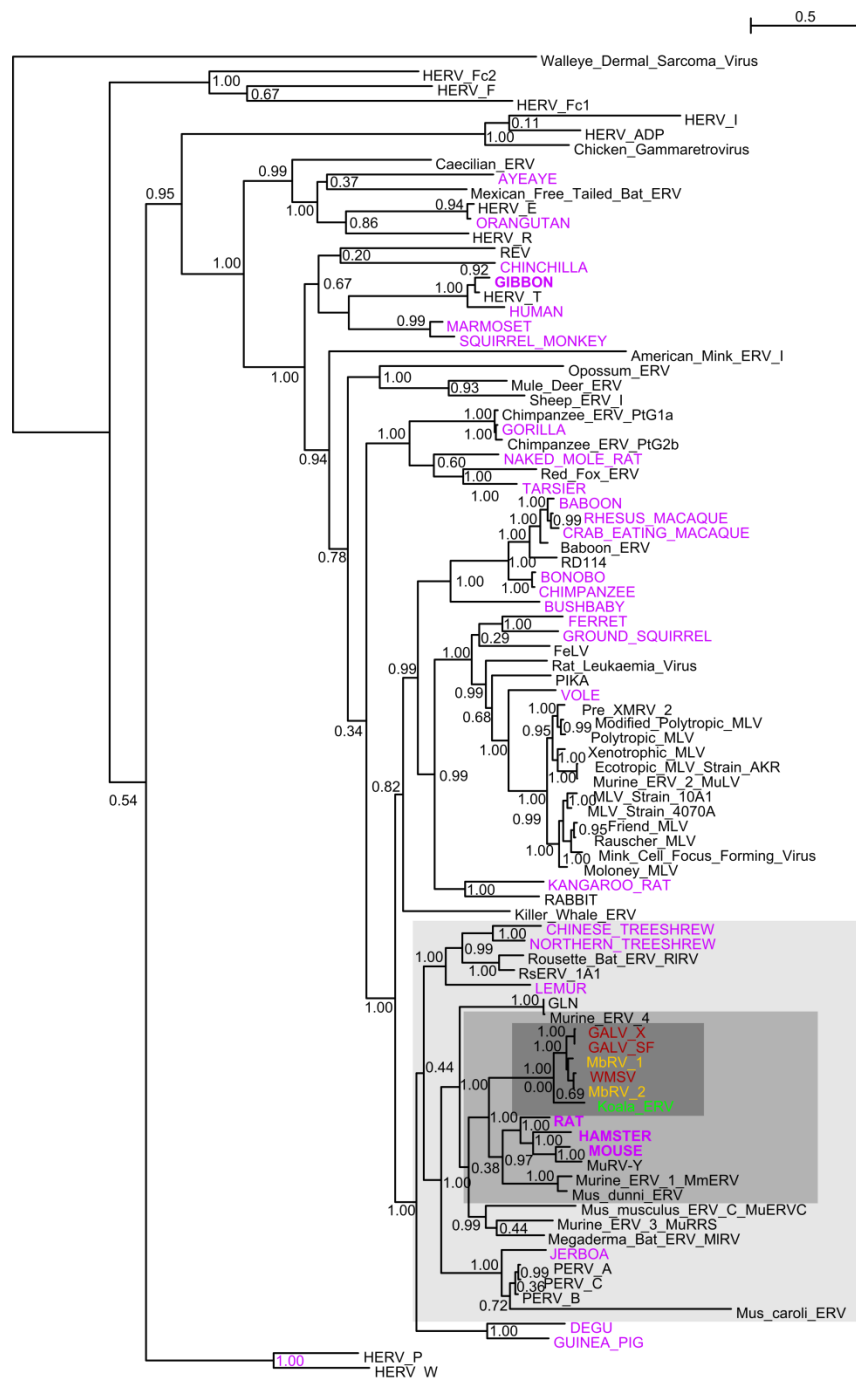


Figure 76: *Pol* gene phylogenetic tree showing the phylogenetic relationship between GALV and its closest relative in each host genome screened. Sequences identified here are shown in pink, GALV strains in red, MBRV in yellow, KoRV in green. Grey squares mark the closest relatives of GALV. Details of previously known sequences are provided in Appendix B.2.

As for GALV-BR, GALV-H and GALV-SEATO only the *env* gene has been sequenced, an *env* phylogeny was built to compare the strains of GALV with their closest relatives identified in Figure 76 (where an *env* sequence was available). This tree is shown in Figure 77. All GALV strains and WMSV form a highly supported monophyletic group and are very similar to each other. KoRV is only slightly distinct from this group. Again, GALV and KoRV are more similar to rodent sequences than to their nearest primate relative, from the lemur.

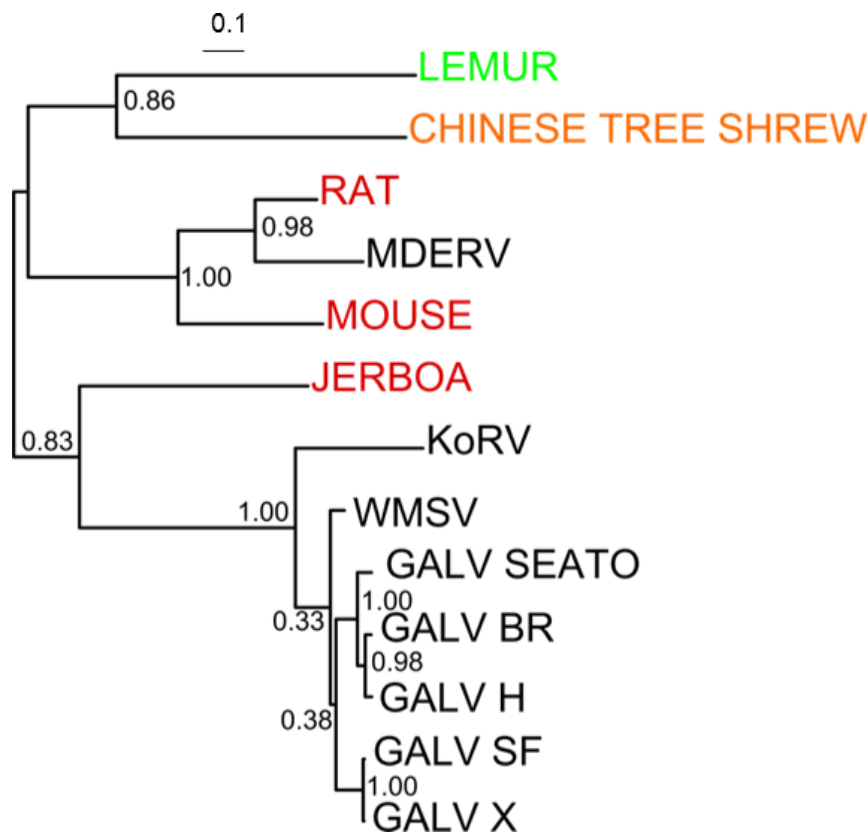


Figure 77: *Env* gene phylogeny showing the relationship between the GALV strains and related sequences from rodents, primates and tree shrews.

New world rodents are shown in red, tree shrews in orange, prosimians in green. Newly identified sequences are coloured by host, previously known sequences are shown in black.

A review of the literature about GALV and of documentation archived during the 1960s and 1970s demonstrated that all of the confirmed cases of GALV were in gibbons which were either in southeast Asia in the mid-1960s or were in contact with gibbons which were in southeast Asia at the time (Table 29).

The exception to this may be the gibbons at COL, however, as the background of these gibbons has not been reported this cannot be confirmed. Although the COL gibbon colony is seemingly distinct from the SFMC colony, a significant portion of the work on GALV had been performed at COL, using gibbons from both SFMC and SEATO, when GALV was identified in this colony in 1973 (Snyder et al., 1973, Kawakami et al., 1972, Kawakami et al., 1973).

Cell lines contaminated with GALV were also traced back to laboratories where significant amounts of work on GALV have been carried out. For example, Okabe et al. (1976) and Chan et al. (1976) included authors at the NIH National Cancer Institute in Maryland, USA, where Lieber et al. (1975) and Todaro et al. (1975) cultured GALV. Burtonboy et al. (1993) and Parent et al. (1998) worked at the University of Louvain, Belgium but the HIV-1 infected cell line from which they isolated GALV was established at the University of California, and some reagents used in this process were obtained from GALV researchers at the NIH National Cancer Institute (Levy et al., 1984). According to the phylogeny in Figure 77, the GALV-X strain appears to be simply be GALV-SF preserved in a cell line and shipped to another laboratory.

Table 29: The gibbon colonies in which GALV was identified, the date the colony was started (where available) and the source and import date of the gibbons in which GALV was isolated.

Colony	Source
SEATO	~80 gibbons purchased in southeast Asia in 1966 (Morris et al., 1966)
San Francisco Medical Center	Colony initiated in the mid-1960s. 6 gibbons purchased together, 2 developed GALV. (Kawakami et al., 1972, Snyder et al., 1973)
Gulf South Primate Center	5 gibbons imported together from southeast Asia in 1968, three tested positive for GALV in 1975 (Todaro et al., 1975).
Comparative Oncology Laboratory	Colony initiated with gibbons from “various sources” with unknown histories. GALV in 2/13 gibbons. (Kawakami et al., 1973)
Hall’s Island	Colony originated with gibbons imported from Thailand in 1970 (IPPL, 1976)

Archived documents from the SEATO medical research laboratory provided further information about the SEATO gibbon colony. The colony was established in 1966 with 71 gibbons (most likely *Hylobates lar*) purchased in Thailand (Morris et al., 1966). SEATO gibbons were used as models for human disease pathogenesis and transmission and were inoculated frequently with blood and tissue from humans, rodents and other gibbons (e.g. (Cadigan et al., 1967, Smith et al., 1968, Bancroft et al., 1975)). A large collection of Asian rodents was held at this facility and also used in these disease studies (Marshall, 1974, Marshall, 1975). SEATO established a free-ranging gibbon colony using some of their laboratory gibbons but all the gibbons had died or were returned to the laboratory by 1975 (Brockelman, 1969). Rats (*Rattus rattus* and *Bandicota indica*) were native to the island where this colony was established (Berkson, 1968). Both of these rat species were screened in the study by Simmons et al. (2014) but did not show any indication of GALV.

6. 4. Discussion

PCR screening of 23 unlinked contemporary captive gibbons in Europe showed no evidence of exogenous or endogenous GALV in these animals. Similarly, veterinary records showed no confirmed deaths from GALV in UK gibbons or woolly monkeys over a period of 44 years, encompassing the period when all confirmed cases of GALV occurred. Gibbons which died of other causes and were examined post-mortem did not show evidence of haemopoietic neoplasm, with the exception of a single gibbon which died in 2006. This gibbon had clinical signs consistent with GALV but was not tested for GALV and, as there has never been a documented case of GALV in Europe and the last documented cases elsewhere were in the 1970s, other causes of lymphadenopathy are likely. These results together suggest that GALV never reached the UK and that the documented prevalence of 11% in the USA, which is still cited today [e.g. (Voevodin and Marx, 2009)] , originally from Kawakami et al. (1973) and Kawakami et al. (1975), was never the case worldwide. The two studies which provide this statistic of 11% were based on 133 gibbons from five US colonies, of which 15 had antibodies reactive to a GALV antigen. However, breaking down this statistic, the 15 gibbons with these antibodies were all from the COL, SEATO and SFMC research colonies, where the prevalence was approximately 15%. None of the remaining 31 gibbons, from colonies elsewhere in the USA, had these antibodies.

As several previous studies have proposed [e.g. (Hayward et al., 2013a, Lieber et al., 1975, Eiden and Taliaferro, 2011, Tarlinton et al., 2008)] , we hypothesise that GALV in gibbons originated as a cross-species transmission from rodents. The presence of a GALV/KoRV like ERV in an Australian rodent (Simmons et al., personal communication) adds strength to this hypothesis. Screening of primates, rodents, lagomorphs and tree shrews confirmed that GALV is more similar to several rodent ERVs than to any ERV found in

primates and that GALV is not endogenous in gibbons. GALV strains are extremely similar to each other genetically and not all strains have been fully sequenced. The divergence of the *env* gene between strains of GALV, which ranges from 85% to 99%, is no greater than the divergence of *env* genes in the viral population within a single HIV-1 infected individual, so they may represent different isolates from a single outbreak (Andréoletti et al., 2007).

The SEATO gibbon colony in the mid-1960s provides an attractive candidate for the location of the overspill event from rodents to gibbons, either through laboratory work or direct contact between animals. The earliest identification of GALV was in four gibbons in this colony which had all been part of the malaria or dengue virus studies at the centre. Both of these studies were long-term, involved many gibbons and used transfusion of blood between gibbons and from humans to gibbons, viruses grown in live rodents and rodent tissues and transmission experiments using mosquitoes fed on infected blood from rodents and gibbons (Johnsen et al., 1969, Halstead, 1964, Diggs and Pavanand, 1969, Muangman, 1971). At least two of the gibbons with confirmed GALV at this colony (identified as gibbons S-76 and S-77) were involved in the same malaria study (De Paoli et al., 1971, Cadigan et al., 1967). The identification numbers of gibbons involved in studies were only sporadically reported, so the exact transmission path of the disease cannot be tracked.

The only confirmed cases of GALV in live gibbons outside of SEATO occurred in the COL and SFMC research colonies and the Hall's Island colony. The SFMC gibbons are known to have been exported from southeast Asia during the period in which GALV was circulating, while the origin of the COL gibbons is less clear. However, De Paoli et al. (1971) reported in 1971 that GALV virus isolation from SEATO gibbons was performed at COL. The Hall's island gibbon from which GALV was isolated were shipped from southeast Asia in 1968. Although we cannot confirm that any of these gibbons originated at SEATO, or were in contact with SEATO gibbons, SEATO gibbons appear to have been exported to US primate laboratories at UC Davis and the NIH

National Cancer Institute on at least two occasions during the 1960s and 1970s (U.S. Fish and Wildlife Service, 1974, U.S. Fish and Wildlife Service, 1975, IPPL, 1978). During this period the origin and movement of primates was not well tracked and, as primate dealing was lucrative in Thailand at this time, capture of gibbons from the free ranging SEATO colony for export is also not unlikely.

The GALV-Br strain also identified in gibbons exported from southeast Asia during this period, which may have originated at SEATO. Alternatively, the cell lines with which the gibbon brain tissue was co-cultured may have been contaminated with GALV, as this work was again carried out at the NIH National Cancer Institute (Todaro et al., 1975). Similarly, the GALV-X strains expressed by cultured cells have links to this institute, where GALV has repeatedly been isolated and cultured [e.g. by Lieber et al. (1975)]. The XMRV controversy discussed in section 1.4.3.6 demonstrated the propensity of MLV-like retroviruses to become laboratory contaminants, including cell culture contaminants, so it is feasible that cells could become infected through this route.

These results together suggest that the outbreak of GALV was a single spillover event from rodents in the 1960s. Phylogenetic analysis confirms the absence of endogenous GALV in gibbons or any other primate. Laboratory analysis confirms the absence of GALV in contemporary European gibbons and analysis of veterinary records suggest it was absent in the UK when the virus was circulating elsewhere. The closest endogenous relatives to GALV are rodent ERVs, particularly an ERV in a rodent native to Australia and Papua New Guinea.

We propose that the spillover event from rodents to gibbons occurred either at the SEATO research colony or elsewhere in southeast Asia during the mid to late 1960's. Gibbons from this region were exported to a few gibbon colonies, at COL, SFMC and Hall's Island, where GALV was again detected. GALV became a common cell culture contaminant in the laboratories where this

work was carried out, so was later detected in cell lines linked to these laboratories. The lack of documented cases of GALV worldwide since 1978, along with these results, suggest that this virus is no longer an active pathogen of gibbons and that it was never widespread amongst the gibbon population.

Chapter 7. Endogenous Epsilon-Like Retroviruses in

Primates

Please note this chapter is identical to the following publication:

Brown, K., Emes, R. E., Tarlinton, R. E. (2014). Endogenous epsilon-like retroviruses in primates. *Journal of Virology*. In press.

Abstract

Several types of cancer in fish are caused by retroviruses, including those responsible for major outbreaks of disease, such as walleye dermal sarcoma virus and salmon swim bladder sarcoma virus. These viruses form a phylogenetic group often described as the “epsilonretrovirus” genus. Epsilon-like retroviruses have become endogenous retroviruses (ERVs) on several occasions, integrating into germline cells to become part of the host genome, and sections of fish and amphibian genomes are derived from epsilon-like retroviruses. However, epsilon-like ERVs have been identified in very few mammals.

We have developed a pipeline to screen full genomes for ERVs and using this pipeline, we have located over 800 endogenous epsilon-like ERV fragments in primate genomes. Genomes from 32 species of mammals and birds were screened and epsilon-like ERV fragments were found in all primate and tree shrew genomes but no others. These viruses appear to have entered the genome of a common ancestor of old and new world monkeys between 42 million and 65 million years ago.

Based on these results, there is an ancient evolutionary relationship between epsilon-like retroviruses and primates. Clearly, these viruses had the potential to infect the ancestors of primates and were at some point a common pathogen in these hosts. Therefore, this result raises questions about the potential of epsilonretroviruses to infect humans and other primates and about the evolutionary history of these retroviruses.

7. 1. Introduction

Epsilonretroviruses are a genus of retrovirus usually associated with fish (Sinzelle et al., 2011). Several common proliferative diseases in commercially important fish species are caused by these viruses. In the walleye (*Sander vitreus*), a species of perch which is an important source of sport fishing revenue in Canada and the northern United States (VanDeValk et al., 2002), up to 30% of some populations are affected annually by skin lesions resulting from the epsilonretrovirus walleye dermal sarcoma virus (WDSV) and up to 10% by skin lesions resulting from the epsilonretrovirus walleye epidermal hyperplasia virus (WEHV) (Rovnak and Quackenbush, 2010). Outbreaks of sarcoma in the Atlantic salmon (*Salmo salar*), a species which makes up almost 2.5% of worldwide aquaculture production, have been attributed to Atlantic salmon swim bladder sarcoma virus (SSSV), which is genetically similar to the epsilonretroviruses (Statistics and Information Service of the Fisheries and Aquaculture Department, 2012, Paul et al., 2006). Other diseases in fish and amphibians have also been provisionally linked to epsilon-like retroviruses (Lepa and Siwicki, 2011, Masahito et al., 1995). However, no epsilon-like retroviruses causing disease in mammals or birds have been identified.

To date, evidence from ERVs has confirmed these viruses as primarily infections of fish. ERVs are retroviruses which have integrated into germline, rather than somatic, cells and are therefore transmitted vertically from parents to offspring and can become a permanent part of the genome of their host. ERVs are degraded over time by mutation and become inactive, but remain detectable in their host genome millions of years after integration. This means they provide valuable insight into the retroviruses a species has been exposed to, deep in its evolutionary history. Epsilon-like ERVs have been found in a diverse range of fish and amphibian genomes, suggesting a long-standing relationship with both these groups (Basta et al., 2009, Betancur et al., 2013,

Herniou et al., 1998). These retroviruses are thought to be the result of multiple integration events taking place over many millions of years, including several relatively recent insertions (Basta et al., 2009, Betancur et al., 2013, Herniou et al., 1998).

Genome-wide screening for all genera of retroviruses has been performed in many species of mammals and birds (Polavarapu et al., 2006b, Stocking and Kozak, 2008, Nellaker et al., 2006) revealing a rich diversity of gammaretroviruses, a genus closely related to epsilonretroviruses. However, epsilon-like ERVs have not been identified in most mammals. Some epsilon-like insertions have previously been found in the human genome. Tristem (2000) identified a group of approximately 70 highly degenerate sequences clustering with non-mammalian retroviruses in the human genome, named as the HERV.HS49C23 group and later subdivided into the HERV-L(b), HERV-R(c), HERV(AC0956774) and ERV(AC018462) families (Katzourakis and Tristem, 2005). These insertions were described as being more closely related to WDSV than to the gammaretroviruses. Oja et al. (2005) identified twelve epsilon-like insertions in the human genome and in our previous work (Brown et al., 2012) we characterised a group of epsilon-like ERVs in the horse genome, using a newly developed bioinformatics pipeline.

We have now screened 32 species of primates, rodents, lagomorphs (rabbits and pikas) and birds for epsilon-like ERVs using this pipeline and, unexpectedly, we have identified several groups of epsilon-like ERVs which appear to be ubiquitous in primates. The integration patterns and phylogeny of these primate epsilon-like (PE) ERVs suggest that they entered the genome of a common ancestor of old and new world monkeys at least 40 million years ago. These results raise several important questions about the origin and evolutionary history of the epsilonretroviruses and their relatives, their relationship with gammaretroviruses and their potential for cross species transmission.

7. 2. Materials and Methods

7.2.1. Genome Screening

A database of 382 *gag*, 670 *pol* and 356 *env* amino acid sequences was built to represent the diversity of known exogenous and ERVs. The viruses included in this dataset are listed in full in Appendix B.10. Details of the genomes screened in this analysis are listed in Table 8. All genomes were downloaded on 08-Mar-2013 from RefSeq release 57, NCBI Genome or Ensembl release 70. For genomes not assembled into chromosomes, scaffolds were concatenated into approximately chromosome-length strings for ease of analysis and later traced back to their original scaffold. Candidate ERV regions were identified using the Exonerate algorithm (Slater and Birney, 2005) and formatted using the Perl pipeline available at <https://github.com/ADAC-UoN/predict.genes.by.exonerate.pipeline>, under the protein2genome model with a minimum hit length of 200 amino acids without introns. When predicted genes overlapped, the gene with the highest Exonerate score was selected.

ERV DNA fragments predicted by Exonerate were verified using a TBLASTX (Altschul et al., 1990) search of the untranslated version of the input database described above. Sequences producing an alignment greater than 100 amino acids in length and with greater than 40% amino acid identity with a sequence in the input database [thresholds based on (Coffin JM et al., 1997)] were classified as ERVs. These sequences were aligned individually to each of the original untranslated input sequences listed in Appendix B.10 using EMBOSS water (Rice et al., 2000) which is based on the Smith-Waterman algorithm (Smith and Waterman, 1981) and finds regions of local similarity amongst otherwise dissimilar sequences. Sequences were categorised into genera according to their highest alignment score. Sequences which showed highest similarity to the epsilon and epsilon-like retroviruses were assigned to a

provisional epsilon-like dataset. All sequences in this dataset were divided by host and their nucleotide sequences were aligned to those of 34 known epsilon and epsilon-like retroviruses and 41 diverse gammaretroviruses using the localpair setting of MAFFT (Kato et al., 2002) with 1000 iterations (these sequences are highlighted in Appendix B.10). This alignment technique and settings were also used for all subsequent multiple sequence alignments. Maximum likelihood phylogenetic trees were built for these alignments using PHYML (Guindon and Gascuel, 2003) under the GTR model with aLRT branch support, no invariable sites, optimised across site rate variation and optimised tree topology. PHYML and these settings were also used for all subsequent tree building. Only sequences clustering within a monophyletic group of epsilon and epsilon-like retroviruses, distinct from the gammaretroviruses, with branch support greater than 75%, were kept in the dataset.

7.2.2. Comparison between Primate Genomes

The Compara EPO six primate alignment (C6P) (Ensembl release 74), an alignment of the DNA sequence of human, chimpanzee, gorilla, orangutan, rhesus macaque and marmoset genomes, was screened for loci containing an epsilon-like ERV pol fragment in at least one host and sequences from these loci were extracted. If there was at least 75% sequence identity between the ERV sequence and the sequence of any host within the ERV region, excluding gaps, the ERV was considered to be present in this host. All ERV sequences for each locus were extracted to form a dataset of epsilon-like ERV fragments in these six primates. Sequences from all hosts at each locus were aligned and PHYML phylogenetic trees were built for each locus. A consensus supertree representing all loci was built using CLANN (Creevey and McInerney, 2005). This analysis was repeated with loci divided according to the families described below.

Consensus nucleotide sequences for each locus from the C6P were generated using the alignments above and the *ambigcons* function of EMBOSS (21). Ambiguous characters were then replaced in equal proportions with each of the bases represented by the character. Sites with gaps in the majority of sequences were excluded from the consensus. This method was also used to build all subsequent consensus sequences. All consensus sequences were combined into a 7426 base pair multiple DNA alignment (including multiple gaps due to the degeneracy of the sequences). This alignment was used to build a phylogenetic tree and sequences were grouped according to this phylogeny. Each group was aligned and used to build a group consensus sequence. All group consensus DNA sequences were aligned with 38 known epsilon and epsilon-like retroviruses, with human ERV I, the closest known gammaretrovirus to the epsilon retroviruses (Herniou et al., 1998) as the outgroup, forming a 5510 base pair multiple alignment. A phylogeny was built from this alignment.

Candidate Exonate sequences from species outside of the six primate species in the Compara six primate alignment were aligned one by one to these group consensus sequences using EMBOSS *water* and assigned to a group according to their highest alignment score.

7.2.3. Genome Characterisation

To isolate LTRs, 8000 bp on either side of the *pol* gene region from each host at each locus was extracted. The regions from the two sides were then aligned to each other using EMBOSS *water* (Rice et al., 2000) which was then used to identify the subsection of this alignment with the highest alignment score. Sequences within this subsection from either side of the *pol* gene which shared 75% sequence similarity, were between 6000 and 15000 bp apart and were between 300 and 1500 bp in length were isolated as candidate LTRs. These thresholds are based on the range of retroviral genome sizes and LTR lengths listed in Bannert et al. (2010). These candidate regions were classified using

CENSOR (Jurka et al., 1996). Sequence pairs classified as ERV LTRs were then used as query sequences and aligned back to all the 8000 base pair regions flanking *pol* genes, again using EMBOSS water, and any new sequences identified were added to the dataset. Loci were dated using the equation $t = k/2N$, where t is time, k is divergence (number of sites at which the LTRs differ over LTR alignment length), and N is the neutral substitution rate of the host, assumed here to be the human neutral substitution rate of 4.5×10^{-9} substitutions per site per year. This is a common ERV dating technique (used for example in Sinzelle et al., 2011, Polavarapu et al., 2006a, Gifford et al., 2008). For loci with recognisable LTRs, human sequences were extracted and aligned to each other and clustered using a PHYLML phylogenetic tree. The human LTRs identified here were used as probes for a genome-wide BLAT search (Kent, 2002) of the human genome, using the UCSC server and a threshold of greater than or equal to 75% sequence identity and 300 base pair length (as above).

For the loci with recognisable LTRs, the 5' and 3' limits of the LTR provide the full span of the ERV, meaning other features of the ERVs could be identified and characterised. The regions between the LTRs were translated in all six reading frames to identify any potential open reading frames (ORFs). The regions between the LTRs and the *pol* regions were also compared using BLASTX (Altschul et al., 1990) to the UNIPROT database to identify any candidate *gag* or *env* genes and to a local database containing the WSDV accessory gene sequences (from GenBank accession NC_001867) to identify sequences resembling these genes. All regions showing significant similarity to any Gag, Env or accessory gene sequences were examined individually, aligned to the appropriate gene from WSDV and aligned to each other to establish if any degenerate ERV derived sequences were present.

7.2.4. Comparison with Other Mammals

The *pol* gene locations in humans and chimpanzees of loci with recognisable LTRs identified in all six primate species were compared to the Compara 37 mammalian genome alignment (C37M) (Ensembl release 74) to ascertain if these loci were conserved in non-simian primates or outside the primates (as described above for the C6P alignment). The regions of all genomes aligning to the human and chimpanzee epsilon-like *pol* gene fragments were extracted. For each host, the percentage of sites in each genome with an identical base to the ERV was calculated. For each species where no ERV was apparent, a 16,000 base pair fragment of the alignment was isolated from each locus, encompassing the site where the ERV was expected and the flanking sequence. A TBLASTN analysis was performed on these fragments using the consensus LTR sequences, *pol* gene sequences and *env* sequence as probes, to identify solo-LTRs or any other ERV fragments which may suggest deletion of the ERV.

7. 3. Results

Our analysis identified 854 *pol* gene sequences (821 using the Exonerate pipeline and 33 more in the locus-by-locus analysis) which form a reliable phylogenetic cluster within the epsilon and epsilon-like retroviruses. The sequences ranged from 568 to 2798 nucleotides in length, with a mean of 993 bp. These sequences were all found in primates and tree shrews (Table 30). Primates are generally divided into four major groups as follows: apes (humans, chimpanzees, gorillas, orangutans and gibbons), old world monkeys (monkeys native to Africa and Asia), new world monkeys (monkeys native to central and south America) and prosimians (tarsiers, lemurs, bushbabies and lorises) (Perelman et al., 2011). Tree shrews are the closest living relatives to modern primates (Perelman et al., 2011). Epsilon-like insertions were identified in all of these groups (Table 30). No epsilon-like insertions were found in rodents, lagomorphs or birds.

Table 30: The number of epsilon-like ERVs of each type (primate epsilon 1 to primate epsilon 3, PE1 to PE3) identified in each host species.

Details of hosts and genome builds can be found in Table 8. Highlighted species are those included in the Compara 6 Primate alignment.

Species	Group	PE1	PE2	PE3	Total
Human	Ape	50	25	6	81
Bonobo	Ape	33	26	4	63
Chimpanzee	Ape	45	23	6	74
Gorilla	Ape	46	22	5	73
Orangutan	Ape	38	20	6	64
Gibbon	Ape	19	26	4	49
Baboon	Old World Monkey	29	26	2	57
Crab-Eating Macaque	Old World Monkey	21	23	3	47
Rhesus Macaque	Old World Monkey	39	20	6	65
Marmoset	New World Monkey	31	15	4	50
Squirrel Monkey	New World Monkey	21	13	2	36
Tarsier	Prosimian	1	8	0	9
Aye-aye	Prosimian	39	49	25	113
Lemur	Prosimian	16	15	8	39
Bushbaby	Prosimian	0	3	3	6
Chinese Treeshrew	Tree Shrew	5	11	0	16
Northern Treeshrew	Tree Shrew	8	4	0	12
TOTAL	-	441	329	84	854

The C6P alignment allows comparison between specific loci in the genomes of six of the 15 species of primate screened here: four apes, one old world monkey and one new world monkey. The 407 epsilon-like ERV sequences we identified in these six species fell at 87 loci. The retrovirus was found at same position in all six C6P species at 36 of these loci and in three or more species at 75 loci. For the remainder, some species had the retrovirus and some did not, however there was insufficient information to distinguish between empty ERV insertion sites, solo-LTRs and a lack of sequence data, due to poor alignment quality at and around the locus.

For each of the 87 loci identified in the C6P analysis, a consensus sequence representing the locus was produced. Phylogenetic analysis showed that these consensus sequences fall into three clear families, provisionally named primate epsilon-like one to primate epsilon-like three (PE1 to PE3) (Figure 78). A consensus sequence was generated for each family based on this information, then sequences from the non-C6P species were assigned to these families using sequence similarity to this consensus. PE1, PE2 and PE3 were all present in all the major primate groups (Table 30). PE3 was not identified in tree shrews, however the total number of ERVs found in tree shrews was relatively small.

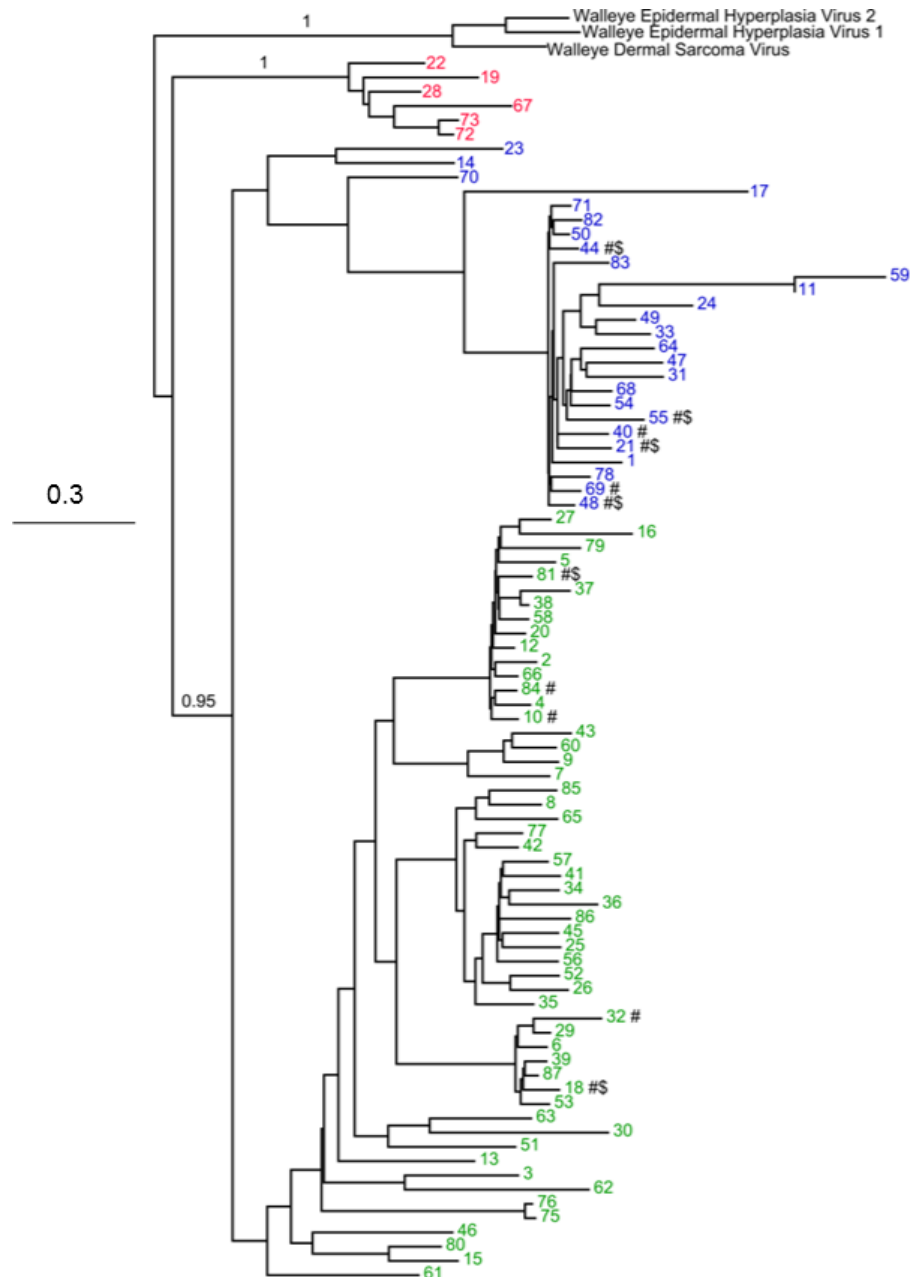


Figure 78: PhyML phylogenetic tree based on a 7426 nucleotide multiple alignment of the consensus sequences for 87 epsilon-like pol gene fragments found in primates, showing the clustering of primate epsilonretroviral loci into three major phylogenetic groups.

PE1 is shown in green, PE2 in blue, PE3 in red. Numbers represent locus numbers, which were assigned arbitrarily. The 11 sequences with recognisable LTRs are labelled with hash symbols (#) and the six sequences with recognisable LTRs which are conserved in the Compara six primate alignment species are labelled with dollar symbols (\$). Walleye dermal sarcoma virus and walleye epidermal sarcoma viruses one and two were used as an outgroup. Details of each locus are provided in Supplementary Table 3. Branch support values are aLRT values calculated in PHYML. Branch support values are only shown for the three major clades.

The majority of previously described epsilon-like ERVs in the human genome were identified using our pipeline and are labelled in Appendix B.11. We identified a total of 81 insertions in the human genome, consistent with the 70 ERVs clustering with non-mammalian ERVs identified by Tristem (2000). Our PE2 group appears to encompass Oja et al.'s (2005) "upper" group of epsilon-like ERVs and our PE1 group their "lower" group. Katzourakis and Tristem's (2005) HERV-AC018462 and HERV-L(b) groups fell into our PE1 group and their HERV-R(c) group into our PE2 group. Three previously described sequences were not identified in our study, the type member of the HERV-AC096774 group described by Katzourakis and Tristem (2005) and the chr1_684233 and chr17_47535521 groups described by Oja et al. (2005) 5000 bp from either side of human chr1_684233 (which corresponds to chr1 594413 in the most recent genome build) were analysed using BLASTX against the nr database and by alignment with known epsilonretroviral *pol* genes but nothing resembling a *pol* gene could be identified. Oja et al.'s chr17_47535521 was in the raw output from Exonerate but fell short of the quality threshold during our BLAST verification step, with the closest match to a known ERV a 64 amino acid segment sharing 54% identity with WDSV. HERV-AC096774 was not identified using Exonerate, however, as stated in Katzourakis and Tristem (2005) this sequence is very degenerate. Both of these sequences are most similar to our PE1 group.

The consensus sequences of PE1, PE2 and PE3 were incorporated into a phylogeny of known epsilon and epsilon-like retroviruses (Figure 79). Mammalian epsilon-like *pol* insertions in this phylogeny are the PE1, PE2 and PE3 consensus sequences, horse epsilon-like ERV fragments from our previous work (Brown et al., 2012), an example epsilon-like virus from Oja et al. and one chimpanzee ERV lineage previously categorised only as "Class I" (Polavarapu et al., 2006a). PE1, PE2 and PE3 form a moderately supported potential phylogenetic cluster with these known mammalian ERVs and the

reptilian epsilon-like ERVs. PE3 seems to be more closely related to the reptile epsilon-like ERVs than to the other mammalian insertions.

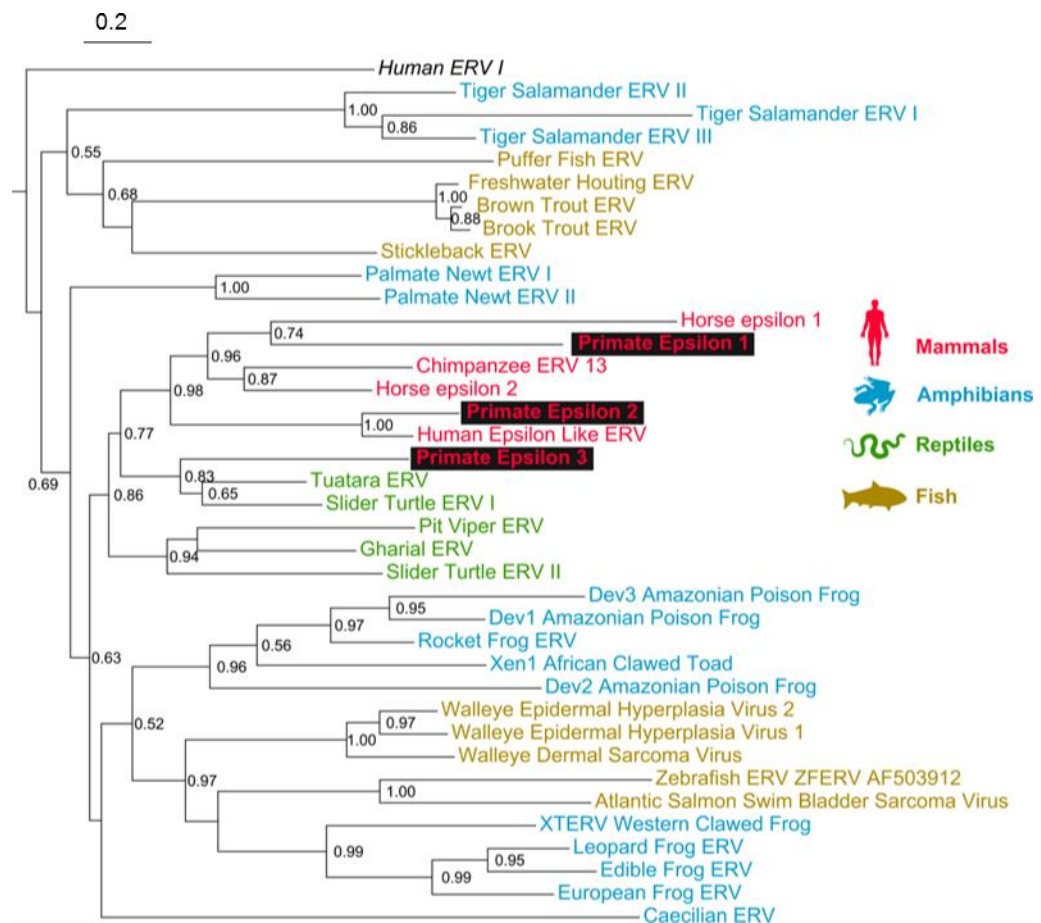


Figure 79: PhyML phylogenetic tree based on a 5510 base pair multiple alignment of the consensus sequences of three phylogenetic groups of primate epsilon-like pol gene fragments and known epsilon and epsilon-like retroviruses.

Mammalian epsilonretroviruses are shown in red, amphibians in blue, reptiles in green and fish in yellow. Newly identified sequences are highlighted. Full details of known epsilonretroviruses in this tree are provided in Appendix B.1. HERV-I is human endogenous retrovirus I, a gammaretrovirus. Branch support values are aLRT values calculated in PHYML, values below 0.5 are not shown.

Potential LTRs were identified flanking 11 of the 87 PE loci, the remainder were too degenerate for reliable LTR sequences to be detected. Dating based on LTR similarity at these loci gave a mean integration date of 34.43 million years ago, with values ranging from 16.48 to 90.49 million years. LTRs clustered into four types, designated type_1 to type_4. PE2 loci had type_1 or type_4 LTRs and PE1 loci type_2 or type_3. No LTRs were identifiable at PE3 loci. These results are summarised in Table 2. Type_4 LTRs were only identified at loci with a median age greater than 34 million years.

Table 31: The phylogenetic group, LTR_type, proportion of sites at which LTRs are not identical to each other and median age of each of the 11 epsilon-like ERV loci flanked by two recognisable LTRs.

Locus	Group	LTR_Type	LTR_Divergence	Median_Age
loc_18	PE1	type_3	0.078	17,319,367
loc_10	PE1	type_1	0.088	19,586,308
loc_81	PE1	type_2	0.100	22,173,007
loc_44	PE2	type_1	0.104	23,052,162
loc_69	PE2	type_1	0.107	23,772,610
loc_48	PE2	type_1	0.117	26,073,350
loc_84	PE1	type_2	0.139	30,939,030
loc_55	PE2	type_4	0.155	34,500,254
loc_21	PE2	type_4	0.176	39,089,995
loc_32	PE1	type_3	0.181	40,322,514
loc_40	PE2	type_4	0.185	41,044,747

Six loci had recognisable LTRs and were identified in all six C6P species. The C37M alignment was used to establish if these specific loci are found in all primates and if they are found outside the primates. The sequences were identifiable at the same positions in all apes, old world monkeys and new world monkeys in the alignment. However, at these positions no ERV sequences were identifiable in prosimian primates or any non-primates, including tree shrews. The C37M alignment is of poor quality for some species

for some regions of the genome, so it is not possible to definitively state that these insertions were absent, but there was no evidence of these insertions at any of the six loci in any of the three prosimian species or one tree shrew species in the alignment. TBLASTN analysis also did not identify any retroviral LTRs, *pol* or *env* gene fragments in these regions or the surrounding sequence in prosimians or non-primates. Therefore, it appears that the insertion of epsilon-like ERVs at these specific sites occurred after the split between tarsiers and old/new world primate ancestors (65 million years ago) but before the split between the ancestors of old and new world monkeys (42 million years ago) (Perelman et al., 2011). These dates are broadly consistent with the estimates above based on LTR divergence. Given that epsilon-like ERV fragments were absent at these loci in prosimians and tree shrews, the prosimian and tree shrew epsilon-like ERV fragments we identified appear to be the result of separate integration events at different integration sites to those in apes, old world monkeys and new world monkeys.

Using the human LTR sequences identified here as probes against the human genome, 777 further potential LTRs were identified. 14 pairs were identified between 8,000 and 15,000 bp apart, suggesting that the ERV sequence between the LTRs has not been deleted but is too degenerate to recognise. The remaining 749 are likely to be solo-LTRs, the result of recombination between the two LTRs flanking an ERV sequence. This gives a ratio of 749 solo-LTRs to 95 ERV sites which have not recombined in the human genome (including the 81 identified with Exonerate and the 14 pairs encompassing unrecognisable ERVs). In mice, the half life for an ERV to recombine and form a solo-LTR is estimated at 0.8 million years (Nellaker et al., 2012). The recombination rate of mice is around half that of humans per generation (Jensen-Seaman et al., 2004) but the mouse generation time is around one fiftieth of that of humans (Keightley and Eyre-Walker, 2000), giving an estimated ERV to solo-LTR half life of 20 million years in humans. At this rate it would take approximately 60 million years to go from 844 ERV sites to 95 ERV sites and 749 solo LTRs, which is within our predicted range of insertion dates.

For the 11 loci with recognisable LTRs, the 5' and 3' limits of the LTR provided the full span of the ERV, meaning other features of the ERVs could be identified and characterised (Appendix B.12). WDSV is the type species for the epsilonretroviruses (International Committee on Taxonomy of Viruses, 2002) and the only epsilonretrovirus with a reference sequence (Genbank accession NC_001867) and so was used for comparisons. Apart from two endonuclease gene insertions, likely to be the result of later retrotransposition events by non-LTR retrotransposons, in humans at locus 84 and chimpanzees at locus 48, the longest ORF was a 296 amino acid, or 888 base pair fragment at locus 32, starting within the 5'LTR and ending within the region where *gag* would be expected. The protein encoded by this ORF shares no homology to any known retroviral protein (determined using BLASTP) and is considerably shorter than any major retroviral protein (WDSV has a 582 amino acid Gag, 1171 amino acid Pro-Pol and 1225 amino acid Env). Therefore, it is very unlikely that any of these ORFs could produce functional viral proteins. BLAST searching identified small *gag* fragments (less than 400 bp) with homology to WDSV between *pol* and the 5' LTR of loci 18, 21 and 44 and *env* fragments sufficient to combine into a 1330 base pair consensus at loci 10 and 81 (Appendix B.12). These *gag* and *env* sequences were however too degenerate for meaningful phylogenetic analysis. No sequences with homology to the three WDSV accessory genes, *orf-A*, *orf-B* and *orf-C* were identified. A partial genome structure for the PE group was deduced from these results and is shown in Figure 80. If accessory genes are excluded, the length of the PE genome and the position of the *pol* gene and *env* fragment are consistent with WDSV and the gaps between these regions are sufficient for the remainder of a functional epsilon-like ERV to have been present at some point in the evolutionary history of the host.

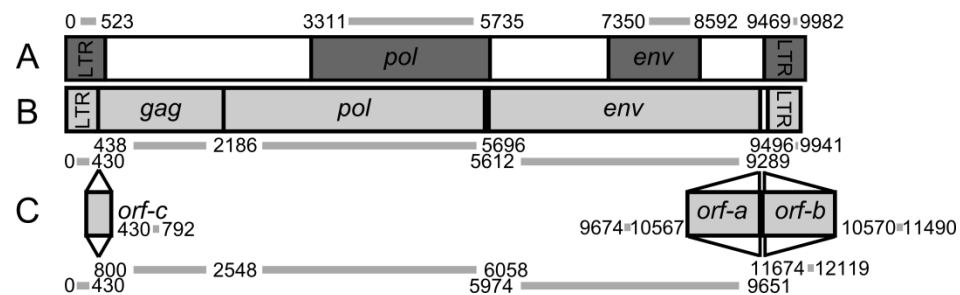


Figure 80: A comparison of identified regions of the PE genome (A) and the reference genome of WDSV (GenBank Accession NC_001867) with orf-a, orf-b and orf-c excluded (B) and included (C) in the genome length and gene position calculations.

Positions for PE are means across all loci with identifiable LTRs.

A supertree representing the evolutionary relationships between sequences from each host at each locus was generated (data not shown). This tree is identical to the consensus host phylogeny, based on 17 host genes, available through the 10k trees project (Arnold et al., 2010). If the loci are divided by family, PE1 and PE2 show this relationship with 100% support for all branches, while PE3 shows ambiguity in the relationship between human, gorilla and chimpanzee, a relationship which is also sometimes ambiguous in evolutionary analyses of the host (Chen and Li, 2001).

7. 4. Discussion

These results confirm the presence of a group of endogenous epsilon-like ERVs in these fourteen primate species and in two species of tree shrew, the closest living relatives of the primates. The sequenced primates are from diverse geographical regions and represent all major primate taxonomic groups, so the identification of PE insertions in all of these hosts suggests that PE is found in all primates. By looking at individual PE loci in six primate species, we have confirmed that PE is likely to have entered the genome of a common ancestor of apes, old world monkeys and new world monkeys, while PE insertions in prosimian primates and tree shrews are likely to represent

separate integration events in ancestors of these species. Many of these ERVs have not been identified previously. This is most likely due to the degree of degeneration of these sequences and the diversity of our input dataset of known retroviruses, which is considerably more comprehensive than those which are generally used.

Mammals, reptiles and birds make up a distinct group in vertebrate phylogeny known as amniotes (Meredith et al., 2011). The phylogenetic tree shown in Figure 79 suggests that all three families of PE insertion may form part of a group of epsilon-like ERVs unique to the amniotes, along with several previously characterised mammalian and reptilian epsilon-like ERVs. The known human epsilon-like ERVs (Katzourakis and Tristem, 2005, Tristem, 2000, Oja et al., 2005) seem to represent members of our PE1 and PE2 families and chimpanzee endogenous retrovirus lineage 13 (Polavarapu et al., 2006a) appears to be a member of PE1. PE3 clusters robustly with a group of reptilian ERVs. Our previously identified horse epsilon-like ERVs (Brown et al., 2012) fall within this provisional amniote ERV group.

The shared insertion sites in new and old world monkeys provide a minimum age for circulation of the exogenous versions of these epsilon-like ERVs of 42 million years ago, and the absence of these shared insertion sites with tarsiers provides a maximum age of 65 million years (Perelman et al., 2011). All known endogenous fish epsilon-like ERVs are considerably more modern than this, with the oldest estimated at 3.79 million years old (Basta et al., 2009). Only one amphibian epsilon-like ERV currently has an estimated integration date, an insertion in *Xenopus tropicalis* dated at 41 million years old (Sinzelle et al., 2011). This date is consistent with the relationships between amphibian retroviruses shown in Figure 80. Therefore, amniote and amphibian retroviruses appear to have been circulating during approximately the same time period while fish endogenous epsilon-like retroviruses are much more recent. The structure of the epsilon-like ERV phylogeny is best explained by a member of a group of circulating amphibian retroviruses 40 to 60 million

years ago entering amphibian genomes multiple times and forming two distinct phylogenetic groups, and a single strain crossing into amniotes and then diversifying to infect different amniote species.

This long gap between the ancient amphibian / amniote viruses and the modern fish viruses raises questions about the evolution of epsilon-like ERVs. The degeneration seen in amphibian and primate endogenous epsilon-like ERVs means they are unlikely to have had the potential to produce functional viral particles recently enough to be responsible for these integrations into fish. If exogenous members of the PE or horse epsilon-like ERV families had remained infectious through this period, there would most likely be more modern integrations detectable in our genome screens, though the possibility remains that other mammals have as yet unidentified epsilon-like ERVs, particularly as horses and primates are quite divergent host species. The remaining explanation is that exogenous epsilon-like retroviruses have been circulating throughout this period in another host or group of hosts and later crossed into fish. Significantly more screening would be needed to identify this host. The three distinct groups of fish/amphibian insertions in Figure 79 suggest that cross-species transmissions into fish have occurred at least three times. As all three phylogenetic groups of fish epsilon and epsilon-like retroviruses are more similar to amphibian ERVs than amniote ERVs, then amphibians could be a candidate. Screening of amphibians for ERVs to date has also been minimal. It is also possible that epsilon-like retroviruses have been circulating amongst fish throughout this time and that there are considerably more epsilon-like ERVs in fish which are yet to be discovered.

The exogenous fish epsilonretroviruses WDSV and WEHV encode three accessory proteins, Rv-cyclin (encoded by *orf a*), Orf-B and Orf-C (Rovnak and Quackenbush, 2010) (Figure 80). We did not identify the genes encoding these proteins at any PE locus or in the horse epsilon-like ERVs. Rv-cyclin and Orf-B are involved in tumour development while Orf-C is involved in apoptosis and tumour regression and tumour development (Rovnak and Quackenbush,

2010). These genes are essential for WDSV proliferation and dissemination (Rovnak and Quackenbush, 2010). However, these genes are not universal in fish retroviruses, for example, they are absent in zebrafish ERV (Shen and Steiner, 2004) and Atlantic salmon swim bladder sarcoma virus (Paul et al., 2006) so they are likely to represent a later acquisition in the lineage leading to WDSV and the WEHVs.

We did not identify any epsilon-like ERVs in any of the 11 rodent species or two lagomorphs we screened. Rodents and lagomorphs are known to carry many endogenous and exogenous gammaretroviruses and appear to have a high vulnerability to retroviral infection (Stocking and Kozak, 2008, Baillie et al., 2004, McCarthy and McDonald, 2004) so it is surprising that their closest sequenced relatives have endogenous epsilon-like ERVs but they do not. One possible explanation for this is that one of the diverse gammaretroviruses infecting rodents offered a protective effect against epsilon-like retroviruses. The use of ERVs as restriction factors against exogenous pathogens is a known mechanism used by some hosts (Arnaud et al., 2007). Alternatively, epsilon-like retroviral host range may depend on a combination of host restriction factors and viral accessory genes in a fashion similar to simian immunodeficiency viruses (SIVs). Finally, it is possible that rodents and lagomorphs lack a receptor which epsilon-like retroviruses require and which is present in primates and horses. The two bird species screened here also lacked epsilon-like ERVs. Birds have an unusual complement of ERVs compared to mammals, which again might have acted as a barrier to epsilon-like retrovirus infection. It is also possible that there are epsilon-like ERVs in other bird species which were not analysed here.

As fish still have active epsilonretroviruses and primate ancestors have clearly been susceptible to epsilon-like retroviruses in the past, it is not inconceivable that fish epsilonretroviruses could enter the human genome again. Further research is needed to establish if the lack of modern infections in mammals is due to a restriction factor or if mammals remain vulnerable to epsilon or

epsilon-like retroviruses. Any restriction factor identified may be of interest to the aquaculture industry in terms of its potential in the control of WDSV and WEHV. The degree to which all the identified PE insertions have degenerated and the lack of functional *gag* and *env* genes make it very improbable that these loci could generate an active epsilon-like retrovirus even by recombination.

In conclusion, epsilon-like ERVs appear to be common to all primate genomes and are likely to be widespread amongst mammals, although they are absent in rodents and lagomorphs. Amniote epsilon-like ERVs may form a distinct group within the epsilon and epsilon-like retrovirus phylogeny and are most likely to be the result of diversification of a cross-species transmissions of viruses circulating 40 to 65 million years ago. Epsilon-like retroviruses appear to have continued to circulate since this time and have most recently invaded the genomes of fish but further research is needed to establish whether these viruses originated in fish or other hosts.

Chapter 8. General Discussion

This study into the ERV content of the Euarchontoglires using a novel analysis pipeline has yielded several important results.

The Exonerate pipeline described in Chapter 2 is clearly an appropriate tool for ERV characterisation and analysis. Numerous ERVs were identified in all genomes screened, including those which have been identified previously and many novel insertions. This approach allows ERVs to be identified quickly and characterised in some depth and methodologies for detecting other key features of ERVs which may be of interest have been successfully used. Using this pipeline, almost 200,000 ERV-like fragments have been identified. This represents one of the most detailed studies into the ERVs of primates and related species to date.

Several themes became apparent when reviewing these results as a whole, which will be discussed in this chapter.

8. 1. Vector Species and Cross-Species Transmissions

One of the most notable properties of the ERVs identified here was the degree to which vector species appear to have been involved in the movement of retroviruses. The transmission of a retrovirus from one host to another via a third species has never been confirmed to occur in nature. However, it has been demonstrated via human intervention, as the two lineages of avian REV in ducks and chickens are closely related to ERVs found in Malagasy carnivores (Niewiadomska and Gifford, 2013). This transmission is thought to be the result of experimental infection with parasites cultured at a zoo where an infected exotic animal was present (Niewiadomska and Gifford, 2013). In

other cases, a vector species cannot be confirmed but is the most feasible explanation for transmission of a retrovirus. For example, the transmission of GALV to koalas must have occurred via at least one intermediate host and one of these hosts has now been potentially identified (Simmons et al., 2014).

Here, a similar case was identified for the primate lentiviruses discussed in Chapter 5. The newly identified endogenous lentivirus in the bushbaby *G. moholi* is almost identical to the endogenous lentiviruses previously identified in lemurs, but these lemurs have not been in contact with *G. moholi* for at least 50 million years. Unless the virus has remained almost unchanged throughout this time and yet not entered any of the 12 other prosimian primate species screened, or all copies in these species have been independently deleted, a vector species provides the only remaining explanation for this transmission. The nature of this vector species is not yet known, however, as is the case for many ERVs, rodents and bats are amongst the most likely candidates for this transmission. In the HERV-F like group, members of the HERV-H-RTVLH2 and HERV-F subfamilies were identified in aye-ayes but in no other prosimians (Figure 46, Figure 47), despite the fact that aye-ayes are only found in Madagascar, where none of the other hosts are present, suggesting a similar vector-species transmission to the hypothesised event leading to pSIVgml and pSIVfdl. This may also be the case for the betaretroviral HERV-K(HML-2), as the aye-aye has recent HML-2 insertions transmitted from another geographically distant host, probably the baboon. Together, these results suggest that transmission of retroviruses across the Mozambique channel between Madagascar and the mainland may have been more common than was previously thought. As discussed in Chapter 5, waves of migration across this channel have occurred throughout history and aerial vectors such as bats, insects and birds still provide a link between Madagascar and mainland Africa.

Madagascar is not the only region subject to this type of vector transmission events. Primate ERVs in the gammaretroviral HERV-I group showed host tracking but their rodent counterparts and those in reptiles, marsupials and

fish have a very scattered distribution both phylogenetically and geographically, indicative of sporadic cross-over events from another host. Birds are a likely candidate for a vector species here, since avian HERV-I-like ERVs are considerably more numerous and diverse (Figure 15) and birds often have a much wider geographic distribution than mammals.

Cross species transmission events not involving an intermediate host have also occurred. One of the clearest examples discussed here is the transmission of GALV to gibbons (Chapter 6). Despite examining tissue samples from numerous gibbon species, veterinary records and the sequenced gibbon genome, there was no evidence that this virus is established in gibbon populations. Instead, it appears that this virus was transferred to a single gibbon, probably from a rodent vector, then to a limited population of other primates directly in contact with this gibbon. As the SEATO gibbons were regularly experimentally challenged with potential rodent and primate pathogens, a scenario similar to that described for REV, involving a contaminated experimental treatment, is possible. The initial GALV outbreak was widely reported when it occurred, with significant resources devoted to its identification and analysis. GALV is still commonly listed as a threat to gibbons today. This demonstrates the importance of analysis of potential vector species and reservoir hosts during retroviral outbreaks.

8. 2. Host Range and Recombination

Cross-species transmission events can only occur if the retrovirus can replicate in both hosts. Factors determining retroviral host range include its receptor, any restriction factors in the host and the lifestyle and geographic range of any susceptible hosts. If multiple retroviruses are able to enter the same type of cell, this provides an opportunity for recombination events, potentially

opening up a new host range for the virus by circumventing restrictions on viral entry and replication in the new host.

The host range of different ERV groups is highly variable. For example, within the larger HERV-F-like group, the HERV-Fc1 like group was identified in seven hosts, all apes and new world monkeys, while the HERV-Fc2 group was identified in 18 hosts, including apes, old and new world monkeys, prosimians, tree shrews, lagomorphs, old and new world rodents and ferrets. The most dramatic examples of recombination events across multiple hosts were seen in the SERV-like group. The phylogeny shown in Figure 66 demonstrates this clearly, with closely related sequences identified here in many pairs of distantly related hosts. The reason why some retroviruses show a stronger tendency towards recombination and cross-species transmission than others is not well known. Interestingly, the SRVs, BaEV, REV and RD114 viruses in this group all share a receptor (Overbaugh et al., 2001, Koo et al., 1992), despite SRVs having a gammaretroviral *env* gene and REV, BaEV and RD114 a betaretroviral *env* gene (the *env* gene is responsible for receptor interaction). This potentially explains the high recombination frequency in this group as the result of gamma and betaretroviral activity within the same cells. The receptors for SERV, TvERV, SMRV and MusD, or ERVs we have identified in this group in lemurs, guinea pigs, rats and mice are not known but would be of interest in further analysis of this group in the light of this potential mechanism and the tendency for this group to swap genes and hosts so readily.

8. 3. Potentially Active ERVs

A number of ERVs were identified which may have the potential to produce active viral particles. Unexpectedly, the majority of these were found in guinea pigs (Chapter 4). As Figure 36 demonstrates, guinea pigs did not have an unusually high number of regions with recognisable *gag*, *pol* and *env*

fragments, although an above average number were detected for a genome only assembled to a scaffold level. However, where these regions were present they were more likely to contain intact ORFs than those in other hosts.

Two loci were identified in guinea pigs in the REV-like group of non-recombinant gammaretroviruses with full-length ORFs for *gag*, *pol* and *env* (Figure 55). These loci showed evidence of selection to maintain viral function, which suggests they entered the genome relatively recently. The *pol* genes of these loci clustered with approximately 70 other, less intact members of the same group in guinea pigs and with *pol* genes from ERVs in related new world rodents (Figure 54). 17 of the guinea pig loci but none of the loci in related rodents were estimated to be less than one million years old. This suggests a recent burst of activity of members of this group in guinea pig hosts.

Another group of potentially intact ERVs was also identified in guinea pigs in the SERV-like group, with betaretroviral SRV/SERV-like *gag* and *pol* ORFs and gammaretroviral MLV-like *env* ORFs (section 4.6.2). 31 other loci were found in this group in guinea pigs, although these were less intact. Only guinea pigs had this particular pattern of recombination, with other recombinants in this family tending to have *gag* and *pol* genes more similar to SMRV and MusD.

Guinea pigs are not currently known to have active endogenous retroviruses but little research has been carried out on this topic. However, guinea pigs are widespread as pets, food and laboratory subjects, so a thorough understanding of their ERVs is important. Guinea pig cells used in culture are known to release retrovirus-like particles (Dahlberg et al., 1980) and these may stem from one or more of these newly identified loci. It would be worthwhile to analyse these particles using updated methodology and to screen multiple guinea pigs or related hosts for these insertions to see if they are polymorphic, which would suggest that these viruses are still spreading through the guinea pig population and may pose a risk for cross-species transmission.

8. 4. Comparison of Genomes

During this analysis it became apparent that even genomes which have been extensively studied in the past, for example the human genome, contain ERVs which are very poorly known. The epsilon-like retroviruses discussed in Chapter 7 provide the clearest example of this. Although there have been scattered reports of human ERVs clustering with “non-mammalian” retroviruses in the past (Tristem, 2000, Katzourakis and Tristem, 2005), these retroviruses have never been discussed in any detail and have not been characterised in non-human primates. These ERVs are numerous, ubiquitous in primate genomes and provide important clues as to the evolutionary history of the epsilonretroviruses and of retroviruses in general.

Another recurring theme was the presence of diverse and numerous ERVs in the genomes of primates which are usually not analysed in depth, specifically the new world monkeys and prosimians. The most noteworthy result of this project was probably from a prosimian host, with the identification of bushbabies, specifically *G. moholi*, as candidates for the original hosts of SIV in primates and therefore, indirectly, of HIV in humans (Chapter 5). pSIVmb is the oldest known lentivirus of a mainland African primate, estimated to have circulated two to five million years ago. As *G. moholi* has a range which overlaps with that of known SIV reservoirs, it is feasible that this was one of the earliest primate lentivirus hosts and was involved in the transmission of these viruses to the simian primates. Many other groups of prosimian ERVs were identified, including endogenous epsilon-like fragments (Chapter 7), recombinant SRV related ERVs in lemurs (Chapter 4) and a large HERV-W like group in bushbabies (Chapter 4). The ERV content of new world primates was more similar to that of old world primates than has previously been established, for example ERVs clustering with some of the HERV-K lineages were identified in new world primates for the first time. Several groups of HERVs could be traced back to an ancestor prior to the divergence of old and

new world primates. These ERV discoveries highlight the limitations of many ERV studies where only humans and other great apes are considered in searching for potentially conserved functions of ERVs or examining structural differences in genome architecture with respect to repetitive elements like ERVs.

8. 5. Relationship between ERVs and XRVs

This study provides some insight into the relationship between ERVs and XRVs. While it may seem counterintuitive to include slowly evolving genomic “fossil” elements such as ERVs in the same phylogenetic analysis as very rapidly mutating XRVs there are a number of lines of reasoning that justify this.

The basic mutation rate of HIV-1 is estimated at over 1×10^{-3} mutations per site per year (Jenkins et al., 2002) while the human mutation rate is estimated at 2.5×10^{-8} mutations per site per year (Nachman and Crowell, 2000). This is a difference of over five orders of magnitude. Despite this, ERVs which integrated millions of years ago and modern XRVs maintain sufficient sequence similarity to be incorporated into the same phylogenetic analyses.

There are several possible reasons for this. The first is selection, although XRVs evolve extremely quickly, they are still subject to evolutionary constraint. Viruses with mutations which inactivate their progeny will not be able to reproduce successfully, so clearly only certain mutations can be passed through the population. Evolutionary constraint is highly variable for different parts of the genome (Overbaugh and Bangham, 2001). Despite the high genomic plasticity of a typical retrovirus, the proportion of the genome required for the retrovirus to function is likely to be sufficient for at least part of an XRV to remain recognisable over periods of time consistent with the gaps between ERV endogenisation events.

Secondly, the mutation rate of XRVs is often considered to be that of HIV, as it is by far the most studied retrovirus. However, other XRVs change considerably more slowly, for example the receptor binding protein encoded by the *env* genes of FeLV strains isolated on different continents more 10 years apart has been shown to be 98% identical (Overbaugh and Bangham, 2001). Although this is still extremely fast by mammalian standards, given that a similarity level of 25% is not unusually in retroviral phylogenetics, this already allows almost 1000 years of evolution without the *env* gene becoming unrecognisable. Li et al. (1999) found traces of HTLV-I in mummies estimated at 1,500 years old with only a one to two percent sequence divergence from modern HTLV-I in an LTR / accessory gene fragment, suggesting an even lower evolutionary rate in this virus. As ancient exogenous retroviral sequences are unavailable, it is not possible to know how rapidly sequence evolution has really occurred in XRVs beyond very recent history.

Finally, it is possible that switching between an endogenous and exogenous form has been more common in retrovirus evolution than is currently known. Recently integrated ERVs have been demonstrated to sometimes have the potential to become reactivated, either by back mutation or recombination, and resume an exogenous lifestyle. Young et al. (2012) demonstrated the emergence of a replication competent ecotropic MLV strain in immune-deficient mice via recombination between a replication-deficient ERV and a replication-competent ERV unable to replicate in murine cells. Retroviruses undergoing an endogenous period would have a reduced substitution rate, so on becoming exogenous would have greater sequence similarity to ancient retroviruses than their contemporary XRV counterparts. However, the ability to become reactivated is highly dependent on the number of mutations acquired during the endogenous period, meaning that very ancient ERVs are extremely unlikely to be able to become active again.

8. 6. Defining an ERV Group

This study provides some insight into the naming conventions of ERVs, which varies greatly in the literature. Currently, ERVs are often categorised based upon related XRVs, but the degree of relatedness is often not specified, for example Borysenko et al.'s (2008) "MLV-like" chicken ERV is quite distantly related to MLV (Figure 41) and was classified here as HERV-I-Like. Using host name to categorise ERVs is also common, for example Tristem et al. (1996), Martin et al. (1999) and Gifford et al. (2005) defined large numbers of retroviruses based only on their host and genus, the majority of which have never been examined in further detail. If a large number of ERVs are identified in a single host they are usually given the name of the host and a number [e.g. Polavarapu et al. (2006a), Huda et al. (2008), Garcia-Etxebarria et al. (2010)]. MLVs are named as ecotropic, polytropic, modified polytropic or xenotropic according to the types of cells in which they can replicate (section 1.4.3.6). HERVs are traditionally named using the single letter code according to the amino acid specificity of the tRNA to which they bind, so, for example, HERV-Ks bind to a lysine tRNA (Bannert and Kurth, 2006). Even newly discovered ERV groups are not named according to a consistent convention, for example prosimian endogenous lentiviruses are named as pSIV plus the initials of the common name of their host (e.g. pSIVgml in the grey mouse lemur) (Gifford et al., 2008), carnivore endogenous lentiviruses as either Mustelidae endogenous lentivirus (MELV) or endogenous lentivirus (ELV) plus the initials of the scientific name of their host (e.g. MELV/ELVmpf in *Mustela putorius furo*) (Han and Worobey, 2012b, Cui and Holmes, 2012), and the two Lagomorph endogenous lentiviruses as rabbit endogenous lentivirus type K (RELIK) and "hare RELIK" (Katzourakis et al., 2007, Keckesova et al., 2009).

The levels of classification of ERVs are also ambiguous. At the broadest classification level, the majority of older studies and many modern studies use the Class I to Class III system described in section 1.4.1, although classification

according to genus is becoming more common. Some authors discuss ambiguities in the genus-level classification system, describing a continuum between the gammaretroviruses and epsilonretroviruses (Jern et al., 2005) and between the alpharetroviruses and betaretroviruses (Bolisetty et al., 2012). For more specific classifications, the definition of an ERV “lineage” as all the products of a particular integration event (Gifford and Tristem, 2003) is probably the least ambiguous, however, as demonstrated here it can be extremely difficult or impossible to distinguish the products of a single integration event followed by retrotransposition from the products of multiple integration events. This lack of specificity can lead to ambiguity about the particular ERV being discussed. The same ERV may also be described on several occasions under a different naming convention. Phylogenetic analyses are often uninformative due to a lack of understanding of the diversity of known ERVs.

Based on our results, we propose a provisional, modified ERV classification scheme, as outlined in Figure 81. Our results suggest that there is no evidence for ambiguity of ERVs on a genus level. For the alpharetroviruses and betaretroviruses, the distinction was clear in terms of host range, as 93% of alpharetrovirus-like fragments were avian, while 99% of betaretrovirus-like fragments were mammalian. Phylogenetically, the alpharetroviruses and betaretroviruses have been shown to form unambiguous, monophyletic groups [for example by Gifford et al. (2005) and Jern et al. (2005)]. In the publication which discusses these two groups as ambiguous (Bolisetty et al., 2012), no overlap between the groups is evident. Therefore, for the currently known ERVs, the alpha-beta distinction appears to be sufficient. The other two genera which have been referred to as ambiguous are the epsilonretroviruses and gammaretroviruses. However, we did not find any real overlap between these groups. The epsilonretroviruses defined in Chapter 7 were those which clustered unambiguously with the epsilonretroviruses when tested against a mixed epsilon and gamma dataset. Previous work [for example Jern et al. (2005) and Herniou et al. (1998)] has also identified these groups as distinct.

Given these results, the current genus level classifications appear to be adequate.

Below the genus level, division of the gammaretroviruses and betaretroviruses into the broad groups discussed in Chapter 4 is sufficient, at least for *pol* gene based classification of the currently known mammalian and avian retroviruses, as it distinguishes between and yet accounts for 80% of the gammaretroviral *pol* gene sequences in the input dataset and 92% of the betaretroviruses. 76 of the 79 undefined gammaretroviral sequences are from reptiles, amphibians or fish and 11 of the 15 undefined betaretroviruses from marsupials, so hosts which have not been examined in detail are likely to yield further groups. However, all of the new gammaretroviruses and betaretroviruses identified in the Euarchontoglires fall securely into the groups described in Chapter 4 (Figure 81).

This classification system proposed here requires a comprehensive test dataset containing a single, well-defined ERV sequence for each known ERV “type”, with types based on the well-supported, monophyletic groups of ERVs within each genus. These types will be referred to provisionally as ERV supergroups. A newly identified ERV is assigned to an existing supergroup if it is more closely related to the type sequence of that group than to members of any other group and if it falls into the group in phylogenetic analysis, with strong branch support. If a newly identified ERV does not meet these criteria for any existing supergroup a new supergroup can be defined. A third level has been defined which is only appropriate for very large supergroups, such as the HERV-F-like group of gammaretroviruses and the HERV-K-like group of betaretroviruses, provisionally referred to as “subgroup” level classification (Figure 81). This is defined in the same way as a supergroup but using a test dataset containing more closely related sequences from within a single supergroup.

The next level of classification which has been used in this study is that of a “cluster” of ERVs. This refers to monophyletic groups of ERVs from the same host, excluding those from other hosts, within a subgroup. These are

synonymous with the clusters defined in section 3.2. Some of the ERVs identified are the only member of their cluster, while others are members of very large groups (the largest is a cluster of 1374 IAP-like elements in the mouse genome) (Figure 81).

Finally, ERVs can be defined individually by their position in the genome, as a single locus (Figure 81). This is important and differs from the classifications commonly used in the literature, which often define a group of ERVs but not the specific locus being examined. Particularly with the advent of next generation sequencing techniques, the genomic location of ERVs is important as other genomic features, for example epigenetic modifications, transcriptome data and SNP data, are defined by in this way and this consistency allows ERVs analysis to be integrated into other genomic studies.

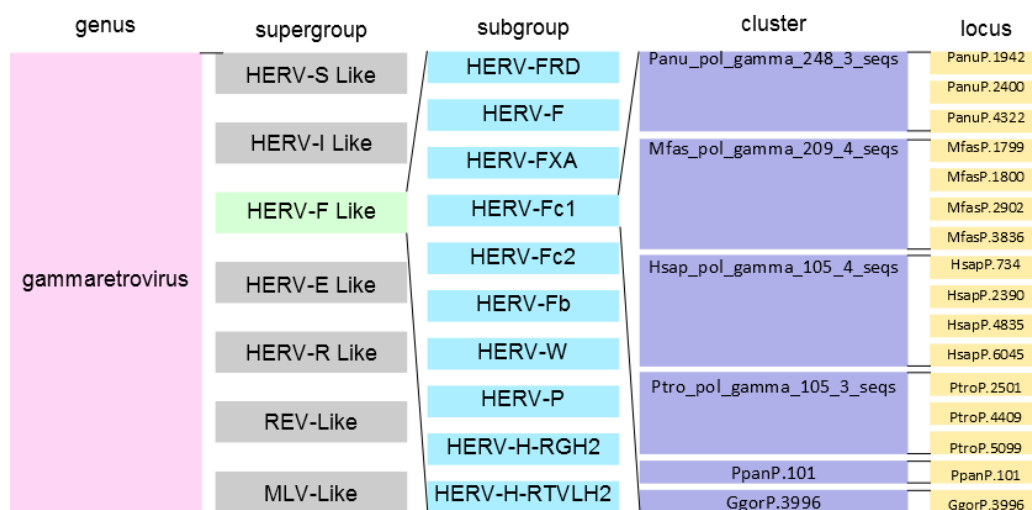


Figure 81: The classification scheme proposed for newly identified ERVs, using the 16 HERV-Fc1 like loci as an example.

This classification scheme is not complete, as many hosts have not been screened comprehensively for ERVs so many new supergroups, subgroups and possibly genera are likely to become apparent in the future. It is also not ideal, as it is still dependent on the quality and size of the test dataset, allowing the user to establish the level of similarity they consider to be acceptable for ERVs to be classified into a single group. It relies upon the existing names of ERVs

and ERVs, despite their inconsistencies. An ideal classification scheme would involve a publicly available, curated, well-designed group of test sequences derived from as many genomes as possible and named using a consistent convention, however this is not currently realistic.

8. 7. Predicting ERV Diversity

Completing this large scale study into the ERV content of various mammalian genomes allows us to examine the evolutionary forces which come into play in determining the outcome of the integration of a retrovirus into the germline and to test various hypotheses about the factors affecting the success of an ERV group.

The two ERV groups whose subgroups were most comprehensively characterised in this study were the eight HERV-K-like and nine HERV-F-like supergroups. All of these subgroups are present in consistent numbers in the genomes of humans, chimpanzees, bonobos and gorillas, have a reliable estimated integration date range in these hosts and are thought to have entered these genomes before they diverged from their last common ancestor. Therefore, these 17 subgroups in these four hosts were used as a test dataset for models predicting ERV diversity.

Firstly, these results show no particular relationship between ERV count and time. The mean ERV count in each subgroup per host in the test dataset and the mean estimated age of each subgroup (Figure 82) provide evidence for this. Pearson's correlation coefficient was calculated for this dataset and showed no significant relationship ($r^2 = 0.183$, $p = 0.497$).

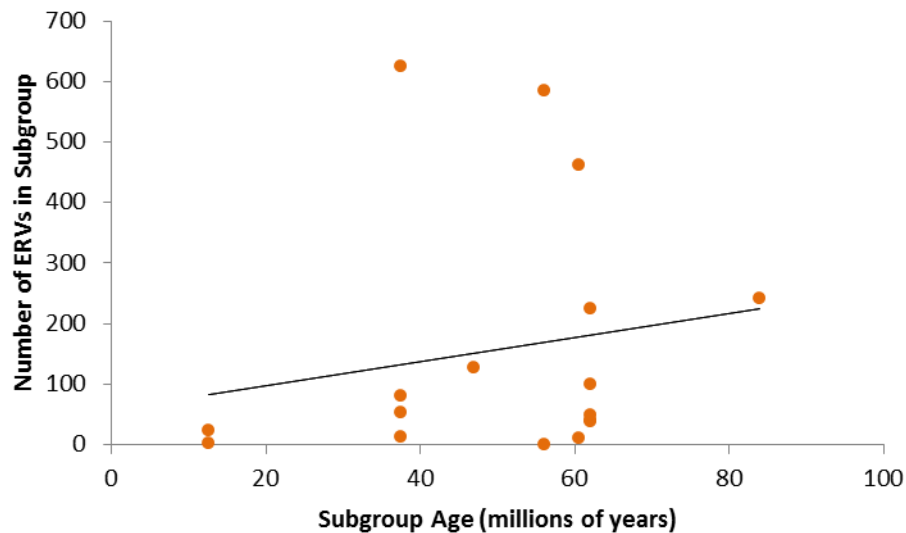


Figure 82: Scatterplot showing the relationship between the estimated integration date of a subgroup of ERVs and the number of ERVs in the subgroup, based on the HERV-F-Like and HERV-K-Like subgroups in chimpanzees, bonobos, humans and gorillas.

This is consistent with the model described by Katzourakis et al. (2005), who demonstrated that even a selectively neutral ERV integration event can have an unpredictable fate in its host genome. In this model, the fate of a neutral ERV depends on three factors (Katzourakis et al., 2005). These are integration rate, recombination rate and rate of inactivating mutations. By incorporating realistic values into this model, the fate of a novel ERV integration over time can be predicted. The integration rate and rate of recombinational deletion have been estimated previously as 3.8×10^{-4} integrations per host per generation (Belshaw et al., 2005a) and 1×10^{-5} recombinational deletions per host per generation (Belshaw et al., 2007) respectively. The probability of an inactivating mutation per ERV insertion can be estimated as the human mutation rate of 2.5×10^{-8} (Nachman and Crowell, 2000), multiplied by the proportion of all mutations which are non-synonymous (76.04%) and the number of bases in the average ERV sequence (8,000) (Bannert et al., 2010). This gives an estimated inactivating mutation rate of 1.52×10^{-4} mutations per ERV per generation.

These values were incorporated into the stochastic model proposed by Katzourakis et al. (2005), implemented in R (R Core Team, 2014) over 100,000 generations (equivalent to two million years with a generation time of 20 years). The results of repeating this simulation 10 times are shown in Figure 83. All of these runs result in exponential growth. The early stages of ERV integration under this model demonstrate a highly variable number of active and inactive ERVs per genome, so for novel integrations this model may be realistic. However, with these parameter values it was not possible to predict ERV families of the age and size found in the test dataset and the predicted rate at which the ERV would spread through the genome would not be sustainable.

In order to account for dynamics of the older ERV lineages, another parameter needs to be incorporated. Therefore, a proportion of integration events eliminated by restriction factors was added, as discussed but not implemented by Katzourakis et al. (2005). Active integration events can be assumed to be considerably more deleterious than inactive integration events, so a 10-fold difference in probability of restriction factor elimination was incorporated. Figure 84 shows the impact of adding a probability of restriction factor elimination per generation ranging from 0.001 to 1. This model generates considerably more realistic dynamics. The test dataset has a median ERV count for a single subfamily of 56 insertions and a range from one to 882 copies. All but the highest of these counts are consistent with some value for the probability of elimination from 0.005 to 0.04 and 56 insertions is most similar to the prediction with a value of 0.02, meaning each active ERV has a 2% probability of restriction factor removal per generation and each inactive ERV a 0.2% probability (Figure 84). This model is much more consistent with the data generated in our study and although it is oversimplified, provides some insight into the many factors and high degree of randomness affecting ERV count at any particular time. The model predicts a large amount of variation in ERV count over time, particularly in the number of inactive ERVs (Figure 84). This mechanism only takes into account reinfection as a

proliferation mechanism, however the most successful groups of ERVs are thought to replicate primarily through retrotransposition in *cis*. The few very large groups of ERVs found in our study, such as the HERV-W subgroup with a median of 584.5 insertions per genome, may be better explained by a model incorporating this factor.

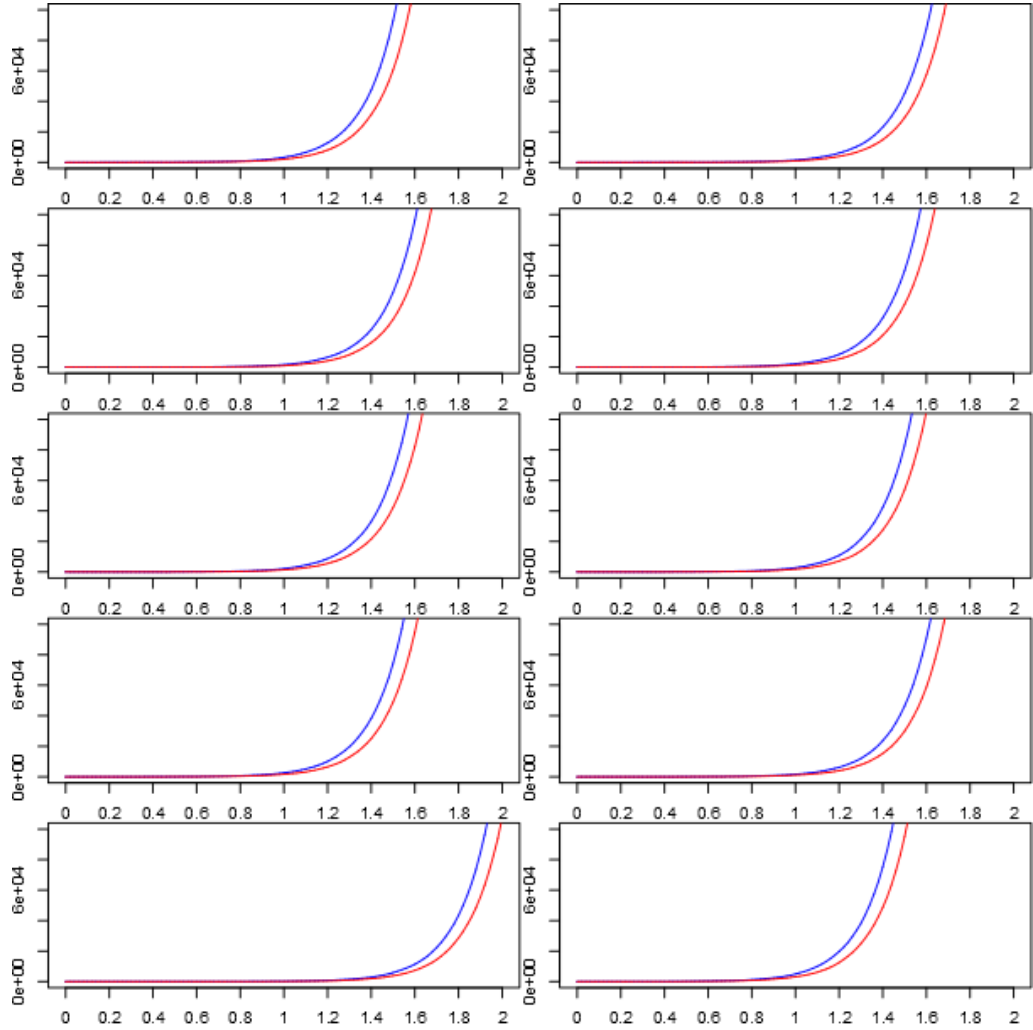


Figure 83: Ten example runs of Katzourakis et al.'s stochastic model of ERV integration rate showing the estimated number of active integrated ERVs (blue) and inactive integrated ERVs (red) over time.

The x-axis denotes time in millions of years and the y-axis predicted ERV count.

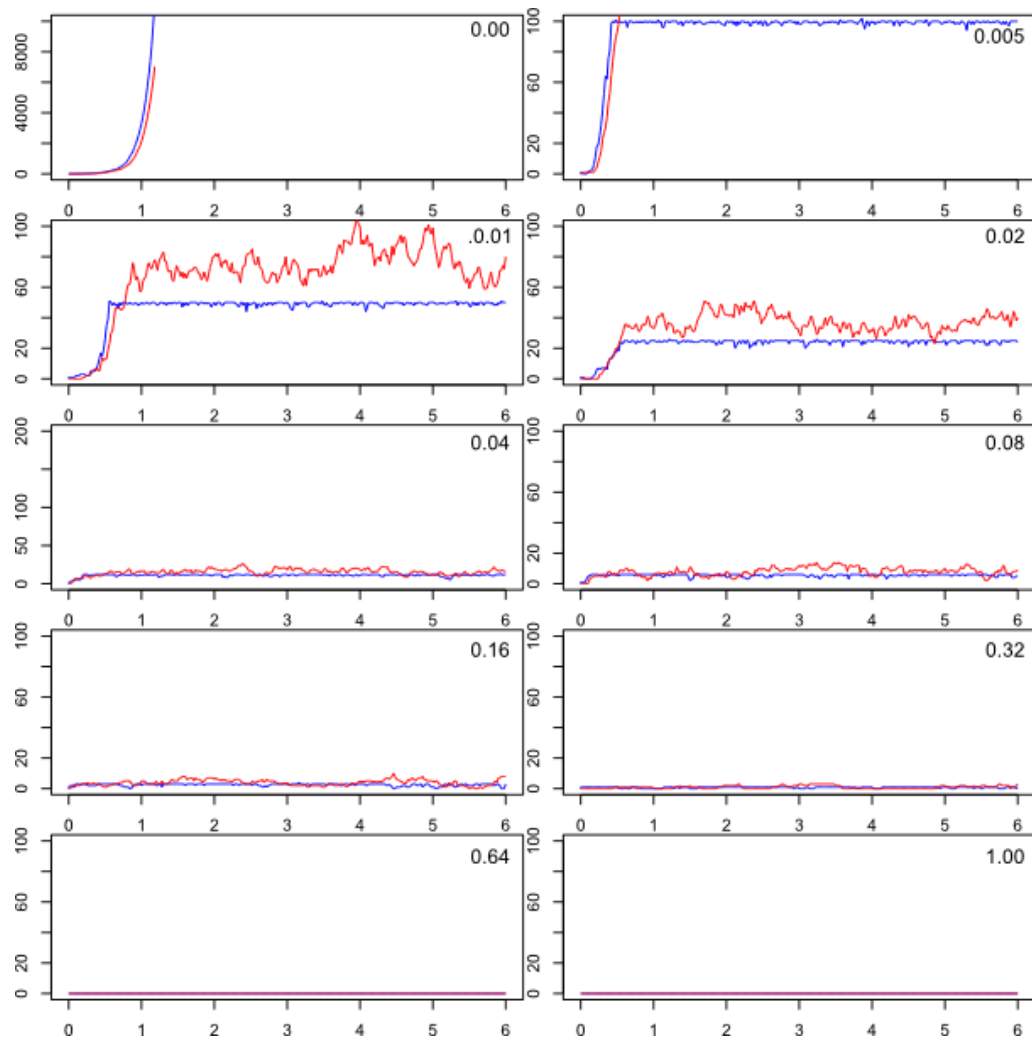


Figure 84: Ten example runs of Katzourakis et al.'s stochastic model of ERV integration rate with varying values of parameter s (top-left of each graph) showing the estimated number of active integrated ERVs (blue) and inactive integrated ERVs (red) over time.

The x-axis denotes time in millions of years and the y-axis predicted ERV count.

Besides these evolutionary factors, other factors have been proposed to affect the diversity of ERVs in the genome, in particular life history traits of the host. Katzourakis et al. (2014) found that host body size explains 37% of variance in ERV integration rate in different hosts and 68% of variance in ERV integration rate. To investigate this relationship, the ERV counts per host identified here were compared to a dataset of 22 life history traits from PanTHERIA (Jones et al., 2009). All traits with data available for at least 14 of the 28 species Euarchontoglires species analysed (no life history data was available for hamster or Chinese tree shrew) were compared to the number of *gag*, *pol* and

env fragments from gamma, beta and spumaviruses, the total number of *gag*, *pol* and *env* fragments and the total number of ERV regions identified in each host using Spearman's rank correlation coefficient.

Several strong positive and negative correlations were identified (Table 32).

Most notably, the number of spumavirus-like *pol* gene fragments correlated strongly with 13 life history traits. Almost all retroviral fragment counts correlated with inter-birth interval and home range size. However, although these results appear significant, there are several confounding factors.

In particular, our dataset is not a random selection of hosts, instead almost all hosts are either primates or rodents. As we have demonstrated, primates and rodents have very different retroviral profiles. Primates and rodents also have very different life histories, in that primates tend to be larger, live longer, mature and undergo life events later, give birth to fewer offspring, disperse over larger areas and live in larger groups. These differences are somewhat accounted for by the distinction between K-selected organisms and r-selected organisms, a common concept in ecology (Pianka, 1970). Briefly, K-selected organisms live in populations close to the carrying capacity of their environment and produce small numbers of large offspring, maximising the probability the offspring will survive (Pianka, 1970). R-selected organisms live close to their maximum reproductive capacity and produce many, smaller offspring, of which relatively few will reach maturity (Pianka, 1970). Primates are largely K-selected and rodents largely r-selected (Pianka, 1970). This leads to a strong correlation between the traits which are consistent with these lifestyles and strong relationships between K-selected traits, primates and primate retroviruses and r-selected traits, rodents and rodent retroviruses. These correlations are therefore not necessarily biologically meaningful. A second confounding factor is the phylogenetic relatedness of the host species, meaning presence of a particular trait in multiple hosts is not necessarily the result of independent acquisition events. For example, all great apes are large, but it is unlikely that each great ape species independently evolved the trait of being large, instead, all great apes are descendants of a

large common ancestor. This effect also confounds retrovirus count, as some retroviral acquisition events will have occurred in a common ancestor while others have occurred independently.

To attempt to account for these factors, the analysis was repeated with primates excluded. Within the primates there is also a gradient in the degree of K-selection from the prosimians to the great apes, concurrent with the variation in ERV profile. This is less apparent within the rodents and lagomorphs, so these were kept in the dataset. After making this correction, considerably fewer strongly correlated traits were observed, with 8 strong correlations identified (compared to 39 with primates included) (Table 33). These eight values had a mean p-value of 0.0181. As 422 comparisons were made, approximately eight significant correlations ($422 * 0.0181$) would be expected by chance at this p-value level, so these are unlikely to be meaningful. These results contrast with those of Katzourakis et al. (2014), who used a more complex model based on multiple regression, taking into account phylogeny. However, given the negative result of this preliminary study, implementation of a more complex model is unlikely to yield more significant results. The limited power of a small dataset of highly correlated traits in a group of highly related hosts may explain the correlation between body size and ERV count identified by Katzourakis et al.

Table 32: Spearman's rank correlation coefficient for the relationship between various life history traits and number of ERV fragments per host.

Correlations greater than or equal to 0.5 are highlighted in pink, correlations less than or equal to -0.5 are shown in blue. Coefficients with a p-value greater than 0.05 have been replaced with zero. Columns and rows with no strong positive or negative correlations have been excluded.

	γ gag	γ pol	γ env	β gag	β env	Spuma gag	Spuma pol	Total gag	Total pol	Total regions
Activity cycle	0.492	0.000	0.000	0.000	0.000	0.659	0.600	0.461	0.000	0.000
Adult body mass	0.418	0.423	0.000	0.000	0.000	0.493	0.698	0.000	0.000	0.000
Adult body length	0.000	0.000	0.000	0.000	0.000	0.642	0.769	0.000	0.000	0.000
Age at first litter	0.000	0.564	0.000	0.000	0.000	0.000	0.804	0.000	0.000	0.000
Gestation Length	0.513	0.523	0.462	0.000	0.000	0.428	0.722	0.000	0.000	0.000
Home range size	0.611	0.576	0.611	0.000	0.000	0.589	0.757	0.544	0.558	0.538
Inter birth interval	0.642	0.696	0.652	0.529	0.620	0.654	0.836	0.620	0.713	0.681
OS per female	0.000	0.000	0.000	0.000	0.000	0.000	-0.553	0.000	0.000	0.000
Max adult age	0.441	0.000	0.000	0.000	0.000	0.000	0.674	0.000	0.000	0.000
Neonate body mass	0.409	0.445	0.000	0.000	0.000	0.432	0.668	0.000	0.000	0.000
Population density	0.000	0.000	0.000	0.000	0.000	0.000	-0.556	0.000	0.000	0.000
Age sexual maturity	0.442	0.000	0.000	0.000	0.000	0.000	0.668	0.000	0.000	0.000
Social group size	0.599	0.000	0.521	0.000	0.000	0.679	0.668	0.000	0.000	0.000
Weaning age	0.446	0.000	0.000	0.000	0.000	0.000	0.696	0.000	0.000	0.000
Weaning body mass	0.000	0.000	0.000	0.000	0.000	0.000	0.745	0.000	0.000	0.000

Table 33: Spearman's rank correlation coefficient for the relationship between various life history traits and number of ERV fragments per host, excluding primates.

Correlations greater than or equal to 0.5 are highlighted in pink, correlations less than or equal to -0.5 are shown in blue. Coefficients with a p-value greater than 0.05 have been replaced with zero. Columns and rows with no strong positive or negative correlations have been excluded.

	β pol	β env	Spuma gag	Spuma pol	Total gag	Total pol
Adult head body length	0.000	-0.750	0.000	0.000	0.000	0.000
Gestation Length	-0.624	0.000	0.000	0.000	0.000	0.000
Population density	0.000	0.000	0.894	0.733	0.783	0.783
Age sexual maturity	-0.691	0.000	0.000	0.000	0.000	0.000
Weaning age	-0.695	0.000	0.000	0.000	0.000	0.000

8. 8. Future Work

Given the volume of data produced by this study, detailed analysis of every ERV fragment identified was unrealistic. Therefore, many results which warrant further study were identified. Several improvements and expansions of this study would also be worthwhile in the future.

In terms of further investigation of the existing results, several hosts had unusual ERV complements which warrant further study. In particular, guinea pigs contained several surprising ERVs, present in very large numbers, and several potentially intact ERVs, as discussed in section 8. 3. A laboratory study is needed to characterise these in depth. If these integrations are modern, they may be polymorphic between populations of guinea pigs. To test this hypothesis a number of samples from guinea pigs from a wide geographic area would be required, plus a samples from a second group of rodents, preferably new world rodents, as a control. The PCR based screening approach described for GALV in Chapter 6, using degenerate gammaretroviral primers, may be sufficient to identify these retroviruses, or a set of specific primers may be needed. It would also be of interest to establish if the retroviral particles released by guinea pig cells in culture originate from these loci and whether these particles are infectious.

For several of the hosts screened here, the currently available genome builds were insufficient for in depth ERV analysis. In particular, tarsiers and aye-ayes had unusual ERV complements which could not be characterised in depth because of the very fragmented quality of the currently available genome sequences. As more data become available it would be worthwhile to rescreen these species and look for intact retroviruses, LTR sequences and recombination events.

In terms of the endogenous lentiviruses and GALV, screening of more host species would be the most productive further study. As of yet, no rodent with a

range overlapping with gibbons has been identified with a virus close enough to GALV to be the source of this outbreak. The only way to identify this virus would be to screen more hosts. For lentiviruses, it is unlikely that *G. moholi* is unique, instead other bushbabies not screened here may harbour this lentivirus. Further characterisation of this lentivirus would also be beneficial, as only a fairly short fragment could be identified here. A full length sequence would provide considerably more information about the origin and age of this ERV. It is possible that this lentivirus is also endogenous in whichever vector host transmitted it from the mainland to Madagascar (or vice versa), so screening the species which are known to have passed between these regions during an appropriate time period would be ideal (although may not be feasible for species which are now extinct).

A simple, logical extension to our study would be to screen more genomes for ERVs. Currently, Ensembl lists 85 available vertebrate full genome sequences. Several of these are of particular interest. For our epsilonretrovirus study, screening fish, reptiles and amphibians would be worthwhile and may help establish the origin of the mammalian epsilonretrovirus-like sequences. Screening marine mammals would also be interesting in detecting the link between epsilonretroviruses currently detected in fish and land mammals. For the GALV study, marsupials are of particular interest due to the presence of the GALV-like KoRV ERV in koalas. The possum sequence in the SRV-like group and echidna sequence in the REV-like group also suggest that the marsupial ERV complement may be worth investigating further. However, as our study has demonstrated, the ERV content of vertebrate genomes is often surprising, so screening any available genome would be worthwhile.

Alongside these specific examples, some recent advances in methodology could be applied to enhance the results. The major drawback of the methodology used in this study is that it is limited by the currently available data, meaning only ERVs with a reasonable degree of similarity to a known retrovirus can be identified. A *de novo* approach to ERV detection would

overcome this limitation but would be difficult to implement given the degraded nature of many ERVs. A supervised machine learning approach may overcome this obstacle to some extent. This would involve providing a training dataset which allows the learning algorithm to identify the link between the input dataset and the desired output. A simulated ERV dataset, containing realistic but hypothetical ERV sequences, would allow such an algorithm to be trained. However, such an approach would be considerably more computationally challenging than the approach described here, and may have little impact on the quality of the results.

Locus-by-locus analysis proved to be a powerful technique in identifying ancient ERVs in this study and this method could be applied across many more loci. Ideally, for each locus in a particular host the orthologues in all other sequenced genomes could be identified, allowing the exact spread of the virus through the genome to be traced. This approach requires high quality whole genome alignments and a large amount of computational power, however this is becoming more and more feasible given the advent of modern sequencing techniques and the availability of high performance parallel computing facilities. Whole genome sequencing of multiple individuals of the same species is also becoming common, with data already available for 2577 humans via the 1000 genomes project (2010) and for at least 10 chimpanzees (Leffler et al., 2013) and 17 gorillas (McManus et al., 2014). Even on a small scale, locus-by-locus analysis within a specific group, for example within the widely studied HERV-K(HML-2) group, within these datasets could answer many of the important questions about their evolution. Shin et al. (2013) discuss 29 human-specific HERV-K(HML-2) loci and identify these as a source of genetic variation between humans and chimpanzees. An approach similar to our locus-by-locus analysis of endogenous epsilonretroviruses (Chapter 7) but using multiple genomes from the same species would help clarify this relationship.

Tools developed for high-throughput genomics may allow further insights from our ERV dataset. For example, as longer reads become available from next generation sequencing technologies, the development of alignment and phylogenetic tools for very large numbers of longer sequences is becoming more popular (Warnow, 2013). Therefore, generation of accurate phylogenetic trees representing a whole genus of retroviruses is becoming more feasible. This would remove the need for the clustering stage of our analysis, which is computationally expensive and involves loss of sequence information. It would also allow large numbers of ERVs from multiple hosts to be directly compared.

Finally, next generation sequencing datasets have an enormous amount of potential in the analysis of ERVs. As we now have considerable information about the ERV content of many genomes, including the positions in the genome at which the ERVs are found, the next step would be to compare this data to that about other genomic features. For example, the DNA methylation level and transcription level of ERVs is of interest in understanding their expression, and MEDip-seq and RNAseq datasets are already available for at least the human and mouse genomes. Comparison of cases and controls for disease phenotypes, especially those thought to be associated with retroviruses, such as schizophrenia, bipolar disorder and MS, in terms of ERV transcription or methylation would be worthwhile and would provide more definitive information about the link between these diseases and ERV expression. As more and more data becomes available, comprehensive understanding of the ERV content of all hosts and the effect of ERVs on host phenotype should become realistic.

8. 9. Conclusions

In conclusion, the pipeline developed during this project provides the means to identify a very large amount of ERV data. Using this pipeline, many of the evolutionary relationships between ERVs in the Euarchontoglires have been elucidated, including those which have been controversial in the past.

Previously neglected host species have been demonstrated to harbour a wide range of ERVs, providing information about viral and host evolution. Rodents and other non-primate hosts have been shown to play a major role in shaping the primate genome through their ability to transfer retroviruses between hosts.

Chapter 9. Appendices

Please note: All Appendices are provided on the attached USB flash drive.

Appendix A. Prior publications

A.1 Characterisation of retroviruses in the horse genome and their transcriptional activity via transcriptome sequencing.

Brown K, Moreton J, Malla S, Aboobaker AA, Emes RD, Tarlinton RE. 2012. Characterisation of retroviruses in the horse genome and their transcriptional activity via transcriptome sequencing. *Virology* 433(1):55-63.

A.2 Characterisation of a group of endogenous gammaretroviruses in the canine genome.

Tarlinton RE, Barfoot HKR, Allen CE, Brown K, Gifford RJ, Emes RD. 2013. Characterisation of a group of endogenous gammaretroviruses in the canine genome. *The Veterinary Journal* 196(1):28-33.

Appendix B. Supplementary Tables

B.1 Figure_sequences.xlsx

The names and accession numbers, Repbase IDs or genome positions (as *genomebuild_chromosome_start_end_strand*) of the previously known retrovirus sequences used to generate the phylogenetic trees shown in the figures in Chapter 1.

B.2 Full_ERV_Nucleotide_Database.xlsx

The full unparsed untranslated dataset of 4124 retrovirus sequences, including name, accession number, ID (Genbank Accession, Repbase ID or chromosome position), gene, genus, sequence and host. Unique names (column H) are used in subsequent tables to allow unambiguous identification of these sequences. Short names are the names used in the phylogenetic analyses in Chapter 4 and Chapter 6. Sequences in these phylogenies which were not in the original input dataset are provided in the Sheet "Additional Tree Sequences".

B.3 Refseq_Retrovirus_Sequences.xlsx

The retrovirus sequences downloaded from RefSeq to test the Exonerate input dataset.

B.4 GROUPED_PREVKNOWN_groups.xlsx

The group in the GROUPED_PREVKNOWN dataset to which each sequence in the PARSED_UT_PREVKNOWN dataset was assigned.

B.5 Test_datasets.xlsx

The abbreviated name (displayed in phylogenetic trees) and full unique name (from Appendix B.2) of the sequences included in the basic phylogenetic test

datasets for betaretrovirus gag, betaretrovirus pol, betaretrovirus env, gammaretrovirus gag, gammaretrovirus pol, gammaretrovirus env, spumavirus gag and spumavirus pol. Each sheet contains details of one dataset. Sequences not in the FULL_UT_PREVKNOWN dataset (usually sequences described after this dataset was created) are highlighted in grey.

B.6 Euarchontoglires_accessions.xlsx

Accession numbers for 15 nuclear genes for the screened Euarchontoglires used to build the host phylogeny.

B.7 Output_db.xls

Details of each sequence in the RAW_EXO_OUT, PARSED_EXO_OUT, CLU_EXO_OUT and GROUPED_EXO datasets, including ID, chromosome or scaffold name, start and end positions within that chromosome/scaffold, fragment length, strand, whether the sequence passed the quality control check to be incorporated into the PARSED_EXO_OUT dataset, the representative in CLU_EXO_OUT for each sequence, the most similar previously known sequence in PARSED_UT_PREVKNOWN to the sequence and the group the representative sequence from CLU_EXO_OUT was placed into in the GROUPED_EXO dataset.

B.8 ERV_Regions.xlsx

Details of the genome regions identified with fragments of one or more ERV genes, including the IDs of the ERV fragments (as listed in Appendix B.7), the genes identified, the number of different genes identified, chromosome / scaffold and start and end positions of the ERV region, region length and strand.

B.9 Lemur_Tree.xlsx

Sequences included in the lentiviral phylogeny. IDs are Genbank accessions unless otherwise specified.

B.10 Epsilonretrovirus_Input.xlsx

Database of known ERVs. Sequences highlighted in pink were included as gammaretroviruses and sequences highlighted in green as epsilonretroviruses in phylogenetic analysis. Where genome positions are given they are in format Genome_chr[chromosome number]_[start position]_[end position], Repbase IDs 6 are given as Repbase_[Repbase_ID]_[start position]_[end position], otherwise Genbank accessions are provided. Where amino acid sequences were required sequences were translated in each reading frame and the reading frame with the least stop codons was used, sequences with >5 stop codons were excluded.

B.11 Epsilonretrovirus_loci.xlsx

Details of the 87 PE loci identified here including locus ID, details of any previous description of the locus, family, and position in the genome of each host.

B.12 Epsilonretrovirus_positions.xlsx

PE genome details in all hosts for the 11 loci with LTRs, including estimated ages, positions in the genome and the actual and relative start and end positions of the identified LTRs, pol and env genes.

Appendix C. Fasta files

C.1 FULL_PREVKNOWN.fas

The full unparsed untranslated dataset of 4124 previously known retrovirus sequences, as described in Appendix B.1.

C.2 PARSED_UT_PREVKNOWN.fas

The parsed, untranslated version of the database of previously known retroviruses, consisting of 1590 nucleotide sequences, as described in Appendix B.1.

C.3 PARSED_T_PREVKNOWN.fas

The parsed, translated version of the database of previously known retroviruses used as an input to the Exonerate pipeline, consisting of 1361 amino acid sequences, as described in Appendix B.1.

C.4 RAW_EXO_OUT.fas

The raw output from the Exonerate algorithm for all host genomes, containing 190,196 candidate ERV fragments. Details of these sequences, including their chromosome locations, are provided in Appendix B.7.

C.5 PARSED_EXO_OUT.fas

The parsed output from the Exonerate algorithm for all host genomes, containing only the 169,424 sequences which passed the quality control step described in section 2.2.1. Sequence names are prefixed as described in the “prefix” column of Table 7. Details of these sequences, including their chromosome locations, are provided in Appendix B.7.

C.6 CLU_EXO_OUT.fas

The clustered output from the Exonerate algorithm for all host genomes, consisting of 47,896 sequences with consensus sequences representing highly similar sequences from the same host. Sequence names are prefixed as described in the “prefix” column of Table 7. The sequences which make up each consensus are listed in Appendix B.7.

Appendix D. Python and R scripts

D.1 make_chromosomes.py

Python script to generate artificial chromosomes from unassembled contigs or scaffolds.

Runs as: python make_chromosomes.py *genome_fasta* *n_chromosomes* *prefix*

Genome_fasta: path to fasta file containing the contigs or scaffolds

n_chromosomes: number of artificial chromosomes required

prefix: prefix for the chromosome files

Input

Genome fasta file

Output

“chroms” directory containing the chromosome fasta files

“lists” directory containing the positions of each scaffold or contig in each chromosome.

D.2 reciprocal_blast.py

Python script to perform a BLASTN comparison between each pair of sequences in a multiple FASTA file. This script requires blastall (Altschul et al., 1990) to run.

Runs as: python reciprocal_blast.py *fasta*

fasta: Multiple sequence FASTA file of candidate ERV fragments

Input

Multiple FASTA file

Output

Directory containing a BLAST output file for each input sequence.

D.3 distance.R

R script to cluster sequences based on a distance matrix.

Runs as: distance.R in the same directory as the input directories

Input

Directory “fastas” of FASTA files and directory “matrices” of distance matrices for these FASTA files.

Output

Directory of FASTA files containing all the sequences in each cluster.

D.4 make_cons.py

Python script for generating a consensus sequence for each file in a directory of aligned input FASTA files. The FASTA file must have each sequence on a single line. This code requires EMBOSS (Rice et al., 2000) to run.

Runs as: python make_cons.py *alignment_dir*

Alignment_dir: the directory containing the alignments

Input

A directory of aligned FASTA files.

Output

A directory, “consensus”, of consensus sequences for these alignments.

D.5 classify_sets.py

Python script generating lists of regions of chromosomes containing at least 2 different genes no more than 5,000 bp apart.

Runs as: `python classify_sets.py dir`

dir: directory of chromosome maps

Input

Chromosome maps sorted by start position, listing the ID, gene, start position, end position and strand of each ERV fragment in PARSED_EXO.

Output

Directory for each chromosome containing lists of fragments from different genes found within 5,000 bp of each other and a list of fragments not within 5,000 bp of another fragment.

Appendix E. Documents concerning gibbon transportation

E.1 1974BANGKO17800.pdf

Correspondence between the American Embassy in Bangkok and the US Secretary of State entitled SEATO: ALLEGED SMUGGLING OF GIBBONS, dated November 1974.

E.2 1975BANGKO15111_b.pdf

Correspondence between the American Embassy in Bangkok and the US Secretary of State entitled ENFORCEMENT OF U.S. WILDLIFE LAWS, dated July 1975.

E.3 1974BANGKO19028_b.pdf

Correspondence between the American Embassy in Bangkok and the US Secretary of State entitled ENFORCEMENT OF WILDLIFE LAWS, dated December 1974.

E.4 1974STATE244644_b.pdf

Correspondence between the American Embassy in Bangkok, the American Embassy in Taipei and the US Secretary of State entitled ALLEGED SMUGGLING OF MONKEYS, dated November 1974.

E.5 1974TAIPEI06749_b.pdf

Correspondence between the American Embassy in Bangkok and the US Secretary of State entitled ALLEGED SMUGGLING OF MONKEYS, dated November 1974.

E.6 1974STATE260770_b.pdf

Correspondence between the American Consulate in Melbourne and the US Secretary of State entitled ENFORCEMENT OF WILDLIFE LAWS - GIBBONS, dated November 1974.

E.7 1974BANGKO17734_b.pdf

Correspondence between the American Embassy in Bangkok, the American Embassy in Taipei and the US Secretary of State entitled ALLEGED SMUGGLING OF GIBBONS, dated November 1974.

E.8 1974STATE260768_b.pdf

Correspondence between the American Embassy in Bangkok and the US Secretary of State entitled ENFORCEMENT OF WILDLIFE LAWS - GIBBONS, dated November 1974.

References

- 1000 GENOMES PROJECT CONSORTIUM. 2010. A map of human genome variation from population-scale sequencing. *Nature*, 467, 1061-1073.
- AAGAARD, L., VILLESEN, P., KJELDBJERG, A. L. & PEDERSEN, F. S. 2005. The ~30-million-year-old ERVPb1 envelope gene is evolutionarily conserved among hominoids and Old World monkeys. *Genomics*, 86, 685-691.
- AHOLA, V., AITOKALLIO, T., VIHINEN, M. & UUSIPAUKKA, E. 2006. A statistical score for assessing the quality of multiple sequence alignments. *BMC Bioinformatics*, 7, 484.
- ALI, J. R. & HUBER, M. 2010. Mammalian biodiversity on Madagascar controlled by ocean currents. *Nature*, 463, 653-656.
- ALTER, H. J., MIKOVITS, J. A., SWITZER, W. M., RUSCETTI, F. W., LO, S.-C., KLIMAS, N., KOMAROFF, A. L., MONTOYA, J. G., BATEMAN, L., LEVINE, S., PETERSON, D., LEVIN, B., HANSON, M. R., GENFI, A., BHAT, M., ZHENG, H., WANG, R., LI, B., HUNG, G.-C., LEE, L. L., SAMEROFF, S., HENEINE, W., COFFIN, J., HORNIG, M. & LIPKIN, W. I. 2012. A Multicenter Blinded Analysis Indicates No Association between Chronic Fatigue Syndrome/Myalgic Encephalomyelitis and either Xenotropic Murine Leukemia Virus-Related Virus or Polytropic Murine Leukemia Virus. *mBio*, 3, e00266-12.
- ALTSCHUL, S. F., GISH, W., MILLER, W., MYERS, E. W. & LIPMAN, D. J. 1990. Basic local alignment search tool. *Journal of Molecular Biology*, 215, 403-410.
- ANDERSSON, A. C., SVENSSON, A. C., ROLNY, C., ANDERSSON, G. & LARSSON, E. 1998. Expression of human endogenous retrovirus ERV3 (HERV-R) mRNA in normal and neoplastic tissues. *International Journal of Oncology*, 12, 309-13.
- ANDRÉOLETTI, L., RÉVEIL, B., MORET, H., BRODARD, V., PHILBERT, F., TABARY, T. & COHEN, J. H. M. 2007. Significant genetic and antigenic variability within the env gene of systemic human immunodeficiency virus type 1 group O populations during the natural course of a heterosexual infection: a pilot study. *Journal of Clinical Microbiology*, 45, 1319-1321.
- ANISIMOVA, M. & GASCUEL, O. 2006. Approximate Likelihood-Ratio Test for Branches: A Fast, Accurate, and Powerful Alternative. *Systematic Biology*, 55, 539-552.
- ANTOINE, P.-O., MARIVAUX, L., CROFT, D. A., BILLET, G., GANERØD, M., JARAMILLO, C., MARTIN, T., ORLIAC, M. J., TEJADA, J., ALTAMIRANO, A. J., DURANTHON, F., FANJAT, G., ROUSSE, S. & GISMONDI, R. S. 2012. Middle Eocene rodents from Peruvian Amazonia reveal the pattern and timing of caviomorph origins and biogeography. *Proceedings of the Royal Society B: Biological Sciences*, 279, 1319-1326.
- ANTONY, J. M., VAN MARLE, G., OPII, W., BUTTERFIELD, D. A., MALLET, F., YONG, V. W., WALLACE, J. L., DEACON, R. M., WARREN, K. & POWER, C. 2004. Human endogenous retrovirus glycoprotein-mediated induction of redox reactants causes oligodendrocyte death and demyelination. *Nature Neuroscience*, 7, 1088-1095.
- ARHEL, N. 2010. Revisiting HIV-1 uncoating. *Retrovirology*, 7, 96.

- ARNAUD, F., CAPORALE, M., VARELA, M., BIEK, R., CHESSA, B., ALBERTI, A., GOLDBERGER, M., MURA, M., ZHANG, Y.-P., YU, L., PEREIRA, F., DEMARTINI, J. C., LEYMASTER, K., SPENCER, T. E. & PALMARINI, M. 2007. A Paradigm for Virus–Host Coevolution: Sequential Counter-Adaptations between Endogenous and Exogenous Retroviruses. *PLoS Pathogens*, 3, e170.
- ARNOLD, C., MATTHEWS, L. J. & NUNN, C. L. 2010. The 10kTrees website: A new online resource for primate phylogeny. *Evolutionary Anthropology: Issues, News, and Reviews*, 19, 114-118.
- AYINDE, D., MAUDET, C., TRANSY, C. & MARGOTTIN-GOGUET, F. 2010. Limelight on two HIV/SIV accessory proteins in macrophage infection: Is Vpx overshadowing Vpr? *Retrovirology*, 7, 35.
- BABA, K., NAKAYA, Y., SHOJIMA, T., MUROI, Y., KIZAKI, K., HASHIZUME, K., IMAKAWA, K. & MIYAZAWA, T. 2011. Identification of Novel Endogenous Betaretroviruses Which Are Transcribed in the Bovine Placenta. *Journal of Virology*, 85, 1237-1245.
- BAILES, E., GAO, F., BIBOLLET-RUCHE, F., COURGNAUD, V., PEETERS, M., MARX, P. A., HAHN, B. H. & SHARP, P. M. 2003. Hybrid Origin of SIV in Chimpanzees. *Science*, 300, 1713.
- BAILLIE, G. J., VAN DE LAGEMAAT, L. N., BAUST, C. & MAGER, D. L. 2004. Multiple Groups of Endogenous Betaretroviruses in Mice, Rats, and Other Mammals. *Journal of Virology*, 78, 5784-5798.
- BAILLIE, G. J. & WILKINS, R. J. 2001. Endogenous Type D Retrovirus in a Marsupial, the Common Brushtail Possum (*Trichosurus vulpecula*). *Journal of Virology*, 75, 2499-2507.
- BALAJ, L., LESSARD, R., DAI, L., CHO, Y.-J., POMEROY, S. L., BREAKFIELD, X. O. & SKOG, J. 2011. Tumour microvesicles contain retrotransposon elements and amplified oncogene sequences. *Nature Communications*, 2-180.
- BALVAY, L., LASTRA, M. L., SARGUEIL, B., DARLIX, J.-L. & OHLMANN, T. 2007. Translational control of retroviruses. *Nature Reviews of Microbiology*, 5, 128-140.
- BANCROFT, W. H., SNITBHAN, R., ROZMIAREK, H. & TINGAPALAPONG, M. 1975. Experimental Infection of Gibbons with a Group B Arbovirus. In: US ARMY MEDICAL COMPONENT (ed.) *SEATO Medical Research Laboratory Annual Progress Reports*.
- BANNERT, N., FIEBIG, U. & HOHN, O. 2010. Retroviral Particles, Proteins and Genomes. In: KURTH, R. & BANNERT, N. (eds.) *Retroviruses. Molecular Biology, Genomics and Pathogenesis*. Norfolk: Caister Academic Press.
- BANNERT, N. & KURTH, R. 2006. The Evolutionary Dynamics of Human Endogenous Retroviral Families. *Annual Review of Genomics and Human Genetics*, 7, 149-173.
- BASTA, H. A., CLEVELAND, S. B., CLINTON, R. A., DIMITROV, A. G. & MCCLURE, M. A. 2009. Evolution of Teleost Fish Retroviruses: Characterization of New Retroviruses with Cellular Genes. *Journal of Virology*, 83, 10152-10162.
- BELSHAW, R., DAWSON, A. L. A., WOOLVEN-ALLEN, J., REDDING, J., BURT, A. & TRISTEM, M. 2005a. Genomewide Screening Reveals High Levels of Insertional Polymorphism in the Human Endogenous Retrovirus Family HERV-K(HML2): Implications for Present-Day Activity. *Journal of Virology*, 79, 12507-12514.

- BELSHAW, R., KATZOURAKIS, A., PAČES, J., BURT, A. & TRISTEM, M. 2005b. High Copy Number in Human Endogenous Retrovirus Families is Associated with Copying Mechanisms in Addition to Reinfection. *Molecular Biology and Evolution*, 22, 814-817.
- BELSHAW, R., PEREIRA, V., KATZOURAKIS, A., TALBOT, G., PAČES, J., BURT, A. & TRISTEM, M. 2004. Long-term reinfection of the human genome by endogenous retroviruses. *Proceedings of the National Academy of Sciences of the United States of America*, 101, 4894-4899.
- BELSHAW, R., WATSON, J., KATZOURAKIS, A., HOWE, A., WOOLVEN-ALLEN, J., BURT, A. & TRISTEM, M. 2007. Rate of Recombinational Deletion among Human Endogenous Retroviruses. *Journal of Virology*, 81, 9437-9442.
- BENACHENHOU, F., JERN, P., OJA, M., SPERBER, G., BLIKSTAD, V., SOMERVUO, P., KASKI, S. & BLOMBERG, J. 2009. Evolutionary Conservation of Orthoretroviral Long Terminal Repeats (LTRs) and *ab initio* Detection of Single LTRs in Genomic Data. *PLoS ONE*, 4, e5179.
- BÉNIT, L., CALTEAU, A. & HEIDMANN, T. 2003. Characterization of the low-copy HERV-Fc family: evidence for recent integrations in primates of elements with coding envelope genes. *Virology*, 312, 159-168.
- BENIT, L., DESSEN, P. & HEIDMANN, T. 2001. Identification, Phylogeny, and Evolution of Retroviral Elements Based on Their Envelope Genes. *Journal of Virology*, 75, 11709-11719.
- BENIT, 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, 3301-3308.
- BERKSON, G. 1968. Ko Klet Kaeo: Habitat description *In: US ARMY MEDICAL COMPONENT* (ed.) *SEATO Medical Research Laboratory Annual Progress Reports*.
- BEST, S., TISSIER, P. L., TOWERS, G. & STOEY, J. P. 1996. Positional cloning of the mouse retrovirus restriction gene Fv1. *Nature*, 382, 826-829.
- BETANCUR, R. R., BROUGHTON, R. E., WILEY, E. O., CARPENTER, K., LOPEZ, J. A., LI, C., HOLCROFT, N. I., ARCILA, D., SANCIANGCO, M., CURETON II, J. C., ZHANG, F., BUSER, T., CAMPBELL, M. A., BALLESTEROS, J. A., ROA-VARON, A., WILLIS, S., BORDEN, W. C., ROWLEY, T., RENEAU, P. C., HOUGH, D. J., LU, G., GRANDE, T., ARRATIA, G. & ORTI, G. 2013. The tree of life and a new classification of bony fishes. *PLoS Currents Tree of Life*, 5.
- BITANYI, S., BJORNSTAD, G., ERNEST, E. M., NESJE, M., KUSILUKA, L. J., KEYYU, J. D., MDEGELA, R. H. & ROED, K. H. 2011. Species identification of Tanzanian antelopes using DNA barcoding. *Molecular Ecology Resources*, 11, 442-449.
- BJERREGAARD, B., HOLCK, S., CHRISTENSEN, I. J. & LARSSON, L. I. 2006. Syncytin is involved in breast cancer-endothelial cell fusions. *Cellular and Molecular Life Sciences CMLS*, 63, 1906-1911.
- BLAISE, S., DE PARSEVAL, N., BÉNIT, L. & HEIDMANN, T. 2003. Genomewide screening for fusogenic human endogenous retrovirus envelopes identifies syncytin 2, a gene conserved on primate evolution. *Proceedings of the National Academy of Sciences*, 100, 13013-13018.

References

- BLAISE, S., DE PARSEVAL, N. & HEIDMANN, T. 2005. Functional characterization of two newly identified Human Endogenous Retrovirus coding envelope genes. *Retrovirology*, 2-19.
- BLANGA-KANFI, S., MIRANDA, H., PENN, O., PUPKO, T., DEBRY, R. & HUCHON, D. 2009. Rodent phylogeny revised: analysis of six nuclear genes from all major rodent clades. *BMC Evolutionary Biology*, 9-71.
- BLUMENTHAL, R., DURELL, S. & VIARD, M. 2012. HIV Entry and Envelope Glycoprotein-mediated Fusion. *Journal of Biological Chemistry*, 287, 40841-40849.
- BOLISETTY, M., BLOMBERG, J., BENACHENHOU, F., SPERBER, G. & BEEMON, K. 2012. Unexpected Diversity and Expression of Avian Endogenous Retroviruses. *mBio*, 3, e00344-12.
- BORYSENKO, L., STEPANETS, V. & RYNDITCH, A. V. 2008. Molecular characterization of full-length MLV-related endogenous retrovirus ChiRV1 from the chicken, *Gallus gallus*. *Virology*, 376, 199-204.
- BOWSER, P. R. & CASEY, J. W. 1993. Retroviruses of fish. *Annual Review of Fish Diseases*, 3, 209-224.
- BRADY, T., LEE, Y. N., RONEN, K., MALANI, N., BERRY, C. C., BIENIASZ, P. D. & BUSHMAN, F. D. 2009. Integration target site selection by a resurrected human endogenous retrovirus. *Genes & Development*, 23, 633-642.
- BRESLIN, PAUL A. S. 2013. An Evolutionary Perspective on Food and Human Taste. *Current Biology*, 23, R409-R418.
- BROCKELMAN, W. Y. 1969. Behavior and Ecology of Gibbons on Kled Kaeo Island In: US ARMY MEDICAL COMPONENT (ed.) *SEATO Medical Research Laboratory Annual Progress Reports*.
- BROWN, J., REST, J., GARCIA-MORENO, J., SORENSON, M. & MINDELL, D. 2008. Strong mitochondrial DNA support for a Cretaceous origin of modern avian lineages. *BMC Biology*, 6, 6.
- BROWN, K., MORETON, J., MALLA, S., ABOOBAKER, A. A., EMES, R. D. & TARLINTON, R. E. 2012. Characterisation of retroviruses in the horse genome and their transcriptional activity via transcriptome sequencing. *Virology*, 433, 55-63.
- BUKRINSKY, M. 2004. A hard way to the nucleus. *Molecular Medicine*, 10, 1-5.
- BURMEISTER, T. 2001. Oncogenic retroviruses in animals and humans. *Reviews in Medical Virology*, 11, 369-380.
- BURTONBOY, G., DELFERRIERE, N., MOUSSET, B. & HEUSTERSPREUTE, M. 1993. Isolation of a C-type retrovirus from an HIV infected cell line. *Archives of Virology*, 130, 289-300.
- CÁCERES, M., PROGRAM, N. C. S. & THOMAS, J. W. 2006. The Gene of Retroviral Origin Syncytin 1 is Specific to Hominoids and is Inactive in Old World Monkeys. *Journal of Heredity*, 97, 100-106.
- CADIGAN, M. A. J., CHAICUMPA, V. & PUHOMCHAREON, S. 1967. Effect of *P. falciparum* infection on Serum Biochemistry Values of the Gibbon In: US ARMY MEDICAL COMPONENT (ed.) *SEATO Medical Research Laboratory Annual Progress Reports*.

- CALISHER, C. H., CHILDS, J. E., FIELD, H. E., HOLMES, K. V. & SCHOUNTZ, T. 2006. Bats: Important Reservoir Hosts of Emerging Viruses. *Clinical Microbiology Reviews*, 19, 531-545.
- CALLAHAN, R. & SMITH, G. H. 2000. MMTV-induced mammary tumorigenesis: gene discovery, progression to malignancy and cellular pathways. *Oncogene*, 19, 992-1001.
- CARRE-EUSEBE, D., COUDOUÉL, N. & MAGRE, S. 2009. OVEX1, a novel chicken endogenous retrovirus with sex-specific and left-right asymmetrical expression in gonads. *Retrovirology*, 6-59.
- CENTRAL INTELLIGENCE AGENCY 2013. The World Factbook 2013-14.
- CHAN, E., PETERS, W. P., SWEET, R. W., OHNO, T., KUFE, D. W., SPIEGELMAN, S. O. L., GALLO, R. C. & GALLAGHER, R. E. 1976. Characterisation of a virus (HL23V) isolated from cultured acute myelogenous leukaemic cells. *Nature*, 260, 266-268.
- CHEN, F.-C. & LI, W.-H. 2001. Genomic divergences between humans and other hominoids and the effective population size of the common ancestor of humans and chimpanzees. *The American Journal of Human Genetics*, 68, 444-456.
- CHERKASOVA, E., MALINZAK, E., RAO, S., TAKAHASHI, Y., SENCHENKO, V. N., KUDRYAVTSEVA, A. V., NICKERSON, M. L., MERINO, M., HONG, J. A., SCHRUMP, D. S., SRINIVASAN, R., LINEHAN, W. M., TIAN, X., LERMAN, M. I. & CHILDS, R. W. 2011. Inactivation of the von Hippel-Lindau tumor suppressor leads to selective expression of a human endogenous retrovirus in kidney cancer. *Oncogene*, 30, 4697-4706.
- CHIU, Y.-L. & GREENE, W. C. 2008. The APOBEC3 Cytidine Deaminases: An Innate Defensive Network Opposing Exogenous Retroviruses and Endogenous Retroelements. *Annual Review of Immunology*, 26, 317-353.
- CHONG, A. Y.-Y., ATKINSON, S. J., ISBERG, S. & GONGORA, J. 2012. Strong purifying selection in endogenous retroviruses in the saltwater crocodile (*Crocodylus porosus*) in the Northern Territory of Australia. *Mobile DNA*, 3, 20.
- CINGÖZ, O. & COFFIN, J. M. 2011. Endogenous Murine Leukemia Viruses: Relationship to XMRV and Related Sequences Detected in Human DNA Samples. *Advances in Virology*, 2011, 10.
- COFFEE, L. L., CASEY, J. W. & BOWSER, P. R. 2013. Pathology of Tumors in Fish Associated With Retroviruses: A Review. *Veterinary Pathology Online*, 50, 390-403.
- COFFIN JM, HUGHES SH & VARMUS HE 1997. *Retroviruses*, New York, Cold Spring Harbor Laboratory Press.
- CONTICELLO, S. G., HARRIS, R. S. & NEUBERGER, M. S. 2003. The Vif Protein of HIV Triggers Degradation of the Human Antiretroviral DNA Deaminase APOBEC3G. *Current Biology*, 13, 2009-2013.
- CONTRERAS-GALINDO, R., KAPLAN, M. H., CONTRERAS-GALINDO, A. C., GONZALEZ-HERNANDEZ, M. J., FERLENGHI, I., GIUSTI, F., LORENZO, E., GITLIN, S. D., DOSIK, M. H., YAMAMURA, Y. & MARKOVITZ, D. M. 2012. Characterization of Human Endogenous Retroviral Elements in the Blood of HIV-1-Infected Individuals. *Journal of Virology*, 86, 262-276.

- CONTRERAS-GALINDO, R., KAPLAN, M. H., LEISSNER, P., VERJAT, T., FERLENGHI, I., BAGNOLI, F., GIUSTI, F., DOSIK, M. H., HAYES, D. F., GITLIN, S. D. & MARKOVITZ, D. M. 2008. Human Endogenous Retrovirus K (HML-2) Elements in the Plasma of People with Lymphoma and Breast Cancer. *Journal of Virology*, 82, 9329-9336.
- CORNELIS, G., HEIDMANN, O., BERNARD-STOECKLIN, S., REYNAUD, K., VÉRON, G., MULOT, B., DUPRESSOIR, A. & HEIDMANN, T. 2012. Ancestral capture of syncytin-Car1, a fusogenic endogenous retroviral envelope gene involved in placentation and conserved in Carnivora. *Proceedings of the National Academy of Sciences*, 109, E432–E441.
- CORNELIS, G., HEIDMANN, O., DEGRELLE, S. A., VERNOCHE, C., LAVIALLE, C., LETZELTER, C., BERNARD-STOECKLIN, S., HASSANIN, A., MULOT, B., GUILLMOT, M., HUE, I., HEIDMANN, T. & DUPRESSOIR, A. 2013. Captured retroviral envelope syncytin gene associated with the unique placental structure of higher ruminants. *Proceedings of the National Academy of Sciences*, 110, E828–E837.
- CORPET, F. 1988. Multiple sequence alignment with hierarchical clustering. *Nucleic Acids Research*, 16, 10881-10890.
- CREEVEY, C. J. & MCINERNEY, J. O. 2005. Clann: investigating phylogenetic information through supertree analyses. *Bioinformatics*, 21, 390-392.
- CUI, J. & HOLMES, E. C. 2012. Endogenous lentiviruses in the ferret genome. *Journal of Virology*, 86, 3383-3385.
- CUI, J., TACHEDJIAN, G., TACHEDJIAN, M., HOLMES, E. C., ZHANG, S. & WANG, L. F. 2012. Identification of diverse groups of endogenous gammaretroviruses in mega- and microbats. *Journal of General Virology*, 93, 2037-45.
- DAHLBERG, J. E., TRONICK, S. R. & AARONSON, S. A. 1980. Immunological relationships of an endogenous guinea pig retrovirus with prototype mammalian type B and type D retroviruses. *Journal of Virology*, 33, 522-530.
- DARRIBA, D., TABOADA, G. L., DOALLO, R. & POSADA, D. 2012. jModelTest 2: more models, new heuristics and parallel computing. *Nature Methods*, 9, 772-772.
- DAVIS, A. & NAYAK, D. 1977. Expression of endogenous retroviral genes in leukemic guinea pig cells. *Journal of Virology*, 23, 263-271.
- DE PAOLI, A., NOLL, W. W. & JOHNSON, D. O. 1971. Leukaemia in the Gibbon. In: US ARMY MEDICAL COMPONENT (ed.) *SEATO Medical Research Laboratory Annual Progress Reports*.
- DEB-RINKER, P., KLEMPAN, T. A., O'REILLY, R. L., TORREY, E. F. & SINGH, S. M. 1999. Molecular Characterization of a MSRV-like Sequence Identified by RDA from Monozygotic Twin Pairs Discordant for Schizophrenia. *Genomics*, 61, 133-144.
- DEPIL, S., ROCHE, C., DUSSART, P. & PRIN, L. 2002. Expression of a human endogenous retrovirus, HERV-K, in the blood cells of leukemia patients. *Leukemia*, 16, 254-259.
- 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, 1548-1556.

- DIGGS, C. L. & PAVANAND, K. 1969. *P. falciparum* in small rodents. In: US ARMY MEDICAL COMPONENT (ed.) *SEATO Medical Research Laboratory Annual Progress Reports*.
- DIMCHEFF, D. E., DROVETSKI, S. V., KRISHNAN, M. & MINDELL, D. P. 2000. Cospeciation and Horizontal Transmission of Avian Sarcoma and Leukosis Virus gag Genes in Galliform Birds. *Journal of Virology*, 74, 3984-3995.
- DIMMIC, M. W., REST, J. S., MINDELL, D. P. & GOLDSTEIN, R. A. 2002. rtREV: an amino acid substitution matrix for inference of retrovirus and reverse transcriptase phylogeny. *Journal of Molecular Evolution*, 55, 65-73.
- DOOLITTLE, R. F., HUNKAPILLER, M. W., HOOD, L. E., DEVARE, S. G., ROBBINS, K. C., AARONSON, S. A. & ANTONIADES, H. N. 1983. Simian sarcoma virus onc gene, v-sis, is derived from the gene (or genes) encoding a platelet-derived growth factor. *Science*, 221, 275-277.
- DUDLEY, J. P., MERTZ, J. A., BHADRA, S., PALMARINI, M. & KOZAK, C. 2011. Endogenous Retroviruses and Cancer. In: DUDLEY, J. (ed.) *Retroviruses and Insights into Cancer*. New York: Springer.
- DUNLAP, K. A., PALMARINI, M., VARELA, M., BURGHARDT, R. C., HAYASHI, K., FARMER, J. L. & SPENCER, T. E. 2006. Endogenous retroviruses regulate periimplantation placental growth and differentiation. *Proceedings of the National Academy of Sciences*, 103, 14390-14395.
- DUNNUM, J. L. & SALAZAR-BRAVO, J. 2010. Molecular systematics, taxonomy and biogeography of the genus *Cavia* (Rodentia: Caviidae). *Sistemática molecular, taxonomía y biogeografía del género Cavia* (Rodentia: Caviidae). *Journal of Zoological Systematics and Evolutionary Research*, 48, 376-388.
- DUPRESSOIR, A., MARCEAU, G., VERNOCHE, C., BÉNIT, L., KANELLOPOULOS, C., SAPIN, V. & HEIDMANN, T. 2005. Syncytin-A and syncytin-B, two fusogenic placenta-specific murine envelope genes of retroviral origin conserved in Muridae. *Proceedings of the National Academy of Sciences of the United States of America*, 102, 725-730.
- EDGAR, R. C. 2004. MUSCLE: multiple sequence alignment with high accuracy and high throughput. *Nucleic Acids Research*, 32, 1792-1797.
- EGGINK, D., BERKHOUT, B. & SANDERS, R. W. 2010. Inhibition of HIV-1 by fusion inhibitors. *Current Pharmaceutical Design*, 16, 3716-3728.
- EHRlich, M. 2002. DNA methylation in cancer: too much, but also too little. *Oncogene*, 21, 5400-5413.
- EIDEN, M. V. & TALIAFERRO, D. L. 2011. Emerging Retroviruses and Cancer. In: DUDLEY, J. (ed.) *Retroviruses and Insights into Cancer*. New York: Springer.
- ELLEDER, D., KIM, O., PADHI, A., BANKERT, J. G., SIMEONOV, I., SCHUSTER, S. C., WITTEKINDT, N. E., MOTAMENY, S. & POSS, M. 2012. Polymorphic Integrations of an Endogenous Gammaretrovirus in the Mule Deer Genome. *Journal of Virology*, 86, 2787-2796.
- ENGELMAN, A. 2010. Reverse transcription and integration. In: BANNERT, N. & KURTH, R. (eds.) *Retroviruses. Molecular Biology, Genomics and Pathogenesis*. Norfolk: Caister Academic Press.

- ERLWEIN, O., KAYE, S., MCCLURE, M. O., WEBER, J., WILLS, G., COLLIER, D., WESSELY, S. & CLEARE, A. 2010. Failure to Detect the Novel Retrovirus XMRV in Chronic Fatigue Syndrome. *PLoS ONE*, 5, e8519.
- ESNAULT, C., PRIET, S., RIBET, D., HEIDMANN, O. & HEIDMANN, T. 2008. Restriction by APOBEC3 proteins of endogenous retroviruses with an extracellular life cycle: ex vivo effects and in vivo "traces" on the murine IAPV and human HERV-K elements. *Retrovirology*, 5, 75.
- FAUCI, A. 1993. Multifactorial nature of human immunodeficiency virus disease: implications for therapy. *Science*, 262, 1011-1018.
- FLICEK, P., AMODE, M. R., BARRELL, D., BEAL, K., BRENT, S., CARVALHO-SILVA, D., CLAPHAM, P., COATES, G., FAIRLEY, S., FITZGERALD, S., GIL, L., GORDON, L., HENDRIX, M., HOURLIER, T., JOHNSON, N., KÄHÄRI, A. K., KEEFE, D., KEENAN, S., KINSELLA, R., KOMOROWSKA, M., KOSCIELNY, G., KULESHA, E., LARSSON, P., LONGDEN, I., MCLAREN, W., MUFFATO, M., OVERDUIN, B., PIGNATELLI, M., PRITCHARD, B., RIAT, H. S., RITCHIE, G. R. S., RUFFIER, M., SCHUSTER, M., SOBRAL, D., TANG, Y. A., TAYLOR, K., TREVANION, S., VANDROVCOVA, J., WHITE, S., WILSON, M., WILDER, S. P., AKEN, B. L., BIRNEY, E., CUNNINGHAM, F., DUNHAM, I., DURBIN, R., FERNÁNDEZ-SUAREZ, X. M., HARROW, J., HERRERO, J., HUBBARD, T. J. P., PARKER, A., PROCTOR, G., SPUDICH, G., VOGEL, J., YATES, A., ZADISSA, A. & SEARLE, S. M. J. 2012. Ensembl 2012. *Nucleic Acids Research*, 40, D84-D90.
- FLOCKERZI, A., RUGGIERI, A., FRANK, O., SAUTER, M., MALDENER, E., KOPPER, B., WULLICH, B., SEIFARTH, W., MULLER-LANTZSCH, N., LEIB-MOSCH, C., MEESE, E. & 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, 354.
- FOWLER, M. E. & MILLER, R. E. 2008. Occupational Exposure to Zoonotic Simian Retroviruses: Health and Safety Implications. *Zoo and Wild Animal Medicine: Current Therapy*. St Louis: Saunders Elsevier.
- FRANK, O., JONES-BRANDO, L., LEIB-MÖSCH, C., YOLKEN, R. & SEIFARTH, W. 2006. Altered Transcriptional Activity of Human Endogenous Retroviruses in Neuroepithelial Cells after Infection with *Toxoplasma gondii*. *Journal of Infectious Diseases*, 194, 1447-1449.
- FULLER, S., SCHWARZ, M. & TIERNEY, S. 2005. Phylogenetics of the allodapine bee genus *Braunsapis*: historical biogeography and long-range dispersal over water. *Journal of Biogeography*, 32, 2135-2144.
- GARCIA-ETXEBARRIA, K. & JUGO, B. M. 2010. Genome-Wide Detection and Characterization of Endogenous Retroviruses in *Bos taurus*. *Journal of Virology*, 84, 10852-10862.
- GARCIA-ETXEBARRIA, K. & JUGO, B. M. 2012. Detection and characterization of endogenous retroviruses in the horse genome by in silico analysis. *Virology*, 434, 59-67.
- GARCIA-MONTOJO, M., DOMINGUEZ-MOZO, M., ARIAS-LEAL, A., GARCIA-MARTINEZ, Á., DE LAS HERAS, V., CASANOVA, I., FAUCARD, R., GEHIN, N., MADEIRA, A., ARROYO, R., CURTIN, F., ALVAREZ-LAFUENTE, R. & PERRON, H. 2013. The DNA Copy Number of Human Endogenous Retrovirus-W (MSRV-Type) Is Increased in Multiple Sclerosis Patients and Is Influenced by Gender and Disease Severity. *PLoS ONE*, 8, e53623.

- GELMAN, I. H., ZHANG, J., HAILMAN, E., HANAFUSA, H. & MORSE, S. S. 1992. Identification and evaluation of new primer sets for the detection of lentivirus proviral DNA. *AIDS research and human retroviruses*, 8, 1981-1989.
- GENTLES, A. J., WAKEFIELD, M. J., KOHANY, O., GU, W., BATZER, M. A., POLLOCK, D. D. & JURKA, J. 2007. Evolutionary dynamics of transposable elements in the short-tailed opossum *Monodelphis domestica*. *Genome Research*, 17, 992-1004.
- GIFFORD, R., KABAT, P., MARTIN, J., LYNCH, C. & TRISTEM, M. 2005. Evolution and Distribution of Class II-Related Endogenous Retroviruses. *Journal of Virology*, 79, 6478-6486.
- GIFFORD, R. & TRISTEM, M. 2003. The evolution, distribution and diversity of endogenous retroviruses. *Virus Genes*, 26, 291-315.
- GIFFORD, R. J., KATZOURAKIS, A., TRISTEM, M., PYBUS, O. G., WINTERS, M. & SHAFER, R. W. 2008. A transitional endogenous lentivirus from the genome of a basal primate and implications for lentivirus evolution. *Proceedings of the National Academy of Sciences of the United States of America*, 105, 20362-20367.
- GILBERT, C., MAXFIELD, D. G., GOODMAN, S. M. & FESCHOTTE, C. 2009. Parallel Germline Infiltration of a Lentivirus in Two Malagasy Lemurs. *PLoS Genetics*, 5, e1000425.
- GOERING, W., RIBARSKA, T. & SCHULZ, W. A. 2011. Selective changes of retroelement expression in human prostate cancer. *Carcinogenesis*, 32, 1484-1492.
- GOFF, S. P. 2007. Retroviridae: The Retroviruses and Their Replication. In: KNIPE, D. M. & HOWLEY, P. M. (eds.) *Fields Virology*. Philadelphia: Lippincott Williams & Wilkins.
- GOLDSTONE, D. C., ENNIS-ADENIRAN, V., HEDDEN, J. J., GROOM, H. C. T., RICE, G. I., CHRISTODOULOU, E., WALKER, P. A., KELLY, G., HAIRE, L. F., YAP, M. W., DE CARVALHO, L. P. S., STOEY, J. P., CROW, Y. J., TAYLOR, I. A. & WEBB, M. 2011. HIV-1 restriction factor SAMHD1 is a deoxynucleoside triphosphate triphosphohydrolase. *Nature*, 480, 379-382.
- GONG, J., SHEN, X. H., QIU, H., CHEN, C. & YANG, R. G. 2011. Rhesus monkey TRIM5alpha has distinct HIV-1 restriction activity among different mammalian cell lines. *Current Microbiology*, 63, 531-537.
- GOTOH, O. 1982. An improved algorithm for matching biological sequences. *Journal of Molecular Biology*, 162, 705-708.
- GÖTTLINGER, H. G. & WEISSENHORN, W. 2010. Assembly and Release. In: BANNERT, N. & KURTH, R. (eds.) *Retroviruses. Molecular Biology, Genomics and Pathogenesis*. Norfolk: Caister Academic Press.
- GRANT, R. F., WINDSOR, S. K., MALINAK, C. J., BARTZ, C. R., SABO, A., BENVENISTE, R. E. & TSAI, C.-C. 1995. Characterization of Infectious Type D Retrovirus from Baboons. *Virology*, 207, 292-296.
- GREEN, P. 1996. SWAT and Cross_Match. <http://www.phrap.org/phredphrap/swat.html>.
- GREENWOOD, A. D., STENGEL, A., ERFLE, V., SEIFARTH, W. & LEIB-MÖSCH, C. 2005. The distribution of pol containing human endogenous retroviruses in non-human primates. *Virology*, 334, 203-213.

- GRIFFITHS, D. J., VENABLES, P. J., WEISS, R. A. & BOYD, M. T. 1997. A novel exogenous retrovirus sequence identified in humans. *Journal of Virology*, 71, 2866-2872.
- GRIFFITHS, D. J., VOISSET, C., VENABLES, P. J. W. & WEISS, R. A. 2002. Novel Endogenous Retrovirus in Rabbits Previously Reported as Human Retrovirus 5. *Journal of Virology*, 76, 7094-7102.
- GUINDON, S., DUFAYARD, J.-F., LEFORT, V., ANISIMOVA, M., HORDIJK, W. & GASCUEL, O. 2010. New Algorithms and Methods to Estimate Maximum-Likelihood Phylogenies: Assessing the Performance of PhyML 3.0. *Systematic Biology*, 59, 307-321.
- GUINDON, S. & GASCUEL, O. 2003. A simple, fast, and accurate algorithm to estimate large phylogenies by maximum likelihood. *Systematic Biology*, 52, 696-704.
- GUNTAKA, R. V. 1993. Transcription termination and polyadenylation in retroviruses. *Microbiology Reviews*, 57, 511-21.
- HAASE, A. T. 2005. Perils at mucosal front lines for HIV and SIV and their hosts. *Nat Rev Immunology*, 5, 783-792.
- HAHN, B. H., SHAW, G. M., DE, K. M., COCK & SHARP, P. M. 2000. AIDS as a Zoonosis: Scientific and Public Health Implications. *Science*, 287, 607-614.
- HALSTEAD, S. B. 1964. Growth of Dengue Viruses in Tissue Culture. In: US ARMY MEDICAL COMPONENT (ed.) *SEATO Medical Research Laboratory Annual Progress Reports*.
- HAN, G.-Z. & WOROBEY, M. 2012a. An Endogenous Foamy-like Viral Element in the Coelacanth Genome. *PLoS Pathogens*, 8, e1002790.
- HAN, G.-Z. & WOROBEY, M. 2012b. Endogenous lentiviral elements in the weasel family (Mustelidae). *Molecular Biology and Evolution*.
- HAN, K., KONKEL, M. K., XING, J., WANG, H., LEE, J., MEYER, T. J., HUANG, C. T., SANDIFER, E., HEBERT, K., BARNES, E. W., HUBLEY, R., MILLER, W., SMIT, A. F. A., ULLMER, B. & BATZER, M. A. 2007. Mobile DNA in Old World Monkeys: A Glimpse Through the Rhesus Macaque Genome. *Science*, 316, 238-240.
- HART, D., FRERICHS, G. N., RAMBAUT, A. & ONIONS, D. E. 1996. Complete nucleotide sequence and transcriptional analysis of snakehead fish retrovirus. *Journal of Virology*, 70, 3606-16.
- HAYWARD, A., GRABHERR, M. & JERN, P. 2013a. Broad-scale phylogenomics provides insights into retrovirus–host evolution. *Proceedings of the National Academy of Sciences*, 110, 20146-20151.
- HAYWARD, J., TACHEDJIAN, M., CUI, J., FIELD, H., HOLMES, E., WANG, L.-F. & TACHEDJIAN, G. 2013b. Identification of diverse full-length endogenous betaretroviruses in megabats and microbats. *Retrovirology*, 10, 35.
- HEDGES, S. B., DUDLEY, J. & KUMAR, S. 2006. TimeTree: a public knowledge-base of divergence times among organisms. *Bioinformatics*, 22, 2971-2972.
- HEIDMANN, O., VERNOCHE, C., DUPRESSOIR, A. & HEIDMANN, T. 2009. Identification of an endogenous retroviral envelope gene with fusogenic activity and placenta-specific expression in the rabbit: a new "syncytin" in a third order of mammals. *Retrovirology*, 6-107.

- HERNIOU, E., MARTIN, J., MILLER, K., COOK, J., WILKINSON, M. & TRISTEM, M. 1998. Retroviral Diversity and Distribution in Vertebrates. *Journal of Virology*, 72, 5955-5966.
- HIMATHONGKHAM, S. & LUCIW, P. A. 1996. Restriction of HIV-1 (Subtype B) Replication at the Entry Step in Rhesus Macaque Cells. *Virology*, 219, 485-488.
- HINDMARSH, P. & LEIS, J. 1999. Retroviral DNA integration. *Microbiology and Molecular Biology Reviews*, 63, 836-43.
- HOHN, O., KRAUSE, H., BARBAROTTO, P., NIEDERSTADT, L., BEIMFORDE, N., DENNER, J., MILLER, K., KURTH, R. & BANNERT, N. 2009. Lack of evidence for xenotropic murine leukemia virus-related virus(XMRV) in German prostate cancer patients. *Retrovirology*, 6-92.
- HOLDER, M. & LEWIS, P. O. 2003. Phylogeny estimation: traditional and Bayesian approaches. *Nature Reviews of Genetics*, 4, 275-284.
- HOLT, M., SHEVACH, E. & PUNKOSDY, G. 2013. Endogenous Mouse Mammary Tumor Viruses (Mtv): New Roles for an Old Virus in Cancer, Infection and Immunity. *Frontiers in Oncology*, 3.
- HORVATH, J. E., WEISROCK, D. W., EMBRY, S. L., FIORENTINO, I., BALHOFF, J. P., KAPPELER, P., WRAY, G. A., WILLARD, H. F. & YODER, A. D. 2008. Development and application of a phylogenomic toolkit: resolving the evolutionary history of Madagascar's lemurs. *Genome Research*, 18, 489-499.
- HRECKA, K., HAO, C., GIERSEWSKA, M., SWANSON, S. K., KESIK-BRODACKA, M., SRIVASTAVA, S., FLORENS, L., WASHBURN, M. P. & SKOWRONSKI, J. 2011. Vpx relieves inhibition of HIV-1 infection of macrophages mediated by the SAMHD1 protein. *Nature*, 474, 658-661.
- HU, W.-S. & HUGHES, S. H. 2012. HIV-1 reverse transcription. *Cold Spring Harbor perspectives in medicine*, 2.
- HUANG, W., LI, S., HU, Y., YU, H., LUO, F., ZHANG, Q. & ZHU, F. 2011. Implication of the env gene of the human endogenous retrovirus W family in the expression of BDNF and DRD3 and development of recent-onset schizophrenia. *Schizophrenia Bulletin*, 37, 988-1000.
- HUANG, X. & MADAN, A. 1999. CAP3: A DNA Sequence Assembly Program. *Genome Research*, 9, 868-877.
- HUDA, A., POLAVARAPU, N., JORDAN, I. K. & MCDONALD, J. 2008. Endogenous retroviruses of the chicken genome. *Biology Direct*, 3, 9.
- HUE, S., GRAY, E., GALL, A., KATZOURAKIS, A., TAN, C., HOULDCROFT, C., MCLAREN, S., PILLAY, D., FUTREAL, A., GARSON, J., PYBUS, O., KELLAM, P. & TOWERS, G. 2010. Disease-associated XMRV sequences are consistent with laboratory contamination. *Retrovirology*, 7, 111.
- HÜTTER, G., NOWAK, D., MOSSNER, M., GANEPOLA, S., MÜßIG, A., ALLERS, K., SCHNEIDER, T., HOFMANN, J., KÜCHERER, C., BLAU, O., BLAU, I. W., HOFMANN, W. K. & THIEL, E. 2009. Long-Term Control of HIV by CCR5 Delta32/Delta32 Stem-Cell Transplantation. *New England Journal of Medicine*, 360, 692-698.

References

- IKEDA, H. & SUGIMURA, H. 1989. Fv-4 resistance gene: a truncated endogenous murine leukemia virus with ecotropic interference properties. *Journal of Virology*, 63, 5405-5412.
- INTERNATIONAL COMMITTEE ON TAXONOMY OF VIRUSES 2002. *ICTVdB - The Universal Virus Database, version 4*.
- IPPL 1976. The Hall's Island Gibbon Project. *International Primate Protection League Newsletter*.
- IPPL 1978. Summary of the Disposition of Gibbons by SEATO Medical Research Laboratory, Bangkok, Thailand. *International Primate Protection League Newsletter*.
- ISSEL, C. J., RUSHLOW, K., FOIL, L. D. & MONTELARO, R. C. 1988. A perspective on equine infectious anemia with an emphasis on vector transmission and genetic analysis. *Veterinary Microbiology*, 17, 251-286.
- JARATLERDSIRI, W., RODRIGUEZ-ZARATE, C. J., ISBERG, S. R., DAMAYANTI, C. S., MILES, L. G., CHANSUE, N., MORAN, C., MELVILLE, L. & GONGORA, J. 2009. Distribution of Endogenous Retroviruses in Crocodilians. *Journal of Virology*, 83, 10305-10308.
- JARMUZ, A., CHESTER, A., BAYLISS, J., GISBOURNE, J., DUNHAM, I., SCOTT, J. & NAVARATNAM, N. 2002. An Anthropoid-Specific Locus of Orphan C to U RNA-Editing Enzymes on Chromosome 22. *Genomics*, 79, 285-296.
- JENKINS, G. M., RAMBAUT, A., PYBUS, O. G. & HOLMES, E. C. 2002. Rates of Molecular Evolution in RNA Viruses: A Quantitative Phylogenetic Analysis. *Journal of Molecular Evolution*, 54, 156-165.
- JENSEN-SEAMAN, M. I., FUREY, T. S., PAYSEUR, B. A., LU, Y., ROSKIN, K. M., CHEN, C.-F., THOMAS, M. A., HAUSSLER, D. & JACOB, H. J. 2004. Comparative Recombination Rates in the Rat, Mouse, and Human Genomes. *Genome Research*, 14, 528-538.
- JERN, P. & COFFIN, J. M. 2008. Effects of Retroviruses on Host Genome Function. *Annual Review of Genetics*, 42, 709-732.
- JERN, P., SPERBER, G. & BLOMBERG, J. 2005. Use of Endogenous Retroviral Sequences (ERVs) and structural markers for retroviral phylogenetic inference and taxonomy. *Retrovirology*, 2-50.
- JERN, P., SPERBER, G. O. & BLOMBERG, J. 2004. Definition and variation of human endogenous retrovirus H. *Virology*, 327, 93-110.
- JERN, P., SPERBER, G. O. & BLOMBERG, J. 2006. Divergent Patterns of Recent Retroviral Integrations in the Human and Chimpanzee Genomes: Probable Transmissions between Other Primates and Chimpanzees. *Journal of Virology*, 80, 1367-1375.
- JERN, P., STOEY, J. P. & COFFIN, J. M. 2007. Role of APOBEC3 in Genetic Diversity among Endogenous Murine Leukemia Viruses. *PLoS Genetics*, 3, e183.
- JOEL, R. & SANDRA, L. Q. 2010. Walleye Dermal Sarcoma Virus: Molecular Biology and Oncogenesis. *Viruses*, 2.
- JOHNSEN, D. O., TANTICAHROENYOS, P. & PULLIAM, J. D. 1969. Gibbon Leukemia. In: US ARMY MEDICAL COMPONENT (ed.) *SEATO Medical Research Laboratory Annual Progress Reports*.

- JONES, D. T., TAYLOR, W. R. & THORNTON, J. M. 1992. The rapid generation of mutation data matrices from protein sequences. *Computer applications in the biosciences : CABIOS*, 8, 275-282.
- JONES, K. E., BIELBY, J., CARDILLO, M., FRITZ, S. A., O'DELL, J., ORME, C. D. L., SAFI, K., SECHREST, W., BOAKES, E. H., CARBONE, C., CONNOLLY, C., CUTTS, M. J., FOSTER, J. K., GRENYER, R., HABIB, M., PLASTER, C. A., PRICE, S. A., RIGBY, E. A., RIST, J., TEACHER, A., BININDA-EMONDS, O. R. P., GITTLEMAN, J. L., MACE, G. M., PURVIS, A. & MICHENER, W. K. 2009. PanTHERIA: a species-level database of life history, ecology, and geography of extant and recently extinct mammals. *Ecology*, 90, 2648-2648.
- JOUVENET, N., LAINE, S., VIVARES, L. P. & MOUGEL, M. 2011. Cell biology of retroviral RNA packaging. *RNA Biology*, 8, 572-580.
- JUKES, T. H. & CANTOR, C. R. 1969. Evolution of protein molecules. *Mammalian protein metabolism*, 21-132.
- JURKA, J., KAPITONOV, V. V., PAVLICEK, A., KLONOWSKI, P., KOHANY, O. & WALICHIEWICZ, J. 2005. Repbase Update, a database of eukaryotic repetitive elements. *Cytogenetic and Genome Research*, 110, 462-467.
- JURKA, J., KLONOWSKI, P., DAGMAN, V. & PELTON, P. 1996. CENSOR--a program for identification and elimination of repetitive elements from DNA sequences. *Computational Chemistry*, 20, 119-21.
- KAISER, S. M., MALIK, H. S. & EMERMAN, M. 2007. Restriction of an Extinct Retrovirus by the Human TRIM5{alpha} Antiviral Protein. *Science*, 316, 1756-1758.
- KAMBOL, R. & TRISTEM, M. 2005. The Diversity and Distribution of Piscine Endogenous Retrovirus in Piscine Hosts. *Beyond Genome: Harnessing the Potential. Proceedings of the 6th National Congress on Genetics*, 248-254.
- KANTERS, S., MILLS, E. J., THORLUND, K., BUCHER, H. C. & IOANNIDIS, J. P. A. 2014. Antiretroviral therapy for initial human immunodeficiency virus/AIDS treatment: critical appraisal of the evidence from over 100 randomized trials and 400 systematic reviews and meta-analyses. *Clinical Microbiology and Infection*, 20, 114-122.
- KATOH, K., MISAWA, K., KUMA, K. & MIYATA, T. 2002. MAFFT: a novel method for rapid multiple sequence alignment based on fast Fourier transform. *Nucleic Acids Research*, 30, 3059-3066.
- KATZOURAKIS, A., MAGIORKINIS, G., LIM, A. G., GUPTA, S., BELSHAW, R. & GIFFORD, R. 2014. Larger Mammalian Body Size Leads to Lower Retroviral Activity. *PLoS Pathogens*, 10, e1004214.
- KATZOURAKIS, A., RAMBAUT, A. & PYBUS, O. G. 2005. The evolutionary dynamics of endogenous retroviruses. *Trends in Microbiology*, 13, 463-8.
- KATZOURAKIS, A. & TRISTEM, M. 2005. Phylogeny of Human Endogenous and Exogenous Retroviruses. In: SVERDLOV, E. D. (ed.) *Retroviruses and Primate Genome Evolution*. Austin: Landes Bioscience.
- KATZOURAKIS, A., TRISTEM, M., PYBUS, O. G. & GIFFORD, R. J. 2007. Discovery and analysis of the first endogenous lentivirus. *Proceedings of the National Academy of Sciences*, 104, 6261-6265.

References

- KAUFMANN, S., SAUTER, M., SCHMITT, M., BAUMERT, B., BEST, B., BOESE, A., ROEMER, K. & MUELLER-LANTZSCH, N. 2010. Human endogenous retrovirus protein Rec interacts with the testicular zinc-finger protein and androgen receptor. *Journal of General Virology*, 91, 1494-1502.
- KAWAKAMI, T., BUCKLEY, P. M., MCDOWELL, T. S. & DEPAOLI, A. 1973. Antibodies to simian C-type virus antigen in sera of gibbons (*Hylobates* sp.). *Nature New Biology*, 246, 105-7.
- KAWAKAMI, T. G., BUCKLEY, P. M., DEPAOLI, A., NOLL, W. & BUSTAD, L. K. 1975. Studies on the prevalence of type C virus associated with gibbon hematopoietic neoplasms. *Bibliotheca Haematologica*, 385-9.
- KAWAKAMI, T. G., HUFF, S. D., BUCKLEY, P. M., DUNGWORTH, D. L., SYNDER, S. P. & GILDEN, R. V. 1972. C-type virus associated with gibbon lymphosarcoma. *Nature New Biology*, 235, 170-1.
- KECKESOVA, Z., YLINEN, L. M. J. & TOWERS, G. J. 2004. The human and African green monkey TRIM5 α genes encode Ref1 and Lv1 retroviral restriction factor activities. *Proceedings of the National Academy of Sciences of the United States of America*, 101, 10780-10785.
- 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, 7-11.
- KEIGHTLEY, P. D. & EYRE-WALKER, A. 2000. Deleterious Mutations and the Evolution of Sex. *Science*, 290, 331-333.
- KENT, W. J. 2002. BLAT—The BLAST-Like Alignment Tool. *Genome Research*, 12, 656-664.
- KENT, W. J., SUGNET, C. W., FUREY, T. S., ROSKIN, K. M., PRINGLE, T. H., ZAHLER, A. M. & HAUSSLER, D. 2002. The human genome browser at UCSC. *Genome Research*, 12, 996-1006.
- KHAN, A. S., GALVIN, T. A., JENNINGS, M. B., GARDNER, M. B. & LOWENSTINE, L. J. 1991. SIV of stump-tailed macaque (SIVstm) is a divergent Asian isolate. *Journal of Medical Primatology*, 20, 167-71.
- KIM, F. J., BATTINI, J.-L., MANEL, N. & SITBON, M. 2004. Emergence of vertebrate retroviruses and envelope capture. *Virology*, 318, 183-191.
- KIM, H.-S., TAKENAKA, O. & CROW, T. J. 1999. Isolation and phylogeny of endogenous retrovirus sequences belonging to the HERV-W family in primates. *Journal of General Virology*, 80, 2613-2619.
- KIM, H.-S., YI, J.-M., HIRAI, H., HUH, J.-W., JEONG, M.-S., JANG, S.-B., KIM, C.-G., SAITOU, N., HYUN, B.-H. & LEE, W.-H. 2006. Human Endogenous Retrovirus (HERV)-R family in primates: Chromosomal location, gene expression, and evolution. *Gene*, 370, 34-42.
- KLYMIUK, N., MULLER, M., BREM, G. & AIGNER, B. 2003. Characterization of endogenous retroviruses in sheep. *Journal of Virology*, 77, 11268-11273.

- KOO, H. M., GU, J., VARELA-ECHAVARRIA, A., RON, Y. & DOUGHERTY, J. P. 1992. Reticuloendotheliosis type C and primate type D oncoretroviruses are members of the same receptor interference group. *Journal of Virology*, 66, 3448-3454.
- KOZIREVA, S., LEJNIECE, S., BLOMBERG, J. & MUROVSKA, M. 2001. Human retrovirus type 5 sequences in non-Hodgkin's lymphoma of T cell origin. *AIDS Research and Human Retroviruses*, 17, 953-956.
- KRAKOWER, J. M., TRONICK, S. R., GALLAGHER, R. E., GALLO, R. C. & AARONSON, S. A. 1978. Antigenic characterization of a new gibbon ape leukemia virus isolate: seroepidemiologic assessment of an outbreak of gibbon leukemia. *International Journal of Cancer*, 22, 715-720.
- KUMAR, S. & HEDGES, S. B. 1998. A molecular timescale for vertebrate evolution. *Nature*, 392, 917-920.
- LA MANTIA, G., MAGLIONE, D., PENGUE, G., DI CRISTOFANO, A., SIMEONE, A., LANFRANCONE, L. & LANIA, L. 1991. Identification and characterization of novel human endogenous retroviral sequences preferentially expressed in undifferentiated embryonal carcinoma cells. *Nucleic Acids Research*, 19, 1513-1520.
- LAGUETTE, N., SOBHIAN, B., CASARTELLI, N., RINGEARD, M., CHABLE-BESSIA, C., SEGERAL, E., YATIM, A., EMILIANI, S., SCHWARTZ, O. & BENKIRANE, M. 2011. SAMHD1 is the dendritic- and myeloid-cell-specific HIV-1 restriction factor counteracted by Vpx. *Nature*, 474, 654-657.
- LAMERE, S. A., ST. LEGER, J. A., SCHRENZEL, M. D., ANTHONY, S. J., RIDEOUT, B. A. & SALOMON, D. R. 2009. Molecular Characterization of a Novel Gammaretrovirus in Killer Whales (*Orcinus orca*). *Journal of Virology*, 83, 12956-12967.
- LARKIN, M. A., BLACKSHIELDS, G., BROWN, N. P., CHENNA, R., MCGETTIGAN, P. A., MCWILLIAM, H., VALENTIN, F., WALLACE, I. M., WILM, A., LOPEZ, R., THOMPSON, J. D., GIBSON, T. J. & HIGGINS, D. G. 2007. Clustal W and Clustal X version 2.0. *Bioinformatics*, 23, 2947-2948.
- LASKA, M. J., BRUDEK, T., NISSEN, K. K., CHRISTENSEN, T., MOLLER-LARSEN, A., PETERSEN, T. & NEXO, B. A. 2012. Expression of HERV-Fc1, a human endogenous retrovirus, is increased in patients with active multiple sclerosis. *Journal of Virology*, 86, 3713-22.
- LASKA, M. J., NISSEN, K. K. & NEXØ, B. A. 2013. (Some) Cellular Mechanisms Influencing the Transcription of Human Endogenous Retrovirus, HERV-Fc1. *PLoS ONE*, 8, e53895.
- LAVIALLE, C., CORNELIS, G., DUPRESSOIR, A., ESNAULT, C., HEIDMANN, O., VERNOCHE, 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.
- LAVIE, L., MEDSTRAND, P., SCHEMPF, 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, 8788-8798.
- LE TORTOREC, A., WILLEY, S. & NEIL, S. J. 2011. Antiviral inhibition of enveloped virus release by tetherin/BST-2: action and counteraction. *Viruses*, 3, 520-40.

- 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.
- LEE, D., DAS GUPTA, J., GAUGHAN, C., STEFFEN, I., TANG, N., LUK, K.-C., QIU, X., URISMAN, A., FISCHER, N., MOLINARO, R., BROZ, M., SCHOCHETMAN, G., KLEIN, E. A., GANEM, D., DERISI, J. L., SIMMONS, G., HACKETT, J., JR., SILVERMAN, R. H. & CHIU, C. Y. 2012. In-Depth Investigation of Archival and Prospectively Collected Samples Reveals No Evidence for XMRV Infection in Prostate Cancer. *PLoS ONE*, 7, e44954.
- LEE, J. W. & KIM, H. S. 2006. Endogenous retrovirus HERV-I LTR family in primates: sequences, phylogeny, and evolution. *Archives of Virology*, 151, 1651-1658.
- LEE, S.-Y., HOWARD, T. M. & RASHEED, S. 1998. Genetic Analysis of the Rat Leukemia Virus: Influence of Viral Sequences in Transduction of the c-rasProto-Oncogene and Expression of Its Transforming Activity. *Journal of Virology*, 72, 9906-9917.
- LEE, Y. N. & BIENIASZ, P. D. 2007. Reconstitution of an Infectious Human Endogenous Retrovirus. *PLoS Pathogens*, 3, e10.
- LEFFLER, E. M., GAO, Z., PFEIFER, S., SÉGUREL, L., AUTON, A., VENN, O., BOWDEN, R., BONTROP, R., WALL, J. D., SELLA, G., DONNELLY, P., MCVEAN, G. & PRZEWORSKI, M. 2013. Multiple Instances of Ancient Balancing Selection Shared Between Humans and Chimpanzees. *Science*, 339, 1578-1582.
- LEHMANN, M. J., SHERER, N. M., MARKS, C. B., PYPAERT, M. & MOTHES, W. 2005. Actin- and myosin-driven movement of viruses along filopodia precedes their entry into cells. *Journal of Cell Biology*, 170, 317-25.
- LEMONS DE MATOS, A., VAN DER LOO, W., AREAL, H., LANNING, D. K. & ESTEVES, P. J. 2011. Study of Sylvilagus rabbit TRIM5alpha species-specific domain: how ancient endoviruses could have shaped the antiviral repertoire in Lagomorpha. *BMC Evolutionary Biology*, 11, 294.
- LENASI, T., CONTRERAS, X. & PETERLIN, B. M. 2010. Transcription, splicing and transport of retroviral RNA. In: BANNERT, N. & KURTH, R. (eds.) *Retroviruses. Molecular Biology, Genomics and Pathogenesis*. Norfolk: Caister Academic Press.
- LEPA, A. & SIWICKI, A. 2011. Retroviruses of wild and cultured fish. *Polish Journal of Veterinary Sciences*, 14, 703.
- LERCHE, N. W. & OSBORN, K. G. 2003. Simian Retrovirus Infections: Potential Confounding Variables in Primate Toxicology Studies. *Toxicologic Pathology*, 31, 103-110.
- LERCHE, N. W., SWITZER, W. M., YEE, J. L., SHANMUGAM, V., ROSENTHAL, A. N., CHAPMAN, L. E., FOLKS, T. M. & HENEINE, W. 2001. Evidence of Infection with Simian Type D Retrovirus in Persons Occupationally Exposed to Nonhuman Primates. *Journal of Virology*, 75, 1783-1789.
- LEVY, J. A., HOFFMAN, A. D., KRAMER, S. M., LANDIS, J. A., SHIMABUKURO, J. M. & OSHIRO, L. S. 1984. Isolation of lymphocytopathic retroviruses from San Francisco patients with AIDS. *Science*, 225, 840-2.
- LI, H.-C., FUJIYOSHI, T., LOU, H., YASHIKI, S., SONODA, S., CARTIER, L., NUNEZ, L., MUNOZ, I., HORAI, S. & TAJIMA, K. 1999. The presence of ancient human T-cell lymphotropic virus type I provirus DNA in an Andean mummy. *Nature Medicine*, 5, 1428-1432.

- LI, M., KAO, E., GAO, X., SANDIG, H., LIMMER, K., PAVON-ETERNOD, M., JONES, T. E., LANDRY, S., PAN, T., WEITZMAN, M. D. & DAVID, M. 2012. Codon-usage-based inhibition of HIV protein synthesis by human schlafen 11. *Nature*, 491, 125-128.
- LIANG, Q., XU, Z., XU, R., WU, L. & ZHENG, S. 2012. Expression Patterns of Non-Coding Spliced Transcripts from Human Endogenous Retrovirus HERV-H Elements in Colon Cancer. *PLoS ONE*, 7, e29950.
- LIEBER, M. M., SHERR, C. J., TODARO, G. J., BENVENISTE, R. E., CALLAHAN, R. & COON, H. G. 1975. Isolation from the Asian mouse *Mus caroli* of an endogenous type C virus related to infectious primate type C viruses. *Proceedings of the National Academy of Sciences*, 72, 2315-2319.
- LIM, E. S., FREGOSO, O. I., MCCOY, C. O., MATSEN, F. A., MALIK, H. S. & EMERMAN, M. 2012. The ability of primate lentiviruses to degrade the monocyte restriction factor SAMHD1 preceded the birth of the viral accessory protein Vpx. *Cell Host and Microbe*, 11, 194-204.
- LO, S.-C., PRIPUZOVA, N., LI, B., KOMAROFF, A. L., HUNG, G.-C., WANG, R. & ALTER, H. J. 2010. Detection of MLV-related virus gene sequences in blood of patients with chronic fatigue syndrome and healthy blood donors. *Proceedings of the National Academy of Sciences*, 107, 15874-15879.
- LOMBARDI, V. C., RUSCETTI, F. W., DAS GUPTA, J., PFOST, M. A., HAGEN, K. S., PETERSON, D. L., RUSCETTI, S. K., BAGNI, R. K., PETROW-SADOWSKI, C., GOLD, B., DEAN, M., SILVERMAN, R. H. & MIKOVITS, J. A. 2009. Detection of an Infectious Retrovirus, XMRV, in Blood Cells of Patients with Chronic Fatigue Syndrome. *Science*, 326, 585-589.
- LUBAN, J. 2010. Retroviral restriction factors. In: KURTH, R. & BANNERT, N. (eds.) *Retroviruses*. Norfolk: Caister Academic Press.
- 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, 206-211.
- MAEDA, N. & KIM, H. S. 1990. Three independent insertions of retrovirus-like sequences in the haptoglobin gene cluster of primates. *Genomics*, 8, 671-83.
- MAGIORKINIS, G., GIFFORD, R. J., KATZOURAKIS, A., DE RANTER, J. & BELSHAW, R. 2012. Env-less endogenous retroviruses are genomic superspreaders. *Proceedings of the National Academy of Sciences*, 109, 7385-7390.
- MALIK, H. S. & EICKBUSH, T. H. 1999. Modular Evolution of the Integrase Domain in the Ty3/Gypsy Class of LTR Retrotransposons. *Journal of Virology*, 73, 5186-5190.
- MALIK, H. S., HENIKOFF, S. & EICKBUSH, T. H. 2000. Poised for Contagion: Evolutionary Origins of the Infectious Abilities of Invertebrate Retroviruses. *Genome Research*, 10, 1307-1318.
- MANG, R., GOUDSMIT, J. & VAN DER KUYL, A. C. 1999. Novel Endogenous Type C Retrovirus in Baboons: Complete Sequence, Providing Evidence for Baboon Endogenous Virus gag-pol Ancestry. *Journal of Virology*, 73, 7021-7026.

- MANSFIELD, K. G., LERCH, N. W., GARDNER, M. B. & LACKNER, A. A. 1995. Origins of simian immunodeficiency virus infection in macaques at the New England Regional Primate Research Center. *Journal of Medical Primatology*, 24, 116-22.
- MARIANI, R., CHEN, D., SCHRÖFELBAUER, B., NAVARRO, F., KÖNIG, R., BOLLMAN, B., MÜNK, C., NYMARK-MCMAHON, H. & LANDAU, N. R. 2003. Species-Specific Exclusion of APOBEC3G from HIV-1 Virions by Vif. *Cell*, 114, 21-31.
- MARSHALL, J. T. 1974. Study of Vertebrate Reservoirs of Disease In: US ARMY MEDICAL COMPONENT (ed.) *SEATO Medical Research Laboratory Annual Progress Reports*.
- MARSHALL, J. T. 1975. Vertebrate Reservoirs of Disease In: US ARMY MEDICAL COMPONENT (ed.) *SEATO Medical Research Laboratory Annual Progress Reports*.
- MARTIN, J., HERNIOU, E., COOK, J., O'NEILL, R. W. & TRISTEM, M. 1999. Interclass Transmission and Phyletic Host Tracking in Murine Leukemia Virus-Related Retroviruses. *Journal of Virology*, 73, 2442-2449.
- MARTIN, J., HERNIOU, E., COOK, J., WAUGH O'NEILL, R. & TRISTEM, M. 1997. Human endogenous retrovirus type I-related viruses have an apparently widespread distribution within vertebrates. *Journal of Virology*, 71, 437-43.
- MARTIN, R. 2008. Colugos: obscure mammals glide into the evolutionary limelight. *Journal of Biology*, 7, 13.
- MARTÍNEZ BARRIO, Á., EKERLJUNG, M., JERN, P., BENACHENHOU, F., SPERBER, G. O., BONGCAM-RUDLOFF, E., BLOMBERG, J. & ANDERSSON, G. 2011. The First Sequenced Carnivore Genome Shows Complex Host-Endogenous Retrovirus Relationships. *PLoS ONE*, 6, e19832.
- MASAHITO, P., NISHIOKA, M., UEDA, H., KATO, Y., YAMAZAKI, I., NOMURA, K., SUGANO, H. & KITAGAWA, T. 1995. Frequent Development of Pancreatic Carcinomas in the *Rana nigromaculata* Group. *Cancer Research*, 55, 3781-3784.
- MATSUOKA, M. & JEANG, K.-T. 2007. Human T-cell leukaemia virus type 1 (HTLV-1) infectivity and cellular transformation. *Nature Reviews Cancer*, 7, 270-280.
- MAYER, J., SAUTER, M., RACZ, A., SCHERER, D., MUELLER-LANTZSCH, N. & MEESE, E. 1999. An almost-intact human endogenous retrovirus K on human chromosome 7. *Nature Genetics* 21, 257-258.
- MCCARTHY, E. & MCDONALD, J. 2004. Long terminal repeat retrotransposons of *Mus musculus*. *Genome Biology*, 5, R14.
- MCGIRR, K. & BUEHURING, G. 2006. Tax & Rex: Overlapping Genes of the Deltaretrovirus Group. *Virus Genes*, 32, 229-239.
- MCMANUS, K. F., KELLEY, J. L., SONG, S., VEERAMAH, K. R., WOERNER, A. E., STEVISON, L. S., RYDER, O. A., APE GENOME PROJECT, G., KIDD, J. M., WALL, J. D., BUSTAMANTE, C. D. & HAMMER, M. F. 2014. Inference of Gorilla Demographic and Selective History from Whole-Genome Sequence Data. *Molecular Biology and Evolution*.
- MEDSTRAND, P., MAGER, D. L., YIN, H., DIETRICH, U. & BLOMBERG, J. 1997. Structure and genomic organization of a novel human endogenous retrovirus family: HERV-K (HML-6). *Journal of General Virology*, 78, 1731-1744.

- MEISLER, M. H. & TING, C.-N. 1993. The Remarkable Evolutionary History of the Human Amylase Genes. *Critical Reviews in Oral Biology & Medicine*, 4, 503-509.
- MEREDITH, R. W., JANEČKA, J. E., GATESY, J., RYDER, O. A., FISHER, C. A., TEELING, E. C., GOODBLA, A., EIZIRIK, E., SIMÃO, T. L. L., STADLER, T., RABOSKY, D. L., HONEYCUTT, R. L., FLYNN, J. J., INGRAM, C. M., STEINER, C., WILLIAMS, T. L., ROBINSON, T. J., BURK-HERRICK, A., WESTERMAN, M., AYOUB, N. A., SPRINGER, M. S. & MURPHY, W. J. 2011. Impacts of the Cretaceous Terrestrial Revolution and KPg Extinction on Mammal Diversification. *Science*, 334, 521-524.
- MEYER, A. & ZARDOYA, R. 2003. Recent advances in the (molecular) phylogeny of vertebrates. *Annual Review of Ecology, Evolution, and Systematics*, 34, 311-338.
- MICHAEL, N. L. & MOORE, J. P. 1999. HIV-1 entry inhibitors: evading the issue. *Nature Medicine*, 5, 740-2.
- MIKOVITS, J. A., LOMBARDI, V. C., PFOST, M. A., HAGEN, K. S. & RUSCETTI, F. W. 2010. Detection of an infectious retrovirus, XMRV, in blood cells of patients with chronic fatigue syndrome. *Virulence*, 1, 386-390.
- MIYAUCHI, K., MARIN, M. & MELIKYAN, G. B. 2011. Visualization of retrovirus uptake and delivery into acidic endosomes. *Biochemistry Journal*, 434, 559-69.
- MIYAZAWA, T., YOSHIKAWA, R., GOLDBER, M., OKADA, M., STEWART, H. & PALMARINI, M. 2010. Isolation of an infectious endogenous retrovirus in a proportion of live attenuated vaccines for pets. *Journal of Virology*, 84, 3690-3694.
- MOIR, S., CHUN, T.-W. & FAUCI, A. S. 2011. Pathogenic Mechanisms of HIV Disease. *Annual Review of Pathology: Mechanisms of Disease*, 6, 223-248.
- MONAGHAN, M. T., GATTOLLIAT, J.-L., SARTORI, M., ELOUARD, J.-M., JAMES, H., DERLETH, P., GLAIZOT, O., DE MOOR, F. & VOGLER, A. P. 2005. Trans-oceanic and endemic origins of the small minnow mayflies (Ephemeroptera, Baetidae) of Madagascar. *Proceedings of the Royal Society B: Biological Sciences*, 272, 1829-1836.
- MORRIS, J. H., NEGUS, N. C. & SPERTZEL, R. O. 1966. Ecology of the Gibbon (*Hylobates lar* lar). In: US ARMY MEDICAL COMPONENT (ed.) *SEATO Medical Research Laboratory Annual Progress Reports*.
- MOTHES, W. & UCHIL, P. 2010. Retroviral entry and uncoating. In: BANNERT, N. & KURTH, R. (eds.) *Retroviruses. Molecular Biology, Genomics and Pathogenesis*. Norfolk: Caister Academic Press.
- MOYES, D., GRIFFITHS, D. J. & VENABLES, P. J. 2007. Insertional polymorphisms: a new lease of life for endogenous retroviruses in human disease. *Trends in genetics : TIG*, 23, 326-333.
- MUANGMAN, D. 1971. Laboratory Studies on the Interaction between Dengue Virus and the *Aedes aegypti* Mosquito Vector In: US ARMY MEDICAL COMPONENT (ed.) *SEATO Medical Research Laboratory Annual Progress Reports*.
- MURPHY, F. A., GIBBS, E. P. J., HORZINEK, M. C. & STUDDERT, M. J. 1999. *Veterinary Virology*, San Diego, Academic Press.
- MURPHY, H. W., MILLER, M., RAMER, J., TRAVIS, D., BARBIERS, R., WOLFE, N. D. & SWITZER, W. M. 2006. Implications of simian retroviruses for captive primate population

- management and the occupational safety of primate handlers. *Journal of Zoo and Wildlife Medicine*, 37, 219-33.
- MURPHY, K., TRAVERS, P. & WALPORT, M. 2008. *Janeway's Immunobiology*, New York, T & F Informa.
- NACHMAN, M. W. & CROWELL, S. L. 2000. Estimate of the Mutation Rate per Nucleotide in Humans. *Genetics*, 156, 297-304.
- NADEL, E., BANFIELD, W., BURSTEIN, S. & TOUSIMIS, A. 1967. Virus particles associated with strain 2 guinea pig leukemia (L2C/NB). *Journal of the National Cancer Institute*, 38, 979-982.
- NAKANO, K. & WATANABE, T. 2012. HTLV-1 Rex: the courier of viral messages making use of the host vehicle. *Frontiers of Microbiology*, 3, 330.
- NANDI, J. S., TIKUTE, S. A., CHHANGANI, A. K., POTDAR, V. A., TIWARI-MISHRA, M., ASHTEKAR, R. A., KUMARI, J., WALIMBE, A. & MOHNOT, S. M. 2003. Natural infection by simian retrovirus-6 (SRV-6) in Hanuman langurs (*Semnopithecus entellus*) from two different geographical regions of India. *Virology*, 311, 192-201.
- NEIL, S. J. D., ZANG, T. & BIENIASZ, P. D. 2008. Tetherin inhibits retrovirus release and is antagonized by HIV-1 Vpu. *Nature*, 451, 425-430.
- NELLAKER, C., KEANE, T., YALCIN, B., WONG, K., AGAM, A., BELGARD, T. G., FLINT, J., ADAMS, D., FRANKEL, W. & PONTING, C. 2012. The genomic landscape shaped by selection on transposable elements across 18 mouse strains. *Genome Biology*, 13, R45.
- NELLAKER, C., YAO, Y., JONES-BRANDO, L., MALLET, F., YOLKEN, R. H. & KARLSSON, H. 2006. Transactivation of elements in the human endogenous retrovirus W family by viral infection. *Retrovirology*, 3, 44.
- NEWBERNE, J. W. & ROBINSON, V. B. 1960. Spontaneous tumors in primates-a report of two cases with notes on the apparent low incidence of neoplasms in subhuman primates. *American Journal of Veterinary Research*, 21, 150-5.
- NIEWIADOMSKA, A. M. & GIFFORD, R. J. 2013. The Extraordinary Evolutionary History of the Reticuloendotheliosis Viruses. *PLoS Biology*, 11, e1001642.
- OAKES, B., TAI, A., CINGOZ, O., HENEFIELD, M., LEVINE, S., COFFIN, J. & HUBER, B. 2010. Contamination of human DNA samples with mouse DNA can lead to false detection of XMRV-like sequences. *Retrovirology*, 7-109.
- OJA, M., SPERBER, G. O., BLOMBERG, J. & KASKI, S. 2005. Self-organizing map-based discovery and visualization of human endogenous retroviral sequence groups. *International Journal of Neural Systems*, 15, 163-179.
- OKABE, H., GILDEN, R. V., HATANAKA, M., STEPHENSON, J. R., GALLAGHER, R. E., GALLO, R. C., TRONICK, S. R. & AARONSON, S. A. 1976. Immunological and biochemical characterisation of type C viruses isolated from cultured human AML cells. *Nature*, 260, 264-266.
- OPLER, S. R. 1967. Observations on a new virus associated with guinea pig leukemia: preliminary note. *Journal of the National Cancer Institute*, 38, 797-800.

- OVERBAUGH, J. & BANGHAM, C. R. M. 2001. Selection Forces and Constraints on Retroviral Sequence Variation. *Science*, 292, 1106-1109.
- OVERBAUGH, J., MILLER, A. D. & EIDEN, M. V. 2001. Receptors and Entry Cofactors for Retroviruses Include Single and Multiple Transmembrane-Spanning Proteins as well as Newly Described Glycophosphatidylinositol-Anchored and Secreted Proteins. *Microbiology and Molecular Biology Reviews*, 65, 371-389.
- PALMARINI, M., MURA, M. & SPENCER, T. E. 2004. Endogenous betaretroviruses of sheep: teaching new lessons in retroviral interference and adaptation. *Journal of General Virology*, 85, 1-13.
- PANDREA, I. & APETREI, C. 2010. Where the Wild Things Are: Pathogenesis of SIV Infection in African Nonhuman Primate Hosts. *Current HIV/AIDS Reports*, 7, 28-36.
- PARENT, I., QIN, Y., VANDENBROUCKE, A. T., WALON, C., DELFERRIÈRE, N., GODFROID, E. & BURTONBOY, G. 1998. Characterization of a C-type retrovirus isolated from an HIV infected cell line: complete nucleotide sequence. *Archives of Virology*, 143, 1077-1092.
- PAUL, T. A., QUACKENBUSH, S. L., SUTTON, C., CASEY, R. N., BOWSER, P. R. & CASEY, J. W. 2006. Identification and Characterization of an Exogenous Retrovirus from Atlantic Salmon Swim Bladder Sarcomas. *Journal of Virology*, 80, 2941-2948.
- PAVLÍČEK, A., PAČES, J., ELLEDER, D. & HEJNAR, J. 2002. Processed Pseudogenes of Human Endogenous Retroviruses Generated by LINEs: Their Integration, Stability, and Distribution. *Genome Research*, 12, 391-399.
- PAYNE, L. N. & NAIR, V. 2012. The long view: 40 years of avian leukosis research. *Avian Pathology*, 41, 11-19.
- PECON-SLATTERY, J., TROYER, J. L., JOHNSON, W. E. & O'BRIEN, S. J. 2008. Evolution of feline immunodeficiency virus in Felidae: implications for human health and wildlife ecology. *Vet Immunology and Immunopathology*, 123, 32-44.
- PEDERSEN, F. S. & SØRENSEN, A. B. 2010. Pathogenesis of Oncoviral Infections. In: KURTH, R. & BANNERT, N. (eds.) *Retroviruses*. Norfolk: Caister Academic Press.
- PERELMAN, P., JOHNSON, W. E., ROOS, C., SEUÁNEZ, H. N., HORVATH, J. E., MOREIRA, M. A. M., KESSING, B., PONTIUS, J., ROELKE, M., RUMPLER, Y., SCHNEIDER, M. P. C., SILVA, A., O'BRIEN, S. J. & PECON-SLATTERY, J. 2011. A Molecular Phylogeny of Living Primates. *PLoS Genetics*, 7, e1001342.
- PEREZ-CABALLERO, D., SOLL, S. J. & BIENIASZ, P. D. 2008. Evidence for Restriction of Ancient Primate Gammaretroviruses by APOBEC3 but Not TRIM5 α Proteins. *PLoS Pathogens*, 4, e1000181.
- PERRON, H., BERNARD, C., BERTRAND, J.-B., LANG, A. B., POPA, I., SANHADJI, K. & PORTOUKALIAN, J. 2009. Endogenous retroviral genes, Herpesviruses and gender in Multiple Sclerosis. *Journal of the Neurological Sciences*, 286, 65-72.
- PERRON, H., HAMDANI, N., FAUCARD, R., LAJNEF, M., JAMAIN, S., DABAN-HUARD, C., SARRAZIN, S., LEGUEN, E., HOUENOU, J., DELAVEST, M., MOINS-TEISERENC, H., BENGOUFA, D., YOLKEN, R., MADEIRA, A., GARCIA-MONTOJO, M., GEHIN, N., BURGELIN, I., OLLAGNIER, G., BERNARD, C., DUMAINE, A., HENRION, A., GOMBERT, A., LE DUDAL, K., CHARRON, D., KRISHNAMOORTHY, R., TAMOUZA, R. & LEBOYER, M.

2012. Molecular characteristics of Human Endogenous Retrovirus type-W in schizophrenia and bipolar disorder. *Translational Psychiatry*, 2.
- PERRON, H. & LANG, A. 2010. The Human Endogenous Retrovirus Link between Genes and Environment in Multiple Sclerosis and in Multifactorial Diseases Associating Neuroinflammation. *Clinical Reviews in Allergy & Immunology*, 39, 51-61.
- PERSAUD, D., GAY, H., ZIEMNIAK, C., CHEN, Y. H., PIATAK, M., CHUN, T.-W., STRAIN, M., RICHMAN, D. & LUZURIAGA, K. 2013. Absence of Detectable HIV-1 Viremia after Treatment Cessation in an Infant. *New England Journal of Medicine*, 369, 1828-1835.
- PERTEL, T., HAUSMANN, S., MORGER, D., ZUGER, S., GUERRA, J., LASCANO, J., REINHARD, C., SANTONI, F. A., UCHIL, P. D., CHATEL, L., BISIAUX, A., ALBERT, M. L., STRAMBIO-DE-CASTILLIA, C., MOTHES, W., PIZZATO, M., GRUTTER, M. G. & LUBAN, J. 2011. TRIM5 is an innate immune sensor for the retrovirus capsid lattice. *Nature*, 472, 361-365.
- PETERLIN, B. M. 2002. Charting HIV's remarkable voyage through the cell: basic science as a passport to future therapy. *Nature medicine*, 8, 673.
- PIANKA, E. R. 1970. On r- and K-Selection. *The American Naturalist*, 104, 592-597.
- PINCETIC, A. & LEIS, J. 2009. The Mechanism of Budding of Retroviruses from Cell Membranes. *Advances in Virology*, 2009, 9.
- PINCUS, T., ROWE, W. P. & LILLY, F. 1971. A MAJOR GENETIC LOCUS AFFECTING RESISTANCE TO INFECTION WITH MURINE LEUKEMIA VIRUSES: II. APPARENT IDENTITY TO A MAJOR LOCUS DESCRIBED FOR RESISTANCE TO FRIEND MURINE LEUKEMIA VIRUS. *The Journal of Experimental Medicine*, 133, 1234-1241.
- POLANI, S., ROCA, A. L., ROSENSTEEL, B. B., KOLOKOTRONIS, S.-O. & BAR-GAL, G. K. 2010. Evolutionary dynamics of endogenous feline leukemia virus proliferation among species of the domestic cat lineage. *Virology*, 405, 397-407.
- POLAVARAPU, N., BOWEN, N. J. & MCDONALD, J. F. 2006a. Identification, characterization and comparative genomics of chimpanzee endogenous retroviruses. *Genome Biology*, 7, R51.
- POLAVARAPU, N., BOWEN, N. J. & MCDONALD, J. F. 2006b. Newly Identified Families of Human Endogenous Retroviruses. *Journal of Virology*, 80, 4640-4642.
- POLI, G. & ERFLE, V. 2010. Pathogenesis of Immunodeficiency Virus Infections. In: KURTH, R. & BANNERT, N. (eds.) *Retroviruses*. Norfolk: Caister Academic Press.
- PONTIUS, J. U., MULLIKIN, J. C., SMITH, D. R., TEAM, A. S., LINDBLAD-TOH, K., GNERRE, S., CLAMP, M., CHANG, J., STEPHENS, R., NEELAM, B., VOLFOVSKY, N., SCHÄFFER, A. A., AGARWALA, R., NARFSTRÖM, K., MURPHY, W. J., GIGER, U., ROCA, A. L., ANTUNES, A., MENOTTI-RAYMOND, M., YUHKI, N., PECON-SLATTERY, J., JOHNSON, W. E., BOURQUE, G., TESLER, G., PROGRAM, N. C. S. & O'BRIEN, S. J. 2007. Initial sequence and comparative analysis of the cat genome. *Genome Research*, 17, 1675-1689.
- POSADA, D. & CRANDALL, K. A. 2001. Selecting Models of Nucleotide Substitution: An Application to Human Immunodeficiency Virus 1 (HIV-1). *Molecular Biology and Evolution*, 18, 897-906.

- PRUITT, K., BROWN, G., TATUSOVA, T. & MAGLOTT, D. 2012a. Chapter 18: The Reference Sequence (RefSeq) Database. In: MCENTYRE, J. & OSTELL, J. (eds.) *The NCBI Handbook*. Bethesda, USA: National Center for Biotechnology Information.
- PRUITT, K. D., TATUSOVA, T., BROWN, G. R. & MAGLOTT, D. R. 2012b. NCBI Reference Sequences (RefSeq): current status, new features and genome annotation policy. *Nucleic Acids Research*, 40, D130-5.
- PUTNAM, N. H., BUTTS, T., FERRIER, D. E. K., FURLONG, R. F., HELLSTEN, U., KAWASHIMA, T., ROBINSON-RECHAVI, M., SHOGUCHI, E., TERRY, A., YU, J.-K., BENITO-GUTIERREZ, E., DUBCHAK, I., GARCIA-FERNANDEZ, J., GIBSON-BROWN, J. J., GRIGORIEV, I. V., HORTON, A. C., DE JONG, P. J., JURKA, J., KAPITONOV, V. V., KOHARA, Y., KUROKI, Y., LINDQUIST, E., LUCAS, S., OSOEGAWA, K., PENNACCHIO, L. A., SALAMOV, A. A., SATOU, Y., SAUKA-SPENGLER, T., SCHMUTZ, J., SHIN-I, T., TOYODA, A., BRONNER-FRASER, M., FUJIYAMA, A., HOLLAND, L. Z., HOLLAND, P. W. H., SATOH, N. & ROKHSAR, D. S. 2008. The amphioxus genome and the evolution of the chordate karyotype. *Nature*, 453, 1064-1071.
- R CORE TEAM 2014. R: A language and environment for statistical computing. *R Foundation for Statistical Computing*.
- REBOLLO, R., MICELI-ROYER, K., ZHANG, Y., FARIVAR, S., GAGNIER, L. & MAGER, D. 2012. Epigenetic interplay between mouse endogenous retroviruses and host genes. *Genome Biology*, 13, R89.
- REDELSPERGER, F., CORNELIS, G., VERNOCHE, C., TENNANT, B. C., CATZEFLIS, F., MULOT, B., HEIDMANN, O., HEIDMANN, T. & DUPRESSOIR, A. 2014. Capture of syncytin-Mar1, a fusogenic endogenous retroviral envelope gene involved in placentation in the Rodentia squirrel-related clade. *Journal of Virology*.
- REISS, D. & MAGER, D. L. 2007. Stochastic epigenetic silencing of retrotransposons: Does stability come with age? *Gene*, 390, 130-135.
- REITZ, M. S. & GALLO, R. C. 2010. HTLV and HIV. In: KURTH, R. & BANNERT, N. (eds.) *Retroviruses. Molecular Biology, Genomics and Pathogenesis*. Norfolk: Caister Academic Press.
- REITZ, M. S., JR., WONG-STAAAL, F., HASELTINE, W. A., KLEID, D. G., TRAINOR, C. D., GALLAGHER, R. E. & GALLO, R. C. 1979. Gibbon ape leukemia virus-Hall's Island: new strain of gibbon ape leukemia virus. *Journal of Virology*, 29, 395-400.
- REUS, K., MAYER, J., SAUTER, M., ZISCHLER, H., MÜLLER-LANTZSCH, N. & MEESE, E. 2001. HERV-K(OLD): Ancestor Sequences of the Human Endogenous Retrovirus Family HERV-K(HML-2). *Journal of Virology*, 75, 8917-8926.
- RHODES, D. R. & CHINNAIYAN, A. M. 2005. Integrative analysis of the cancer transcriptome. *Nature Genetics*, 37 Suppl, S31-7.
- RIBET, D., HARPER, F., DEWANNIEUX, M., PIERRON, G. & HEIDMANN, T. 2007. Murine MusD Retrotransposon: Structure and Molecular Evolution of an "Intracellularized" Retrovirus. *Journal of Virology*, 81, 1888-1898.
- RIBET, D., HARPER, F., ESNAULT, C., PIERRON, G. & HEIDMANN, T. 2008. The GLN Family of Murine Endogenous Retroviruses Contains an Element Competent for Infectious Viral Particle Formation. *Journal of Virology*, 82, 4413-4419.

- RICE, P., LONGDEN, I. & BLEASBY, A. 2000. EMBOSS: The European Molecular Biology Open Software Suite. *Trends in Genetics*, 16, 276-277.
- RICHARDS, J. R. 2005. Feline immunodeficiency virus vaccine: Implications for diagnostic testing and disease management. *Biologicals*, 33, 215-217.
- ROBINSON, M., ERLWEIN, O., KAYE, S., WEBER, J., CINGOZ, O., PATEL, A., WALKER, M., KIM, W.-J., UIPRASERTKUL, M., COFFIN, J. & MCCLURE, M. 2010. Mouse DNA contamination in human tissue tested for XMRV. *Retrovirology*, 7, 108.
- ROMANO, C. M., RAMALHO, R. F. & ZANOTTO, P. M. D. A. 2006. Tempo and mode of ERV-K evolution in human and chimpanzee genomes. *Archives of Virology*, 151, 2215-2228.
- ROSS, S. R. 2010. Mouse Mammary Tumor Virus Molecular Biology and Oncogenesis. *Viruses*, 2, 2000-2012.
- ROVNAK, J. & QUACKENBUSH, S. L. 2010. Walleye dermal sarcoma virus: molecular biology and oncogenesis. *Viruses*, 2, 1984-1999.
- ROZAS, J. & ROZAS, R. 1995. DnaSP, DNA sequence polymorphism: an interactive program for estimating population genetics parameters from DNA sequence data. *Computer applications in the biosciences : CABIOS*, 11, 621-625.
- RUPRECHT, K., MAYER, J., SAUTER, M., ROEMER, K. & MUELLER-LANTZSCH, N. 2008. Endogenous retroviruses and cancer. *Cellular and Molecular Life Sciences*, 65, 3366-3382.
- SÁEZ-CIRIÓN, A., BACCHUS, C., HOCQUELOUX, L., AVETTAND-FENOEL, V., GIRAULT, I., LECUROUX, C., POTARD, V., VERSMISSE, P., MELARD, A., PRAZUCK, T., DESCOURS, B., GUERGNON, J., VIARD, J.-P., BOUFASSA, F., LAMBOTTE, O., GOUJARD, C., MEYER, L., COSTAGLIOLA, D., VENET, A., PANCINO, G., AUTRAN, B., ROUZIOUX, C. & THE, A. V. S. G. 2013. Post-Treatment HIV-1 Controllers with a Long-Term Virological Remission after the Interruption of Early Initiated Antiretroviral Therapy ANRS VISCONTI Study. *PLoS Pathogens* 9, e1003211.
- SAKUMA, R., NOSER, J. A., OHMINE, S. & IKEDA, Y. 2007. Rhesus monkey TRIM5[alpha] restricts HIV-1 production through rapid degradation of viral Gag polyproteins. *Nature Medicine*, 13, 631-635.
- SAMJI, H., CESCION, A., HOGG, R. S., MODUR, S. P., ALTHOFF, K. N., BUCHACZ, K., BURCHELL, A. N., COHEN, M., GEBO, K. A., GILL, M. J., JUSTICE, A., KIRK, G., KLEIN, M. B., KORTUIS, P. T., MARTIN, J., NAPRAVNIK, S., ROURKE, S. B., STERLING, T. R., SILVERBERG, M. J., DEEKS, S., JACOBSON, L. P., BOSCH, R. J., KITAHATA, M. M., GOEDERT, J. J., MOORE, R., GANGE, S. J., FOR THE NORTH AMERICAN, A. C. C. O. R. & DESIGN OF IE, D. E. A. 2013. Closing the Gap: Increases in Life Expectancy among Treated HIV-Positive Individuals in the United States and Canada. *PLoS ONE*, 8, e81355.
- SAMONDS, K. E., GODFREY, L. R., ALI, J. R., GOODMAN, S. M., VENCES, M., SUTHERLAND, M. R., IRWIN, M. T. & KRAUSE, D. W. 2013. Imperfect Isolation: Factors and Filters Shaping Madagascar's Extant Vertebrate Fauna. *PLoS ONE*, 8, e62086.
- SATO, E., FURUTA, R. & MIYAZAWA, T. 2010. An Endogenous Murine Leukemia Viral Genome Contaminant in a Commercial RT-PCR Kit is Amplified Using Standard Primers for XMRV. *Retrovirology*, 7, 110.

- SAWYER, S. L., EMERMAN, M. & MALIK, H. S. 2004. Ancient Adaptive Evolution of the Primate Antiviral DNA-Editing Enzyme APOBEC3G. *PLoS Biology*, 2, e275.
- SAWYER, S. L., WU, L. I., EMERMAN, M. & MALIK, H. S. 2005. Positive selection of primate TRIM5 α identifies a critical species-specific retroviral restriction domain. *Proceedings of the National Academy of Sciences of the United States of America*, 102, 2832-2837.
- SCHMIDT, M., WIRTH, T., KRÖGER, B. & HORAK, I. 1985. Structure and genomic organization of a new family of murine retrovirus-related DNA sequences (MuRRS). *Nucleic Acids Research*, 13, 3461-3470.
- SCHMITT, K., REICHRATH, J., ROESCH, A., MEESE, E. & MAYER, J. 2013. Transcriptional Profiling of Human Endogenous Retrovirus Group HERV-K(HML-2) Loci in Melanoma. *Genome Biology and Evolution*, 5, 307-328.
- SCHÖN, U., SEIFARTH, W., BAUST, C., HOHENADL, C., ERFLE, V. & LEIB-MÖSCH, C. 2001. Cell Type-Specific Expression and Promoter Activity of Human Endogenous Retroviral Long Terminal Repeats. *Virology*, 279, 280-291.
- SCHWARTZ, A. M., MCCRACKIN, M. A., SCHINAZI, R. F., HILL, P. B., VAHLENKAMP, T. W., TOMPKINS, M. B. & HARTMANN, K. 2014. Antiviral efficacy of nine nucleoside reverse transcriptase inhibitors against feline immunodeficiency virus in feline peripheral blood mononuclear cells. *American Journal of Veterinary Research*, 75, 273-281.
- SEIFARTH, W., BAUST, C., MURR, A., SKLADNY, H., KRIEG-SCHNEIDER, F., BLUSCH, J., WERNER, T., HEHLMANN, R. & LEIB-MÖSCH, C. 1998. Proviral Structure, Chromosomal Location, and Expression of HERV-K-T47D, a Novel Human Endogenous Retrovirus Derived from T47D Particles. *Journal of Virology*, 72, 8384-8391.
- SEIFARTH, W., BAUST, C., SCHON, U., REICHERT, A., HEHLMANN, R. & LEIB-MOSCH, C. 2000. HERV-IP-T47D, a novel type C-related human endogenous retroviral sequence derived from T47D particles. *AIDS Research and Human Retroviruses*, 16, 471-80.
- 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, 6408-6416.
- SELLON, R. K. & HARTMANN, K. 2006. Feline Immunodeficiency Virus Infection. In: GREENE, C. E. (ed.) *Infectious Diseases of the Dog and Cat*. Missouri: Elsevier.
- SHCHELOKOVSKYY, P., TRISTRAM-NAGLE, S. & DIMOVA, R. 2011. Effect of the HIV-1 fusion peptide on the mechanical properties and leaflet coupling of lipid bilayers. *New Journal of Physics*, 13, 25004.
- SHEEHY, A. M., GADDIS, N. C., CHOI, J. D. & MALIM, M. H. 2002. Isolation of a human gene that inhibits HIV-1 infection and is suppressed by the viral Vif protein. *Nature*, 418, 646-650.
- SHEN, C.-H. & STEINER, L. A. 2004. Genome Structure and Thymic Expression of an Endogenous Retrovirus in Zebrafish. *Journal of Virology*, 78, 899-911.
- SHIN, W., LEE, J., SON, S.-Y., AHN, K., KIM, H.-S. & HAN, K. 2013. Human-Specific HERV-K Insertion Causes Genomic Variations in the Human Genome. *PLoS ONE*, 8, e60605.

- SIMMONS, G., CLARKE, D., MCKEE, J., YOUNG, P. & MEERS, J. 2014. Discovery of a Novel Retrovirus Sequence in an Australian Native Rodent (*Melomys burtoni*): A Putative Link between Gibbon Ape Leukemia Virus and Koala Retrovirus. *PLoS ONE*, 9, e106954.
- SIMMONS, G., GLYNN, S. A., KOMAROFF, A. L., MIKOVITS, J. A., TOBLER, L. H., HACKETT, J., TANG, N., SWITZER, W. M., HENEINE, W., HEWLETT, I. K., ZHAO, J., LO, S.-C., ALTER, H. J., LINNEN, J. M., GAO, K., COFFIN, J. M., KEARNEY, M. F., RUSCETTI, F. W., PFOST, M. A., BETHEL, J., KLEINMAN, S., HOLMBERG, J. A., BUSCH, M. P. & GROUP, F. T. B. X. S. R. W. 2011. Failure to Confirm XMRV/MLVs in the Blood of Patients with Chronic Fatigue Syndrome: A Multi-Laboratory Study. *Science*, 334, 814-817.
- SINGH, S., KAYE, S., FRANCIS, N., PESTON, D., GORE, M., MCCLURE, M. & BUNKER, C. 2013. Human endogenous retrovirus K (HERV-K) rec mRNA is expressed in primary melanoma but not in benign naevi or normal skin. *Pigment Cell & Melanoma Research*, 26, 426-428.
- SINZELLE, L., CARRADEC, Q., PAILLARD, E., BRONCHAIN, O. J. & POLLET, N. 2011. Characterization of a *Xenopus tropicalis* Endogenous Retrovirus with Developmental and Stress-Dependent Expression. *Journal of Virology*, 85, 2167-2179.
- SLATER, G. & BIRNEY, E. 2005. Automated generation of heuristics for biological sequence comparison. *BMC Bioinformatics*, 6, 31.
- SMIT, A. F. 1996. The origin of interspersed repeats in the human genome. *Current Opinion in Genetics and Development*, 6, 743-748.
- SMITH, P. C., YUILL, T. M. & BUCHANAN, R. D. 1968. Effect of *P. falciparum* infection on Serum Biochemistry Values of the Gibbon In: US ARMY MEDICAL COMPONENT (ed.) *SEATO Medical Research Laboratory Annual Progress Reports*.
- SMITH, R. 2010. Contamination of clinical specimens with MLV-encoding nucleic acids: implications for XMRV and other candidate human retroviruses. *Retrovirology*, 7, 112.
- SMITH, T. F. & WATERMAN, M. S. 1981. Identification of common molecular subsequences. *Journal of Molecular Biology*, 147, 195-197.
- SNYDER, S. P., DUNGWORTH, D. L., KAWAKAMI, T. G., CALLAWAY, E. & LAU, D. T. 1973. Lymphosarcomas in two gibbons (*Hylobates lar*) with associated C-type virus. *Journal of the National Cancer Institute*, 51, 89-94.
- SOEHARSONO, S., WILCOX, G., PUTRA, A., HARTANINGSIH, N., SULISTYANA, K. & TENAYA, M. 1995. The transmission of Jembrana disease, a lentivirus disease of *Bos javanicus* cattle. *Epidemiology and infection*, 115, 367-374.
- SOMMERFELT, M. A. 1999. Retrovirus receptors. *Journal of General Virology*, 80, 3049-3064.
- SOMMERFELT, M. A., HARKESTAD, N. & HUNTER, E. 2003. The endogenous langur type D retrovirus PO-1-Lu and its exogenous counterparts in macaque and langur monkeys. *Virology*, 315, 275-282.
- SONIGO, P., BARKER, C., HUNTER, E. & WAIN-HOBSON, S. 1986. Nucleotide sequence of Mason-Pfizer monkey virus: An immunosuppressive D-type retrovirus. *Cell*, 45, 375-385.

- SPERBER, G., LOVGREN, A., ERIKSSON, N.-E., BENACHENHOU, F. & BLOMBERG, J. 2009. RetroTector online, a rational tool for analysis of retroviral elements in small and medium size vertebrate genomic sequences. *BMC Bioinformatics*, 10, S4.
- SPERBER, G. O., AIROLA, T., JERN, P. & BLOMBERG, J. 2007. Automated recognition of retroviral sequences in genomic data—RetroTector®. *Nucleic Acids Research*, 35, 4964-4976.
- ST.CYR COATS, K., PRUETT, S. B., NASH, J. W. & COOPER, C. R. 1994. Bovine immunodeficiency virus: incidence of infection in Mississippi dairy cattle. *Veterinary Microbiology*, 42, 181-189.
- STATISTICS AND INFORMATION SERVICE OF THE FISHERIES AND AQUACULTURE DEPARTMENT 2012. FAO yearbook. Fishery and Aquaculture Statistics.
- STAUFFER, Y., THEILER, G., SPERISEN, P., LEBEDEV, Y. & JONGENEEL, C. V. 2004. Digital expression profiles of human endogenous retroviral families in normal and cancerous tissues. *Cancer Immunology*, 4, 2.
- STEVENSON, M. 2003. HIV-1 pathogenesis. *Nature Medicine*, 9, 853-60.
- STOCKING, C. & KOZAK, C. 2008. Endogenous retroviruses. *Cellular and Molecular Life Sciences*, 65, 3383-3398.
- STOYE, J. P. 2012. Studies of endogenous retroviruses reveal a continuing evolutionary saga. *Nature Reviews of Microbiology*, 10, 395-406.
- STOYE, J. P. & COFFIN, J. M. 1987. The four classes of endogenous murine leukemia virus: structural relationships and potential for recombination. *Journal of Virology*, 61, 2659-2669.
- STOYE, J. P., MORONI, C. & COFFIN, J. M. 1991. Virological events leading to spontaneous AKR thymomas. *Journal of Virology*, 65, 1273-1285.
- STRAUSS, J. H. & STRAUSS, E. G. 2008. *Viruses and Human Disease*, Oxford, UK, Elsevier.
- STREMLAU, M., OWENS, C. M., PERRON, M. J., KIESSLING, M., AUTISSIER, P. & SODROSKI, J. 2004. The cytoplasmic body component TRIM5[alpha] restricts HIV-1 infection in Old World monkeys. *Nature*, 427, 848-853.
- STRICK, R., ACKERMANN, S., LANGBEIN, M., SWIATEK, J., SCHUBERT, S., HASHEMOLHOSSEINI, S., KOSCHECK, T., FASCHING, P., SCHILD, R., BECKMANN, M. & STRISSEL, P. 2007. Proliferation and cell–cell fusion of endometrial carcinoma are induced by the human endogenous retroviral Syncytin-1 and regulated by TGF-β. *Journal of Molecular Medicine*, 85, 23-38.
- SUBRAMANIAN, R., WILDSCHUTTE, J., RUSSO, C. & COFFIN, J. 2011. Identification, characterization, and comparative genomic distribution of the HERV-K (HML-2) group of human endogenous retroviruses. *Retrovirology*, 8, 90.
- SUZUKI, Y. & CRAIGIE, R. 2007. The road to chromatin - nuclear entry of retroviruses. *Nature Reviews of Microbiology*, 5, 187-96.
- SWITZER, W., JIA, H., HOHN, O., ZHENG, H., TANG, S., SHANKAR, A., BANNERT, N., SIMMONS, G., HENDRY, R. M., FALKENBERG, V., REEVES, W. & HENEINE, W. 2010. Absence of evidence of Xenotropic Murine Leukemia Virus-related virus infection in persons

- with Chronic Fatigue Syndrome and healthy controls in the United States. *Retrovirology*, 7, 57.
- TAKEUCHI, J., PERCHE, B., MIGRAINE, J., MERCIER-DELARUE, S., PONSCARME, D., SIMON, F., CLAVEL, F. & LABROSSE, B. 2013. High level of susceptibility to human TRIM5alpha conferred by HIV-2 capsid sequences. *Retrovirology*, 10, 50.
- TANG, M. & SHAFER, R. 2012. HIV-1 Antiretroviral Resistance. *Drugs*, 72, e1-e25.
- 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, 783-787.
- TARLINTON, R., MEERS, J. & YOUNG, P. 2008. Endogenous retroviruses. *Cellular and Molecular Life Sciences*, 65, 3413-3421.
- TARLINTON, R. E., BARFOOT, H. K. R., ALLEN, C. E., BROWN, K., GIFFORD, R. J. & EMES, R. D. 2012. Characterisation of a group of endogenous gammaretroviruses in the canine genome. *The Veterinary Journal*.
- TARLINTON, R. E., MEERS, J. & YOUNG, P. R. 2006. Retroviral invasion of the koala genome. *Nature*, 442, 79-81.
- TARUSCIO, D. & MANTOVANI, A. 1996. Eleven chromosomal integration sites of a human endogenous retrovirus (HERV 4-1) map close to known loci of thirteen hereditary malformation syndromes. *Teratology*, 54, 108-10.
- TARUSCIO, D. & MANUELIDIS, L. 1991. Integration site preferences of endogenous retroviruses. *Chromosoma*, 101, 141-156.
- TERZIAN, C., PELISSON, A. & BUCHETON, A. 2001. Evolution and phylogeny of insect endogenous retroviruses. *BMC Evolutionary Biology*, 1, 3.
- THEILEN, G. H., GOULD, D., FOWLER, M. & DUNGWORTH, D. L. 1971. C-type virus in tumor tissue of a woolly monkey (*Lagothrix* spp.) with fibrosarcoma. *Journal of the National Cancer Institute*, 47, 881-9.
- THEODOROU, V., KIMM, M. A., BOER, M., WESSELS, L., THEELEN, W., JONKERS, J. & HILKENS, J. 2007. MMTV insertional mutagenesis identifies genes, gene families and pathways involved in mammary cancer. *Nature Genetics*, 39, 759-69.
- TIPTON, K. F. 1994. Nomenclature Committee of the International Union of Biochemistry and Molecular Biology (NC-IUBMB). Enzyme nomenclature. Recommendations 1992. Supplement: corrections and additions. *European Journal of Biochemistry*, 223, 1-5.
- TODARO, G. J., LIEBER, M. M., BENVENISTE, R. E. & SHERR, C. J. 1975. Infectious primate type C viruses: Three isolates belonging to a new subgroup from the brains of normal gibbons. *Virology*, 67, 335-43.
- TÖNJES, R. R. & NIEBERT, M. 2003. Relative Age of Proviral Porcine Endogenous Retrovirus Sequences in *Sus scrofa* Based on the Molecular Clock Hypothesis. *Journal of Virology*, 77, 12363-12368.
- TOWERS, G. 2007. The control of viral infection by tripartite motif proteins and cyclophilin A. *Retrovirology*, 4, 40.

- TRISTEM, M. 2000. Identification and Characterization of Novel Human Endogenous Retrovirus Families by Phylogenetic Screening of the Human Genome Mapping Project Database. *Journal of Virology*, 74, 3715-3730.
- TRISTEM, M., KABAT, P., LIEBERMAN, L., LINDE, S., KARPAS, A. & HILL, F. 1996. Characterization of a novel murine leukemia virus-related subgroup within mammals. *Journal of Virology*, 70, 8241-8246.
- 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. *Current Biology*, 11, 1531-1535.
- U.S. FISH AND WILDLIFE SERVICE 1974. ENFORCEMENT OF WILDLIFE LAWS - GIBBONS 1974STATE260768_b.
- U.S. FISH AND WILDLIFE SERVICE 1975. ENFORCEMENT OF WILDLIFE LAWS - GIBBONS 1975BANGKO14759_b.
- URISMAN, A., MOLINARO, R. J., FISCHER, N., PLUMMER, S. J., CASEY, G., KLEIN, E. A., MALATHI, K., MAGI-GALLUZZI, C., TUBBS, R. R., GANEM, D., SILVERMAN, R. H. & DERISI, J. L. 2006. Identification of a Novel Gammaretrovirus in Prostate Tumors of Patients Homozygous for R462Q RNASEL Variant. *PLoS Pathogens*, 2, e25.
- VAN DER KUYL, A. 2012. HIV infection and HERV expression: a review. *Retrovirology*, 9, 1-10.
- VAN DER KUYL, A., MANG, R., DEKKER, J. & GOUDSMIT, J. 1997. Complete nucleotide sequence of simian endogenous type D retrovirus with intact genome organization: evidence for ancestry to simian retrovirus and baboon endogenous virus. *Journal of Virology*, 71, 3666-3676.
- VAN DER KUYL, A. C. 2011. Characterization of a Full-Length Endogenous Beta-Retrovirus, EqERV-Beta1, in the Genome of the Horse (*Equus caballus*). *Viruses*, 3, 620-8.
- VAN DER KUYL, A. C., DEKKER, J. T. & GOUDSMIT, J. 1995. Distribution of baboon endogenous virus among species of African monkeys suggests multiple ancient cross-species transmissions in shared habitats. *Journal of Virology*, 69, 7877-87.
- VAN DER KUYL, A. C., DEKKER, J. T. & GOUDSMIT, J. 1999. Discovery of a New Endogenous Type C Retrovirus (FcEV) in Cats: Evidence for RD-114 Being an FcEVGag-Pol/Baboon Endogenous Virus BaEVEEnv Recombinant. *Journal of Virology*, 73, 7994-8002.
- VAN KUPPEVELD, F. J. M., JONG, A. S. D., LANKE, K. H., VERHAEGH, G. W., MELCHERS, W. J. G., SWANINK, C. M. A., BLEIJENBERG, G., NETEA, M. G., GALAMA, J. M. D. & VAN DER MEER, J. W. M. 2010. Prevalence of xenotropic murine leukaemia virus-related virus in patients with chronic fatigue syndrome in the Netherlands: retrospective analysis of samples from an established cohort. *British Medical Journal*, 340.
- VANDEVALK, A. J., ADAMS, C. M., RUDSTAM, L. G., FORNEY, J. L., BROOKING, T. E., GERKEN, M. A., YOUNG, B. P. & HOOPER, J. T. 2002. Comparison of Angler and Cormorant Harvest of Walleye and Yellow Perch in Oneida Lake, New York. *Transactions of the American Fisheries Society*, 131, 27-39.
- VANDEWOUDE, S. & APETREI, C. 2006. Going Wild: Lessons from Naturally Occurring T-Lymphotropic Lentiviruses. *Clinical Microbiology Reviews*, 19, 728-762.

- VERNOCHET, C., HEIDMANN, O., DUPRESSOIR, A., CORNELIS, G., DESSEN, P., CATZEFLIS, F. & HEIDMANN, T. 2011. A syncytin-like endogenous retrovirus envelope gene of the guinea pig specifically expressed in the placenta junctional zone and conserved in Caviomorpha. *Placenta*, 885-892.
- VILLESEN, P., AAGAARD, L., WIUF, C. & PEDERSEN, F. 2004. Identification of endogenous retroviral reading frames in the human genome. *Retrovirology*, 1, 32.
- VOEVODIN, A. F. & MARX, P. A. 2009. *Simian Virology*, Iowa, USA, Blackwell Publishing.
- VOGEL, M., SCHWARZE-ZANDER, C., WASMUTH, J. C., SPENGLER, U., SAUERBRUCH, T. & ROCKSTROH, J. K. 2010. The treatment of patients with HIV. *Deutsches Ärzteblatt International*, 107, 507-515.
- VOISSET, C. & ANDRAWISS, M. 2000. Retroviruses at a glance. *Genome Biology*, 1, 4015.1-4015.4.
- WAINBERG, MARK A. & JEANG, K.-T. 2011. XMRV as a Human Pathogen? *Cell Host & Microbe*, 9, 260-262.
- WALKER, J. D., GEISSMAN, J. W., BOWRING, S. A. & BABCOCK, L. E. 2012. *Geologic Time Scale v. 4.0: Geological Society of America*, The Geological Society of America.
- WALLACE, I. M., ORLA, O. S. & HIGGINS, D. G. 2005. Evaluation of iterative alignment algorithms for multiple alignment. *Bioinformatics*, 21, 1408-1414.
- WANG-JOHANNING, F., FROST, A. R., JIAN, B., AZEROU, R., LU, D. W., CHEN, D.-T. & JOHANNING, G. L. 2003. Detecting the expression of human endogenous retrovirus E envelope transcripts in human prostate adenocarcinoma. *Cancer*, 98, 187-197.
- WANG-JOHANNING, F., FROST, A. R., JOHANNING, G. L., KHAZAELI, M. B., LOBUGLIO, A. F., SHAW, D. R. & STRONG, T. V. 2001. Expression of Human Endogenous Retrovirus K Envelope Transcripts in Human Breast Cancer. *Clinical Cancer Research*, 7, 1553-1560.
- WANG-JOHANNING, F., LIU, J., RYCAJ, K., HUANG, M., TSAI, K., ROSEN, D. G., CHEN, D.-T., LU, D. W., BARNHART, K. F. & JOHANNING, G. L. 2007. Expression of multiple human endogenous retrovirus surface envelope proteins in ovarian cancer. *International Journal of Cancer*, 120, 81-90.
- WANG, L., YIN, Q., HE, G., ROSSITER, S. J., HOLMES, E. C. & CUI, J. 2013. Ancient invasion of an extinct gammaretrovirus in cetaceans. *Virology*, 441, 66-69.
- WANG, X.-S., ZHANG, Z., WANG, H.-C., CAI, J.-L., XU, Q.-W., LI, M.-Q., CHEN, Y.-C., QIAN, X.-P., LU, T.-J., YU, L.-Z., ZHANG, Y., XIN, D.-Q., NA, Y.-Q. & CHEN, W.-F. 2006. Rapid Identification of UCA1 as a Very Sensitive and Specific Unique Marker for Human Bladder Carcinoma. *Clinical Cancer Research*, 12, 4851-4858.
- WANG, Y., LISKA, F., GOSELE, C., SEDOVA, L., KREN, V., KRENOVA, D., IVICS, Z., HUBNER, N. & IZSVAK, Z. 2010. A novel active endogenous retrovirus family contributes to genome variability in rat inbred strains. *Genome Research*, 20, 19-27.
- WARNOW, T. 2013. Large-Scale Multiple Sequence Alignment and Phylogeny Estimation. In: CHAUVE, C., EL-MABROUK, N. & TANNIER, E. (eds.) *Models and Algorithms for Genome Evolution*. Springer London.

- WARRILOW, D., TACHEDJIAN, G. & HARRICH, D. 2009. Maturation of the HIV reverse transcription complex: putting the jigsaw together. *Reviews in Medical Virology*, 19, 324-337.
- WEISS, R. 2006. The discovery of endogenous retroviruses. *Retrovirology*, 3, 67.
- WEISS, R. A. & VOGT, P. K. 2011. 100 years of Rous sarcoma virus. *The Journal of Experimental Medicine*, 208, 2351-2355.
- WENTZENSEN, N., COY, J. F., KNAEBEL, H.-P., LINNEBACHER, M., WILZ, B., GEBERT, J. & VON KNEBEL DOEBERITZ, M. 2007. Expression of an endogenous retroviral sequence from the HERV-H group in gastrointestinal cancers. *International Journal of Cancer*, 121, 1417-1423.
- WHITE, T. E., BRANDARIZ-NUNEZ, A., VALLE-CASUSO, J. C., AMIE, S., NGUYEN, L. A., KIM, B., TUZOVA, M. & DIAZ-GRIFFERO, F. 2013. The retroviral restriction ability of SAMHD1, but not its deoxynucleotide triphosphohydrolase activity, is regulated by phosphorylation. *Cell Host and Microbe*, 13, 441-51.
- WOLGAMOT, G., BONHAM, L. & MILLER, A. D. 1998. Sequence analysis of Mus dunni endogenous virus reveals a hybrid VL30/gibbon ape leukemia virus-like structure and a distinct envelope. *Journal of Virology*, 72, 7459-7466.
- WU, T., YAN, Y. & KOZAK, C. A. 2005. Rmcf2, a xenotropic provirus in the Asian mouse species Mus castaneus, blocks infection by polytropic mouse gammaretroviruses. *Journal of Virology*, 79, 9677-84.
- XIONG, Y. & EICKBUSH, T. H. 1990. Origin and evolution of retroelements based upon their reverse transcriptase sequences. *EMBO Journal*, 9, 3353-62.
- YANDELL, M. & ENCE, D. 2012. A beginner's guide to eukaryotic genome annotation. *Nature Reviews of Genetics*, 13, 329-342.
- YAP, M. W., COLBECK, E., ELLIS, S. A. & STOYE, J. P. 2014. Evolution of the Retroviral Restriction Gene Fv1: Inhibition of Non-MLV Retroviruses. *PLoS Pathogens*, 10, e1003968.
- YAP, M. W. & STOYE, J. P. 2013. Apparent effect of rabbit endogenous lentivirus type K acquisition on retrovirus restriction by lagomorph Trim5αs. *Philosophical Transactions of the Royal Society B: Biological Sciences*, 368.
- YI, J.-M. & KIM, H.-S. 2007. Expression and phylogenetic analyses of human endogenous retrovirus HC2 belonging to the HERV-T family in human tissues and cancer cells. *Journal of Human Genetics*, 52, 285-296.
- YI, J.-M., SCHUEBEL, K. & KIM, H.-S. 2007. Molecular genetic analyses of human endogenous retroviral elements belonging to the HERV-P family in primates, human tissues, and cancer cells. *Genomics*, 89, 1-9.
- YI, J. M. & KIM, H. S. 2006. Molecular evolution of the HERV-E family in primates. *Archives of Virology*, 151, 1107-1116.
- YODER, A. D., BURNS, M. M., ZEHR, S., DELEFOSSE, T., VERON, G., GOODMAN, S. M. & FLYNN, J. J. 2003. Single origin of Malagasy Carnivora from an African ancestor. *Nature*, 421, 734-737.

- YOHN, C. T., JIANG, Z., MCGRATH, S. D., HAYDEN, K. E., KHAITOVICH, P., JOHNSON, M. E., EICHLER, M. Y., MCPHERSON, J. D., ZHAO, S., PÄÄBO, S. & EICHLER, E. E. 2005. Lineage-specific expansions of retroviral insertions within the genomes of African great apes but not humans and orangutans. *PLoS Biology*, 3, e110.
- YOSHIKAWA, R., SATO, E., IGARASHI, T. & MIYAZAWA, T. 2010. Characterization of RD-114 virus isolated from a commercial canine vaccine manufactured using CRFK cells. *Journal of Clinical Microbiology*, 48, 3366-3369.
- YOSHIKAWA, R., SATO, E. & MIYAZAWA, T. 2011a. Contamination of infectious RD-114 virus in vaccines produced using non-feline cell lines. *Biologicals*, 39, 33-37.
- YOSHIKAWA, R., SATO, E. & MIYAZAWA, T. 2011b. Presence of Infectious RD-114 Virus in a Proportion of Canine Parvovirus Isolates. *Journal of Veterinary Medicine and Science*, 74, 347-350.
- YOSHIKAWA, R., YASUDA, J., KOBAYASHI, T. & MIYAZAWA, T. 2012. Canine ASCT1 and ASCT2 are functional receptors for RD-114 virus in dogs. *Journal of General Virology*, 93, 603-607.
- YOUNG, G. R., EKSMOND, U., SALCEDO, R., ALEXOPOULOU, L., STOEY, J. P. & KASSIOTIS, G. 2012. Resurrection of endogenous retroviruses in antibody-deficient mice. *Nature*, 491, 774-778.
- YU, P., LÜBBEN, W., SLOMKA, H., GEBLER, J., KONERT, M., CAI, C., NEUBRANDT, L., PRAZERES DA COSTA, O., PAUL, S., DEHNERT, S., DÖHNE, K., THANISCH, M., STORSBERG, S., WIEGAND, L., KAUFMANN, A., NAIN, M., QUINTANILLA-MARTINEZ, L., BETTIO, S., SCHNIERLE, B., KOLESNIKOVA, L., BECKER, S., SCHNARE, M. & BAUER, S. 2012a. Nucleic Acid-Sensing Toll-like Receptors Are Essential for the Control of Endogenous Retrovirus Viremia and ERV-Induced Tumors. *Immunity*, 37, 867-879.
- YU, S. L., JUNG, W. Y., JUNG, K. C., CHO, I. C., LIM, H. T., JIN, D. I. & LEE, J. H. 2012b. Characterization of porcine endogenous retrovirus clones from the NIH miniature pig BAC library. *Journal of biomedicine & biotechnology*, 2012, 482568.
- ZAKHAROV, E. V., SMITH, C. R., LEES, D. C., CAMERON, A., VANE-WRIGHT, R. I. & SPERLING, F. A. H. 2004. INDEPENDENT GENE PHYLOGENIES AND MORPHOLOGY DEMONSTRATE A MALAGASY ORIGIN FOR A WIDE-RANGING GROUP OF SWALLOWTAIL BUTTERFLIES. *Evolution*, 58, 2763-2782.
- ZHANG, J. & WEBB, D. M. 2004. Rapid evolution of primate antiviral enzyme APOBEC3G. *Human Molecular Genetics*, 13, 1785-1791.
- ZSÍROS, J., JEBBINK, M. F., LUKASHOV, V. V., VOÛTE, P. A. & BERKHOUT, B. 1999. Biased Nucleotide Composition of the Genome of HERV-K Related Endogenous Retroviruses and Its Evolutionary Implications. *Journal of Molecular Evolution*, 48, 102-111.