University of Glasgow

Mining genome data for endogenous viral elements and interferon stimulated genes: insights into host virus co-evolution

by
Tristan Philip Wesley Dennis
BSc (Hons)

Submitted in fulfilment of the requirements of the degree of Doctor of Philosophy (PhD).

MRC-University of Glasgow Centre for Virus Research
College of Medical, Veterinary and Life Sciences

July 2018

Abstract

Paleovirology is the study of viruses over evolutionary timescales. Contemporary paleovirological analyses often rely on sequence data, derived from organism genome assemblies. These sequences are the germline inherited remnants of past viral infection, in the form of endogenous viral elements and the host immune genes that are evolving to combat viruses. Their study has found that viruses have exerted profound influences on host evolution, and highlighted the conflicts between viruses and host immunity. As genome sequencing technology cheapens, the accumulation of genome data increases, furthering the potential for paleovirological insights. However, data on ERVs, EVEs and antiviral gene evolution, are often not captured by automated annotation pipelines. As such, there is scope for investigations and tools that investigate the burgeoning bulk of genome data for virus and antiviral gene sequence data in the search of paleovirological insight. I used DIGS, a similarity-searched based framework, in heuristic investigations of host:virus co-evolution in three contexts: ERVs, EVEs and interferon stimulated genes (ISGs). These three projects were unified by the approach and theme: the heuristic, similarity search based mining of genome sequences for insight into host:virus co-evolution. In the first project, I identified and characterised five novel murine ERV lineages, and used murine ERV sequences to provide insight into the recent acquisition and functionalisation of murine ERVs. In the second project, I performed a comparative analysis of animal endogenous circoviruses (CVe), increasing the known past host range of family *Circoviridae* and finding that genus *Cyclovirus* is more likely to be a clade of arthropod viruses than vertebrate viruses, in conflict with recent metagenomics analyses. In the third project, I screened mammalian genomes in an attempt to identify lineage specific patterns of ISG expansion and loss, finding three novel instances of ISG loss in cetaceans and felid carnivores. Together, these datasets will be of great utility in future studies of host:virus co-evolution, and underscore the utility of searching WGS data for paleovirological insight.

Table of contents

Abstract 1
Table of contents 2
List of figures 6
List of tables 8
Acknowledgements 9
Author's Declaration 10
Publications included in this thesis 11
Abbreviations 12
1 – Introduction 15
1.1 – Direct paleovirology – Endogenous retroviruses 16
1.1.1 - Retroviral genome structure 17
1.1.2 - Retroviral life cycle 18
1.1.3 - Retroviral gene products 19
1.1.4 – Accessory genes 22
1.2 – ERV structure, replication and classification 24
1.2.1 - ERV structure 24
1.2.2 - ERV amplification 24
1.2.3 - Retroviral taxonomy 26
1.2.4 - Class I retroviruses 28
1.2.5 - Class II retroviruses 29
1.2.6 - Class III retroviruses 30
1.3 – ERVs and the host 31
1.3.1 – Retroviruses and disease 31
1.3.2 – Restriction factors 32
1.3.3 - Exapted sequences 32
1.3.4 – Epigenetics 33
1.3.5 - LTR based gene regulation 34
1.4 – Murine ERVs 36
1.4.1 - Class I murine ERVs 37
1.4.2 - Class II murine ERVs 38
1.4.3 - Class III murine ERVs 39
1.4.4 - ERV-like transposons 39
1.5 – Direct paleovirology: EVEs 40

1.5.1 – Circoviruses 41
1.5.2 - Circovirus classification 42
1.5.3 - Circovirus diversity, host range, and role in disease 43
1.6 - Indirect paleovirology – antiviral genes 45
1.6.1 - Interferon stimulated genes (ISGs). 45
1.6.2 – Virus:ISG coevolution 45
1.6.3 – Antiviral gene dynamism 46
1.7 - Using paleovirology to study host:virus co-evolution 47
1.7.1 - Phylogenetic analysis 48
1.7.2 - Sequence analysis 50
1.7.3 – The timeline of host:virus co-evolution 51
1.7.4 – Functional genomics analyses 52
1.7.5 – Study of gene evolution 53
1.8 – Mining genome data for paleovirological insight 53
1.8.1 – Similarity searches 53
1.8.2 - Introduction to DIGS 56
1.8.3 - Why use DIGS? 57
1.9 - Research scope and aims 58
2 – Materials and Methods 59
2.1 - Materials 60
2.1.1 – Sequence data 59
2.1.2 – Software and tools 62
2.2 – Methods 65
2.2.1 – DIGS 65
2.2.2 - GLUE 65
2.2.3 - Host evolutionary timelines 70
2.2.4 – Recovery of proviruses 70
2.2.5 – Provirus screening and consolidation 70
2.2.6 – Sequence analysis 71
2.2.7 – Pairwise LTR dating 71
2.2.8 – Enrichment of genomic features at ERV loci 72
2.2.9 – Chromosomal distribution of ERVs 73
2.2.10 – Orthologous ERV and EVE integrations 73
2.2.11 – CVe sequence analyses 73
2.2.12 – PCR of CVe-Pseudomyrmex –Peter Flynn 73
2.2.13 – ISG screen design 74

2.2.14 – ISG data processing and display 74
2.2.15 – Shared genomic organisation 75
3 – Discovery and characterisation of murine endogenous retroviruses 76
3.1 – Introduction 76
3.2 – Results 79
3.2.1 - Assessment of murine ERV diversity using phylogenetic screening 79
3.2.2 - Identification of additional murine ER`V lineages 82
3.2.3 - Evolutionary history of Gag and Env 85
3.2.4 - A survey of murine ERV counts 87
3.2.5 - Analysis of novel ERVs 89
3.2.6 - Evolutionary analysis of murine ERVs 93
3.2.7 - Functional analysis of murine ERVs 102
3.2.9 – Dating ERV integrations 106
3.3 – Discussion 109
3.3.1 – Generation of a murine ERV dataset 109
3.3.2 – A transition in murine:ERV co-evolutionary history 109
3.3.3 – Identification of ERVs potentially involved in physiological processes 111
3.3.4 – Conclusions 113
4 - Paleovirological analyses of endogenous circoviruses 114
4.1 – Introduction 115
4.2 - Results 118
4.2.1 - Identification and characterisation of animal CVe 118
4.3.2 – Identification of orthologs and construction of a timeline of CVe evolution 124
4.3.3 - Phylogenetic analysis of Circoviridae Rep 127
4.3.4 - Phylogenetic analysis of Circoviridae Cap 131
4.3.5 - Species breakdown 133
4.3.5.1 - CVe in jawless fish 133
4.3.5.2 - CVe in Actinoperygii 134
4.3.5.3 - CVe in amphibians 135
4.3.5.4 - CVe in reptiles 136
4.3.5.5 - CVe in birds 136
4.3.5.6 - CVe in mammals 137
4.3.5.6 - CVe in mites 140

4.3.5.7 - CVe in ants 140
4.3.5.8 - CVe in other invertebrates 141
4.4 – Discussion 143
4.4.1 – Assessment of the methods 143
4.4.2 - CVe provide information about circovirus evolution 143
4.4.3 - Mapping the host range of circoviruses 144
4.4.4 - Impact of CVe on host genome evolution 147
4.4.5 – Conclusions 148
5 – Searching for ISG expansion and loss in mammals 150
5.1 – Introduction 150
5.2 – Results 152
5.2.1 - Initial screening and results 152
5.2.2 – IFIT 157
5.2.3 - XAF1 161
5.2.4 - SERTAD1 163
5.2.5 - Other potentially deleted genes 163
5.3 – Discussion 165
5.3.1 – Assessment of the methods 165
5.3.2 – Deletions in mammalian xaf1 and ifit 166
5.3.3 – Conclusions 168
6 - General discussion 169
6.1 – Assessment of the methods 169
6.2 - Future directions 171
6.3 – Conclusions 172
Bibliography 174
Appendices 200

List of figures

Figure 1.1 - Canonical retroviral genome structure 17
Figure 1.2 – Reverse transcription 21
Figure 1.3 - ERV structures 25
Figure 1.4 - ERV amplification strategies 26
Figure 1.5 - Unrooted pol NJ phylogeny of Retroviridae 27
Figure 1.6 - Distinguishing structural features of retrovirus groups 28
Figure 1.7 - KAP1 (TRIM28) repression of ERV expression 35
Figure 1.8 - Phylogeny of known murine ERV lineages 36
Figure 1.9 - Canonical circovirus genome structure. 43
Figure 1.10 - ERV amplification dynamics reflected in their phylogenies. 50
Figure 1.11 - ERV invasion leads to orthologous integrations 52
Figure 1.12 - Strategies for similarity searches of genomes 55
Figure 1.12 - Basic DIGS searching strategy 56
Figure 2.1 – Genome screening in DIGS. 68
Figure 2.2 - Phylogenetic screening with DIGS 68
Figure 2.3 - Entity-Relationship diagram of database generated by DIGS 69
Figure 3.1 – The mouse genome contains 13 distinct ERV lineages. 81
Figure 3.2 - Phylogenetic relationships between murine and nonmurine gammaretrovirus Pol sequences 84
Figure 3.3 - Phylogenetic relationships between murine and nonmurine Betaretroviruses 85
Figure 3.4 - Phylogenetic relationships between murine and nonmurine gammaretrovirus Gag sequences 86
Figure 3.5: Phylogenetic relationships between murine and nonmurine gammaretrovirus TM sequences 87
Figure 3.6 - Genome structures of novel murine ERV lineages 92
Figure 3.7 – ML phylogenies of murine ERVs 98
Figure 3.8 - Enrichment of ERVs relative to transcription start sites 103
Figure 3.9 - Enrichment of TF binding and H methylation at ERV loci 103
Figure 3.10 – Paired LTR estimation of murine ERV integration dates 107
Figure 3.12 – Orthologs of four murine ERV lineages 108

Figure 4.1 - Genome structures of 24 endogenous circovirus (CVe) elements identified in animal genomes. 124
Figure 4.2 - Orthologs of four CVe 125
Figure 4.3 - Evolutionary relationships of vertebrate species in which CVe have been identified, and timeline of CVe evolution. 126
Figure 4.4 - Phylogenetic relationships between Circoviridae and CRESS 129
Figure 4.5 - Maximum likelihood phylogeny of aligned amino acid Circoviridae Rep sequences. 130
Figure 4.6 - Maximum likelihood phylogeny of aligned amino acid Circoviridae Cap sequences 132
Figure 4.7 - Maximum likelihood phylogeny of CVe Rep recovered from the genome of the inshore hagfish. 133
Figure 4.8 - Phylogeny of aligned CVe Rep amino acid sequences recovered from fish and amphibian genomes. 135
Figure 4.9 - Phylogeny of CVe Rep amino acid sequences recovered from carnivore genome assemblies. 139
Figure 4.10 - PCR confirmation of CVe-Pseudomyrmex presence in three populations of Pseudomyrmex gracilis 142
Figure 5.1 – A landscape view of the heatmap showing normalised results of DIGS searches of ISGs in mammalian genomes 154
Figure 5.2 - UCSC genome browser view of the Bos taurus ifit2 and ifit3 locus
Figure 5.3 - ifit3 has been deleted in cetaceans. 158
Figure 5.4 - ifit2 has been pseudogenised in cetaceans. Alignment view of translated in-frame cetacean pseudo-ifit2 160
Figure 5.5 - Genome browser view of cetacean IFIT2 and IFIT3 loci 160
Figure 5.6 - Screenshot of the UCSC genome browser centred on the dog xaf1 containing region. 161
Figure 5.7 - Alignment view of dog XAF1 intron and exon (indicated in red line) sequences. 162
Figure 5.8 - Genome browser schematic of the carnivore XAF1 162

List of tables

Table 2.1 - Functional genomics datasets used in this thesis 60
Table 2.2 – Reference seqs used in DIGS screening for ERVs and EVEs 60
Table 3.1 - Murine ERV RT counts 80
Table 3.2 - Murine ERV proviruses and solo LTRs 89
Table 3.3 – Novel MLV proviruses 93
Table 3.4 – IAP proviruses are significantly enriched proximal to gene groups involved in reproduction 104
Table 3.5 – MusD proviruses are significantly enriched proximal to gene groups involved in reproduction 104
Table 3.6 – Mus.g2 solo LTRs enriched proximal to protein coding genes. 105
Table 3.7 – Orthologous murine ERV integrations 108
Table 4.1 - CVe and CVl detected in vertebrate genome assemblies. 120
Table 4.2 - NCBI reference seqs used in screening and initial alignment 122
Table 4.3 – Expression predictions and accessions for CVe 124
Table 5.1 - Reported core mammal ISG deletions in Figure 1 155
Table 5.2 - Results for BLASTn queries of cetacean genomes with queries of cow genomic regions overlapping IFIT3 159
Table 5.3 - Results for BLASTn queries of cetacean genomes with queries of the cow IFIT2 sequence 159
Table 5.4 - Results for BLASTn sequences of felid genomes with queries of the cat xaf1 containing region 159

Acknowledgements

My long-winded acknowledgements begin of course with my long-suffering supervisors: Dr Rob Gifford, thank you for giving me the opportunity to study paleovirology in your group. To Dr Sam Wilson, thank you for guiding me through a long, and at times turbulent PhD, and for helping me when I really needed it. To members of the Gifford and Wilson groups, past and present: my gratitude for your advice, tolerance, and lessons in science, Spanish, Chinese and Basque. I hope we will continue to work together in the future.

To my collaborators: Peter Flynn and Corrie Moreau at the Chicago Field Museum, and to William de Souza and Henan Zhu. Thank you for the fruitful collaborations, great discussions and support, and lasting friendships.

To my friends, I could not have done this (or indeed anything) without you. I would especially like to thank: Antoine for keeping it cool; Catrina and Dave for being my lifeline for a year and a half; Meredith (!), Poppy, Arne, Felix and Ilaria, for providing a transnational scientific support group and drinking society; Meg, Barney, Ros and Sophie as always; my old friends James, Alistair, Olivia, Charlotte and Elin; Lana, for being my collaborator in escapism, past and future, and everyone at The Kitchen Window, for helping me through the final few months of purgatory. The list is long, and the above and the rest are thanked for your counsel, humour, discussions scientific or otherwise, and support.

Thank you to my past teachers and lecturers: Jan Owen, Neil Pockett, Richard Raistrick, Roger Poland and Stuart Neil. All are unified by their patience and skill for transmitting their passion for their subject to a reluctant student.

Thank you to Ayley, my partner and best friend. Your confidence, humour and compassion shines through easy days and hard days. I truly couldn't have made it through without you. Thank you also to Shona for giving me a home, and the rest of the Benton / Urquhart clan for giving me a second family in Scotland.

Finally, to my sister and parents: Rosie, Ian and Rachael. You have my endless gratitude for your hundredfold reciprocated love and confusingly immutable belief in my success, despite evidence often existing to the contrary.

Author’s Declaration

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

Printed Name: Tristan Dennis

Signature:

Publications included in this thesis:

Dennis TPW, de Souza WM, Marsile-Medun S, Singer JB, Wilson SJ, Gifford RJ. The evolution, distribution and diversity of endogenous circoviral elements in vertebrate genomes. *Virus Res*. 2018;(March). doi:10.1016/j.virusres.2018.03.014.

Dennis, TPW, Flynn, PJ, de Souza, WM, Singer JB, Moreau CS, Wilson, SJ, Gifford RJ. Insights into circovirus host range from the genomic fossil record. *Journal of Virology*. 2018;(June). doi:10.1128/JVI.00145-18.

Shaw, AE, Hughes J, Gu Q., Behdenna A, Singer JB, Dennis TPW, Orton RJ, Varela M, Gifford RJ, Wilson SJ, Palmarini, M. Fundamental properties of the mammalian innate immune system revealed by multispecies comparison of type I interferon responses. *PLoS Biol*. 2017;15(12):1-23. doi:10.1371/journal.pbio.2004086.

Zhu H, Dennis TPW, Hughes J, Gifford RJ. Database-integrated genome screening (DIGS): exploring genomes heuristically using sequence similarity search tools and a relational database. *BioRXIV*. 2018. doi:http://dx.doi.org/10.1101/246835.

Publications contributed to, but not in this thesis:

De Souza WM, Dennis TPW, Fumagalli MJ, et al. Novel parvoviruses from wild and domestic animals in Brazil provide new insights into parvovirus distribution and diversity. *Viruses*. 2018;10(4). doi:10.3390/v10040143.

Abbreviations

BLAST	Basic local alignment search tool
CA	Capsid
CERV	Chimpanzee endogenous retrovirus
ChIP	Chromatin immunoprecipitation
Cp	Capsid
CRESS-DNA	Circular, rep encoding ssDNA viruses
CSF	Cerebrospinal fluid
CVe	Endogenous circoviral element
CVl	Circovirus like
CyCV-VN	cyclovirus-Viet Nam
DIGS	Database integrated genome screening
dsDNA	Double-stranded DNA
DU	dUTPase
ECDC	European Centre for Disease Control
Env	Envelope
ERV	Endogenous retrovirus
ESC	Embryonic stem cell
ESCRT	Endosomal sorting complexes required for transport
ETn	Early transposon
EVE	Endogenous viral element (non-retroviral)
Gag	Group specific antigen
GaLV	Gibbon ape leukemia virus
GEO	Gene expression omnibus
GLN	Mus musculus retrovirus utilising tRNAGLN
GLUE	Genes Linked by Underlying Evolution
GREAT	Genome regions enriched for annotations
gRNA	Genomic RNA
HMM	Hidden Markov model
IAP	Intracisternal a-type particle
IFN	Interferon
IN	Integrase
IR	Intergenic region
ISD	Immunosuppressive domain

ISG	Interferon stimulated gene
Kb	Kilobase
KoRV	Koala retrovirus
KRAB-ZFP	Krüeppel-activated-box zinc finger protein
LINE1	Long interspersed nuclear element
lncRNA	Long non-coding RNA
LTR	Long terminal repeat
MA	Matrix
MaLR	Mammalian apparent retrotransposon
Mb	Megabase
MdEV	Mus dunni endogenous retrovirus
mESC	Murine embryonic stem cell
MMERV	Mus musculus endogenous retrovirus
MMTV	Mouse mammary tumour virus
MPMV	Mason-Pfizer monkey virus
mpmv	Modified polytropic
mRNA	Messenger RNA
MSA	Multiple sequence alignment
MuERV-C	Murine endogenous retrovirus C
MuERV-L	Murine endogenous retrovirus utilising tRNAlys
MuRRS	Murine retrovirus like sequence
MURVY	Murine retrovirus repeated on the Y chromosome
MURVY	Murine endogenous retrovirus present on the Y chromosome
MUSCLE	MUltiple Sequence Comparison by Log-Expectation
MusD	Mus musculus type-D retrovirus
Mya	Millions of years ago
NC	Nucleocapsid
NJ	Neighbour joining
ORF	Open reading frame
OriR	Origin of replication
PAMPs	Pathogen associated molecular patterns
PAUP	Phylogenetic analysis using parsimony
PBS	Primer binding site
PERV	Porcine endogenous retrovirus
PIC	Pre-integration complex
pmv	Polytropic

Pol	Polymerase
PPT	Polypurine tract
PR	Protease
PSI-BLAST	Position specific iterated BLAST
PSSM	Position specific scoring matric
PWMS	Postweaning multisystemic wasting syndrome
RCR	Rolling circle replication
RDBMS	Relational database management system
Rep	Replication associated protein
RT	Reverse transcriptase
Sag	Superantigen
SQL	Structured query language
ssDNA	Single stranded DNA
SU	Subunit
TE	Transposable elements
TF	Transcription factor
TM	Transmembrane
TRIM28	Tripartite-motif-containing protein 28
TRIM28	Tripartite-motif-containing protein 28
TSD	Target site duplications
TSD	Target site duplication
VL30	Virus like retrotransposon 30
WGS	Whole genome sequence
WGS	Whole genome sequence
xmv	Xenotropic
XRV	Exogenous retrovirus

Viruses are the most abundant biological entities on earth (**Edwards and Rohwer, 2005**), but despite their ubiquity, it is difficult to study co-evolution of viruses and their hosts over deep timescales, as they leave negligible physical remnants of their existence behind. In addition, viruses evolve extremely rapidly, so it is forgivable to propose that they are only a few thousand years old using methods that rely on the study of sequence evolution. (**Holmes, 2011**). Until the discovery of the integrated provirus (**Temin, 1964**), and the subsequent identification of endogenous retroviruses (**Weiss and Payne, 1971**), the study of viruses over evolutionary timescales was impossible. Endogenisation protects the virus, as the endogenous viral element (EVE) will (mostly) evolve at the neutral substitution rate of the host, preserving the integrated viral material from mutational and physical degradation. This process of endogenisation, although rare, has taken place many times over evolutionary timescales, resulting in widespread integration of virus sequences in organism germlines (Aiewsakun **and Katzourakis, 2015**). The study of EVEs, and the signatures of host co-evolution, is therefore currently the only way to study virus evolution in deep time.

The study of viruses over deep timescales is known as 'paleovirology'. (**Patel *et al*, 2011**). The field of paleovirology is split into two basic methodologies: direct and indirect (**Katzourakis, 2013; Patel *et al*, 2011**). Direct paleovirology largely relies on the *in silico* identification and study of integrated virus sequences in host genomes. However, germline incorporation of viral sequences is rarely followed by fixation of those sequences. Because of this, the study of antiviral gene evolution is extremely informative where virus fossils are absent. This is known as 'indirect paleovirology'. Indirect paleovirology entails the study of antiviral gene evolution, and is particularly valuable when interaction between a gene product and a virus is known (**Sawyer *et al*, 2005; Busnadiego *et al*, 2014**).

Virus endogenisation is rare. Viral genetic material must first integrate into the germline, and genomic integration is not an obligate step in the life cycle of most viruses. Then, the germ cell containing the EVE must develop into a successful organism that is able to breed. Finally, the EVE must the

increase in frequency until it has reached 'fixation' – where it is present in every individual of a species. Given that most EVEs are likely to be purged from the germline within a few generations unless they confer a selective advantage, or are preserved by linkage or chance, the diversity of integrated EVE DNA in the genomic fossil record likely represents a fraction of past viral diversity. The study of EVEs is made even more disjointed by the eventual fate of almost all organisms: extinction. Because of this, it is difficult to make definitive inferences about the past diversity of viruses based on the presence or absence of EVEs in contemporary host genomes. (**Gifford, 2012**). Despite the rarity of virus endogenisation, germline invasion by viruses has taken place often enough for the virus fossil record to be a valuable resource for studying host:virus co-evolution.

As a consequence of their replication strategy, retroviruses are the most common form of EVE; integration is an obligate stage in the retroviral life cycle, and endogenous retroviruses (ERVs) can undergo amplification in their hosts (**Mager and Stoye, 2015**). Non-retroviral EVEs are therefore rarer. Nevertheless, the diversity of EVEs in (from divergent genomes) spans the known diversity of viruses (**Aiewsakun and Katzourakis, 2012**). Subsequently in this thesis, non-retroviral endogenous viral elements will be referred to as EVEs, whereas endogenous retroviruses (or retroviral elements) will be referred to as ERVs.

As genome sequencing technologies cheapen, the bulk of published genome data increases. ERVs, EVEs, and data on antiviral gene:virus co-evolution, are often not captured by automated annotation pipelines. This thesis conducts paleovirological analyses of host:virus co-evolution– through the indirect paleovirological study of antiviral interferon stimulated genes (ISGs) and direct paleovirological study of EVEs and ERVs.

1.1 – Direct paleovirology – Endogenous retroviruses

Retroviruses (family *Retroviridae*) are reverse transcribing RNA viruses that often cause immunosuppressive and/or neoplastic diseases. They are marked by their ability, as an obligate part of their life cycle, to reverse

transcribe their RNA genomes into double stranded DNA (dsDNA), followed by stable integration of this intermediate into the host genome.

1.1.1 - Retroviral genome structure

Retroviruses (family *Retroviridae*) possess ~10kb genomes that encode a variety of protein products. (**Figure 1.1a**). ERV genomes consist of integrated retroviral proviruses (or remnants thereof). As this corresponds to the reverse transcribed DNA intermediate of the retrovirus genome (as opposed to the viral gRNA), the structure and organisation of retroviruses will be discussed as such. The genome consists of coding gene sequences flanked by identical long terminal repeats (LTRs), and other auxiliary sequences that have other roles. The coding gene sequences typically consist of *gag, pro, pol* and *env* (**Coffin, 1997**). Although all retroviral genomes follow the canonical order of LTR-*gag*-*pol*-*env*-LTR, genomic organisation and translation/transcription strategies can vary, and are detailed in **Figure 1.1C.**

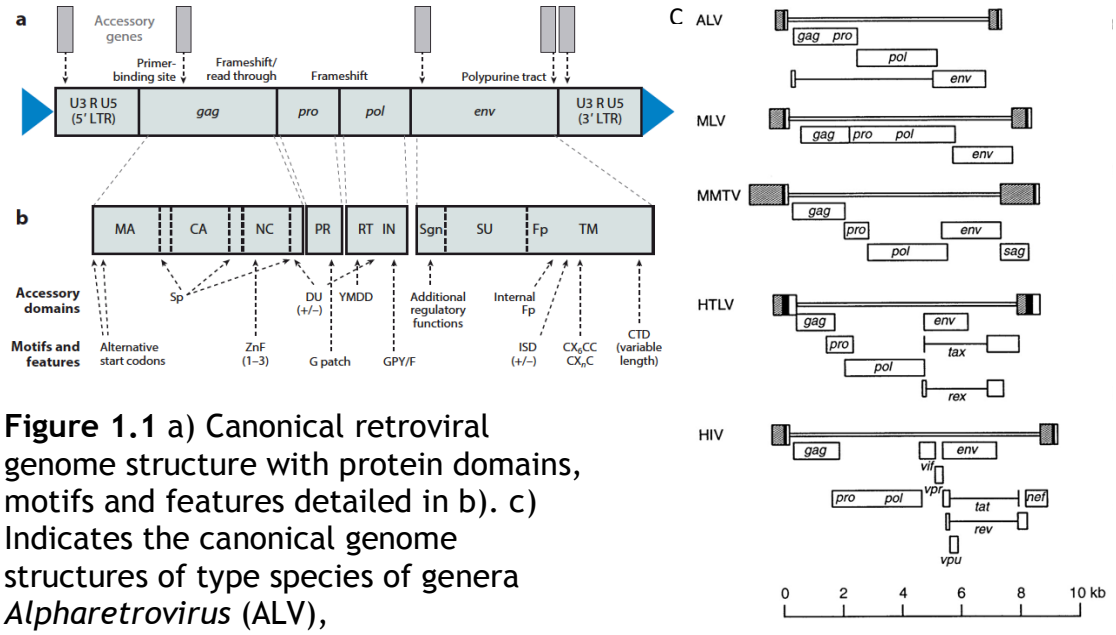

Figure 1.1 a) Canonical retroviral genome structure with protein domains, motifs and features detailed in b). c) Indicates the canonical genome structures of type species of genera *Alpharetrovirus* (ALV), *Gammaretrovirus* (MLV), *Betaretrovirus* (MMTV), *Deltaretrovirus* (HTLV) and *Lentivirus* (HIV). From **Johnson, 2015** (a, b) and **Coffin, 1997** (c).

1.1.2 - Retroviral life cycle

To enter the host cell, the retroviral Env proteins interact with their cognate receptors on the surface of the cell, resulting in fusion of the cell and virion membranes. The contents of the viral particle are released into

the cytoplasm of the host cell. The viral gRNA is reverse transcribed by reverse transcriptase (RT) into dsDNA copies, which are trafficked to the nucleus with viral integrase (IN). After entering the nucleus, the dsDNA intermediate is integrated into the chromosomal DNA of the host cell. It is transcribed by cellular DNA dependent RNA polymerase II into messenger RNAs (mRNA), which in some cases are spliced, then translated into the various viral protein products. Env glycoproteins are embedded into the host cell membrane, in preparation for the emerging viral particles, which assemble (driven by Gag), and bud off from the host cell (**Coffin, 1997**).

1.1.3 - Retroviral gene products

The Gag protein comprises the structural components of the retrovirus virion. Gag is expressed as a polyprotein that is proteolytically cleaved into constituent parts. These parts are the matrix (MA), capsid (CA) and nucleocapsid (NC), and are ordered MA-CA-NC in the uncleaved Gag polyprotein (**Figure 1.1b**). Retroviral Gag proteins are typically myristoylated at the amino terminus of the MA domain. In instances where this moiety is absent, retroviruses are unable to bud, and Gag precursors accumulate in the cell. CA acts as a mediator for protein interactions involved in virion assembly, and create the capsid shell that surrounds the viral core (**Coffin, 1997**). NC binds tightly to genomic RNA, forming ribonucleoprotein complexes in the virion core. This interaction is coordinated by one or two Cys-His zinc fingers (**Chance *et al*, 1992**). The Cys-His zinc finger has the consensus structure CX2CX4 HX4C (CCHC), where the residues designated by Xs are not conserved either among retroviruses or between the two zinc fingers of a single NC. Deletion or mutation of the CCHC results in either aberrant composition, or absence, of genomic RNA in virions (**Coffin, 1997**). NC also contains the late domain. The late domain is involved in virus budding late in the budding process (**Göttlinger *et al*, 1991; Wills *et al*, 1994**). Late domains can be encoded by one or more of the following motifs: PPPY, P(T/S)AP, or YPXnL. Respectively, these motifs interact with the following components of the cellular endosomal sorting complexes required for transport (ESCRT): Nedd4, TSG101, and ALIX (**Demirov and Freed 2004; Freed *et al*, 2002**)

In most retroviral genera, Gag-Pro-Pol is translated as a single polyprotein, which is then processed following attachment to the membrane and prior to budding (**Oroszlan and Luftig, 1990**). In the majority of cases, translation of the gag-pro-pol containing mRNA terminates at the end of the *gag* gene. This termination can be overcome by either ribosomal frameshifting prior to the *pro* or *pol* genes, or readthrough of the end-*gag* stop codon. (**Coffin, 1997**). This produces Gag-Pro or Gag-Pro-Pol fusion proteins respectively (**Figure 1.1b**). In both cases, autocatalysis of the fusion protein by Pro produces the mature, separate proteins encoded by *gag, pro* and *pol* (**Coffin *et al*, 1997; Dunn *et al*, 2002**). The *pro* and *pol* genes encode the enzymes responsible for virus replication. *pro* encodes an aspartyl protease that cleaves mature peptides from immature polyprotein precursors. *pol* encodes the reverse transcriptase (RT) and integrase (IN) proteins. (**Figure 1.1b**). Respectively, the RT and IN reverse transcribe the DNA provirus from the gRNA, and integrate it into the genome. (**Coffin, 1997; Jern *et al*, 2005**). RT possesses two catalytic domains: one reverse transcribes DNA from RNA, the second possesses RNaseH activity. In this process, the gRNA is reverse transcribed into proviral DNA using the host's cellular supply of nucleotides.

The exact locations of reverse transcription is so far not known. For *Spumaretroviruses,* reverse transcription proceeds in virions. (**Coffin, 1997**). The viral gRNA acts as a template for reverse transcription (Coffin, 1997). The viral enzymes RNase H and reverse transcriptase (RT) are essential for reverse transcription, the steps of which are detailed in **Figure 1.2**:

1. Minus DNA strand synthesis starts at the 3'-OH of a cellular tRNA acting as a primer that binds the PBS in the virus RNA.
2. The first strand of DNA is synthesized until the RT enzyme reaches the 5' end of the gRNA, generating a small ssDNA fragment of around 100-150 bases (termed 'strong-stop DNA').
3. Viral RNaseH degrades the gRNA template complementary to the strong stop DNA. The R region of the ssDNA anneals to the complementary R region at the 3' of the gRNA. In the first of two template exchanges, RT switches strand, attaching to the R region in the 3' end of the viral RNA.
4. Next, this DNA intermediate is extended towards the 5' end of the viral RNA genome (terminating at the PBS), and the RNA strand used as template is

now degraded by the RNAseH, except for the RNaseH resistant polypurine tract (PPT).

5. The RNA fragment containing the PPT acts as a primer for the generation of the +DNA strand synthesis, which is generated until it reaches the 3'-end of the -DNA strand.

6. Now RNAseH degrades all viral RNA remaining, exposing complementary sequences near the 3'-end of the +DNA strand

7. The +DNA strand is transferred to the 5'-end of the -DNA strand, annealing to the complementary PBS at the 5' end, facilitating the second template exchange reaction.

8. Finally, the synthesis of both DNA strands is completed, with each strand serving as template for the other strand. Once the dsDNA is synthesized, the provirus will be transferred into the nucleus to be integrated into the host genomic DNA. (**Coffin, 1997**).

The second protein encoded by *pol* is IN. IN integrates the reverse-transcribed DNA intermediate into the genome of the host, where it can be transcribed by the host's cellular machinery. After reverse transcription, IN binds both ends of the viral genomic dsDNA, forming a closed, circular molecular complex. This interaction is thought to be partially mediated and stabilised by the integrase zinc finger motif (**Khan *et al*, 1991; Zheng *et al*, 1996**). The resulting preintegration complex (PIC) is trafficked to the nucleus, where it is integrated into the host cell genome. The integration specificity differs amongst retroviruses – Lentiviruses appear to integrate into active transcription units, whereas gammaretroviruses integrate preferentially close to transcription start sites (**Schroder *et al*, 2002; Wu *et al*, 2003**). Additionally integration specificity is directed to some extent by IN (**Lewinski *et al*, 2006**), possibly through the IN GPY/F motif (**Malik and Eickbush, 1999**). The resulting integrated provirus consists of the full retroviral genome, flanked by short 4-6bp target site duplications (TSDs), of the structure 5'LTR-Gag-Pro-Pol-Env-3'LTR. This provirus is now replicated during host cell division, and (assuming no selection pressures), evolves at the host background rate of neutral substitution.

The *env* gene encodes the Env protein that is displayed on the surface of the virus membrane. Its key function is enabling entry of the virion into a host

cell by binding to the host cell receptor and mediating fusion of the virus and host cell membranes (**Coffin, 1997**). Env therefore enables virion infectivity and determines receptor tropism. The Env polyprotein is cleaved into two mature peptides: subunit (SU) and transmembrane (TM), by host cellular proteases. The mature Env complex is a multimer consisting of three SU and TM subunits, where SU mediates receptor binding and TM triggers virus:host cell membrane fusion. Of SU and TM, SU is typically highly divergent, and TM is relatively conserved (**Benit *et al*, 2001**). In gammaretroviruses, the TM subunit possesses a disulphide motif (CX_nCC) involved in SU:TM binding, and a highly conserved immunosuppressive domain (ISD). (**Benit *et al*, 2001**; **Coffin, 1997**).

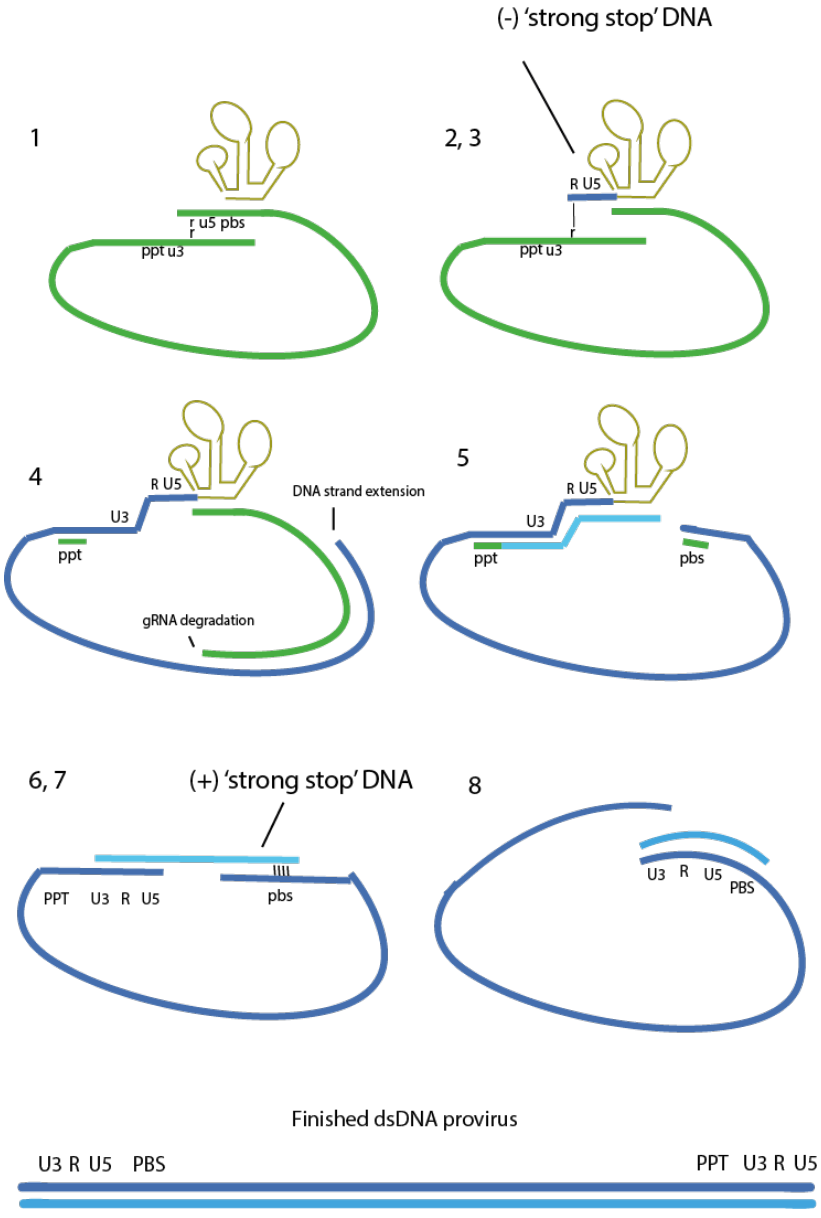

Figure 1.2 - Reverse transcription, according to the steps 1-8 detailed above. Figure adapted from Flint, 2007.

In the integrated provirus, retroviral genes are always flanked by identical LTRs. The LTR consists of repeated DNA sequences 150-1500bp in length. They contain regulatory elements for proviral integration and transcription as well as retroviral mRNA processing. LTRs are divided into three sections: U3, R and U5 (**Figure 1.1a**). In viral gRNA, the 5' end is structured R-U5 and the 3' end is structured U3-R, where R is a repeated element common to both ends, R-U5 is the 5' most sequence of the gRNA, and U3-R is the 3'-most sequence of the gRNA. The LTRs of the provirus are identical, with the structure U3-R-U5. Because the LTRs are synthesised *de novo* with each replication cycle, they are almost always identical. (**Coffin, 1997**).

1.1.4 – Accessory genes

In addition to the *gag, pol* and *env* genes common to all retroviruses, some retrovirus lineages encode additional accessory proteins that can act as adjuncts; increasing infectivity or affecting various aspects of virus replication. In this instance, we give three examples of accessory genes relevant to this thesis.

Most gammaretroviruses express an accessory protein called glycogag. Glycogag is encoded from unspliced RNA through an alternative start codon (CUG) that is present upstream of, and in frame with, the normal Gag start codon. This leader sequence is an addition to the normal Gag N-terminus. The resulting protein is integrated into the cell membrane, glycosylated, and cleaved so that only half of the conventional Gag sequence remains (**Fujisawa *et al*, 1997)**. Glycogag (and HIV-1 Nef) was recently shown to enhance gammaretroviral infectivity through interaction with SERINC3 and SERINC5 (**Rosa *et al*, 2015; Usami *et al*, 2015**).

The superantigen (*sag*) is an ORF encoded by type-B retroviruses, notably the murine mammary tumour virus (MMTV), which encodes *sag* adjacent to the 3' LTR (**Figure 1.1C)**. Sag is expressed on the cell surface, and interacts with specific Vβ chains of the T-cell receptor. This induces an activation signal from the T-cell to the infected cell. When Sag is expressed

from an endogenous MMTV provirus, it leads to the depletion of certain T-cell subgroups. (**Coffin, 1997**).

The *dut* gene is present in viruses of genus *Betaretotrovirus* and some *Lentiviruses*. Where present, it is located either in between Gag and Pol, or in between RT and IN (**Figure 1.1b**). It encodes a dUTPase that catalyses the hydrolysis of dUTP into dUMP and pyrophosphate. This effect can benefit viral replication through reducing the incorporation of genotoxic uracil into the viral cDNA during reverse transcription. This effect is mediated through the reduction of cellular dUTP concentrations. (**Lerner *et al*, 1995**).

1.2 – ERV structure, replication and classification

1.2.1 - ERV structure

Structurally, ERVs take four different forms in the genome. (**Figure 1.2**). (**Mager and Stoye, 2015**). 1: Complete provirus. These consist of paired LTRs of ~150-1000bp flanking identifiable *gag, pol,* and *env* ORFs. These proviruses can range from replication competent, intact integrations, to heavily degraded sequences marred by stop codons and frameshifts. Examples of these are MLV in mice, and ERV-Fc in mammals, respectively (**Jern *et al*, 2007; Diehl *et al*, 2016).** 2: Altered or 'slimmed down' ERV. Altered ERVs possess the general canonical structure of a complete ERV, but are lacking one or more ORFs (typically the *env* ORF). Examples of altered ERVs include the MusD and MuERV-L retrotransposons found in mice (**Mager and Freeman, 2000; Benit *et al*, 1997**). 3: Substituted ERVs. In some cases, recombination can produce ERV-like elements possessing LTRs and an internal region possessing no homology to known retroviruses. (**Mager and Stoye, 2015**). These elements can be replicated in a similar manner to retroviral vectors: for example, the early transposon (ETn) and virus-like-30 (VL30) transposons of mice share LTRs with, and are replicated with the assistance of, MusD and MMERV respectively (**Baust *et al*, 2003; Wolgamot *et al*, 1998**). 4: Solo LTR. Ultimately, the single most abundant ERV element in animal genomes is the solo LTR that arises by homologous recombination between the paired LTRs of complete proviruses (**Belshaw *et al*, 2007) (Figure 1.3**).

1.2.2 - ERV amplification

ERVs can amplify using three pathways: reinfection, retrotransposition, and complementation *in trans* (**Figure 1.4**). Additionally, by 'hitch-hiking' on another transposable or duplicating genomic element, ERVs and EVEs can increase their copy number, like the MURVY ERV amplified on the mouse Y chromosome (**Hutchison and Eicher, 1989**). ERVs are capable of reinfection and retrotransposition, dependent on their structure. Reinfecting ERVs require an envelope and a functional receptor (or *in trans* complementation from a helper virus) and form infectious particles that infect new cells within

the host (**Figure 1.4**). For example, functional envelopes have been found in the GLN and MLV lineages of murine ERV, and infectious particle formation has been observed (**Ribet *et al*, 2008; Kozak *et al*, 1984).** A mechanism by which the host can restrict this is through evolution of the cognate receptor, leading to xenotropism, whereby ERVs are unable to infect the cells of the species hosting it. (**Kozak, 2013**). Altered ERVs have no envelope, and have an entirely intracellular replication cycle, whereby the genome of the ERV is transcribed and reintegrated into the cell's genome. (**Figure 1.3**). Examples of these are MuERV-L, IAP and MusD: although intracellular virus-like particles have been observed for these lineages, they are not known to replicate through reinfection. (**Ribet *et al*, 2008; Ribet *et al*, 2008; Ristevski *et al*, 1999**). There is a striking correlation between the loss of the *env* gene, and intragenomic proliferation; not only are *env*-less ERVs undergoing massive amplifications in mammalian genomes but *env*-loss appears to be a necessary factor for such amplification, possibly due to the burden placed on the host by a circulating infectious retrovirus, and the immunosuppressive properties of *env* genes. (**Cantrell *et al*, 2005; Costas *et al*, 2003; Ribet *et al*, 2008; Magiorkinis *et al*, 2012**).In addition to *env* loss, the alteration of the Gag MA domain appears to be indicative of intracellularisation. In IAP sequences, polymorphisms in MA are associated with the targeting of nascent particles to intracellular compartments **(Dewannieux *et al*, 2004).** Similarly, addition of the deleted myristoylation signal in Gag retargets the MusD retrotransposon towards the cell membrane, suggesting that Gag may also play a role in the intracellularisation of replicating ERVs. (**Ribet *et al*, 2007**)

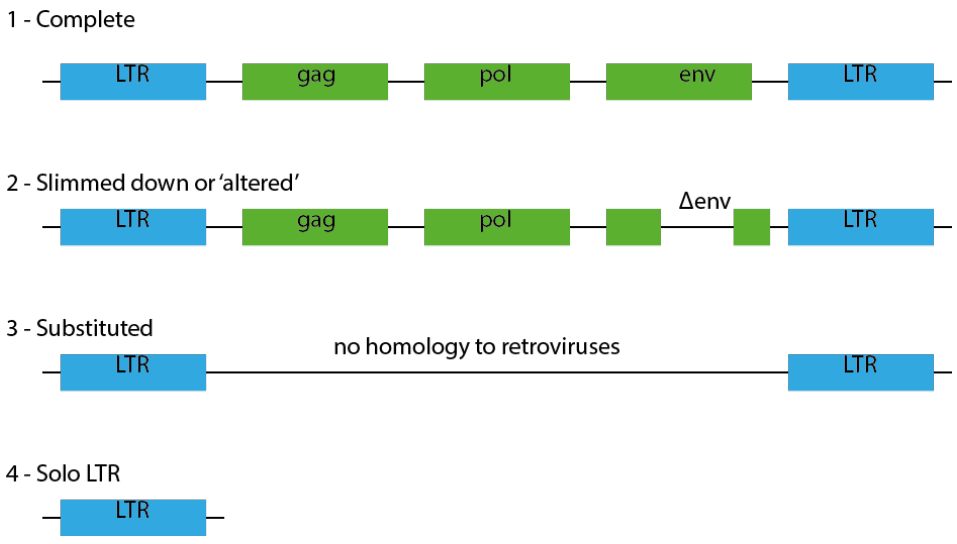

Figure 1.3: ERV structures. ORFs labelled. Figure adapted from **Mager and Stoye, 2015.**

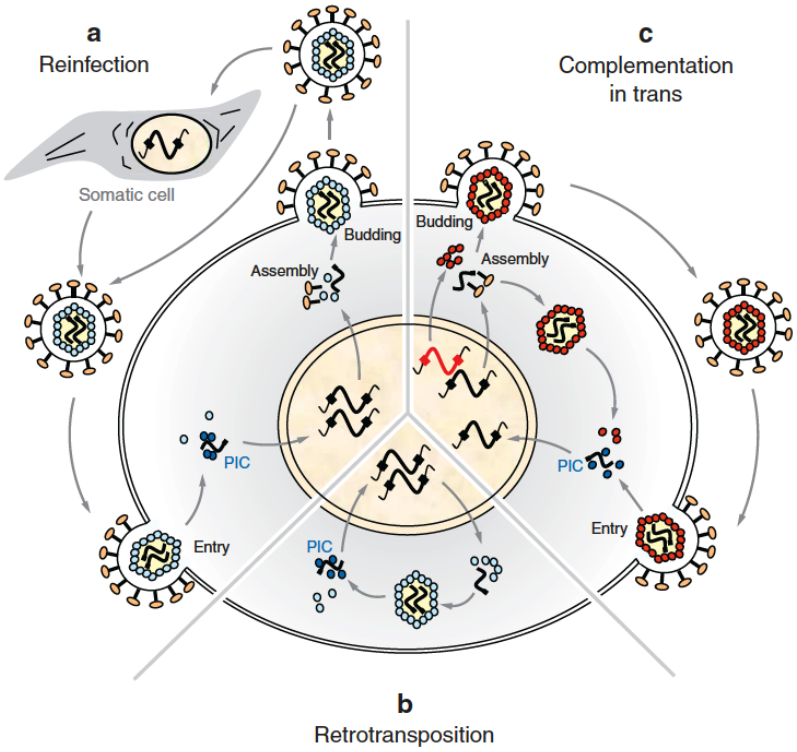

Figure 1.4: **ERV amplification strategies.** A) Reinfection by a complete retrovirus encoding intact genes and *cis* acting elements. B) Indicates retrotransposition of an intracellularised retrovirus and C) Indicates complementation *in trans* of a replication incompetent transposon. **Figure from Bannert and Kurth, 2006.**

1.2.3 - Retroviral taxonomy

ERVs are members of family *Retroviridae*. *Retroviridae*, as a family of reverse transcribing viruses, is a member family of order *Ortervirales*. *Ortervirales* contains all seven families of reverse transcribing viruses; united by the monophyly of the RT peptide. (**Krupovic *et al*, 2018**). Classification of retroviruses often relies on phylogenetic analysis of the *pol* sequence (**Figure 1.4**). This is due to the higher level of sequence conservation of *pol* relative to other retroviral genes. (**Xiong and Eickbush, 1990; Llorens *et al*, 2008).** The TM domain of the Env peptide is also used due to its relative conservation. (**Benit *et al*, 2001**). Structural features and sequence motifs of retroviral genomes (presence and type of Gag late domains, for example) are

also used to assist in classification of retroviruses (**Jern *et al*, 2005**). (**Figure 1.5).**

Retroviridae is split into two subfamilies: *Orthoretrovirinae* and *Spumavirinae,* which contain three robustly supported groups, called 'classes'. Class I and II belong to *Orthoretrovirinae* and Class III groups within *Spumavirinae*. **(Figure 1.5).** 'Class' is a misleading term – implying the taxonomic grouping that is several levels above the level of family. Nevertheless, the diversity of exogenous and endogenous retroviruses is clustered into these three groups in phylogenies. (**Herniou *et al*, 1998; Llorens *et al*, 2008).**

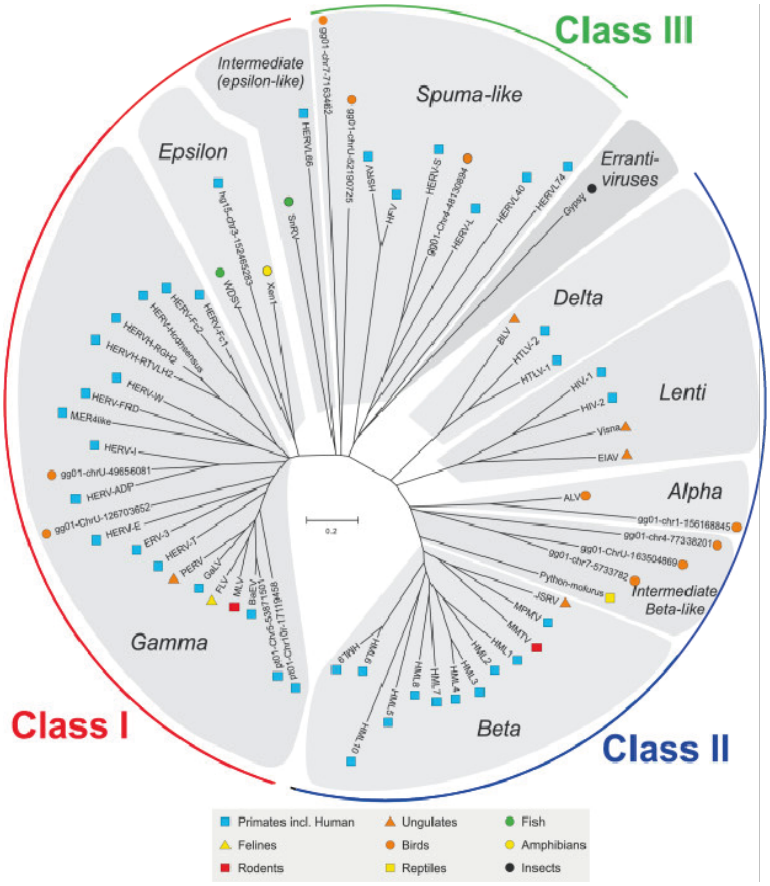

Figure 1.5) **Unrooted *pol* NJ phylogeny of *Retroviridae*** with class indicated, genus indicated by grey shaded group, host range indicated by coloured symbol. Note that *Gammaretrovirus* is used to describe all of the class I ERVs in the phylogeny. Figure adapted from **Jern *et al*, 2005.**

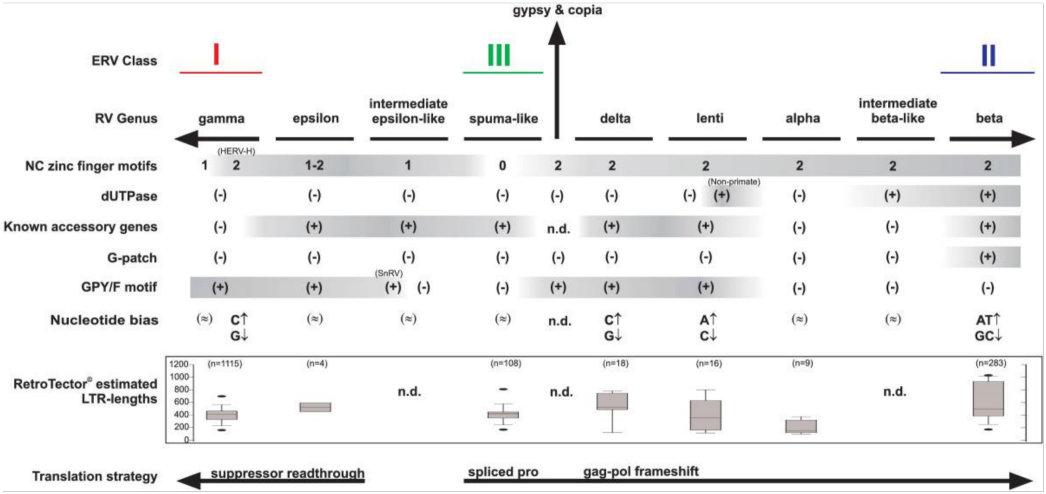

Figure 1.6: Distinguishing structural features of retrovirus groups (stated on the top of the table as RV-genus). Structural feature listed on the left side with LTR length and translation strategy detailed below. From **Jern *et al*, 2005**.

1.2.4 - Class I retroviruses

Retroviridae class I contains the *Gammaretrovirus* and *Epsilonretrovirus* genera. The class I retroviruses possess structural characteristics that distinguish them from other retroviruses. Class I retroviruses encode Gag-Pol through translation readthrough (as opposed to frameshifting of class II retroviruses, and the alternative splicing of class III viruses), and encode a GPY/F motif in IN. Also distinguishing class I ERVs is the encoding of a single NC zinc finger (with the exception of HERV-H, which has two) (**Jern *et al*, 2005**). Studies of exogenous gammaretroviruses have found that they possess a charged assembly helix that assists in coordinating NC:nucleic acid interactions in lieu of a second NC zinc finger (**Cheslock *et al*, 2003**). (**Figure 1.4; Figure 1.5**)

The class I retroviruses are found as exogenous and endogenous retroviruses in a diverse range of animals. The *Epsilonretrovirus* genus contains retroviruses of fish (**Gifford *et al*, 2003**). Viruses of genus *Gammaretrovirus* are simple retroviruses with a type-C particle morphology. They are found in mammals and birds, and are particularly well-represented in rodents, with the murine leukemia virus being a well-studied rodent exogenous gammaretrovirus (**Hayward *et al*, 2013**; **Kozak, 2015**).

Gammaretroviruses are common as mammalian ERVs – well known gammaretroviruses include the murine leukemia virus (MLV), RD114 and porcine endogenous gammaretrovirus (PERV). (**Mager and Stoye, 2015**). Notably, the mouse possesses numerous endogenous gammaretroviruses: MdEV, McERV, MuRRS, MMERV, GLN, MURVY and MuERV-C (discussed in further detail later in this introduction) (**Johnson, 2015; Mager and Stoye, 2015; Stocking and Kozak, 2008).** By contrast, the diversity of endogenous gammaretroviruses in the human genome is comparatively sparse, consisting solely of HERV-T.

Gammaretroviruses tend to occur in derived positions in phylogenies based on ERV polymerase sequences. They grade into groups of ERVs with a widespread distribution amongst vertebrates (**Herniou *et al*, 1999; Martin *et al*, 1999)**. These include viruses related to ERV-Fc, HERV-H, HERV-W, ERV-9, ERV-Rb, ERV-I and ERV-Lb (**Tristem, 2000; Herniou *et al*, 1999**). Although these viruses are often termed Gammaretroviruses (**Hayward *et al*, 2015; Jern *et al*, 2005**), in phylogenies, they group outwith the established exogenous diversity of genus *Gammaretrovirus* (**Figure 1.4**). (**Tristem *et al*, 2000; Benit *et al*, 2001; Herniou *et al*, 1999).** They appear to largely have degraded ORFs, precluding them from activation as retrotransposons or exogenous retroviruses (**Widegren *et al*, 1996; Diehl *et al*, 2016; Tristem, 2000**). Based on their more basal position within the class I ERV phylogeny, their demonstrable age, their widespread nature within animals, and the lack of evidence for exogenous viruses within these lineages, it is possible that these viruses represent a more ancient complement of retrovirus genera, of which the genus *Gammaretrovirus* comprises a more modern relative. (**Figure 1.4; Figure 1.5).**

1.2.5 - Class II retroviruses

Retroviridae class II contains the *Beta-, Alpha-, Delta-* and *Lentivirus* genera. (**Figure 1.4**). In addition to displaying a B-type (round eccentric core) or d-type (rough cylindrical core) particle morphologies, they display multiple peptide characteristics that differentiate them from other retrovirus groups. All class II retroviruses possess two NC zinc fingers, and produce a Gag-Pol polyprotein through one or two ribosomal frameshifting sites, as opposed to

the readthrough mechanism. **(Coffin, 1997) (Figure 1.5**). In addition, non-primate lentiviruses and betaratroviruses encode a dUTPase within *pol*. Finally, Betaretroviruses can be distinguished from other retroviruses based on the presence of a glycine rich 'G-patch' in the C-terminus of Pro. (**Jern *et al*, 2005).** (**Figure 1.4; Figure 1.5)**

With the exception of pyERV, identified in boid snakes (**Huder *et al*, 2002**), class II retroviruses are exclusively found in mammals and birds. The *alpharetrovirus* genus consists of exogenous and endogenous retroviruses of birds, including avian leucosis virus: the first ERV to be characterised (**Weiss and Payne, 1971; Gifford *et al*, 2005**). By contrast, genera *Betaretrovirus, Deltaretrovirus* and *Lentivirus* are found exclusively as exogenous and endogenous viruses of mammals (**Gifford *et al* 2005; Katzourakis *et al*, 2007; Hron *et al*, 2018**). (**Figure 1.4; Figure 1.5)**

Genus *Betaretrovirus* consists of the type-B and type-D retroviruses, of which the mouse mammary tumour virus (MMTV) and Mason-Pfizer monkey virus (MPMV) are type species respectively. **The type-B and type-D retroviruses are differentiated by their particle morphology and envelope taxonomy.** Betaretroviruses are found as exogenous and endogenous viruses in mammals. (**Gifford *et al*, 2003).** Notably, this genus contains MMTV – a well-studied murine exogenous and endogenous retrovirus. In addition, several intracellularised retrotransposons of murid rodents: MusD, IAP and mysTR (**Mager and Freeman, 2000**; **Ribet *et al*, 2008**; **Cantrell *et al*, 2005),** display close phylogenetic relationships to the *Betaretrovirus* genus. (**Figure 1.4**).

1.2.6 - Class III retroviruses

Class III retroviruses bear close phylogenetic relationship to genus *Spumaretrovirus* (foamy viruses) (**Figure 1.4**). In addition to the foamy viruses, class III contains the HERV-U, HERV-S and HERV-L lineages of ERV (**Cordonnier *et al*, 1995; Tristem, 2000**). ERVs of these lineages are common to many mammalian groups (**Herniou *et al*, 1999**) and are likely extremely ancient: in the case of ERV-L, predating the divergence of placental mammals (**Lee *et al*, 2014**). Despite the advanced age of ERV-L, it

has undergone the transition to a retrotransposon in many mammalian lineages (**Magiorkinis *et al*, 2012**). The ERV-L lineage occurs most notably in the mouse as MuERV-L, where its evolution and impact upon the murine host has been studied extensively (**Costas *et al*, 2003; Benit *et al*, 1997**). (**Figure 1.4; Figure 1.5)**

1.3 – ERVs and the host

1.3.1 – Retroviruses and disease

Retroviruses have been extensively studied as the causes of immunosuppressive disease. The best characterised of this is human immunodeficiency virus 1 (HIV-1), a lentivirus that has led to pandemic acquired immunodeficiency syndrome (AIDS). Other lentiviruses; feline immunodeficiency virus (FIV), equine infectious anemia virus (EIAV) and simian immunodeficiency virus (SIV) are studied as causes of immunodeficiency in their hosts.

Retroviruses can also be carcinogenic. Indeed, MMTV, MLV, rous sarcoma virus (RSV) and avian leucosis virus (ALV) were all known as the etiologic agents behind neoplastic disease before the elucidation of the integrated provirus (**Green, 1946; Gross, 1953; Rous, 1911; Temin, 1976**). Retroviruses mediate their oncogenic effects through multiple main methods. The first is ectopic activation of a host proto-oncogene through nearby integration, such as MMTV-based activation of *wnt* or *fgf* (**Ross, 2010**), or through IAP/MLV recombination and insertional mutagenesis in AKR mice (**Stoye *et al*, 1991**). The second is through the introduction of a viral oncogene, such as RSV induction of fibrosarcoma in chickens through transduction of the viral sarcoma (*v-src*) gene. (**Vogt, 2012**). Thirdly, retroviruses can mediate mutagenic effects through interruption of normal genomic activity through loss of gene function (integration in or near a gene). (**Coffin, 1997**).

ERVs have also been studied as etiologic agents of disease. In addition to the well-studied effects of MLV, MMTV and RSV, amongst others, as

inherited (ERV) causes of cancer in mice and chickens, ERVs are responsible for a high degree of genomic mutagenesis. In mice, it is estimated that 10% of genomic mutations with an observable phenotype are as a result of ERV activity (**Maksakova *et al*, 2006; Waterston *et al*, 2002**). Furthermore ERVs have been implicated in a diverse range of human disease processes, including motor neurone disease (**Li *et al*, 2015)** and Hodgkins lymphoma (**Lamprecht *et al*, 2010**). ERV pathogenicity has also been studied outwith humans, mice and birds, with the endogenising koala retrovirus (KoRV) being associated with lymphoma (**Tarlinton *et al*, 2006**).

1.3.2 – Restriction factors

A number of restriction factors antagonising ERV replication have been identified, primarily through their anti-HIV activity. **(Malim and Bieniasz, 2012)**. An example of an anti-ERV restriction factor is APOBEC3G. APOBEC3G is known to inhibit retrovirus replication through cytidine deaminase activity (causing mutational catastrophe), and inhibition of reverse transcription (**Malim and Bieniasz, 2012)**. APOBEC has also been shown to inhibit endogenous retroviruses prior to integration: HERV-K, MLV and MuERV-L have shown evidence of APOBEC3G mediated hypermutation (**Lee and Bieniasz, 2007; Perez-Caballero *et al*, 2006; Jern *et al*, 2007; Blanco-Melo *et al*, 2018**). Additionally, the murine restriction factors Fv1 and MOV10 have been identified as factors restricting ERV replication (**Nair and Rein, 2014**; **Frost *et al*, 2010**). Remarkably, these genes show homology to MuERV-L and MLV peptides respectively, indicating a retroviral origin (**Benit *et al*, 1997; Wang *et al*, 2010**).

1.3.3 - Exapted sequences

ERVs have been adopted to serve beneficial purposes. The host can adopt ERV coding sequences, or utilise the regulatory features of ERV sequences, such as the enhancer activity of LTRs. Virus proteins have been exapted to serve as restriction factors: In mice, Fv1 and Mov10 have been exapted from MuERV-L Gag and MLV (**Benit *et al*, 1997; Arjan-Ojedra *et al*, 2012**). The *syncytin* gene is another example of viral coding sequence exaptation. *Syncytin* is a captive retroviral envelope gene that is critical in

placental syncytium formation (**Mi *et al*, 2000**). The utility of *syncytin* is such that retroviral *env* adoption in the form of syncytins has taken place independently, multiple times over the course of mammalian evolution, leading to the hypothesis that this is an example of convergent evolution – the *syncytin* gene is required to form the mammalian placenta. (**Lavialle *et al*, 2013**). Additionally, diversity in captured *syncytin* genes is also associated with diversity of placenta structures **(Vernochet *et al*, 2014**)

1.3.4 - Epigenetics

In the mouse, it has been demonstrated that ERV transcription is suppressed via the remodelling of chromatin structure, presumably as a mechanism to repress ERV activation in unwanted contexts. These epigenetic mechanisms include histone methylation – as demonstrated for IAP, MusD and MLV (**Walsh *et al*, 1998; Rowe *et al*, 2013; Wolf and Goff, 2009**). This state can be established by histone and DNA methyltransferases which are recruited to ERV sequences in response to krüeppel-activated-box zinc finger protein (KRAB-ZFP) binding and subsequent TRIM28 recruitment. (Tripartite-motif-containing protein 28, also known as KAP1). (**Wolf and MacFarlan, 2015; Rowe *et al*, 2010; Rowe *et al*, 2013**). An extensive body of literature has grown around the ZFP mediated recruitment of TRIM28, and its utility in repressing ERVs, particularly during embryonic development. (**Rowe and Trono, 2011**) (**Figure 1.6**). In addition to driving retroviral repression, the KRAB-ZFP/TRIM28 pathway has been adapted to provide an ERV-based gene regulation network. **(Rowe *et al*, 2013**).

In murine embryonic stem cells (ESCs), deletion of TRIM28 resulted in de-repression of MuERV-L transcripts and associated genes, and also in an increase in two-cell-embryo like cells, indicating that murine ERVs play a role in regulating gene transcription and embryonic stem cell fate (**MacFarlan *et al*, 2012**). **Rowe *et al*, 2013** used ChIP-seq studies to highlight the use of ERVs in preventing untimely activation of genes involved in embryonic development, and highlight ERV loci that may be actively involved in these processes. The species-specificity of ERV regulation of gene repression is highlighted by these studies, and by **Brattås *et al*, 2016**, who found that

TRIM28 binds to primate-specific ERVs in human ESCs, resulting in transcriptional control during human neural development.

1.3.5 - LTR based gene regulation

ERVs can also drive gene expression through their LTRs. Retroviral LTRs possess promoter and enhancer sequences, and are present in abundance as solo LTR elements throughout mammalian genomes (**Belshaw *et al*, 2007**). Therefore, LTRs have abundant potential to regulate host gene expression. This is supported by evidence of non-coding transposon sequences evolving under purifying selection close to human genes (**Lowe *et al*, 2007**).

Studies of ERVs and gene regulation have implicated ERVs in gene regulation in many processes, including embryonic development, immunity, and placental physiology. ERVs were found to have contributed up to 25% of OCT4, NANOG and CTCF binding sites in human and murine embryos (**Kunarso *et al*, 2010**). **Schmid and Bucher (2010)** used chromatin immunoprecipitation with sequencing (ChIP-seq) to find inducible STAT1 binding sites in human and mouse ERV LTR sequences. This work was built upon by **Chuong *et al*, (2016),** who showed that a STAT1-bound MER41 LTR (the same type described by **Schmid and Bucher**) regulates the expression of *aim2, apol1, ifi6* and *sectm1,* as well as describing widespread STAT1 bound LTRs in mammalian genomes. In another ChIP-seq based study of species-specific ERV gene regulation, **Chuong *et al*, (2013**) studied the effect of rat and mouse ERVs on gene regulation in the placenta and found that species specific ERV families drive gene expression in placental cells. Finally, in a wide-ranging study, **Sundaram *et al*, (2014),** assayed 26 transcription factor (TF) networks in four human and mouse leukemia and lymphoblast cell lines. This ChIP-seq based study found cell-type and species specific contributions by ERVs to transcription factor binding networks. These studies highlight both the widespread nature and species specificity of ERV contribution of regulatory sequences.

Along with discoveries of co-opted ERV promoter and gene sequences, studies of ERVs in health and disease highlight the utility of studying ERVs in fully understanding host:virus co-evolution, and shows ERVs as dualistic entities: a) as parasites that must be kept in check to avoid disease and

harmful mutation to the host; and b) as symbionts that can be used by the host to their advantage.

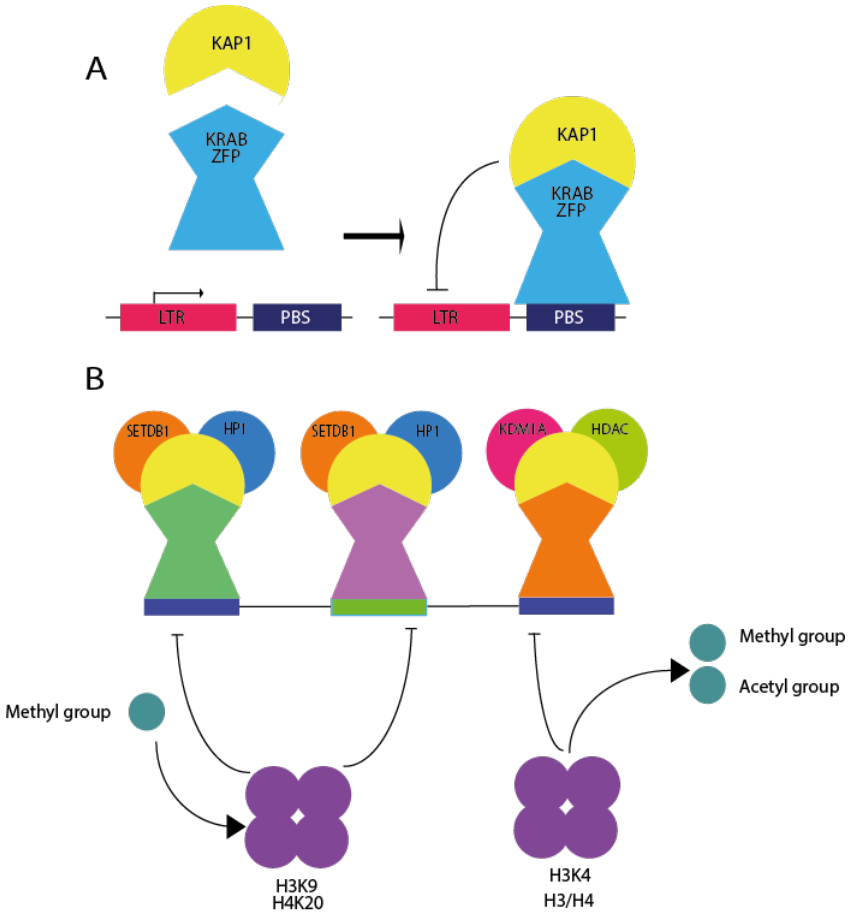

Figure 1.7: KAP1 (TRIM28) repression of ERV expression. A) KRAB-ZFP binds to an ERV PBS. TRIM28 is recruited causing the suppression of expression. B) Individual KRAB-ZFPs recognize different ERVs and recruit distinct KAP1 complexes (shown by different ZFP and sequence colours). SETDB1-containing complexes H3K9 and K4K20 methylation, and KDM1A/HDAC complexes catalyse H3 and H4 demethylation and deacetylation. Figure adapted from **Gifford *et al*, 2013.**

1.4 - Murine ERVs

A large proportion of insights into host:ERV co-evolution have been made in mice. The mouse is an extremely common model organism in medical genetics, and has been extensively characterised and studied. This extends to genomic data: the bulk of genome and annotation data available for the mouse is second only to that of the human. Also, the mouse is used and bred frequently as a model organism, due to a low generation time and small size, making it easy to validate and study ERV-related insights acquired from genome data. Mice possess ERVs of all three classes (**Figure 1.7**).

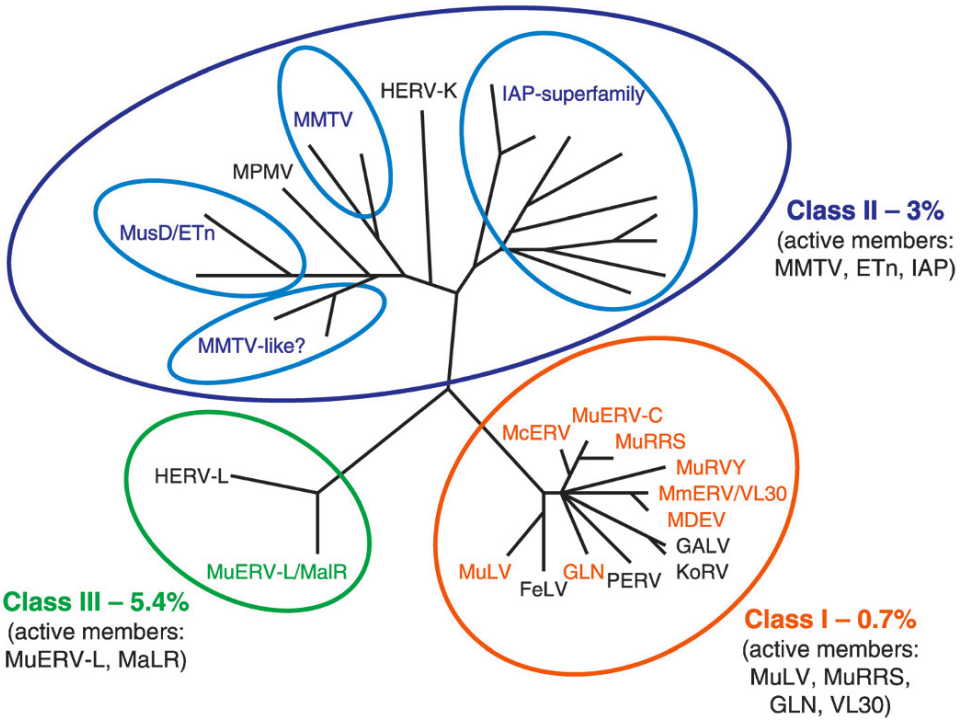

Figure 1.8) Phylogeny of known murine ERV lineages. Class indicated by coloured ring, the same coloured text within indicates murine ERV lineages. '%' indicates % of the murine genome taken up by ERV sequences of the given class. Figure from **Stocking, 2008.**

1.4.1 - Class I murine ERVs

Of the three ERV classes in mice, class I ERVs are the most diverse but lowest in number (**Figure 1.7**). All class I murine ERVs described so far are gammaretroviruses. MLV is the most extensively characterised class I murine ERV (**Gross, 1953**). A endogenous gammaretrovirus with an exogenous counterpart, endogenous MLV is capable of reinfecting the mouse genome, and is polymorphic between different strains of inbred mouse (**Frankel *et al*, 1990)**. MLV possesses a Pro PBS, with 49 loci present in the published genome of the C57BL/6 mouse. About 20 copies are shared between every mouse, (**Jern *et al*, 2007; Kozak, 2015**). It is split into three groups based on their ability to infect cells of the mouse and other species: xenotropic, polytropic and modified polytropic (*xmv, pmv* and *mpmv* respectively – not to be confused with the Mason-Pfizer Monkey Virus). (**Kozak, 2015**).

Mus musculus retrovirus utilising $tRNA^{GLN}$ (GLN) is present at about 80 copies in the murine genome (**Itin and Keshet, 1986; Ribet *et al*, 2008**). GLN proviruses are ~8.4kb long with a 430bp LTR. Two GLN proviruses possess all the requirements for infectious particle formation, which was observed in *in vitro* analyses of GLN (**Ribet *et al*, 2008**).

Loci in the *Mus musculus* endogenous retrovirus (MMERV) (**Bromham *et al*, 2001**) also possess intact envelopes, leading to speculation that it may be present as an endogenous gammaretrovirus of mice. *In silico* searches have identified 191 MMERV loci in the genome of the mouse (**Lee *et al*, 2012**), and this lineage has been shown to possess two PBS specificities (PBS^{pro} and PBS^{gly}). MMERV has been discovered in *Mus caroli* and *Mus dunni* also (named McERV and MdEV). MMERV proviruses are ~8.6kb in length, with ~450bp LTRs, and possess *gag, pol* and *env* ORFs, some of which are completely intact. Like MLV, MMERV has a ~99bp 5' extension of Gag, forming a Glyco-Gag ORF. (**Bromham *et al*, 2001**).

There are three non-intact lineages of class I murine ERV. The first is murine retrovirus like sequence (MuRRS). (**Schmidt *et al*, 1985**), with an estimated copy number of 50-100, 600bp LTRs and a 5.7kb long genome missing *env* and punctuated by stop codons. The second is murine endogenous

retrovirus C (MuERV-C), a similarly degraded gammaretrovirus (**Zhao *et al*, 1999)** with a LINE1 fragment in place of a 5'LTR and Gag c-terminus, a deleted *env* ORF and numerous stop codons and frameshifts. The final degraded class I murine ERV is murine retrovirus on the Y chromosome (MURVY). MURVY has an 8.8kb genome, and 627/628bp long 5'/3' LTRs respectively. Around 500 inactivated copies of the MURVY exist on the Y chromosome as part of a duplicated segment. (**Hutchison and Eicher, 1989; Eicher *et al*, 1989**).

1.4.2 - Class II murine ERVs

Class II ERVs group close to exogenous Betaretroviruses. Indeed, the *Betaretrovirus* type species is MMTV (**Green *et al*, 1946**). Three copies of MMTV have been found in the C57BL/6 mouse genome (**Kozak *et al*, 1987**), all of which are intact and capable of particle formation. (**Figure 1.7**).

The remaining class II murine ERVs are known to be retrotransposons, the most prolific of which is the intracisternal a-type particle (IAP) (**Dalton *et al*, 1961**). IAP has since been characterised as a diverse ERV 'superfamily' (**Qin *et al*, 2010; Magiorkinis *et al*, 2012).** Although an *env*-encoding IAP locus exists and, is capable of infectious particle formation (**Ribet *et al*, 2008**), IAP has undergone *env* loss and explosive proliferation in the genomes of several mammalian species, including that of the mouse (**Magiorkinis *et al*, 2012**). Although IAP displays a diverse range of structures and LTR sequences, they are unified by the presence of intact Gag and Pol encoding genes, flanking LTRs 300-600bp in length, and betaretrovirus-like sequence features (**described in greater detail above**).

The *Mus musculus D-type retrovirus* (MusD) is a retrotransposing class II ERV, discovered due to its similarity to ETn. (**Mager and Freeman, 2000**). MusD is ~6-7.8 kb in length, with some proviruses encoding intact Gag and Pol peptides, and capable of forming intracellular particles. However, in all cases, the Gag peptide is missing the n-terminal myristoylation signal, precluding infectious particle formation. Provision of this signal, and an *in trans* expressed *env* protein allows for rescued particle formation (**Mager and Freeman, 2000; Benit *et al*, 2007**).

The final class II murine ERV is MYSERV. A poorly defined group of murine ERVs, MYSERV is a RepBase designation for an inactive family of ERVs marked by their similarity to *mys* and *mys*TR, a group of ERV-like retrotransposons undergoing explosive amplification in other species of rodent, in a similar manner to IAP (**Wichman *et al*, 1985; Cantrell *et al*, 2005).**

1.4.3 - Class III murine ERVs

The only known class III murine (spumaretrovirus-like) ERV is MuERV-L (**Cordonnier *et al*, 1995**). Part of the ERV-L lineage, MuERV-L is marked by its advanced age, predating the divergence of placental mammals (**Lee *et al*, 2014**). In many lineages, including humans, ERV-L is inactive and degraded. Contrastingly in mice, MuERV-L is an active retroelement, having undergone multiple bursts of expansion in the last 2-10 million years (**Benit *et al*, 1997**; **Costas *et al*, 2003),** and has contributed to mice the antiretroviral *fv1* gene (**Benit *et al*, 1997**). MuERV-L is approximately 6 kb in length, with intact Gag-Pol ORFs and paired LTRs of ~500nt in length and a PBS complementary to $tRNA^{lys}$. (**Figure 1.7**).

1.4.4 - ERV-like transposons

What I describe here as 'ERV-like' transposons corresponds to the 'altered ERVs' described above. These elements possess flanking LTRs homologous to extant ERVs, but possess internal regions that display little or no homology to retroviral genes. In all three instances (ETn, VL30 and MaLR), the elements have LTRs homologous to known ERVs, suggesting that they replicate in a non-autonomous fashion, utilising the proteins of other ERVs to replicate.

Approximately 5kb in length, VL30 (virus-like 30) elements consist of paired LTRs flanking fragmented *gag* and *pol* ORFs. (**Keshet and Itin, 1982; Adams *et al*, 1988**). The *gag* and *pol* like ORFs bear only a faint homology in certain sections to other retroviral proteins. VL30 LTRs were found to be similar to MdEV, an endogenous gammaretrovirus of mice (**Wolgamot *et al*,**

1998): a prelude to the discovery of MMERV (**Bromham *et al*, 2001**). As such, there is speculation that recombination between VL30 and an ancestral gammaretrovirus was responsible for the generation of MdEV/MMERV (**Wolgamot *et al*, 1998**).

Early transposon (ETn) elements were characterised as retrovirus like elements active during murine embryogenesis (**Brulet *et al*, 1983**). The full-length ETn element is 5.6kb in length, is flanked by paired LTRs ~200bp in length. The internal region has no potential ORFs, nor any significant homology to retroviral proteins (**Sonigo *et al*, 1987**). The finding of a small region of homology in ETn to a fragment of betaretrovirus led to the characterisation of MusD – a transposing ERV with identical LTRs to ETn. (**Mager and Freeman, 2000**).

The mammalian apparent retrotransposon (MaLR) is the final non-autonomous ERV-like transposon. Present in the murine genome at ~388,000 copies, it is by far the most numerous retrotransposon. The internal region is unrelated to known ERVs, but it shares ~50% LTR homology with MuERV-L LTRs, suggesting that it replicates by utilising MuERV-L proteins. (**Waterston *et al*, 2002; Smit *et al*, 1993**).

1.5 – Direct paleovirology: EVEs

Compared to ERVs, non-retroviral EVEs (EVEs) are relatively rare. However, a surprising level of diversity of integrated EVEs has been discovered in eukaryotic genomes, starting with the 2004 discovery of endogenous flaviviral elements in the genome of the *Aedes* mosquito. (**Crochu *et al*, 2004**). Subsequently, it has been found that every known group of viruses is capable of endogenisation (**Katzourakis and Gifford, 2010; Aiewsakun and Katzourakis, 2012**).

The mechanism by which retroviruses endogenise is obvious – genomic integration of a dsDNA intermediate is an obligate step in the retroviral life cycle. Mechanisms of EVE integration are less obvious. It has been documented that DNA viruses can integrate into host genomes via non-

homologous dsDNA end-joining (e.g. for hepadnaviruses, **Bill and Summer, 2004**) and telomeric homologous recombination (e.g. for herpesviruses, **Morissette and Flamand, 2010**). For RNA viruses, the mechanisms of integration are even more mysterious, given that existence as a DNA intermediate, presence in the nucleus and chromosomal integration are not a necessary part of their life cycle. It has been proposed that RNA viruses are reverse transcribed and integrated by retroviral or retrotransposon proteins to aid endogenisation (**Katzourakis and Gifford, 2010**). Additionally, recombination between ERVs and exogenous RNA viruses can take place, leading to integration of the RNA virus as an EVE (**Geuking *et al*, 2009**).

For ssDNA viruses of eucaryotes, only the integration of parvoviruses through non-homologous DNA recombination has been studied, due to the potential of adeno-associated virus in gene therapy. (**Kotin *et al*, 1992; Krupovic and Forterre, 2015**). For circular ssDNA viruses, such a circoviruses, nanoviruses and geminiviruses, no integration mechanism has been found, despite evidence of widespread integration of these viruses into eucaryotic germlines (**Katzourakis and Gifford, 2010**). It has been proposed that integration took place during repair of chromosomal DNA strand breaks via a nonhomologous end joining mechanism. Additionally, the rolling-circle method of replication employed by these viruses may be significant, given that rolling circle rep-like endonucleases mediate integration of circular ssDNA molecules in prokaryotic genomes (**Krupovic and Forterre, 2015**).

1.5.1 - Circoviruses

The increasing body of sequence and metagenomics data is bringing attention to the circular ssDNA viruses, particularly circoviruses (family Circoviridae). Circoviruses are small, with 1.8-2.1kb ssDNA genomes. (**Figure 1.8**). All circoviruses encode two genes: *cap* and *rep,* with a third ORF (*orf3*) being encoded by porcine circoviruses (PCVs), and other ORFs being identified in other circovirus species (**Rosario *et al*, 2017**). The *rep* and *cap* genes diverge from one of two intergenic regions (IR) in an ambisense fashion (**Figure 1.8**), and are transcribed as such. In the case of PCVs, alternative splicing takes place, leading to the production of numerous transcripts. However, this has not been studied in other circoviruses (**Cheung *et al*,**

2012). *cap* encodes the Cp protein: the viral structural protein. Cp is significantly divergent, with the only distinguishing motif being an N terminal domain rich in basic amino acids. This is hypothesised to have DNA binding activity, and may be involved in the intracellular distribution of viral proteins during replication (**Cheung and Greenlee, 2011**). *rep* encodes the replication associated protein (Rep), also known as the replicase (referred to as such hereafter). Rep is the most conserved circovirus protein, and possesses features associated with rolling circle replication (RCR) in other ssDNA viruses. The IR between the 5' ends of the genes is characterised by a conserved nonanucleotide motif, where Rep initiates RCR (**Figure 1.9**).

The circovirus life cycle begins with entrance via clathrin-mediated endocytosis (**Cheung *et al*, 2012**). The viral particle then penetrates the nucleus, where it uncoats. In the nucleus, the ssDNA genome is converted to dsDNA by host factors, and transcribed to produce viral mRNAs, which are translated into viral proteins in turn. RCR of the viral genome is initiated via a stem loop structure in the intergenic region which is marked by a conserved nonanucleotide motif. The viral Rep protein 'nicks' the genome at this location, and the exposed ssDNA is bound by host DNA polymerase. The host DNA polymerase synthesises a new copy of the viral ssDNA genome, and the old strand is displaced. The resulting dsDNA (consisting of an original DNA strand and the newly synthesised complementary strand) is again nicked by a viral Rep protein, and the process begins anew, allowing rapid synthesis of a large amount of viral genomes (**Cheung *et al*, 2012**).

1.5.2 - Circovirus classification

Circoviruses are circular, rep encoding ssDNA viruses (CRESS-DNA). A large number of CRESS-DNA sequences are highly divergent, and are ubiquitous. They are commonly discovered in metagenomics samples of the environment and of oceanic invertebrates, as well as stool and tissue samples of animals (**Delwart *et al*, 2012**). This group of loosely-defined viruses represents a pool of incredible diversity (**Simmonds *et al*, 2017; Schulman and Davidson, 2017**), of which established ssDNA virus groups, such as Circoviridae, represent just a part.

Family Circoviridae consists of two genera: *Circovirus* and *Cyclovirus*. The two genera share the basic structural features of Circoviridae: *rep* and *cap* ORFs arranged in an ambisense orientation in a small circular ssDNA genome. Cycloviruses are classified based on phylogeny, and on the putative differences in replication and transcription employed by cycloviruses. First, the 3' IR present in circoviruses is absent or consistently smaller in cycloviruses (**Li *et al*, 2010**). Secondly, the putative origin of replication (nonanucleotide motif) is located on the Rep encoding strand of circoviruses and the Cp encoding strand of cycloviruses. Finally, the presence of introns has been reported for cycloviruses. (**Rosario *et al*, 2012**).

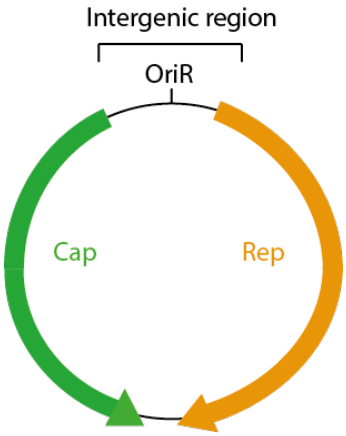

Figure 1.9) Canonical circovirus genome structure. Orange and green ORFs are labelled with direction of expression.

1.5.3 - Circovirus diversity, host range, and role in disease

Members of the genus *Circovirus* have been studied extensively as veterinary pathogens. They have been identified in numerous species of bird, including pigeons, geese (**Todd *et al*, 2001a**), canaries (**Todd *et al*, 2001b**), finches, gulls (**Todd *et al*, 2007**), and corvids (**Stewart *et al*, 2006**). They are well known as the agent causing beak and feather disease (**Raidal *et al*, 2015**). Similarly, PCV-1/2 has been well-studied as the causative agent of postweaning multisystemic wasting disease in swine (**Hamel *et al*, 1998**). As well as well-studied disease associations, circoviruses have been identified in numerous hosts through a series of degenerate PCR and metagenomics studies, including freshwater fish (**Lorincz *et al*, 2011; Lorincz *et al*, 2012**), chimpanzees, humans, (**Li *et al*, 2010**) dogs (**Hsu *et al* 2016**), and Aleutian

mink (**Lian *et al*, 2014**). In addition to discovery in a wide range of hosts and being veterinary pathogens, circoviruses have gained prominence due to their presence as a contaminating agent in the live-attentuated rotavirus (RotaRix) vaccine (**Victoria *et al*, 2010**), although no adverse reactions or human infection by circoviruses as a result of this vaccine have been reported **(FDA.gov**). In addition, despite having no evidence of human infection, circoviruses are studied as possible contamination agents in porcine-human xenotransplantation (**Denner and Mankertz, 2017**).

Circovirus EVEs (CVe) are relatively widespread in animal genomes. Similarity-search based approaches have established a wide past host range of circoviruses in multiple mammalian groups (**Katzourakis and Gifford, 2010**), birds **(Cui *et al*, 2014**), snakes (**Gilbert *et al*, 2014**) and fish (**Feher *et al*, 2013**). In addition, a study by **Liu *et al*, 2011** found numerous circo-like viruses integrated in the genomes of mites and a gastropod (**Liu *et al*, 2011**). These studies have also informed the time-scale of circovirus evolution: an ortholog between different members of genus *Crotalus* set the minimum estimated date of circovirus infection of snakes to at least 10 million years ago (mya). (**Gilbert *et al*, 2012**). Similarly, the discovery of an orthologous CVe between the dog and the cat set the date of circovirus invasion of carnivore germlines to at least 54mya **(Katzourakis and Gifford, 2010**).

In contrast to circoviruses, cyclovirus CVe have yet to be discovered. Cycloviruses have been isolated from invertebrates and vertebrates. Mammals and birds, including a large number of domestic animals – chickens, sheep, cows, goats, horses and cats (**Delwart and Li, 2012**), make up the vertebrate sources of cyclovirus sequences, where arthropods (cockroaches and dragonflies) make up the invertebrate sources. Interestingly, cycloviruses have also been isolated from human samples of sera, cerebrospinal fluid (CSF) and bronchial aspirate (**Phan *et al*, 2014; Smits *et al*, 2013; Tan *et al*, 2013**). These discoveries, have led to speculation that cycloviruses may be the etiological agents behind human diseases. For example, cyclovirus-Viet Nam (CyCV-VN) was identified in the CSF of patients with CNS infections (**Tan *et al*, 2013**). CyCV-VN was subsequently discovered to have a geographical distribution spanning Asia, Africa, Europe and South America. (**Macera *et al*,**

2016; Garigliany *et al*, 2014; Phan *et al*, 2014). These studies have led to speculation that cycloviruses may be an emerging human pathogen, resulting from cross-species transmission of mammalian and arthropod circo- and cycloviruses (**Li *et al*, 2011; Delwart *et al*, 2012**).

1.6 - Indirect paleovirology – antiviral genes

The majority of viruses do not leave behind EVEs, making the viral fossil record as sparse as its geological-paleontological counterpart. The study of antiviral genes can be used to infer the presence and activity of viruses over evolutionary timescales. This approach is called 'indirect paleovirology'. (**Patel *et al*, 2011**) This approach is most powerful where a virus:host conflict is known – i.e. where a gene is known to be antiviral. For example, many studies involving paleovirology have focussed on interferon stimulated genes (ISGs).

1.6.1 - Interferon stimulated genes (ISGs).

ISGs are a component of the innate immune system. They are stimulated by Type I, II and Type III interferons (IFNs), often secreted in response to the detection of conserved pathogen associated molecular patterns (PAMPs). The induction of these genes in response to IFN signalling leads to the expression of hundreds of antiviral factors, resulting in an 'antiviral state' within the cell that is refractory to viral replication. (**Schoggins *et al*, 2011**). ISGs are critical in combating viral infection, and there are numerous examples of direct co-evolutionary conflicts between viruses and antiviral ISGs (**Sawyer *et al*, 2005; Daugherty *et al*, 2014**). These conflicts can be studied as a history of history of host:virus co-evolution, and can be traced through indirect paleovirology studies of ISGs.

1.6.2 – Virus:ISG coevolution

Viral evolution requires the host to co-evolve in order to counteract constantly changing viruses. Variants of host genes are fixed under selection pressure exerted by viral disease. This fixation is more rapid than usual due to the fitness cost associated with inability to repel viral infection – the host

may suffer less due to a variant gene than an inability to combat the virus. The virus then counter-evolves, and the variant then begins to exert a renewed selection pressure. These bouts of selection represent virus:host conflict and co-evolution. (**Daugherty and Malik, 2012).** They can be identified by searching protein coding gene sequences for instances of positive selection – whereby beneficial mutations are increased in frequency until they become fixed in the opoulation. This is assessed by studying the ratios of non-synonymous nucleotide changes that change the amino acid sequence versus synonymous changes that leave the amino acid unchanged due to genetic code redundancy. A greater rate of fixation for non-synonymous changes versus synonymous is mostly explained by the action of positive (aka diversifying) selection. (**Lemey *et al*, 2009**). Examples of genes that are, or have been, under positive selection are *trim5a and tetherin*(**Sawyer *et al*, 2005, McNatt *et al*, 2009; Busnadiego *et al*, 2014**). Signatures of residues under positive selection at the host:virus interface can be used to guide *in vitro* studies of host:virus co-evolution. For example, **Sawyer *et al*, (2005**) successfully used selection analyses to discern specific residues responsible for antiviral activity and species specificity in primate TRIM5a sequences.

1.6.3 – Antiviral gene dynamism

Antiviral gene groups often undergo lineage specific changes in copy number, as these changes are more likely to be selected for in ISGs. This is a potential adaptation by the host to adapt antiviral specificity without losing existing function. This phenomenon, referred to as gene dynamism (**Daugherty *et al*, 2016; Mitchell *et al*, 2015**), is particularly common in ISG families (**Shaw *et al*, 2017**).

Duplication at antiviral ISG loci presumably confers a selective advantage, in that multiple copies of an antiviral gene can then go on to encode different antiviral specificities (**Daugherty *et al*, 2014**). A broad study of the mammalian interferome found that antiviral ISGs were significantly more likely to have undergone expansion (**Shaw *et al*, 2017**). This is mirrored by studies of single genes/gene groups. Gene duplication has been reported at the *trim5* locus in cows, and at the *mxb* locus in the

common ancestor of eutherian mammals. (**Sawyer *et al*, 2007; Mitchell *et al*, 2015**). Relatively extensive duplication has taken place at mammalian *apobec3g* and *parp* family loci. (**Daugherty *et al*, 2014; Munk *et al*, 2012**). The consequences of these gene duplications are evident: *parp* family duplication has given rise to a gene family encoding a diverse range of antiviral specificities. (**Daugherty *et al*, 2014**). This is supported by *in vitro* analyses of IFIT1 and IFIT1B specificities: ancestral duplications at the mammalian *ifit1* locus gave rise to multiple *ifit* proteins encoding different antiviral specificities (**Daugherty *et al*, 2016**).

Numerous instances of gene loss have been observed at antiviral gene loci., The *mx* proteins have been inactivated in toothed whales (**Braun *et al*, 2015**). These losses occur either through pseudogenisation, gene conversion, or deletion of the locus, and they have occurred in the *mx, ifit, parp, trim5* and *apobec3g* gene families (**Mitchell *et al*, 2015; Daugherty *et al*, 2016; Daugherty *et al*, 2014; Sawyer *et al*, 2007; Munk *et al*, 2012**). The selective advantage of gene loss is more difficult to ascertain, especially in contrast to gene duplication – where the duplicate gene can be analysed in *silico* and in the lab. **Braun *et al*, (2015)** speculate that loss of the *mx* genes in odontoceti (toothed whales) is due to viral infection that exploited the *mx* genes. By contrast, it is possible that a relaxation of selection pressures may be responsible for gene inactivation and loss. This could occur either through assumption of the lost gene's function by a paralog or through relaxation of a specific viral selection pressure (**Sawyer *et al*, 2007**).

Studying gene dynamism can be a valuable tool in indirect paleovirological studies of host:virus co-evolution. Gene loss can be used to make inferences on the lineage-specificity of antiviral activity. In addition, the lineage-specific loss of a gene suspected of antiviral activity can provide supporting evidence for its antiviral activity: loss of a gene indicates that a housekeeping role is either unlikely, highly lineage specific, or redundant. Contrastingly, studying ISG expansion can give insights into the range of specificities encoded by the duplicated genes (**Daugherty *et al*, 2014; Daugherty *et al*, 2016**).

1.7 - Using paleovirology to study host:virus co-evolution

Indirect (gene-based) and direct (ERV and EVE based) paleovirological analyses can provide insights into host-virus co-evolution. These analyses are performed on ERV, EVE and ISG nucleotide or amino acid sequences acquired through the mining of genome data, either through *in silico* approaches, such as similarity searching, or through molecular cloning, sequencing and assaying of virus and gene sequences (**Katzourakis and Gifford, 2010; Frankel and Stoye, 1989**).

1.7.1 - Phylogenetic analysis

ERV and EVE sequences can be used to reconstruct the phylogenetic relationships between endogenous and extant exogenous viruses. This can be used as a basis for classification. (**Tristem *et al*, 2000**; **Benit *et al*, 2001**). In addition, it can identify new genera and groups. An example of this is in Retroviridae, where ERV diversity grouping outwith established genera warrants the creation of new genera within Retroviridae (**Johnson *et al*, 2015**). Similarly, the identification of novel parvoviruses by metagenomic and *in silico* screening led to the (tentative) proposition of the *Chapparvovirus* genus. (**de Souza *et al*, 2017).**

In addition, phylogenetic analysis can be used to identify virus past host range. The discovery of an EVE or an ERV in a host species definitively extends the host range of that virus group to the species in which it was discovered. Examples of this are the unexpected discovery of filovirus EVEs in rodent genomes (**Taylor *et al*, 2010**), and the discovery of a deltaretrovirus ERV in the genome of the long-fingered bat (**Hron *et al*, 2018**). Furthermore, where a phylogeny of retroviral peptides from multiple hosts is different to that of their corresponding hosts, it is indicative of host switching having taken place. (**Diehl *et al*, 2016**). These discoveries of host range can underscore the stability or instability of host range in groups of viruses (i.e. how prone they are to crossing species barriers), and could potentially be used as predictive factors for identifying virus reservoirs by guiding virus discovery efforts. (**Taylor *et al*, 2010; Katzourakis and Gifford, 2010**).

The strong conservation of virus polymerases (**Xiong and Eickbush, 2010**), makes them candidates for the phylogenetic analysis of EVE and ERV evolution, and many studies rely on the RT domain for classification and phylogenetic analysis of ERVs (**Tristem, 2000; Jern *et al*, 2005; Hayward *et al*, 2015**). In addition, an apparent conservation and predisposition of viral polymerases to integrate has made them the most appropriate candidates for phylogenetic analysis of EVEs. (**Katzourakis and Gifford, 2010; Beyli *et al*, 2010**). Other peptides can be used: the TM domain of Env has been used in phylogenetic screening and analysis of Retroviridae (**Benit *et al*, 2001**). The propensity of retroviruses to recombine has resulted in the resolution of different evolutionary histories for different protein coding genes within single retrovirus lineages (**Benit *et al*, 2001; Diehl *et al*, 2016).** An example of this is the evolutionary history of recombination between the retroviral envelope and polymerase. Env, which determines host specificity, has a specificity for mammals in the genus *Betaretrovirus*, and for a diverse range of hosts in genus *Gammaretrovirus*. Betaretroviral acquisition of a gammaretrovirus *env* gene can confer a more diverse host range, and result in the reactivation of previously defunct ERVs (**Henzy and Johnson, 2013).**

The amplification processes and history of individual ERV lineages is reflected in phylogenies of ERV sequences (**Katzourakis *et al*, 2005**). As such, they can be used to infer the dynamics of ERV reproduction in a host genome. ERV phylogenies can produce distinct types of topology based on their age. 'Starlike' trees have short internal branches, and long external branches. (**Figure 1.9**). The short internal branches may reflect ancient proliferation events, with the long branches indicating subsequent lengthy residence in the host germline following inactivation. (**Figure 1.9**). These topologies are displayed by ancient ERVs, such as ERV-Fb (**Katzourakis *et al*, 2005**), and HERV-W (**Grandi *et al*, 2016**). 'Comblike' trees are often displayed by youthful, proliferating ERV lineages, such as mysTR (**Cantrell *et al*, 2005**), HERV-K-HML2 (**Katzourakis *et al*, 2005**) and IAP (**Magiorkinis *et al*, 2012**). (**Figure 1.9**). Such topologies can be used to make general inferences as to the evolutionary history and transpositional activity of an ERV.

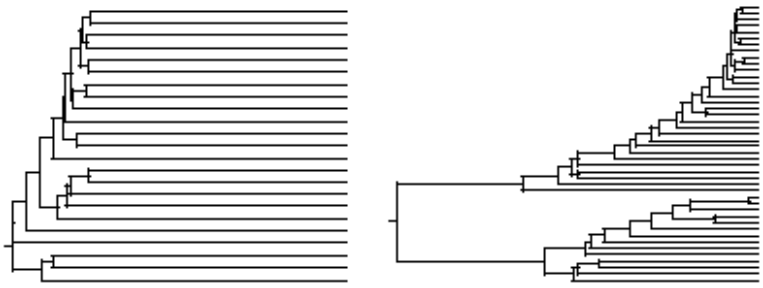

Figure 1.10) ERV amplification dynamics reflected in their phylogenies. Left: a 'star-like' phylogeny commonly displayed by old, inactive ERV lineages (HERV-FB), contrasted with Right: a 'comb-like' phylogeny of the type commonly shown by recently acquired, possibly active ERV lineages (HERV-K-HML2). Figure adapted from **Katzourakis *et al*, 2005.**

1.7.2 - Sequence analysis

Sequence analysis of ERVs and EVEs, and the identification of conserved structural motifs can be used to assist phylogenetic analysis and classification (**Jern *et al*, 2005; Rosario *et al*, 2017**). The presence or absence of features idiosyncratic to a viral lineage can be identified, such as the presence of accessory genes, structural features in protein coding ORFs, non-coding motifs (PBS, PPT, nonanucleotide motif) and replication strategies (e.g. ribosomal frameshifting/termination suppression). These features assist in classification of virus sequences. (**Jern *et al*, 2005;**). (**Figure 1.5**).

Where retroviral genome features have been analysed *in vitro,* their presence, absence and characteristics can be used to make inferences in ERV and EVE biology. For example, the loss of the Gag myristoylation signal in MusD indicates that it is incapable of forming infectious particles – a defect rescued by provision of the signal *in vitro* (**Ribet *et al*, 2007**). Subsequently, the loss of particle-forming activity has been proposed as important in determining *env* loss, intracellularisation and intragenomic proliferation (**Dewannieux *et al*, 2004**; **Magiorkinis *et al*, 2012).** In addition, the activity of innate immune effectors can be identified in ERV sequences: **Jern *et al*, 2007** and **Lee *et al*, 2008** identify evidence of APOBEC3G mediated hypermutation in the proviruses of MLV and HERV-K(HML2) respectively.

Mutations in ERV proviruses can therefore be studied and corrected. Then, the ancestral forms of EVE and ERV sequences can be reconstructed *in silico* and used in *in vitro* studies of the properties of these viral elements and genes, with examples being the reconstruction and assaying of HERV-K and MuERV-L as HERV-K(CON) and ancML (Dewannieux *et al,* 2006; **Lee and Bieniasz, 2007**; **Blanco-Melo *et al*, 2018**).

1.7.3 – The timeline of host:virus co-evolution

EVE and ERV sequence data can be used to calibrate the time-scale of host:virus evolution. They can be dated using orthology - where an integration is shared between two species, having integrated in a common ancestor). This approach uses estimates of host divergence times to propose minimum estimated integration dates of ERV and ERV loci. (**Lee *et al*, 2013; Keckesova *et al*, 2012**). (**Figure 1.10**). Additionally, EVE and ERV sequences can be dated using the co-divergence between the two sequences that has taken place since integration. For example, in duplicated EVEs, or in the paired LTRs of ERVs, the sequences are identical upon integration. The pairwise distance between the sequences can be estimated, and used with the host rate of neutral substitution to estimate the date of duplication or integration (**Gifford *et al*, 2008**). However, recombination or gene conversion that alters LTR sequences can lead to artefactual results when estimating the age of ERV integrations (**Johnson, 2015**). These events can be resolved through phylogenetic analysis of ERV LTR sequences.

These data can provide multiple insights into the timescale of host:virus co-evolution. They have shown that viruses are much older than molecular clock estimates provide –the slowest molecular clocks indicate that viruses are hundres, if not thousands of years old (**Drummond, 2003**). Not only do EVEs and ERVs contradict this estimate quite dramatically – ERV-L orthologs indicate that *Retroviridae* is at least 100 million years old (**Lee *et al*, 2014**), but they also allow for reassessment of the ages of viral groups previously thought to be quite modern. The discovery of lentivirus orthologs in the genomes of primates shows them to be at least 12 million years old (**Gifford *et al*, 2008**). Similarly, hepadnaviruses were thought to be relatively modern viruses, with an age in the order of thousands of years (**Zhou and**

Holmes, 2007), but orthologous hepadnaviruses in bird genomes demonstrates their age to be around 40 million years (**Gilbert and Feschotte, 2010**). ERV and EVE sequence data has provided evidence that, despite rapid mutation, viruses have remained relatively conserved over deep time. This process is described as 'idempotence' – rapid evolution in a constrained mutational space leads to converged evolution to similar, conserved states. (**Gifford, 2012; Holmes *et al*, 2011**). Finally, the identification of dated EVEs can date the association of viruses with certain species groups (e.g. the

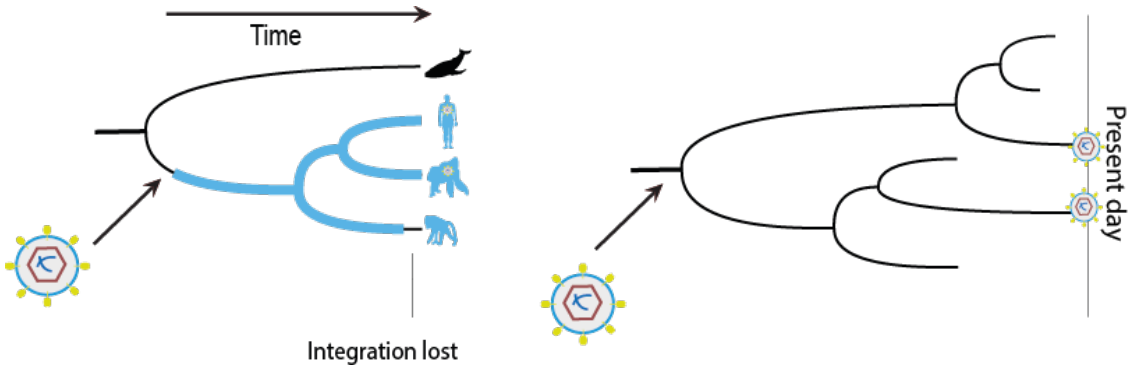

Figure 1.11) ERV invasion leads to orthologous integrations (a). b) Shows how acquired EVE/ERV lineages are lost from the viral fossil record by lineage extinction prior to the present day.

association of snakes with circoviruses is at least 12 million years old (**Gilbert *et al*, 2014**).

1.7.4 - Functional genomics analyses

In many instances, EVE and particularly ERV sequences have been shown to have been functionalised by the host. Typically this involves either the LTR based or epigenetic regulation of genes, or the direct contribution of an ERV/EVE ORF to the repertoire of host protein coding genes. (**Mager and Stoye, 2015**). ChIP-seq studies can discern regulatory factor binding and histone methylation state at or around ERV sequences to determine their role in gene expression and regulation. These studies rely on genome annotation of ERVs and EVEs, which in turn allow study of individual ERV loci and thus giving context to studies of ERV functionalisation. For example, **Schmid *et al*, 2010** find that STAT1 binds MER41 elements in humans. **Chuong *et al*, 2016** build on this to find that MER41 is a primate specific ERV LTR.

1.7.5 – Study of gene evolution

There are numerous cases where ERV genes have been functionalised or 'exapted' by the host, with notable examples being *fv1* and *syncytin*. Additionally, an exapted EVE ORF has also been found in parasitic wasps (**Bezier *et al*, 2009**), and endogenous bornavirus like elements are also under scrutiny for their potential role in host physiology (**Kobayashi *et al*, 2016; Fujino *et al*, 2014**). Identification of potentially exapted ERV ORFs can be undertaken through screening of genome data for intact ORFs, analysis of discovered ORFs for signatures of selection, followed by *in vitro* analysis of the ORF. An example of this is the identification of percomORF – a retroviral Env ORF under purifying selection in spiny-rayed fish (**Henzy *et al*, 2017**).

1.8 – Mining genome data for paleovirological insight

Direct and indirect paleovirology studies rely on gene and ERV/EVE sequences. The largest source of gene and EVE/ERV sequence data is WGS data. (**Hayward *et al*, 2013; Hayward *et al*, 2015**; **Katzourakis and Gifford, 2010**). Sequence data is accumulating at an enormous rate – not just whole genome sequence assemblies, but also transcriptome, ChIP-seq and metagenomics data. ERVs and EVEs are often not captured by conventional automated analysis pipelines. Additionally, information on gene dynamism and birth/loss has the potential to not be captured by annotation pipelines. (**Braun *et al*, 2015**). This presents a requirement for methods that allow the retrieval of ERV, EVE and ISG peptides in a manner that fully leverages the rapidly burgeoning bulk of genome sequence data to allow indirect and direct paleovirological analyses to be performed.

1.8.1 – Similarity searches

Similarity searching can be used to find sequences homologous to a query in a specified database (an organism WGS assembly, for example). (**Lemey *et al*, 2009**). It is frequently used in paleovirological analyses of ERVs, EVEs and ISGs. (**Tristem *et al*, 2000; Katzourakis and Gifford, 2010;**

Daugherty *et al*, 2016 – amongst many others). In some cases, similarity searches can be used to identify genome features that are not conserved (and thus cannot be directly identified using a similarity search), but are identifiable based on their positional relationships to more conserved sequence features (**Figure 1.12c**). This approach is particularly useful when characterizing genomic sequences such as endogenous viruses and transposons that have relatively predictable structures comprising compact arrangements of genes, and is relied upon by the RetroTector **(Sperber *et al*, 2009**) and LTR-Digest (**Steinbiss *et al*, 2009**) ERV detection tools.

Hidden markov models (HMMs) can be used in an iterative approach. Using HMMER (**Prakash *et al*, 2017**), a set of query sequences is used to create a hidden Markov model (HMM). This in turn is used to query sequence databases further and find more distant homologies. Distant homologs are used to update the HMM. This process is repeated until no new sequences are found (**Figure 1.12b)**. HMMs have been used to make inferences into virus evolution in deep time (**Nasir and Anolles, 2015**). However, HMMs tend to rely on intact peptide sequences or complete coding ORFs, making them less amenable to use in investigations involving ancient, non-functional DNA sequences.

One of the most common methods of similarity searching is the basic local alignment search tool (BLAST) (**Camacho *et al*, 2009).** BLAST takes a user inputted query sequence and efficiently searches target databases for sequences exhibiting similarity (homology) to the query (hits) (**Figure 1.12a)**. BLAST based approaches are useful in that they do not rely on existing annotations, they can be used to study features that have not been captured by automated annotation pipelines, and rely less on the presence of a complete open reading frame (ORF). Examples include undiscovered genes, transposable elements, non-coding sequences, and any type of degraded genomic material. In the most basic example, similarity searches can be used to retrieve single sequences from genomic databases for further analysis – using the NCBI BLAST webserver, for example. Similarity searches can be augmented with other strategies to enable recovery of more distantly homologous sequences. Position specific iterated BLAST (PSI-BLAST) use iterative approaches to achieve this. (**Altschul *et al*, 1997**). In PSI-BLAST,

the first search is performed to identify related amino acid sequences in the target database. The MSA constructed using these sequences is used to create a position specific scoring matric (PSSM). The search then continues in a similar manner to HMMER as described above, halting when convergence is reached (no new sequences are discovered) (**Figure 1.12b)**.

BLAST-based approaches are especially useful when investigating genomic features that are not well annotated in public sequence databases, such as small RNAs, pseudogenes, transposable elements, highly dynamic gene families, ERVs and EVEs. These investigations often take the form of a heuristic process (i.e. a trial-and-error discovery process with loosely defined rules). Tools are required that not only implement automated screening pipelines, but also facilitate interrogation, analysis and interpretation of the data they produce.

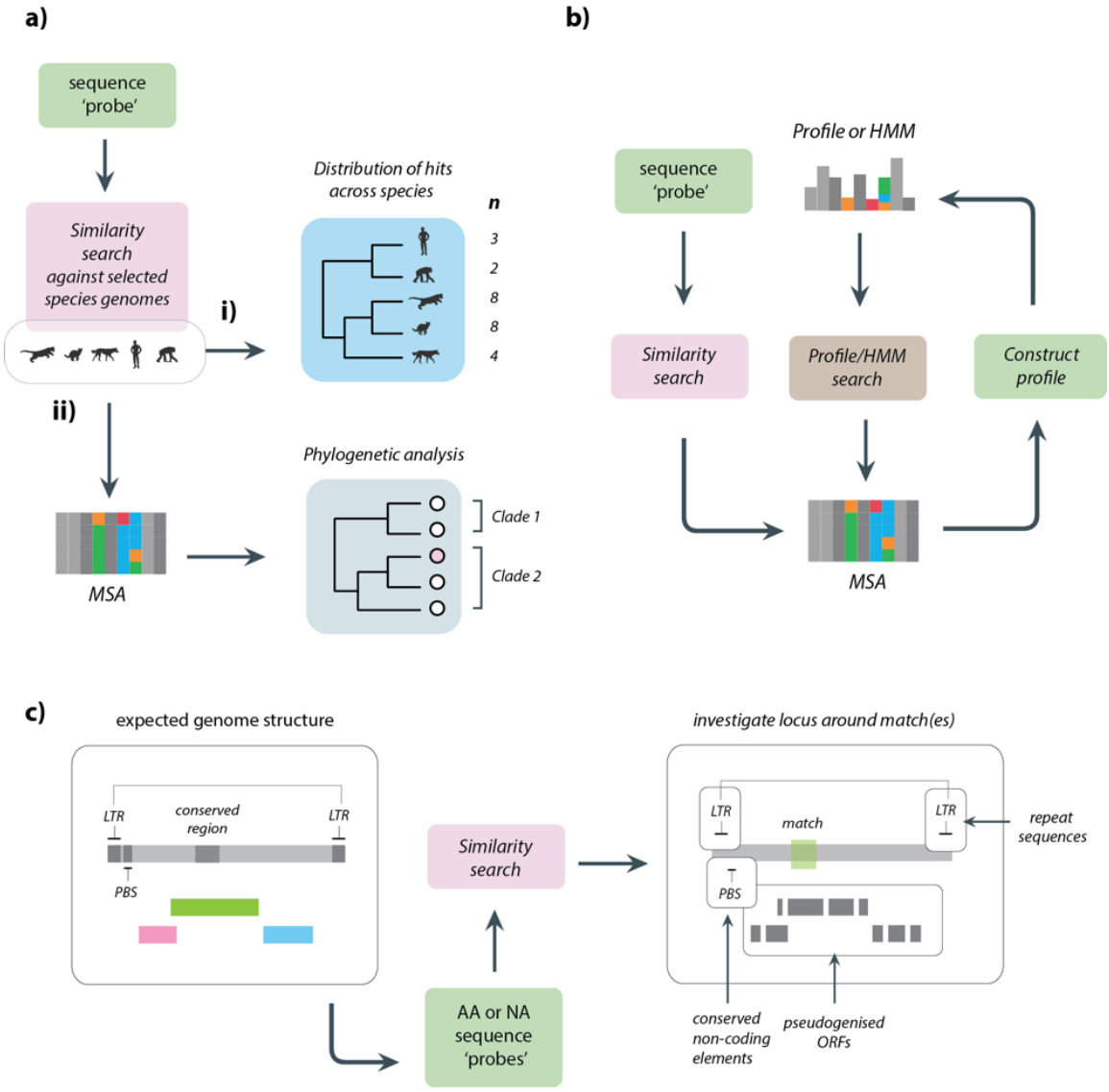

Figure 1.12) Strategies for similarity searches of genomes based

(a) Basic approach: a sequence probe is used to search target genomes, and the results are used to investigate; (i) copy number variation across species genomes; (ii) relationships between a set of aligned, homologous sequences are analysed with phylogenetics.
(b) Iterative approaches to detect remote homologies: in position-specific iterated BLAST (PSI-BLAST), an initial similarity search is performed to identify related sequences in the target database, these are used to derive a PSSM which in turn is used to search for more matches. This can lead to the detection of new homologs, which are used to update the PSSM for another round of screening. This process is repeated until derived, or until convergence (the point at which no new sequences are identified by searching). A similar approach is implemented in HMMR, but using a hidden Markov model (HMM) in place of a PSSM.
(c) Using conserved sequences to detect and characterize novel genes, pseudogenes and mobile genetic elements: characterization of loci detected via similarity-based screening can sometimes lead to the identification of new probes for subsequent rounds of screening. In this instance, the expected structure of a retroviral genome is used to identify conserved features common to Retroviridae.

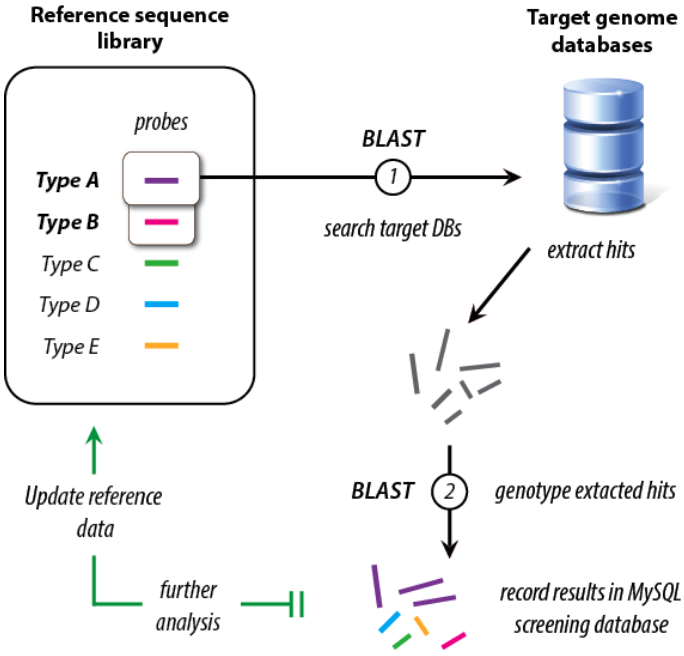

Figure 1.13) Basic DIGS searching strategy involving 1) the searching of target databases with a subset of probes derived from a reference library. This is followed by the 'genotyping' of extracted hits by comparison to the reference library and recording in MySQL (2). This approach can be extended to perform a heuristic program of investigation based on analysis of results (green line). Figure provided by Rob Gifford.

1.8.2 - Introduction to DIGS

The database integrated genome screening (DIGS) tool is an implementation of BLAST coupled with a relational database management system (RDBMS) that can be used to systematically screen organism genomes for sequences of interest. DIGS screens take on two main steps. (**Figure 1.12**), and require two key components (in addition to a genome database): the reference and probe libraries.

The reference library is used to classify hits identified in the first round of screening. It consists of a FASTA file of sequences. The sequences in the reference library typically encompass the genetic diversity of the sequences under investigation. For example, in DIGS for ERVs using RT, we used a reference library that spanned the diversity of Retroviridae – Spuma-, Gamma-, and Betaretroviruses, as well as non-retroviral transposon sequences derived from *gypsy*, *copia*, and LINE1 elements. The inclusion of non-retroviral RT sequences was so that if searches for retroviral RT matched,

however distantly, to non-retroviral elements, they would be classified as such. The probe library is used to screen the genomes under investigation. The probes are often derived from the reference library. Although the entire reference library can be used, a subset is usually sufficient. If the probe library consists of amino acid sequences, tBLASTn is used in BLAST searches. If nucleotide sequences, BLASTn is used.

Screening proceeds in two rounds. First, BLAST is used to search each genome specified in the control file with each sequence contained in the probe library. The results of this screen are stored. In the second round of screening, each result of the first round of screening is BLASTed against the reference library. It is then classified based on the highest significant sequence match. This is called the 'genotyping' stage. The results of this second stage are stored in the results table (**Figure 1.12**).

1.8.3 - Why use DIGS?

In addition to utilising BLAST – a well-established tool, DIGS has a number of advantages. It can be used to implement automated screens in a heuristic investigation, screening the published body of genome data using BLAST to search for any sequences of interest: ERVs, EVEs or ISGs.

This flexibility is reflected in the second 'genotyping' step of screening, where hits are classified according to the reference library. This allows the user to essentially create a bespoke screening programme. For example, if the user is interested in studying Betaretroviruses, they can compile a reference library encompassing a high diversity of betaretroviral sequences for more accurate classification. This reference library can be updated with newly acquired and analysed sequences, allowing for more fine-grained classification of the sequences of interest by repeating the 'genotyping' step of screening with the updated library. This can be used to discover endogenous viral diversity by implementing a 'phylogenetic screening' approach.

The use of an RDBMS is a key feature of DIGS. BLAST reports are large and extensive. The amount of processing and organisation required for

analysis beyond the scope of manually searching the results of the web-tool is not trivial. In addition, web-based and command-line BLAST are limited in the data sets that can be interpolated with the results. Using an RBDMS is an easy way to store BLAST results – the database is interrogated using structured query language (SQL), allowing powerful and flexible interrogations and operations to be performed on the data. In addition, the use of auxiliary data tables allows the layering of multiple different datatypes onto DIGS results (**discussed in Methods**).

1.9 - Research scope and aims

Direct and indirect paleovirological analyses use ERV, EVE and ISG sequences to make discoveries of biological relevance to host:virus co-evolution. These sequences can be retrieved from publicly available WGS databases using similarity-search based approaches. With the increasing ease of genome sequencing, genome data is accumulating at a rate faster than can currently be analysed. There is therefore unprecedented scope for mining genome data in heuristic investigations into host-virus co-evolution. The aim of this PhD was to study host:virus co-evolution by systematically mining genome sequence data. Accordingly, I chose to study endogenous viral elements, endogenous retroviruses, and interferon stimulated genes.

Chapter 2 – Materials and Methods

2.1 - Materials

2.1.1 – Sequence data

Genomes

Animal WGS data was retrieved from NCBI genomes. WGS assembly information, corresponding animal taxonomy, and usage in each chapter, is tabulated in **Appendix 2.1**.

Datasets

Host taxonomy data were retrieved from NCBI taxonomy and tabulated in Appendix 2.1. Functional genomics datasets were retrieved from the NCBI Gene Expression Omnibus (GEO) (**Edgar et al, 2002**). (**Table 2.1**). The functional genomics datasets were generated as part of ChIP-seq surveys of chromatin and genomic features in mouse embryonic stem cells. (see **Table 2.1**).

Probe and reference libraries

Reference libraries for the ERV screens were derived from previously published studies of ERV and exogenous retrovirus (XRV) taxonomy (**Tristem, 2000; Benit et al, 2001; Herniou et al, 1999**), as well as reviews and analyses of murine ERVs (**Stocking, 2008**). Probe libraries were extracted as subsets of the reference. (**Table 2.2**) Similarly, previous studies of Circoviridae taxonomy, as well as literature searches, were used to derive reference libraries spanning the established diversity of Circoviridae, as well as basal CRESS sequences. (**Table 2.2**). For the ISG screens, The probe consisted of the human orthologs of the core mammal gene in question, and the references consisted of the human paralogs and mammalian orthologs of that gene, retrieved from Biomart. (**Smedley et al, 2015**).

Table 2.1: Functional genomics datasets used in this thesis

Cell type	Factor ChIPped	Citation	GEO Accession
mES	TRIM28 dependent mH3K9	Rowe *et al*, 2013	GSE41903.
E14	SOX2	Chen *et al*, 2012	GSE11431
E14	OCT4	Chen *et al*, 2012	GSE11431
E14	NANOG	Chen *et al*, 2012	GSE11431
E14	ESRRB	Chen *et al*, 2012	GSE11431
E14	SMAD1	Chen *et al*, 2012	GSE11431
E14	E2F1	Chen *et al*, 2012	GSE11431
E14	TCFCP2I1	Chen *et al*, 2012	GSE11431
E14	CTCF	Chen *et al*, 2012	GSE11431
E14	ZFX	Chen *et al*, 2012	GSE11431
E14	STAT3	Chen *et al*, 2012	GSE11431
E14	KIF4	Chen *et al*, 2012	GSE11431
E14	c-Myc	Chen *et al*, 2012	GSE11431
E14	n-Myc	Chen *et al*, 2012	GSE11431
E14	GFP/control	Chen *et al*, 2012	GSE11431
E14	p300	Chen *et al*, 2012	GSE11431
E14	Suz12	Chen *et al*, 2012	GSE11431
mES	p53	Li *et al*, 2012	GSE26360 GSE33913
Murine BMM	STAT1	Ng *et al*, 2011	

Table 2.2 – Reference sequences used in DIGS screening for ERVs and EVEs

Full name	Abbrev	Virus family	Accession
Beak and feather disease virus	BFDV	Circoviridae	NC_001944
Columbid circovirus	CoCV	Circoviridae	NC_002361
Starling circovirus	StCV	Circoviridae	NC_008033
Canary circovirus	CaCV	Circoviridae	NC_003410
Starling circovirus	SvCV	Circoviridae	NC_008033
Raven circovirus	RaCV	Circoviridae	NC_008375
Gull circovirus	GuCV	Circoviridae	NC_008521
Finch circovirus	FiCV	Circoviridae	NC_008522
Zebra finch circovirus	ZfCV	Circoviridae	NC_026945
Duck circovirus	DuCV	Circoviridae	NC_007220
Swan circovirus	SwCV	Circoviridae	NC_025247
Barbel circovirus	BarbCV	Circoviridae	NC_015399
Wels catfish circovirus	SgCV	Circoviridae	NC_025246
Porcine circovirus 1	PCV-1	Circoviridae	NC_001792
Porcine circovirus 2	PCV-2	Circoviridae	NC_005148
Canine circovirus 1	CfCV	Circoviridae	NC_020904

Mink circovirus	MiCV	Circoviridae	NC_023885
Mexican free-tailed bat circovirus	TbCV	Circoviridae	NC_028045
Porcine circovirus 3	PCV-3	Circoviridae	NC_031753
Cyclovirus VN isolate cs1	CyCV-VN	Circoviridae	KF031471
Human cyclovirus VS5700009	CyCV-VS5700009	Circoviridae	KC771281
Bat circovirus isolate BtRp-CV-14/GD2012	BtRp-CV-14	Circoviridae	KJ641714
Human stool-associated circular virus NG13	CyCV-NG13	Circoviridae	GQ404856
Dragonfly cyclovirus 5	DfCyV-5	Circoviridae	JX185426
Circoviridae 5-LDMD-2014	5-LDMD-2013	CRESS	NC_025710
Circoviridae 11-LDMD-2014	11-LDMD-2013	CRESS	NC_025716
Circoviridae 13-LDMD-2014	13-LDMD-2013	CRESS	NC_025717
Circoviridae 16-LDMD-2014	16-LDMD-2013	CRESS	NC_025720
Avon-Heatchote Estuary Associated Virus 13	AHEaCV-13	CRESS	NC_026639
Avon-Heatchote Estuary Associated Virus 14	AHEaCV-14	CRESS	NC_026641
Avon-Heatchote Estuary Associated Virus 21	AHEaCV-21	CRESS	NC_026648
Calanoida sp. copepod associated circular virus	COACV	CRESS	NC_027795
Baboon Endogenous Retrovirus	BaEV	Retroviridae	AHZZ010479
Chimpanzee Endogenous Retrovirus 1	CERV-1	Retroviridae	See GLUE pro repo
Chimpanzee Endogenous Retrovirus 2	CERV-2	Retroviridae	See GLUE pro repo
Canis familiaris Endogenous Retrovirus	CfERV	Retroviridae	See GLUE pro repo
Feline Leukemia Virus	FeLV	Retroviridae	M18247.1
Gibbon Ape Leukemia Virus	GaLV	Retroviridae	M26927.1
Human Endogenous Retrovirus T	HERV-T	Retroviridae	AC022143.7*
Koala Retrovirus	KoRV	Retroviridae	AF151794.2
Mus dunni Endogenous Virus	MDEV	Retroviridae	AF053745.1
Murine Leukemia Virus (Moloney)	MLV	Retroviridae	J02255.1
Type C Murine Retroviruslike DNA Sequence	MuRRs	Retroviridae	X02487.1
Reticuloendotheliosis Virus	REV	Retroviridae	NC_006934
Rhinolophus ferrumequinum Retrovirus	RfRV	Retroviridae	JQ303225.1
Human Endogenous Retrovirus W	HERV-W	Retroviridae	AC005187.1*
Endogenous Retrovirus 9	ERV-9	Retroviridae	CT737353.7*
Endogenous Retrovirus E	ERV-E	Retroviridae	AL023280.1*
Endogenous Retrovirus F Type C	ERV-Fc	Retroviridae	See GLUE pro repo
Lemon Shark Endogenous Retrovirus	RV-Lemonshark	Retroviridae	Y07810.1
Komodo Dragon Endogenous Retrovirus	RV Komodo Dragon	Retroviridae	Y07807.1
Human Endogenous Retrovirus - I	HERV-I	Retroviridae	M92067.1
Human Endogenous Retrovirus F type B	HERV-Fb	Retroviridae	AP001629.1*
Human Endogenous Retrovirus X type A	HERV-Xa	Retroviridae	AC114321.2*
Walleye Epidermal Hyperplasia Virus	WEHV	Retroviridae	AF133051.1
Walleye Dermal Sarcoma Virus	WDSV	Retroviridae	AF033822.1
Human T-Cell Lymphotropic Virus 1	HTLV-1	Retroviridae	J02029.1
Human T-Cell Lymphotropic Virus 2	HTLV-2	Retroviridae	M10060.1
Bovine Leukemia Virus	BLV	Retroviridae	K02120.1
Rabbit Endogenous Lentivirus K	RELIK	Retroviridae	FJ493031.1

Grey Mouse Lemur Prosimian Immunodefiency Virus	pSIVgml	Retroviridae	FJ461357.1
Human Immunodeficiency Virus type 1	HIV -1	Retroviridae	K03455.1
Feline Immunodefiency Virus	FIV	Retroviridae	M25381.1
Avian Leukosis Virus	ALV	Retroviridae	M37980.1
Human Endogenous Retrovirus K, type HML2	HERV-K HML2	Retroviridae	KJ155476.1*
Human Endogenous Retrovirus K, type HML4	HERV-K HML4	Retroviridae	AC010617.8*
Human Endogenous Retrovirus K, type HML5	HERV-K HML5	Retroviridae	AC068723.5*
Human Endogenous Retrovirus K, type HML6	HERV-K HML6	Retroviridae	AC026775.6*
Human Endogenous Retrovirus K, type HML9	HERV-K HML9	Retroviridae	AC091849.2*
Human Endogenous Retrovirus K, type 14C	HERV-K14C	Retroviridae	AC069306.6*
Human Endogenous Retrovirus K, type HML8	HERV-K HML8	Retroviridae	AC016955.15
Kangaroo Endogenous Retrovirus	KERV	Retroviridae	AF044909.1
Python molurus Endogenous Retrovirus	pyERVmol	Retroviridae	AF500296.1
Mouse Mammary Tumour Virus	MMTV	Retroviridae	M15122.1
Jaagskiete Sheep Retrovirus	JSRV	Retroviridae	M80216.1
Mason Pfizer Monkey Virus	MPMV	Retroviridae	M12349.1
Mus D Type Endogenous Retrovirus 1	MusD	Retroviridae	AF246632.1
Oryctolagus cuniculus Endogenous Retrovirus H	RERV-H	Retroviridae	AF480924.1
Endogenous Retrovirus L	ERV-L	Retroviridae	Y12713.1
Murine Endogenous Retrovirus L	MuERV-L	Retroviridae	AJ233590.1
Human Endogenous Retrovirus S	HERV-S	Retroviridae	AC004385.1*
Taenniopygia guttata Endogenous Retrovirus Type F	TgERV-f	Retroviridae	XM_0125762
Mallard Retrovirus	RV-Mallard	Retroviridae	See GLUE pro repo
Alligator Retrovirus	RV-Alligator	Retroviridae	See GLUE pro repo
Human Endogenous Retrovirus R type B	HERV-Rb	Retroviridae	AC004045.1*
Opossum Retrovirus	RV-Opossum	Retroviridae	AJ236123.1
Endogenous Retrovirus P	HERV-P	Retroviridae	AC002069.1*
Apteryx australis endogenous virus	RV-Kiwi	Retroviridae	AY820065.1
Zebrafish Retrovirus	ZFERV	Retroviridae	FO704649.2
Snakehead Retrovirus	SNRV	Retroviridae	U26458
Mys Transposable Endogenous Retrovirus	mysTR	Retroviridae	DQ139770.1
Equine Foamy Virus	EFV	Retroviridae	AF201902.1
Feline Foamy Virus	FFV	Retroviridae	Y08851.1
Simian Foamy Virus	SFV1	Retroviridae	NC_001364.1

2.1.2 – Software and tools

Genome screening

The database-integrated genome screening (DIGS) tool (version 1.1) was used to perform systematic screening of whole genome sequence assemblies. (Zhu et al, 2018)

Alignment and phylogenetic analysis

MUltiple Sequence Comparison by Log-Expectation (MUSCLE) is multiple sequence aligner for both nucleotide and protein sequences. (**Edgar, 2004**). MUSCLE v3.8.31 created all multiple sequence alignments (MSA) used in this thesis. MSAs were then manually edited using AliView. (**Larsson et al, 2014**).

Substitution models were estimated using the ModelFinder implementation of the IQ-TREE suite (**Kalyaanamoorthy et al, 2017**). IQ-TREE-omp was used to perform phylogenetic analysis of ERV and EVE peptide sequences, with support evaluated using 1000 ultrafast bootstrap (UFBOOT) replicates (**Minh et al, 2013**).

Annotation and analysis of ERV and EVE sequences

ERV and EVE sequences were recovered from host genomes and used to generate MSAs. Known structural features were searched for in the alignments. LTRharvest and LTRdigest are implemented utilities of the GenomeTools package. GenomeTools v1.5.8 was applied in this study. LTRharvest is a de novo detection tool designed specifically for LTR retrotransposons (**Ellinghaus, et al 2008**). LTRdigest is the annotation tool for characterising the internal coding region defined by LTRharvest (**Steinbiss et al, 2009**). The domain detection function of LTRdigest is performed by using phmmer, a program of the HMMER package.

Phylogenetic analysis using parsimony (PAUP) (**Swofford, 2002**) was used to calculate pairwise distances between LTRs for LTR-based dating of ERV sequences.

DFAM (**Hubley et al, 2015**) was used to search for repetitive elements in genomic sequences.

Functional genomics analyses

Bedtools is a set of utilities that are used for a wide-range of genomics analysis tasks (**Quinlan and Hall, 2010**). Bedtools allows the user to intersect,

merge, count, complement and shuffle genomic intervals in various formats, e.g. BAM, BED, GFF/GTR/VCF.

UCSC Liftover (**Kent et al, 2018**) was used to map locus coordinates between different genome builds.

CCG-ChIP-seq is a suite of tools used for analysing the results of ChIP-seq experiments. It was used to plot enrichment of genomic features (i.e. transcription factor binding) at ERV loci. (**Ambrosini et al, 2016**).

The genome regions enriched for annotations (GREAT) webtool was used to assess uploaded tracks of genomic features for enrichment close to protein coding genes (**McLean et al, 2010**).

Collation of ERV and EVE sequences and auxiliary data

Genes Linked by Underlying Evolution (GLUE) is a bioinformatics environment designed for building projects of alignments linked by evolutionary hypotheses. It was used to store and manipulate all alignments in this thesis. In addition, it was also used to manage and leverage tables of auxiliary data (i.e. genomic features). (**Singer et al, 2018**).

Locus analysis

The UCSC and NCBI genome browsers are resources that we used for interactively visualising genomic data. (**Kent et al, 2002**).

Genomicus (**Louis et al, 2013**) is a webtool for visualising shared genomic organisation across animal species.

Other software and tools

Geneious (**Kearse et al, 2012**) is an all-purpose platform for the organisation and analysis of sequence data.

EMBOSS (**Rice et al, 2011**) is a web suite of tools for the analysis of sequence data. Of EMBOSS, Dottup and Seqret are tool used for the visualisation of pairwise alignments, and the conversion of sequence file formats respectively.

Timetree (**Hedges et al, 2015**) is a public knowledge-base for information on the evolutionary timescale of life. Data from thousands of

published studies are assembled into a searchable tree of life scaled to time. The data can be searched to produce taxa divergence times, timelines for speciation events for individual taxa, and to produce time calibrated phylogenies of the evolution of life.

Structured query language (SQL – pronounced 'sequel') is a programming language for retrieving and operating on data held in relational databases.

Perl and Python are high level, object oriented programming languagea. In a bioinformatics context, they are often used to script basic functionality, as well as perform more in-depth analyses. All in-house scripts used in these studies were written in Perl 5 and Python 2.7.

R is a statistical and graphing programming language commonly used for data analytics and graphing.

2.2 – Methods

2.2.1 – DIGS

DIGS links BLAST (**Camacho et al, 2009**) with the MySQL RDBMS. It can be used to implement systematic genome screens in a heuristic manner. DIGS screens, at a minimum, employ a reference library, a probe library, and a screening target (organism WGS). The reference library is used to classify hits identified in the first round of screening. It consists of a FASTA file of sequences. The sequences in the reference library typically encompass the genetic diversity of the sequences under investigation. For example, in DIGS for ERVs using RT, we used a reference library that spanned the diversity of Retroviridae – Spuma-, Gamma-, and Betaretroviruses, as well as non-retroviral transposon sequences derived from gypsy, copia, and LINE1 elements. The inclusion of non-retroviral RT sequences was so that if searches for retroviral RT matched, however distantly, to non-retroviral elements, they would be classified as such. The probe library is used to screen the genomes under investigation. The probes are often derived from the reference library. Although the entire reference library can be used, a subset is usually sufficient. If the probe library consists of amino acid sequences, tBLASTn is used in BLAST searches. If nucleotide sequences, BLASTn is used.

Screening proceeds in two rounds. First, BLAST is used to search each genome specified in the control file with each sequence contained the probe library. The results of this screen are stored in the database. In the second round of screening, each result of the first round of screening is BLASTed against the reference library. It is then classified based on the highest significant sequence match in the reference library. This is called the 'genotyping' stage. The results of this second stage are stored in the digs_results table (**Figure 2.3**).

Searches in DIGS can take on an iterative aspect. The reference library can be updated to reflect analysis performed on sequences found in previous searches. For example, if searching for ERV sequences in an organism with an undescribed ERV complement: the gulper eel (used here as an example), the hits in the digs_results table will correspond to the previously described sequences in the reference library. Although a good 'first-pass' estimation of ERV diversity, further analysis is needed for fine-grained classification of ERVs in the gulper eel. Hits are retrieved from the database and classified with phylogenetic analysis. The classified hits are added to the reference library. The contents of the results database can be put through the second genotyping stage of DIGS screening again. Sequences bearing significant similarity to the new addition will be updated as such. This process can be repeated numerous times until the contents of the results database reflects more properly the contents of the genomes under investigation. Using this iterative phylogenetic screening approach, DIGS can be used in heuristic investigations of genomes (**Figure 2.2**).

A key aspect of DIGS is the use of relational databases. Relational databases are information storage frameworks that incorporate shared relations between stored pieces of information (**Figure 2.3**). These relationships are based on unique keys that are shared between different data objects in a database (**Figure 2.3**). Relational databases use tables to store data. The tables are arranged in columns. The database consists of multiple tables that share common 'keys'. These shared keys comprise the 'relations' in the relational database. (**Figure 2.3**). Data can be retrieved from the database by searching for specific columns within a table, for

example, to retrieve all data corresponding to 'MuERV-L' from the organism '*Mus musculus*'.

Relational databases have a number of features that make them useful for genome mining. Firstly, relational databases are capable of storing and managing millions of rows of data. In a bioinformatics context, NCBI taxonomy and ENSEMBL are based on MySQL relational databases. Screening many species groups with large numbers of sequences can produce many hundreds of thousands of rows of data – an RBDMS offers a convenient and tidy way to store such large datasets. Relational databases offer methods to retrieve data from these large datasets based on properties. For example, one can easily retrieve all data pertaining to a specific species, gene type, or any other data objects produced in DIGS screening, as the data in the DIGS screening database is essentially list of BLAST reports. Similarly, the data can be flexibly interrogated by joining the digs_results database with user-defined auxiliary data tables. For example, a data table containing virus taxonomy information can be joined (using a shared key) to the digs_results table to retrieve results that only correspond to a specific virus genus, family or order. Auxiliary data tables can be added at the user's leisure, enabling the interpolation of a diverse range of datasets. Finally, SQL commands, data can easily be summarised into tables. For example, the breakdown of ERV sequences by taxonomy in a given species can be achieved, to give detailed data summaries.

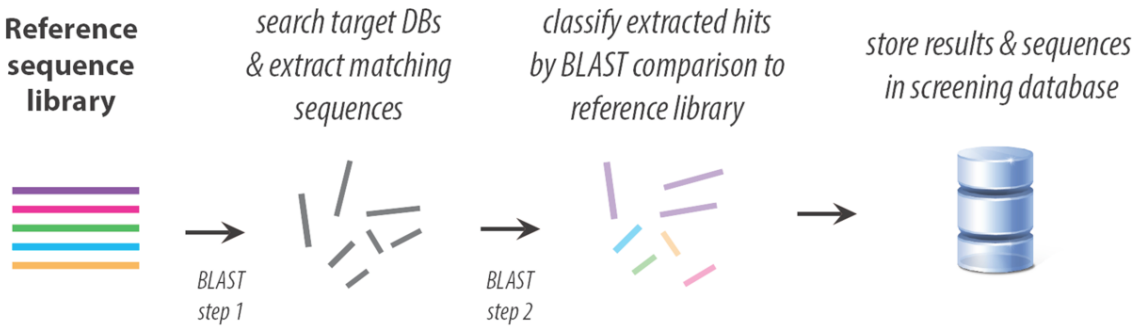

Figure 2.1 – Genome screening in DIGS. Details the searching of target DBs with probe sequences, and subsequent classification. Figure provided by Dr Rob Gifford.

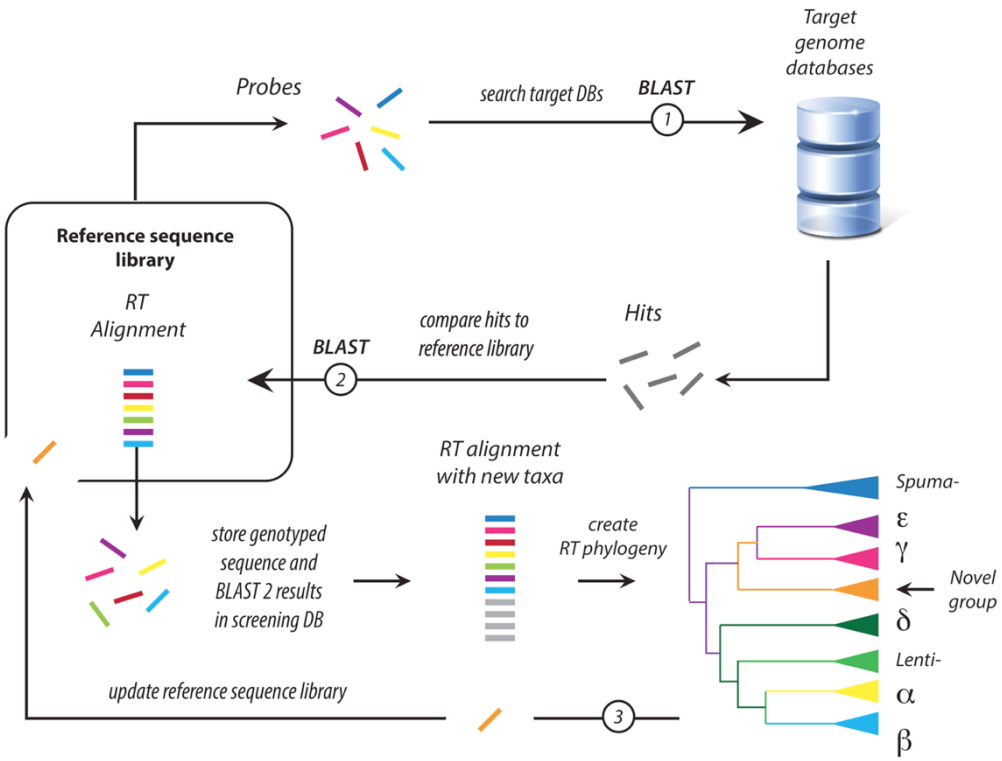

Figure 2.2 -Phylogenetic screening with DIGS. Details the search of target genome databases with a probe set derived from a more comprehensive reference sequence library (1), and subsequent 'genotyping' via comparison of hits to the reference library (2). Then, sequences are classified using phylogenetic analysis and added to the reference library (3). (1) (2) and (3) are then repeated. Figure provided by Dr Rob Gifford.

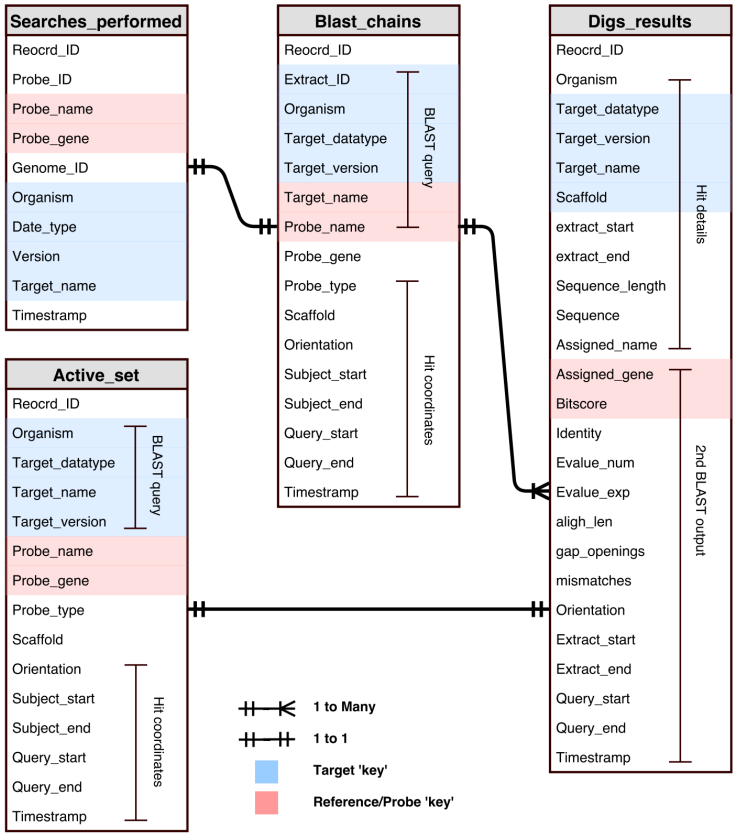

Figure 2.3: Entity-Relationship diagram of MySQL database generated by the DIGS tool.
For each DIGS screening project, the DIGS tool creates a new schema in the MySQL database. Each schema has four tables: BLAST_chains, Digs_results, Seaches_performed, and Active_set. Crossbars show the range of information section in the table; the relationship between each table are linked by relational arrows. Figure kindly provided by Dr Henan Zhu.

2.2.2 -GLUE

All of the ERV and CVe sequences, alignments and metadata were collated in GLUE. GLUE is an open, data-centric bioinformatics environment for organising and leveraging sequence data (**Singer et al, 2018**). The data are published as online repositories at: https://github.com/giffordlabcvr/Retroviridae-GLUE and https://github.com/giffordlabcvr/DIGS-for-EVEs

2.2.3 - Host evolutionary timelines

Timelines and phylogenies of host evolution were retrieved from the Timetree database. Information gained from analyses in this thesis were then

plotted on the time-calibrated phylogenies. For example, where an orthologs was established, it was plotted onto the host phylogeny for that lineage.

2.2.4 – Recovery of proviruses

Where ERV lineages lacked representative proviral sequences, the 10kb regions flanking the RT hit were extracted using blastdbcmd wrapped in a custom perl script. The sequences were analysed using ORFfinder, manual and local sequence alignment, HMMs and positional homology (the identification of sequence features in reference to the RT hit) to identify conserved features common to retroviral genomes. Where possible, LTRs were manually identified using sequence alignment visualised in DOTTUP (**Rice et al, 2000**), followed by identification of 5'/3' TG/CA motifs, target site duplications (TSDs) and U3, R and U5 regions.

2.2.5 – Provirus screening and consolidation

Retroviral ORFs and LTRs were derived from representative provirus sequences of the murine ERV lineages analysed in this thesis, and used to create probe and reference libraries. These libraries were used in DIGS screens of the murine genome. Provirus sequences were identified using the 'consolidate' subroutine of DIGS. Consolidation involves the merging of adjacent or overlapping DIGS hits based on proximity and significance into proviral loci. Briefly, hits are ordered by chromosome, position, and orientation, and merged if they follow the structure LTR-Gag-Pol-(Env)-LTR (where Env is optional) and are < 10,000 nucleotides in length. Consolidated proviruses and solo LTRs were extracted using blastdbcmd and entered into MSAs in the GLUE framework.

2.2.6 – Sequence analysis

Provirus sequences were analysed for sequence features (ORFs, genes, non-coding features) using LTRDigest. LTRdigest uses phmmer to assess the presence of protein motifs and domains. Additionally, it uses similarity searches within the regions bound by LTRs to search for non-coding features

(LTR sections, PBS and PPT). Furthermore, retroviral sequences were investigated through alignment to annotated reference sequences in GLUE.

2.2.7 – Pairwise LTR dating

Non recombinant proviral LTRs were used in pairwise-LTR dating. Bash scripts were used to each pair of LTRs for each provirus into alignments which were then aligned using MUSCLE. The alignments were converted into NEXUS format in EMBOSS Seqret (**Rice et al, 2000**), appended with a NEXUS block stipulating the appropriate model of evolution (GTR+G4), and the pairwise distance calculated using PAUP (**Swofford , 2002**). This output was used to calculate the approximate date of integration for each pair of LTRs (representing a provirus). LTRs are identical upon provirus integration, upon which they become subject to the host background rate of substitution (4.5x10-9 substitutions per year, for the mouse (**Waterston et al, 2002**) . As such, the age of the provirus can be calculated using the formula:

$$\text{Approx. integration date} = \left(\frac{\text{LTR pairwise distance}}{\text{host neutral substitution rate}}\right)/2$$

2.2.8 – Enrichment of genomic features at ERV loci

Enrichment of ERVs close to protein coding genes was assessed by uploading annotation tracks of murine ERVs to GREAT as 'test features' for assessment of enrichment across the whole murine genome (build mm10). GREAT calculates statistics by associating genomic regions with nearby genes and applying the gene annotations to the regions. Association is a two-step process. First, every gene is assigned a regulatory domain. Then, each genomic region is associated with all genes whose regulatory domain it overlaps. In this instance, we selected parameters that searched for the single nearest gene to an ERV locus within 1000kb, including curated regulatory domains (where experimental demonstration of gene regulation extends the regulatory domain to be beyond the 1000kb parameter). Results were downloaded, parsed, and tabulated.

When assessing murine ERV loci for enrichment of transcription factor binding, ChIP-seq datasets were mapped to the mm9 build of the Mus

musculus genome using UCSC Liftover. The converted tracks, along with ChIP-seq BED files were uploaded to the CCG-ChIP seq analysis web server and analysed using ChIP-cor, which reads ChIP-Seq tag positions/annotation files and generates a positional correlation histogram between the two annotation datasets, showing the abundance of the 'target' feature as a function of the distance from the 'reference' feature. Range (distance from '0' from the ERV centre) was given as 10kb, window width (intervals into which counts are binned) was given as 1000, and the results were globally normalised, so that the enrichment as plotted on the histogram represented the value relative to the background level of binding for the genome (mm9). As such, the value on the histogram effectively represents the fold binding increase vs genome background. The raw data used to build the histograms were entered into Prism, and plotted. For clarity, only factors with a fold enrichment of five or greater were included in the plots.

2.2.9 – Chromosomal distribution of ERVs

Chromosomal distribution of ERVs was calculated by deriving murine chromosome lengths using bash script. The fraction of each chromosome of the genome was calculated (by dividing the chromosome size by the total genome size). The expected number of ERV integrations per chromosome was calculated by multiplying the total number of ERV integrations with the chromosome fraction, as done in a previous study for VL30 (**Markopoulos et al, 2016**). The chi-squared test was used to analyse the chromosomal distribution of VL30 elements by comparing the observed with the expected number of Murine ERV elements, assuming a random insertion model.

2.2.10 – Orthologous ERV and EVE integrations

Sequences for which >100bp of unambiguous (i.e. >80% identity) genomic flanking sequence were identified in sister taxa to the host were considered orthologous. In addition, DOTTUP was used to create dotplots of aligned genomic regions. The dotplots were used to visualise flanking genomic regions to in supporting evidence of orthology. For example, where ambiguous genomic flanking regions existed but the alignment of the greater

genomic context supported orthology. Species divergence times were extracted from the Timetree database. (**Hedges *et al*, 2015**).

2.2.11 – CVe sequence analyses

CVe were considered endogenous where >100bp of flanking sequence could be identified. The presence of sequence degradation (following lengthy residence in the host germline) was used as supporting evidence.

2.2.12 – PCR of CVe-Pseudomyrmex – performed by Peter Flynn

Genomic DNA was extracted from ant tissue samples following the Moreau protocol (Moreau, 2014) and a DNAeasy Blood & Tissue Kit (Qiagen). PCR amplification of CVe-Pseudomyrmex was performed using two sets of primer pairs designed with Primer3 (http://bioinfo.ut.ee/primer3-0.4.0/), each comprising one primer anchored in the CVe sequence, and another anchored in the genomic flanking sequence. Primer Pair 1 amplified a sequence that was 694 bp long and Primer Pair 2 amplified a sequence that was 286bp long. Primers were tested using illustra PuReTaq Ready-To-Go PCR Beads (GE Lifesciences). A temperature gradient PCR was performed to assess the optimum annealing temperature for the specific primer pairs. PCR was then performed using the genomic DNA ant extractions. The PCR conditions for this run were: an initial denature stage of 5 minutes at 95°C, 30 cycles of 30 seconds denaturing at 95°C, 30 seconds annealing at 49.7°C for Primer Pair 1 and 62°C for Primer Pair 2, and an extension at 72°C for 1 minute, then after 30 cycles a final extension at 72°C for 5 minutes. Each run included a negative control. Amplification products (800-1000bp) for each PCR reaction were excised and run on agarose gels. Bands of the expected size were excised, purified and sequenced via Sanger sequencing.

2.2.13 – ISG screen design

The ISG reference libraries were compiled on an individual basis – for each gene, all known human paralogs and mammalian orthologs were downloaded, suffixed with their common gene name, so that the FASTA ID consisted of ">ENSEMBL-ID_COMMON-NAME". The probe for each gene

consisted of the longest human isoform of each gene, retrieved from ENSEMBL using BioMart. The genes consisted of the 'core mammal' ISG set as described by (**Shaw et al, 2017**). (**See chapter 5 for more details**). Each mammalian genome as listed in Appendix 2.1 was screened using DIGS. Each individual gene was allotted an individual database. Unlike previous screens, the searches did not take on an iterative 'phylogenetic screening approach', and ceased after the first round of DIGS screening.

2.2.14 - ISG data processing and display

Each gene was subjected to a bitscore cutoff that returned all exons of that gene from the human genome. The data were normalised to the median count for each gene in each species, to get a normalised count in each species for each gene. The resulting normalised data were saved as a tab-delimited text file. To generate a heatmap of normalised ISG counts across Mammalia entered into gitools (version 2.3.1) (http://www.gitools.org/). The data were ordered on the Y axis by 'tree order' as established through Timetree (**Hedges et al, 2015**). The linear colour scale (Parameter = Value = Colour code) was set as Minimum = 0.1 = FEFDFD, Middle = 5 = FEEE00, Maximum = 10 = FF0000, Empty = NULL = A5A4A4. Spurious results for HLA were deleted from the heatmap. Additionally, the data table was entered into Tableau Professional Edition and published in Tableau Public for interactive visualisation at (https://public.tableau.com/profile/tristan.dennis#!/vizhome/ISG_0/Sheet1?publish=yes)

2.2.15 - Shared genomic organisation

Shared genomic organisations were studied using the UCSC and NCBI genome browsers in concert with BLAST and BLAT. Query ERV and ISG sequences were BLASTed against host genomes, and the location of the resulting hits opened in the corresponding genome browser. In addition, Genomicus was used to survey gene organisation across multiple mammalian species.

Chapter 3 – Discovery and characterisation of murine endogenous retroviruses

3.1 – Introduction

Retroviruses (family *Retroviridae*) are enveloped viruses that infect vertebrates, causing neoplastic and immunosuppressive disease. All retroviruses are characterized by a replication strategy in which the viral RNA genome is converted to DNA and stably integrated into the genome of the host cell (a form referred to as 'provirus'). Retroviral infection of germline cells (i.e. sperm, eggs or early embryo) can lead to vertical inheritance of proviral loci as host alleles termed endogenous retroviruses (ERVs). Mammalian genomes typically contain thousands of ERV loci, reflecting a long-term co-evolutionary relationship with retroviruses (**Gifford *et al*, 2003**).

ERV sequences in mammalian genomes typically group into phylogenetically distinct lineages (sometimes referred to as 'families') that are thought to have arisen from a small number of germline colonisation events in which integration of proviral sequences into the germline has been followed by copy number expansion, either through reinfection of germline cells, or retrotransposition within them (**Ribet *et al*, 2008**; **Belshaw *et al*, 2007**). A subset of ERV insertions have been genetically fixed in the host germline, and these sequences constitute a kind of genomic 'fossil record' from which the long-term evolutionary history of retroviruses can be inferred. In addition, recent studies have demonstrated that ERVs sequences have often been co-opted or exapted by host genomes, and this has exerted a profound impact on mammalian evolution and biology (**de Parseval and Heidmann, 2005; Gifford *et al*, 2013**). Characterising ERVs can shed light on the biology of ancient retroviruses, and reveal insights into the co-evolutionary processes through which ERVs have shaped host genomes.

The mouse (*Mus musculus musculus*) is a small mammal in the diverse order Rodentia. The mouse is one of several mammalian species that have formed a synanthropic relationship with humans (**Boursot *et al*,1993**), and as

a result have been introduced to every continent barring Antarctica. The small size and short generation time of the mouse have also led to its widespread adoption as a model organism for mammalian biology, disease and genetics, and has led to the production of a richly annotated reference genome sequence. The genome of the C57BL/6J mouse was the second vertebrate genome to be sequenced after that of the human (**Waterston *et al*, 2002**). In addition to a high-quality reference sequence, numerous functional genomics, and other kinds of genetics studies have been performed on the mouse, and many inbred variants have been created. In addition, the genomes of additional inbred mouse strains, and other species in genus *Mus* have recently become available (**Adams *et al*, 2015**).

The mouse reference genome was generated via Sanger sequencing, reducing biases against the inclusion of repetitive elements (like ERVs) in the assembly (as exist for reference assemblies based on short-read sequencing). Sanger, and long-read HTS technologies generate long stretches of sequence that often overlap the repetitive element, allowing it to be placed in the genome assembly. By contrast, short-read sequencing technologies generate reads for repetitive sequences (such as ERVs) that may map to many regions of the genome (if the ERV is commonplace), making it difficult to place with confidence into the finished assembly. (**Gordon *et al*, 2016; Keane, 2016, personal communication**). A large amount of information on host:ERV co-evolution has been gathered through *in silico* and translational studies in the mouse, especially regarding the regulatory role of ERVs in development and immunity (**MacFarlan *et al*, 2012; Rowe *et al*, 2013**; **Chuong *et al*, 2016; Schmid *et al*, 2010**). Despite this, the discovery and characterisation of murine ERV lineages is largely based on well-characterised lineages identified through hybridisation and *in vitro* studies – for MLV and GLN, for example (**Frankel *et al*, 1989; Itin and Keshet, 1986**), or through single-lineage analysis using similarity searches (**Bromham *et al*, 2001**; **Xhao *et al*, 1999**). Such investigations have focussed largely on individual lineages. Large scale investigation of ERVs in the murine genome has usually involved using approaches that characterise only intact loci (**McCarthy *et al*, 2004**), or has used the RepBase database of repetitive elements (**Jurka *et al*, 2005**). RepBase, although comprehensive, uses numerical designations and classification that do not necessarily reflect the evolutionary provenance of

the element in question (**Mager and Stoye, 2015**). There is therefore scope for the discovery and evolutionary analysis of ERVs in the mouse using a 'phylogenetic screening' approach. This approach, in *Homo sapiens,* identified numerous ancient ERV lineages (**Tristem *et al*, 2000**). Thus, the aim of this chapter was to identify novel murine ERV lineages through a comprehensive survey of murine ERV diversity using DIGS. Subsequently, I wished to analyse recovered murine ERV sequences to make insights into host:ERV co-evolution.

3.2 - Results

3.2.1 - Assessment of murine ERV diversity using phylogenetic screening

In order to assess murine ERV diversity, I used DIGS to perform phylogenetic screening with ERV RT peptides. To gain an initial estimate of the number of distinct ERV lineages in *Mus musculus*, the clustering of aligned murine ERV RTs (**Appendix 3.1)** relative to exogenous retroviruses (XRVs) and ERVs from non-murine species was examined. ERV lineages arise when germline invasion is followed by ERV copy number increase through retrotransposition or reinfection. Since most members of an ERV lineage are separated by relatively few rounds of viral replication, they tend to group relatively closely with one another in phylogenies (**Tristem *et al*, 2000**). Therefore, well-supported clades of closely-related murine ERVs that are separated from one another in RT phylogenies by ERVs or XRVs that occur in non-rodent hosts can generally be assumed to have arisen in independent germline invasion events. Based on the RT phylogeny, thirteen distinct ERV lineages can be distinguished in the *Mus musculus* genome (**Figure 3.1**). The RT phylogeny disclosed three main clades (I-III), corresponding to the three major divergences in the evolution of *Retroviridae* (**Llorens et al, 2008**). Endogenous representatives of each clade were identified in the murine genome. Seven of the thirteen ERV RTs fall within the established diversity of two exogenous genera, *Gammaretrovirus* (n=4) and *Betaretrovirus* (n=3) (**Figure 3.1)**. The remaining six represent retroviruses for which closely-related exogenous representatives have not yet been identified.

Twelve of the seventeen ERVs identified by my analysis have been reported previously. Eight of the twelve previously reported ERVs grouped robustly within the clades defined by exogenous Gamma- and Betaretroviruses. The remaining four include two groups (IAP and MysERV) that cluster with clade II (*Betaretrovirus*-related) ERVs, one (MuERV-L) that clusters with clade III (*Spumavirus*-related) ERVs, and one (MuERV-Fc) that clusters with the basal clade I (*Gammaretrovirus*-related) ERVs. Five of the ERVs identified in this analysis have not previously been described in mice. Novel and previously described murine ERVs were tabulated in **Table 3.1**. In total, I identified 4297 RT sequences in the murine genome.

Table 3.1: Murine ERV RT counts retrieved by phylogenetic screening of the murine (mm10) genome. RT counts are broken down by genus, with the lineage, accession (if applicable) and first description (if applicable).

Genus	Lineage	Accession	First described	RT
Lambda	MuERV-L	Y12713	Cordonnier et al, 1995	846
Sigma	Mus.Sigma.1	This study	This study	1
Zeta	MuERV-Fc	This study	Diehl *et al*, 2016	118
Iota	Mus.Iota.1	This study	This study	1
Gamma	MMERV	AC005743	Bromham et al, 2000	546
Gamma	MuRRS	X02487	Schmidt et al, 1985	209
Gamma	GLN	AC136922	Itin and Keshet, 1986	93
Gamma	MLV	NC_001501	Gross, 1953	60
Gamma	MURVY	X87639	Hutchison and Eicher, 1989	56
Gamma	Mus.Gamma.1	This study	This study	3
Gamma	MuERV-C	AF049340	Zhao et al 1999	**3**
Gamma	Mus.Gamma.2	This study	This study	1
Beta	Mus.Beta.1	This study	This study	18
Beta	MMTV	M15122.1	Green et al 1946	18
Beta-like	IAP	A	Dalton et al 1961	1923
Beta-like	MusD	AF246633	Mager et al 2000	341
Beta-like	MYSERV	NT_165681	Wichman et al 1985	60

I named new sequences based on the genus, established or provisional (**Gifford, unpublished**) in which they grouped. Three previously unreported ERVs were identified that grouped robustly within the clades defined by exogenous Gamma- and Betaretroviruses. The two novel gammaretroviruses were named Mus.Gamma.1 (Mus.g1) and Mus.Gamma.2 (Mus.g2), and the novel betaretrovirus as Mus.Beta.1 (Mus.b1). Additionally, two new ERVs were reported that did not group within the established diversity of exogenous genera. These were one clade I (*Gammaretrovirus*-related) ERV: Mus.Iota.1 (Mus.i1, HERV-I related), and one clade III ERV, referred to here as Mus.Sigma.1 (Mus.s1, HERV-S related). Additionally, I identified a murine ERV

grouping robustly with ERV-Fc. (**Figure 3.1**). Initially, I called it Mus.Zeta.1, and referred to it during this analysis. Upon publication of a study of ERV-Fc evolution in mammals by **Diehl *et al*, 2016**, I found that RT sequences recovered in their study matched those I recovered for Mus.Zeta.1. Thus, I changed the name of Mus.Zeta.1 in this study to MuERV-Fc in order to match their nomenclature.

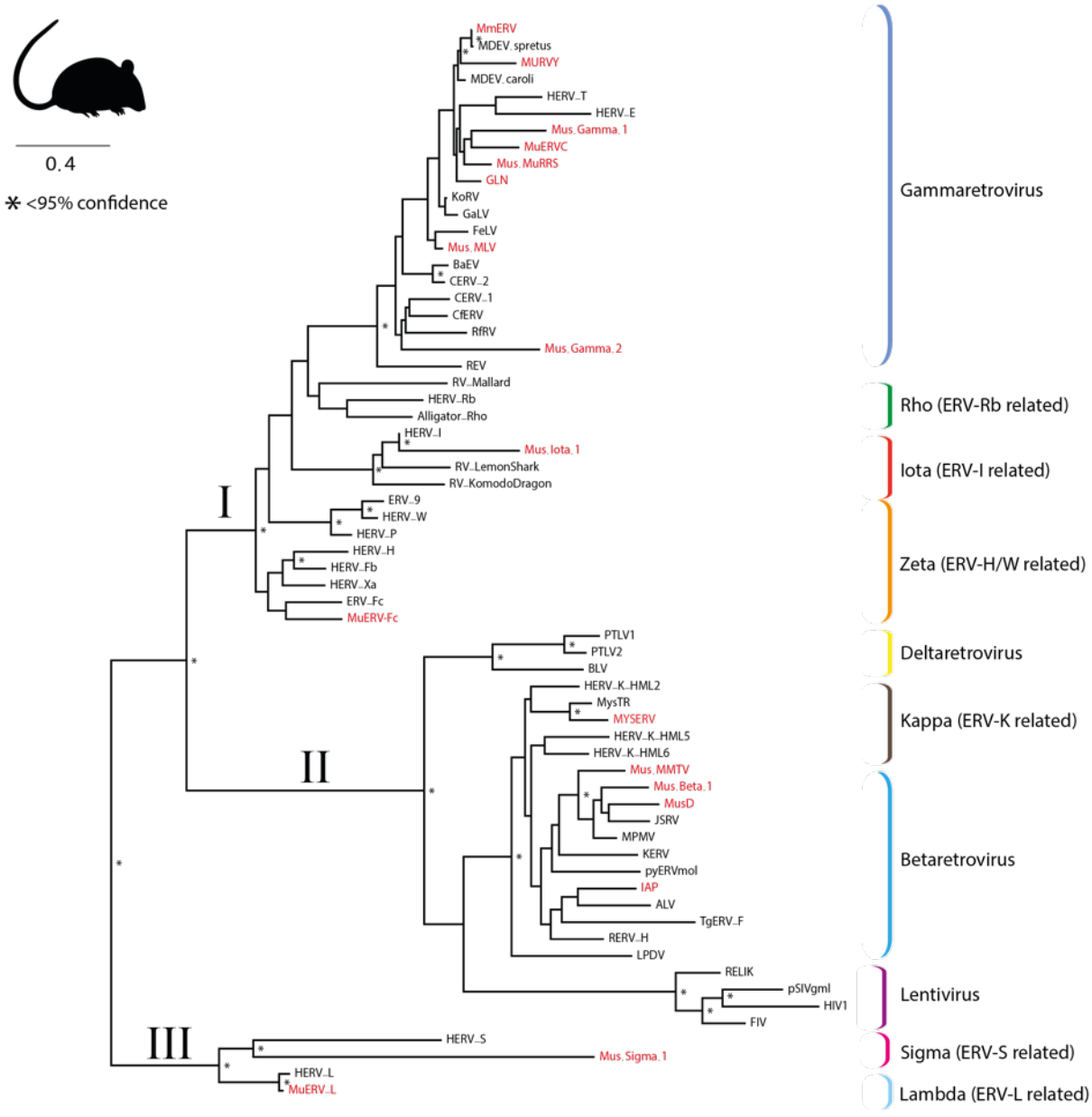

Figure 3.2.1 – The mouse genome contains 13 phylogenetically distinct ERV lineages. Phylogenetic relationships between exogenous and endogenous retroviral RT peptide sequences. The ML tree was produced under the rtREV+G4 model, and support evaluated using 1000 ultrafast bootstrap (UFBOOT) replicates. UFBOOT support of 95% indicated by asterisk (*). Scale bar indicates substitutions per site. Coloured brackets indicate established and provisional retroviral genera. Murine ERV lineages are highlighted in red.

3.2.2 - Identification of additional murine ERV lineages

In the process of recovering representative proviral genome structures I determined that some of the thirteen lineages identified by our initial, RT-based analysis were likely to be derived from two or more independent germline invasions by relatively closely-related viruses, rather than a single, discrete event. This was indicated firstly by RT phylogenies, which revealed well-supported subclade structures within murine ERV lineages, and secondly by LTR sequences, which often differed (i.e. could not be aligned) between members of the subgroups.

Longer (~750 amino acid) Pol sequences were derived from the representative genomes, aligned, and used to reconstruct the evolutionary history of the *Gammaretrovirus* and *Betaretrovirus* genera. The resulting phylogenies were used in concert with LTR sequences to estimate the number of distinct ERV lineages. ERVs with different LTRs were assumed to be derived from different germline invasion events.

The *Gammaretrovirus* Pol phylogeny (**Figure 3.2**) displayed five distinct murine ERV lineages. Murine ERVs clustered robustly in a clade containing predominantly murine gammaretroviruses. Three non-murine lineages emerged from the clade of predominantly murine viruses (KwERV, PERVA/C and KoRV/GaLV). (**Figure 3.2**)

Occupying the most derived position in the tree was the monophyletic sub-clade containing MMERV, *Mus spretus* endogenous retrovirus (MspERV), *Mus dunni* endogenous retrovirus (MdEV), *Mus caroli* endogenous retrovirus (McERV) and MURVY). These ERVs are derived from four species of mouse. These sequences, with the exception of MURVY, follow the host phylogeny for genus *Mus* (**Figure 3.2**). This pattern suggests that this clade is either: a series of orthologous ERVs (**Johnson and Coffin, 1999**), or a family of recently exogenous XRVs that have co-diverged with genus *Mus*. The latter is lent support by the fact that these ERVs have similar genome structures but distinct LTRs. This suggests that either: this clade is derived from multiple

ancestral integration events or that a recombination event took place between the ancestor of these lineages and another LTR containing element, prior to the speciation of *M. musculus*, *M. spretus* and *M. dunni*, but after the divergence between them and *M. caroli*. (this has been suggested by **Wolgamot *et al*, 1998** as the origin of the VL30 retrotransposon).

The second major sub-clade was separated robustly from the MMERV/MURVY containing clade by KoRV and GaLV, and contains GLN, Mus.g2, MuRRS, MuERV-C, and PERV-A/C. LTRs were not identified for MuERV-C. MuRRS, GLN and Mus.g2 had different LTRs. These murine ERVs were considered to have derived in the mouse from at least three germline invasion events. (**Figure 3.2**).

Mus.g1 occupied the most basal position in this large clade, grouping on a long branch with KwERV. Similarly, MLV was the basal-most murine ERV, separated from the large, murine ERV containing clade, by primate gammaretroviruses RD114, baboon endogenous virus (BaEV) and chimpanzee endogenous retrovirus 2 (CERV-2).

Based on the clustering of murine ERVs with one another, and differences in LTR sequences between these ERVs, murine gammaretroviruses were considered to consist of five phylogenetically distinct ERV lineages, derived from at least six germline invasion events.

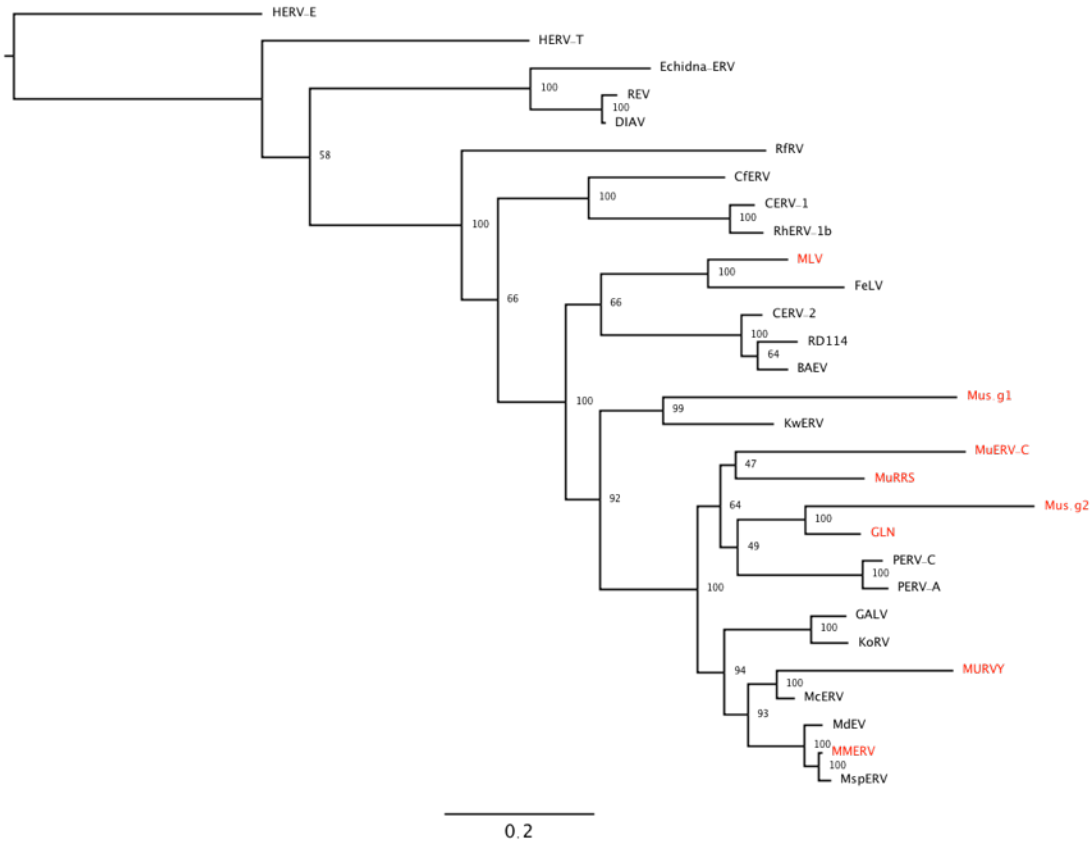

Figure 3.2.2: **Phylogenetic relationships between murine and nonmurine gammaretrovirus Pol sequences.** ML phylogeny was inferred using the rtREV+G4 model, and evaluated using 1000 UFBOOT replicates, with support of >95% indicated by an asterisk. Scale bar indicates substitutions per site. Murine ERV lineages are highlighted in red.

The *Betaretrovirus* Pol phylogeny (**Figure 3.3**) showed five betaretroviral murine ERV lineages. Of these, four group within *Betaretrovirus*– Mus.b1, MusD, IAP and MMTV, and one (MYSERV) groups with HERV-K-HML2. In the tree, murine sequences tended to group basally relative to established non-murine retroviruses, with IAP and MMTV grouping basally in the *Betaretrovirus* genus. All of these taxa possess different LTRs, indicating that these lineages derive from at least five independent germline invasion events.

The RT phylogeny discloses at least 13 murine ERV lineages. When the *Gammaretrovirus* Pol phylogeny, and LTR sequences are taken into account, these appear to be derived from at least 15 germline invasion events.

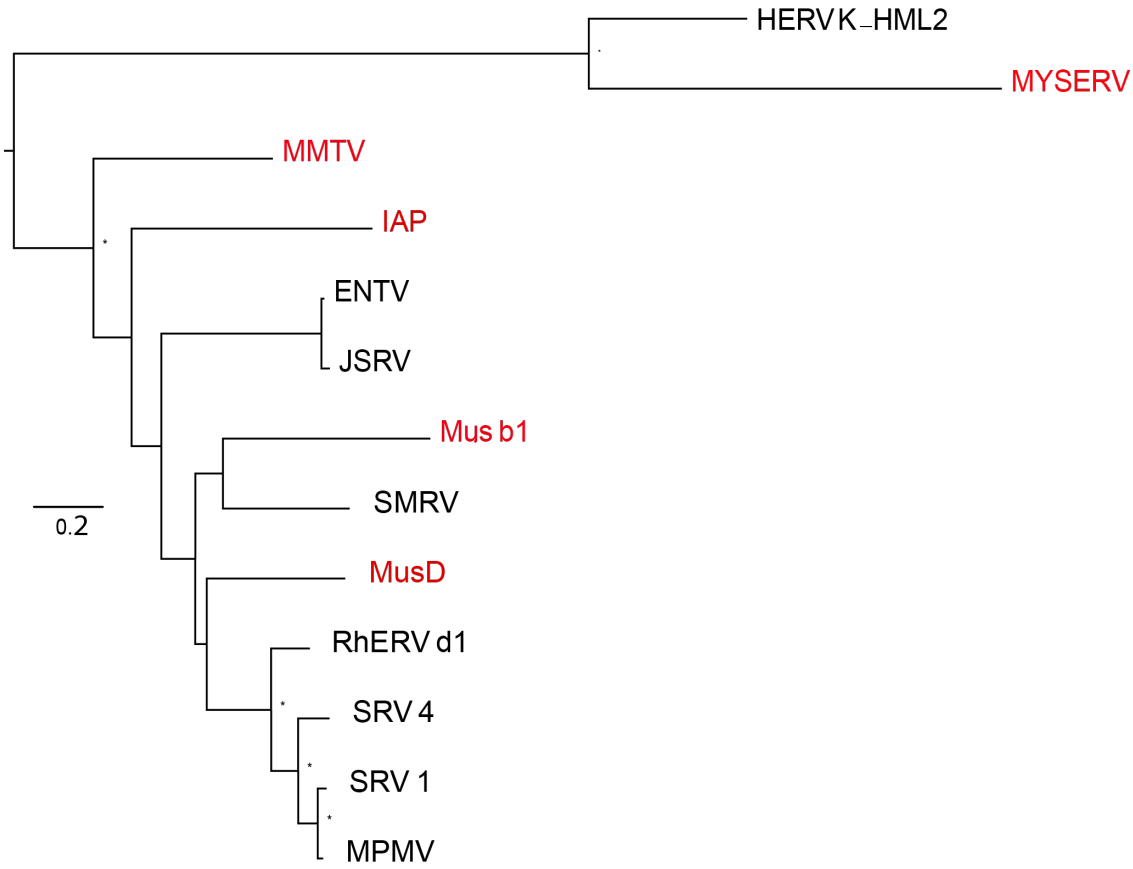

Figure 3.2.3: **Phylogenetic relationships between murine and nonmurine Betaretroviruses.** The ML phylogeny was inferred using the rtREV+G4 model. Support evaluated by 1000 UFBOOT replicates. >95% indicated by an asterisk. Scale bar indicates substitutions per site, and murine ERV lineages are highlighted in red.

3.2.3 - Evolutionary history of Gag and Env

Retroviruses are prone to recombination. Different ORFs within a retroviral lineage may therefore be differently related to one another. Phylogenies of all retroviral peptides can reveal histories of recombination in ERV lineages. (**Diehl *et al*, 2016; Henzy and Johnson, 2013**). I aligned Gammaretroviral Gag and TM peptides and used them to generate ML phylogenies. I used the TM subdomain of *env* due to its relative sequence conservation (**Benit *et al*, 2001**).

Aligning Betaretrovirus Gag to Gammaretroviral Gag proved to be problematic due to sequence divergence, so I analysed only *Gammaretrovirus*-derived Gag sequences. Additionally, most murine Betaretroviral ERVs do not possess an *env* gene, so only Gammaretroviral TM peptides were analysed.

The Gammaretrovirus Gag phylogeny showed a similar topology to Gammaretroviral pol (**Figure 3.4; Figure 3.2**). The main lineage subclades remained relatively constant, and weak support at deeper nodes within the Gag tree made its evolutionary history difficult to confidently determine. Interestingly, KoRV and GaLV split the MMERV containing clade (albeit with limited support ~60%), indicating that recombination has taken place between different members of this lineage, and that the ancestor of KoRV and GaLV lies in this lineage.

The TM phylogeny disclosed two major clades. The first contained MPMV and RD114, and represented the gamma-like *env* peptides of Betaretroviruses. (**Henzy and Johnson, 2013**). (**Figure 3.5**). The second major clade was split into two robustly supported subclades (**Figure 3.5**). The first contained a TM lineage belonging to PERV, KoRV and GaLV, and the MMERV and Mus.g1 lineages. For these ERVs, the TM phylogeny disclosed a similar topology formed as for Gag and Pol, with the absence of KoRV/GaLV and PERV, suggesting that *env* shares a similar evolutionary history with Gag and Pol (**Figure 3.2; Figure 3.4; Figure 3.5**). The second TM lineage contained MLV, Mus.g2 and GLN, with Mus.g2 and GLN grouping with RfRV, a retrovirus of bats, suggesting that the GLN and Mus.g2 *env* is more closely related to RfRV than to other murine ERVs.

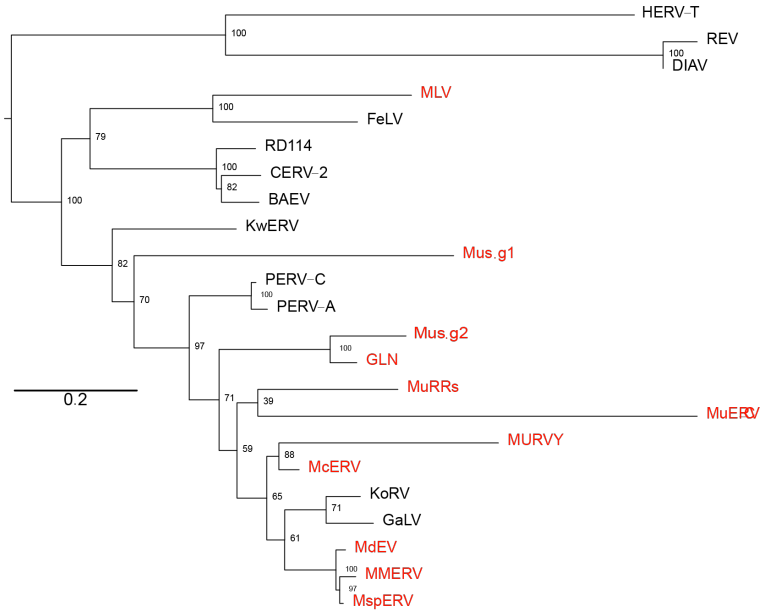

Figure 3.2.4: **Phylogenetic relationships between murine and nonmurine gammaretrovirus Gag sequences.** ML phylogeny was inferred using the rtREV+G4 model, and evaluated using 1000 UFBOOT replicates, with % support indicated at nodes. Scale bar indicates substitutions per site. Murine ERV lineages are highlighted in red.

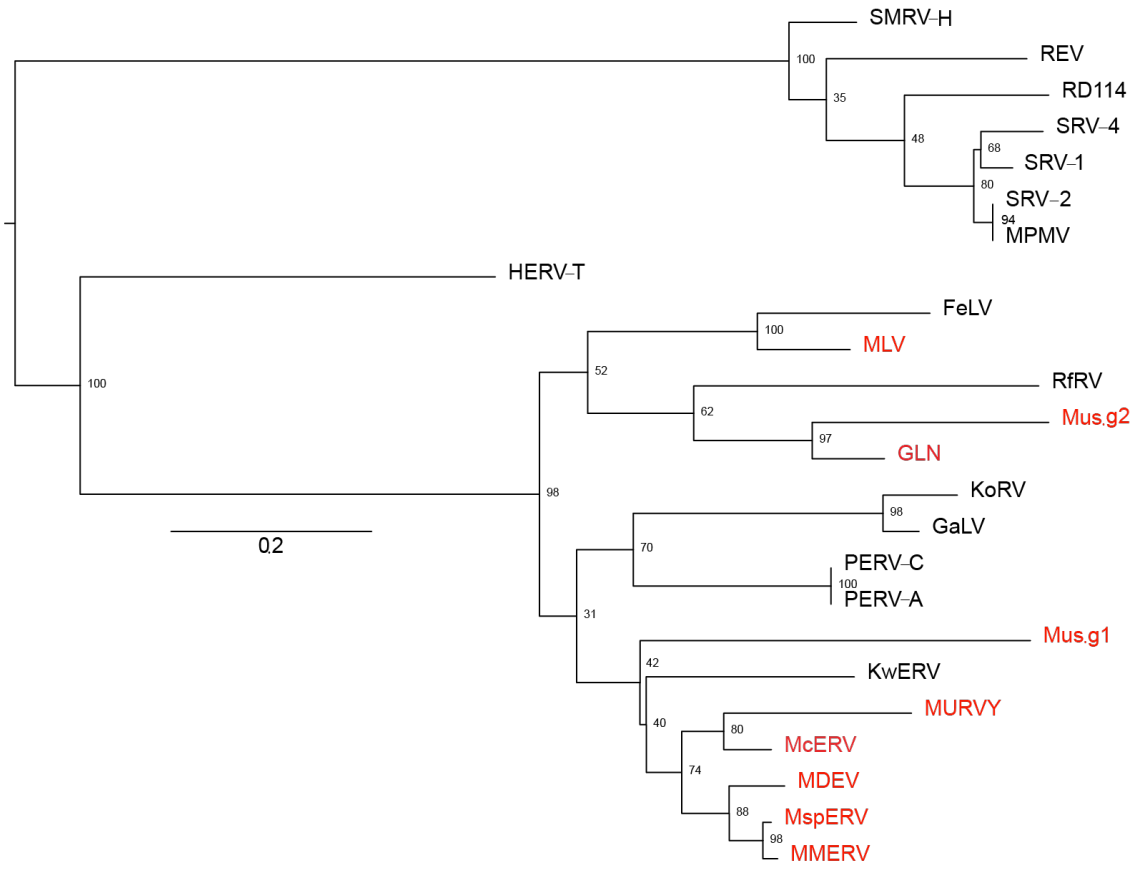

Figure 3.5: **Phylogenetic relationships between murine and nonmurine gammaretrovirus TM sequences.** ML phylogeny was inferred using the rtREV+G4 model, and evaluated using 1000 UFBOOT replicates, with % support indicated at nodes. Scale bar indicates substitutions per site. Murine ERV lineages are highlighted in red.

3.2.4 - A survey of murine ERV counts

Having obtained an overview of murine ERV diversity, I then used DIGS to assess ERV provirus and solo LTR copy number in the murine genome. A secondary aim of this was to collect murine ERV sequences into high-quality alignments for sequence evolutionary analysis, and annotation 'track' generation for interpolation with functional annotation datasets. I used murine ERV representative sequences to generate libraries of translated ERV gene and LTR sequences. These libraries were used in DIGS screens of the murine genome. (**see Methods**). The counts for proviruses and solo LTRs were collated and compared with results from previous analyses. (**Table 3.2**).

A total of 4113 proviruses and 16424 solo LTR elements were recovered. Consistent with previous estimates, I found that class II ERVs made up the highest copy number of murine ERVs (n=2509), consisting largely of IAP

(n=1923) and MusD (n=102), MuERV-L (class III, n=658) had the second highest provirus copy number, followed by MMERV (class I, n=305). (**Table 3.2**).

Generally, solo LTRs outnumbered proviruses. Exceptions to this were MLV and MMTV. This perhaps reflects their status as 'endogenising' ERVs – they have not spent enough time in the murine germline to generate a large number of solo LTRs. Additionally, I did not find any solo LTRs for MURVY – a consequence of its replication solely as part of a duplicating element on the Y chromosome. The ratio at which solo LTRs outnumbered proviruses varied. At the most, Mus.g2 solo LTRs outnumbered proviruses by 609:1. Generally, this ratio was less in 'intermediate' copy number lineages (MuRRS, MusD and GLN proviruses were outnumbered by solo LTRs at ratios of 10:1, 5:1 and 11.9:1 respectively). The highest copy number lineages showed the highest ratio of solo LTRs:proviruses LTRs, with IAP, MuERV-L and MMERV at approximately 4:1, 3:1 and 1:1 respectively **(Table 3.2).**

When mapping ERV counts from other studies onto the counts shown in **Table 3.2**, I found that in every lineage except for MURVY, the provirus counts from this study matched the results from other studies – in the case of MMTV. I excluded IAP from these comparisons, as it has been shown to be highly polymorphic, with copy number varying markedly across different murine genome assemblies (**Ray *et al*, 2011**). Furthermore, manually curating such a large volume of sequences was not feasible.

Table 3.2: Murine ERV proviruses and solo LTRs. Murine ERV proviruses and solo LTRs were retrieved using DIGS and broken down by genus. Additionally, estimates of copy number from published literature are listed on the right hand sides

Genus	Lineage	Provirus	Solo LTR	Previous estimate of prov. copy number
Lambda	MuERV-L	654	1930	50 (Costas *et al*, 2003)
Sigma	Mus.Sigma.1	0	0	n/a
Zeta	MuERV-Fc	1	0	n/a
Iota	Mus.Iota.1	0	0	n/a
Gamma	MMERV	305	173	191 (Lee *et al*, 2012)
Gamma	MuRRS	35	1962	50-100 (Schmidt *et al* 1985)
Gamma	GLN	53	423	50 (Ribet *et al*, 2008)
Gamma	MLV	66	3	54 (Frankel *et al*, 1989
Gamma	MURVY	206	0	500 (Hutchison and Eicher, 1989)
Gamma	Mus.Gamma.1	1	49	n/a
Gamma	MuERV-C	1	0	1 (Zhao *et al*, 1999)
Gamma	Mus.Gamma.2	1	609	n/a
Beta	Mus.Beta.1	1	0	n/a
Beta	MMTV	3	0	4 (Kozak *et al*, 1987)
Beta-like	IAP	2402	6356	polymorphic
Beta-like	MusD	102	893	100 (Ribet *et al*, 2007)
Beta-like	MYSERV	1	0	n/a

3.2.5 - Analysis of novel ERVs

To further characterise novel and partially characterised ERV lineages, I aligned their proviruses to phylogenetically related reference sequences, (i.e. Mus.g1 was aligned to MLV) and sequence comparisons were performed to identify putative ORFs and retroviral genomic features. Interestingly, I identified 12 novel MLV loci, and these were included in my analysis. Additionally, the MuERV-Fc lineage was characterised in greater detail. All of the loci I recovered for these lineages were (apart from MLV) highly degraded, encoding no intact ORFs. As such, I was unable to identify translation strategies. Of the five novel lineages, four possessed identifiable retroviral genes. The remaining two, Mus.i1 and Mus.s1, only possessed fragmented RT sequences.

Mus.g1

I found a single Mus.g1 locus in the genome of the mouse. I was able to identify *gag, pol* and *env* genes that were interrupted by stop codons and frameshifts, as well as flanking LTRs 449bp in length. We could not find any TSDs, presumably due to mutational degradation of the Mus.g1 locus. (**Figure 3.6**), but were able to detect PBS^{PRO} in the region adjoining the 5'LTR and Gag. I was able to recover a partial Gag ORF containing the 5' myristoylation signal of MA, suggesting that Mus.g1 was capable of forming particles upon integration. I characterised features in the Mus.g1 genome that were similar to those of contemporary gammaretroviruses. Firstly, the Gag peptide possessed a single NC zinc finger motif (**Figure 3.6**), as well as a single late domain (**Figure 3.6**). In addition, I found a GPHF motif toward the C-terminal of integrase, as well as a CXCC disulphide motif in *env,* characteristic of gammaretroviruses (**Jern *et al*, 2005**). (**Figure 3.6**)

Mus.g2

Mus.g2, like Mus.g1, possessed a single provirus. I identified fragmented sequences bearing homology to *gag, pol,* and *env* flanked by two relatively short (220bp) LTRs, themselves flanked by 4bp TSDs. **(Figure 3.6**). The 4bp TSDs are indicative of a gammaretrovirus (**Ballandras-Colas *et al*, 2013**). Like Mus.g1, I identified the 5' Gag myristoylation signal (MG) and a patch of basic amino acids at the 5' end of MA, suggesting that this Mus.g2 locus was capable of forming particles upon integration.

Typically for a gammaretrovirus, the Gag ORF encoded a single NC zinc finger, and a single late domain of identical specificity to Mus.g1 (PPPY) (**Figure 3.6**). The Mus.g2 locus also possessed a fragmented IN domain, in which the gammaretroviral GPYL motif could be identified. (**Figure 3.6**). I could also identify a CXCC motif in *env.* (**Figure 3.6**).

MuERV-Fc

MuERV-Fc is class I ERV, also identified in mice by **Diehl *et al*, 2016**, who found degraded Gag sequences, and generated an RT consensus (included in **Figure 3.1**).

I recovered a single, highly fragmented MuERV-Fc provirus from the murine genome. Investigation of the MuERV-Fc sequence revealed sequences homologous to Gag, Pol and Env, but no LTRs. Additionally, I was only able to identify the TM domain of Env. I found two zinc fingers in MuERV-Fc NC, characteristic of non-primate ERV-Fc (**Diehl *et al*, 2016**), as well as a single 'YPRL' late domain. I was also able to find a 5' Gag myristoylation site (**Figure 3.6)**. Additionally, I identified a GPYL motif in IN, a c-terminal zinc finger, and a CXCC motif in *env*.

MYSERV

MYSERV is the name of a RepBase consensus sequence. It is marked by its close relationship to the mysTR family of ERVs undergoing retrotranspositional expansion in other rodent species (**Cantrell *et al*, 2005**). I identified a locus with sequences bearing homology to *gag* and *pol*. These sequences were highly degraded, with numerous stop codons, frameshifts and a truncated *pol* sequence, limiting my ability to characterise the ancestral ORFs of MYSERV. Typically for a class II retrovirus, I identified two zinc fingers in NC. (**Figure 3.6**). Additionally, I identified a fragmented ORF similar to *pol,* bearing homology to RT, RNaseH and a c-terminal IN zinc finger.

MLV

I identified 12 novel MLV loci. (**Table 3.3**). Alignment of these loci to an MLV reference sequence revealed that two of these loci encoded complete proviruses (**Table 3.3**). The remaining ten proviruses had deleted *env* ORFs, of which six had fragmented Gag ORFs, and two had deleted sections in IN. (**Table 3.3**). These loci have remained undiscovered by previous searches of the murine genome for MLV (**Jern *et al*, 2007**; **Frankel *et al*, 1989),** conceivably because these studies have used the *env* ORF as a probe for *in vitro* and *in silico* searches.

Mus.b1, Mus.i1 and Mus.s1

For Mus.b1, I was only able to recover a fragmented Pol peptide. I could recover even less for Mus.i1 and Mus.s1: only finding RT sequences (**Figure 3.6**).

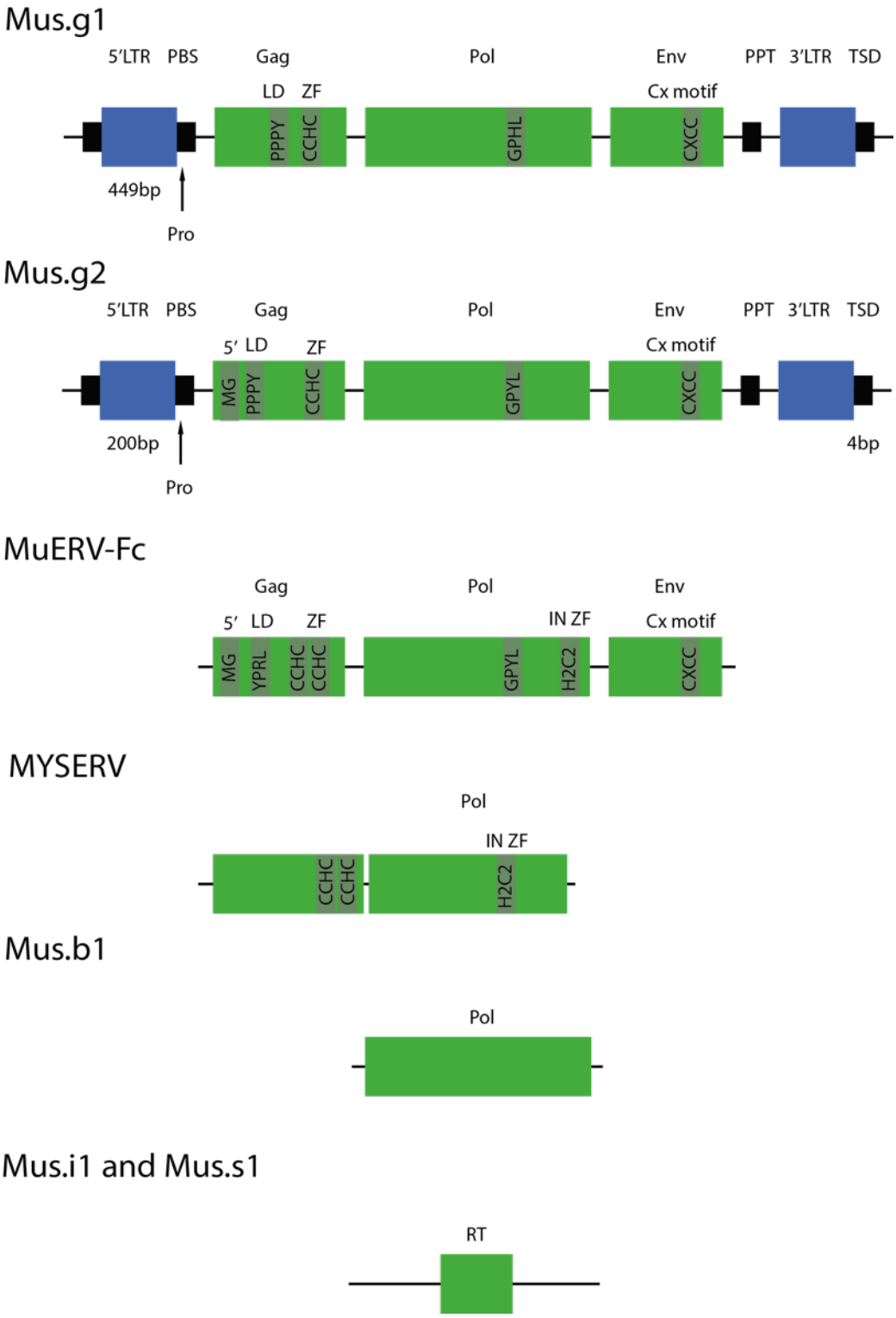

Figure 3.6 – Genome structures of novel murine ERV lineages. Proviruses of novel and partially characterised lineages were extracted from the murine genome and analysed for distinguishing sequence features to assist in phylogenetic and sequence characterisation. Lineage name indicated on the left. Retroviral LTRs indicated by blue bars, and genes by labelled green bars. Shaded regions indicate retroviral sequence motifs as labelled.

Table 3.3 – **Novel MLV proviruses**. Indicates the chromosomal location and distinguishing features of novel murine ERV proviruses.

Scaffold	Start	End	Features
chr10	8542060	8553473	Stop codon in gag, no *env*
chr11	60576721	60588134	Deleted *env*
chr12	54772056	54783499	Intact provirus
chr13	21810284	21821697	Deleted *env*
chr19	38372853	38384266	Stop codon in gag, no *env*
chr4	15233494	15244937	Deletions in *env* and *pol*
chr4	32788395	32799838	Stop codon in gag, no *env*
chr5	122190722	122202135	Stop codon in gag, no *env*
chr6	73284833	73296276	Stop codon in gag, no *env*
chr8	123162591	123173995	Deleted *env*
chr8	123425408	123436851	Intact provirus
chr8	85123082	85134525	Deletions in *env* and *pol*

3.2.6 - Evolutionary analysis of murine ERVs

I used the sequences collected in our screens to generate Pol and solo LTR alignments. I used these alignments to generate phylogenies of ERV lineages. (**Figure 3.7**). Additionally, I mapped PBS specificity as a character onto the phylogenies in an attempt to find instances of PBS switching, which could indicate periods of exogenous replication, or an attempt to adapt to the tRNA abundancies in different murine cell types. (**Tristem *et al*, 2000; Li *et al*, 1994; Whitcomb *et al*, 1995**).

I examined the topologies of the phylogenies and used previous phylogenetic analysis of ERV lineages, and literature regarding the inference of past processes from ERV phylogenies (**Katzourakis *et al*, 2005**) to make inferences as to the past and present activity of murine ERVs. Generally, I found that ERV *pol* phylogenies formed star-like or comb-like topologies. The star-like phylogenies possessed flat, deep branching nodes (indicative of ancestral expansion), followed by long terminal branches (indicating lengthy residence in the host germline). The comb-like phylogenies possessed long

basal branches, with derived clades emerging with shorter branches and lower intra-clade divergence, indicating a subset of potentially active ERV loci (**Katzourakis *et al*, 2005**). I found that demonstrably active (transposing or reinfecting) ERV lineages displayed comb-like topologies in midpoint-rooted phylogenies (**Figure 3.7A-D**), with intact, *env* encoding ERV loci clustering in derived positions (**Figure 3.7** – green dots). Examples of these are MusD, GLN and MuERV-L. MMERV showed a striking two-clade structure, both of which showed comb-like topologies. The two clades were marked by different PBS specificities, with clade one possessing PBS^{GLY} and clade two possessing PBS^{PRO} (**Figure 3.7B**). Additionally, these clades were divergent with respect to one another, perhaps indicating a period of exogenous replication. Inactive ERV lineages tended to display star-like topologies, with flattened deep nodes followed by long terminal branches. I observed this basic topology in MuERV-Fc, MuRRS, MYSERV and Mus.g1/Mus.g2 (solo LTR). This topology likely represents a period of ancestral expansion, followed by lengthy residence in the murine germline (**Katzourakis *et al*, 2005**). (**Figure 3.7G-H).**

IAP and MLV are active ERV lineages that did not display comblike phylogenies. (**Figure 3.7E-F**). They formed highly structured trees comprised of well-defined subclades. In the case of MLV, we found that the subclades corresponded to the three canonical MLV receptor tropisms – xenotropic (xmv), polytropic (pmv) and modified polytropic (mpmv), with PBS specificities corresponding to analyses performed by **Jern *et al*, 2007**. The novel loci we described consisted of xmv (n=2), pmv (n=4) and mpmv (n=6) subtypes. (**Figure 3.7E**). In the case of IAP, the *pol* phylogeny The IAP phylogeny consisted of ten major clades, each separated from another by long branches. The ten clades possess distinct LTR homologies and PBS specificities. It is possible that IAP is an ERV lineage derived from at least eight germline invasion events. Of these clades, one (clade six) contains loci that possess intact env ORFs. In the rest of the sequences, the env ORFs are either fragmented, or missing entirely.

I attempted to recover PBS sequences for murine ERV lineages. For MuERV-Fc, MuRRS, Mus.b1, MYSERV and MuERV-C (Mus.s1 and Mus.i1 notwithstanding), we were unable to recover PBS sequences due to sequence

degradation. For remaining lineages, we mapped PBS specificity as a character onto the trees (**Figure 3.7** - coloured branches). Mapping PBS between groups on the tree shows that PBS mixing is common in IAP clades, with six of the clades containing two PBS specificities, two of then possessing one, and two clades being two degraded to find the PBS. (**Figure 3.7F - clades nine and ten**). When the RepBase designations for IAP elements were mapped to the clades on the tree it was apparent that multiple designations existed for each clade (e.g. clade two being known as IAPEY2 and IAPE3 (**Figure 3.7F** – clade labels). This was not reflected in the clade organisation (e.g. IAPEY2 and IAPEY3 do not group as monophyletic subclades within clade two). With the exception of MMERV and IAP, PBS switching was relatively sporadic, occurring rarely in phylogenies (**Figure 3.8**). These rare instances of PBS switching could conceivably be due to mutation in the host germline. (**Figure 3.7B, 3.7F**). Intriguingly, the MusD phylogeny shows an instance of PBS switching, whereby the ancestral PBS^{LEU} was substituted for PBS^{LYS}, then back to PBS^{LEU} in the clade containing the intact, active proviruses (**Figure 3.7C**).

The LTR phylogenies tended to mimic the basic topology of the *pol* phylogenies, suggesting that the proliferation of solo LTRs is mostly due to proviral recombination into solo LTRs. Proviral LTRs tended to group in derived positions, especially in active ERV lineages – this was not the case for MuRRS or Mus.g2 (**Figure 3.7)**. We observed numerous instances of past expansion and recombinational deletion in murine ERV lineages, whereby expansions have taken place, and become inactive separately from the 'main' lineage of identifiable proviruses. This was apparent when an LTR phylogeny showed large clades of solo LTRs grouping separately to proviral LTRs. An example of this is GLN – with three major LTR clades but only most derived one containing GLN proviruses – indicative of three past genome expansions of which the present is the most recent (**Figure 3.7A**). Similarly, the MusD LTR phylogeny is indicative of three intragenomic expansions, of which one contains the majority of MusD proviruses. (**Figure 3.7C**). For MMERV (**Figure 3.7B**), almost all of the solo LTRs were contained within a single clade. By contrast, the MuERV-L LTR phylogeny disclosed relatively little evidence of expansion distal to the provirus containing clade. (**Figure 3.7D**). Notably, Mus.g2 showed evidence of extensive past expansion followed by inactivation

with a large LTR phylogeny consisting of multiple structured clades and only a single provirus. (**Figure 3.7H**).

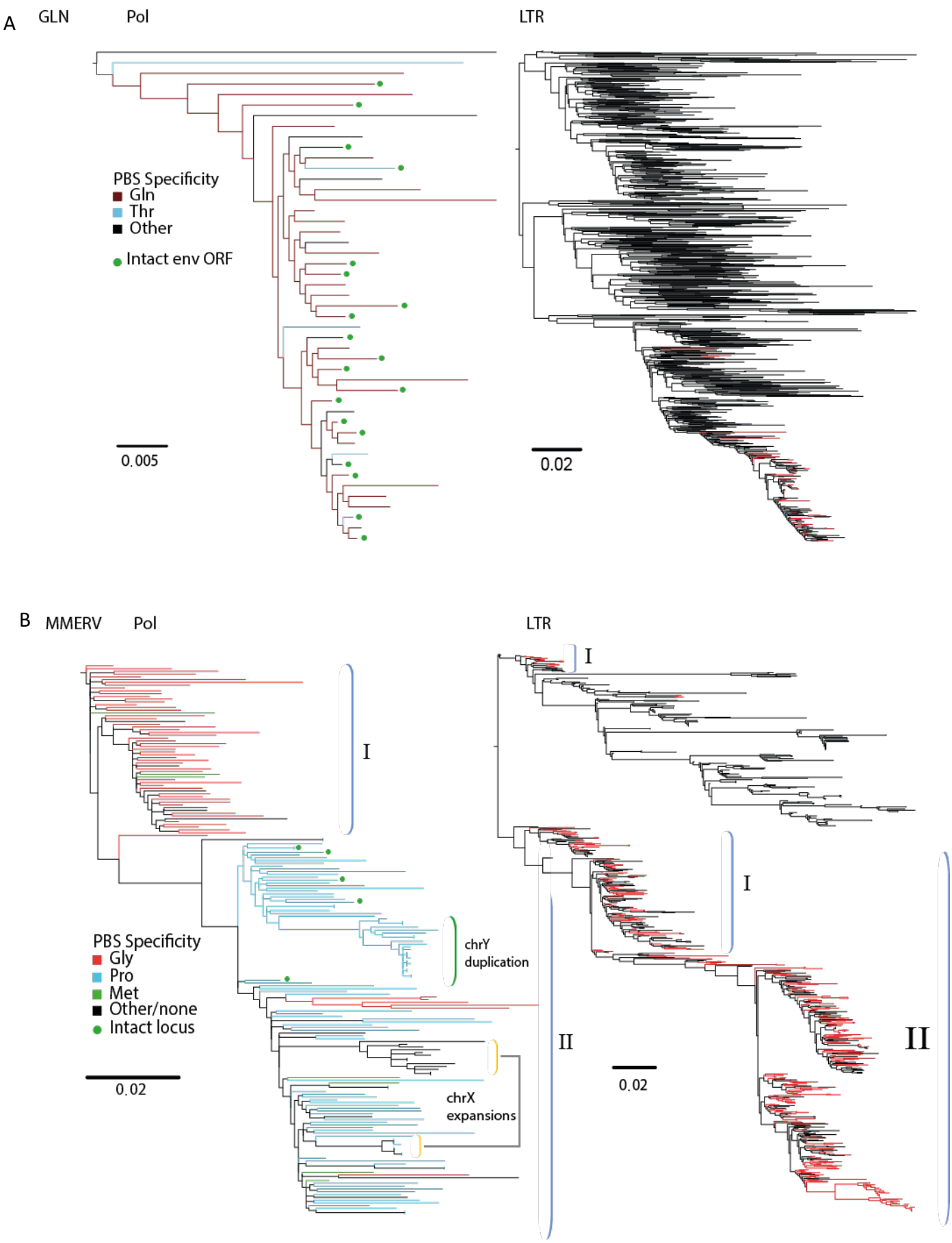

Figure 3.7 – ML phylogenies of GLN (A) and MMERV (B) Pol and LTR sequences. Aligned amino acid (Pol) and nucleotide (LTR) sequences were analysed using IQTREE. Coloured branches indicate PBS specificity on Pol tree (see key). Red branches on LTR trees indicate proviral LTRs.

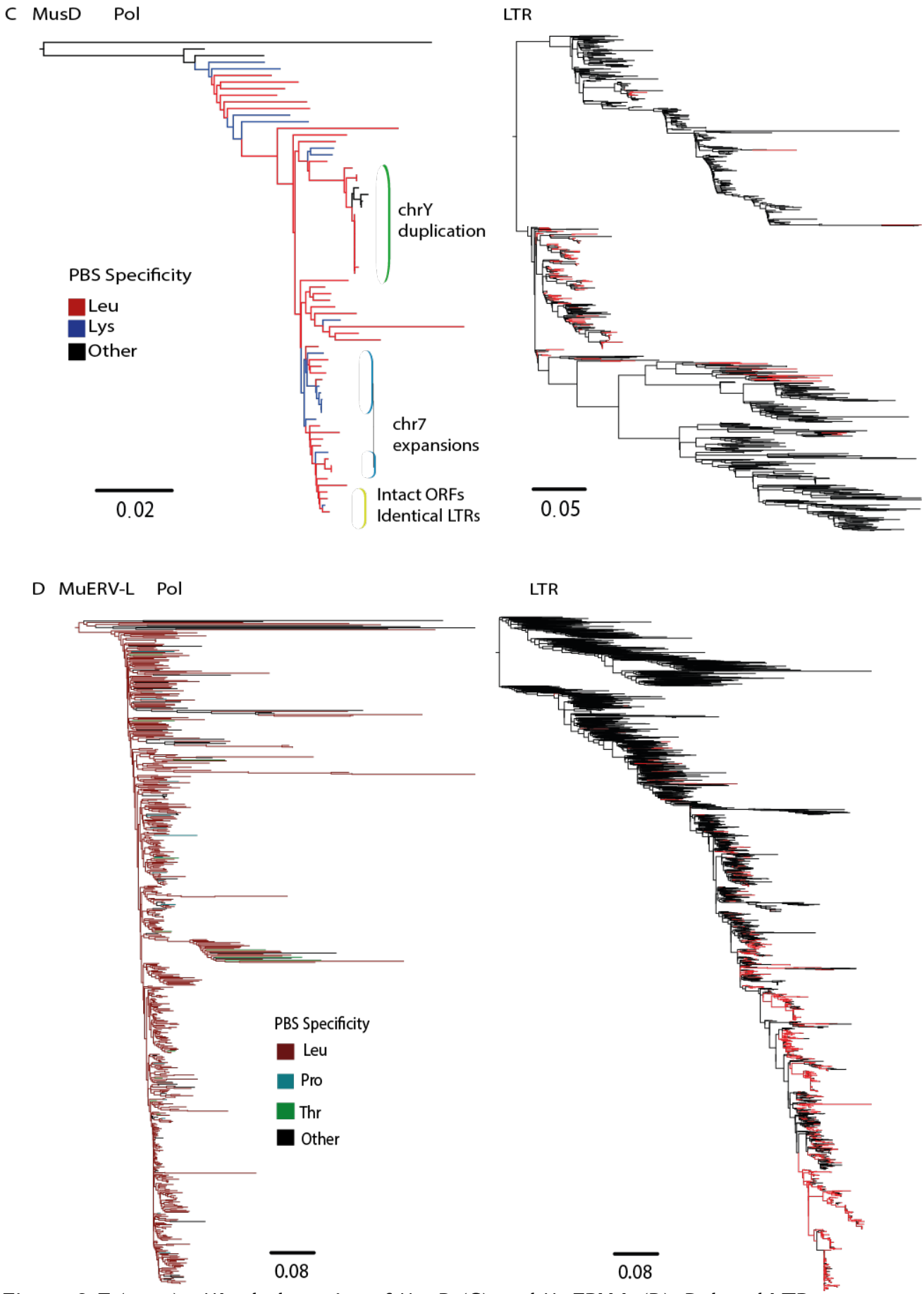

Figure 3.7 (cont) - ML phylogenies of MusD (C) and MuERV-L (D) Pol and LTR sequences. Aligned amino acid (Pol) and nucleotide (LTR) sequences were analysed using IQTREE. Coloured branches indicate PBS specificity on Pol tree (see key). Red branches on LTR trees indicate proviral LTRs.

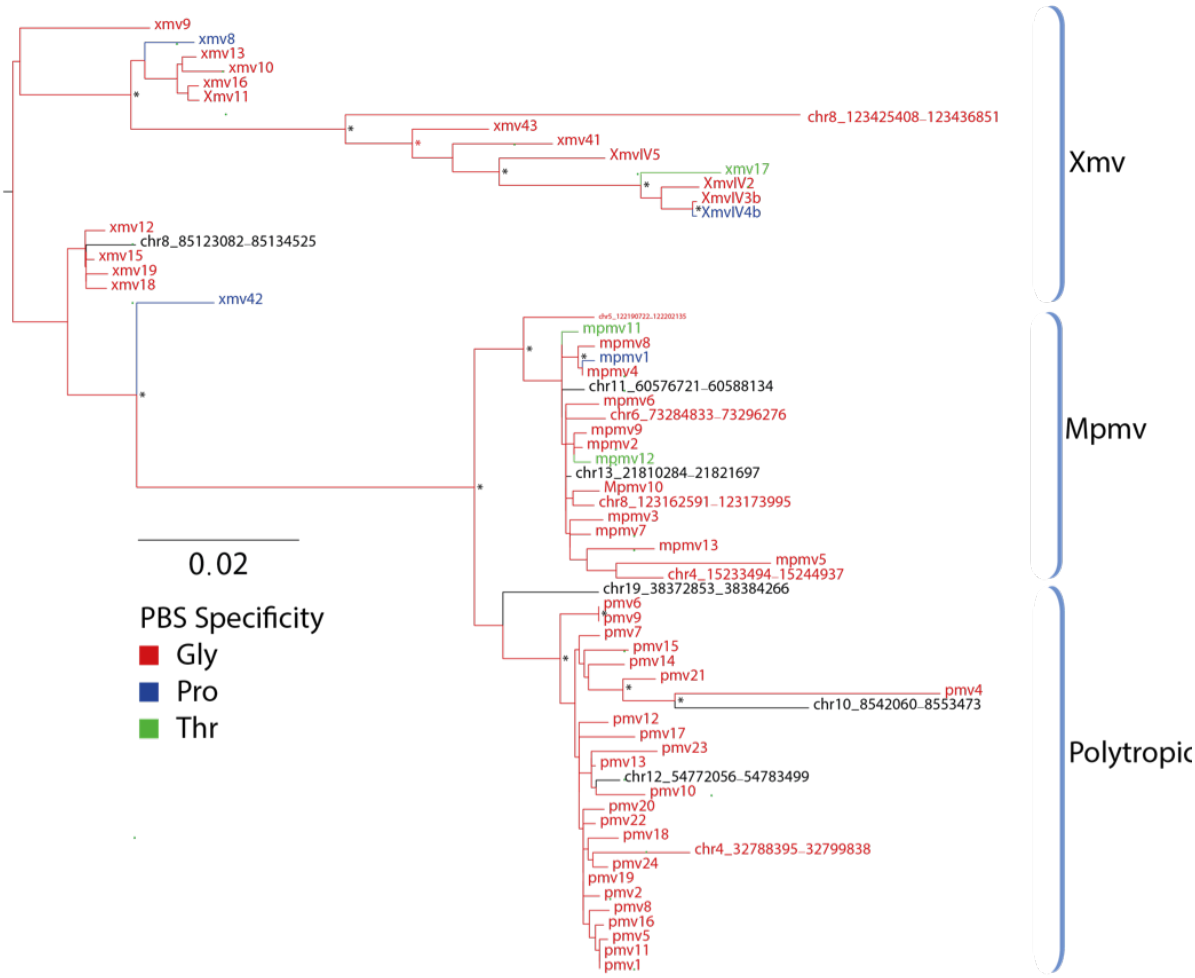

Figure 3.7 (cont) E – ML phylogeny of MLV Pol. Aligned amino acid sequences were analysed using IQTREE. Coloured branches indicate PBS specificity. Scale bar indicates substitutions per site.

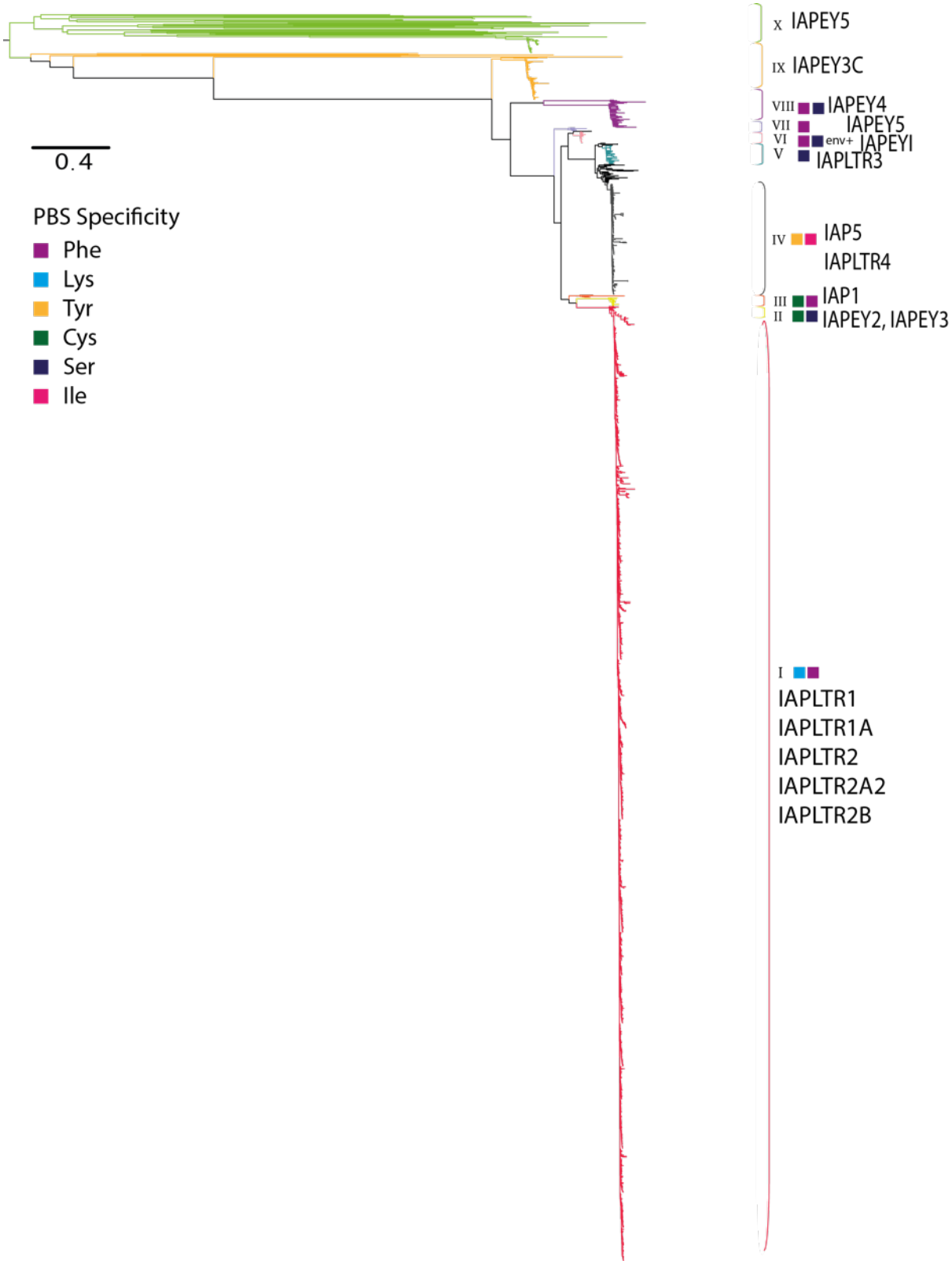

Figure 3.7 (cont) F – ML phylogeny of IAP Pol. Aligned amino acid sequences were analysed using IQTREE. Coloured branches and brackets indicate IAP subclade, with RepBase LTR designation detailed next to the brackets. Coloured boxes indicate PBS

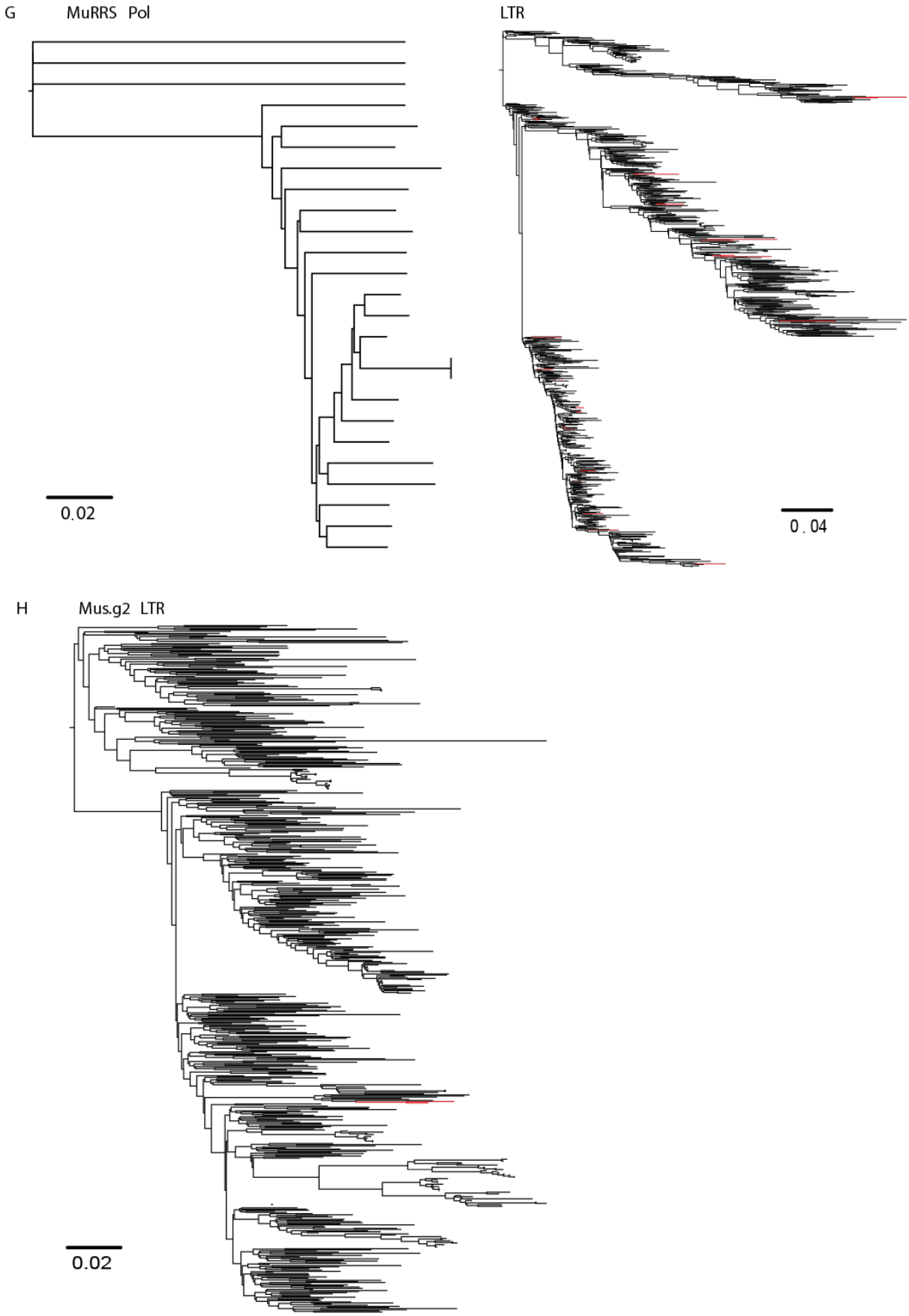

Figure 3.7 (cont) – ML phylogenies of MuRRS (G) Pol and LTR and Mus.g2 (H) LTR sequences. Aligned amino acid (Pol) and nucleotide (LTR) sequences were analysed using IQTREE. Coloured branches indicate PBS specificity on Pol tree (see key). Red branches on LTR trees indicate proviral LTRs.

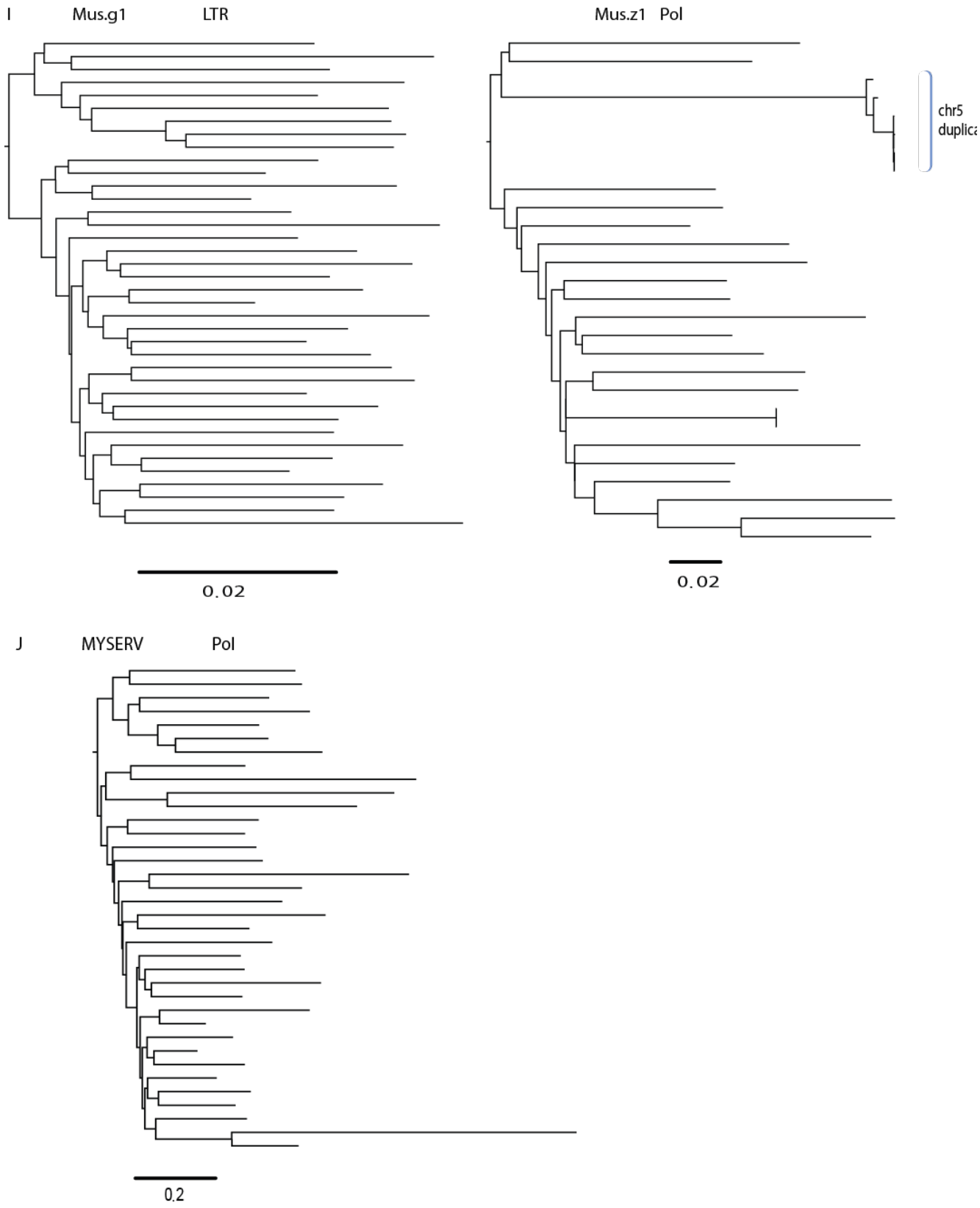

Figure 3.7 (cont) – ML phylogenies of Mus.g1 LTRs, Mu-ERV-Fc Pol (I) and MYSERV pol (J). Phylogenies inferred from aligned amino acid Pol (MuERV-Fc and MYSERV) and nucleotide (Mus.g1) sequences using IQTREE. Scale bars indicate substitutions per site.

3.2.7 - Functional analysis of murine ERVs

Next, I generated annotation tracks of murine ERVs, and used them to assess murine ERV lineages for signatures of functionalisation. First, I uploaded the annotation tracks to GREAT, to see if they were significantly enriched for integration near to genes. For the most part, murine ERVs were located in regions between 50 and 500kb from genes (**Figure 3.8**). However, I found three murine ERV lineages that were significantly associated with gene groups. IAP and MusD were significantly enriched proximal to genes associated with gamete generation and reproductive processes (**Table 3.4; Table 3.5).** Additionally, Mus.g2 (solo LTRs) were found to be significantly enriched with respect to KRAB-ZFPs (**Table 3.6**).

I then searched murine ERV loci for signatures of transcription factor binding and histone methylation (**see Methods**). I selected ChIP-seq datasets from studies of: transcription factors involved in embryonic development (**Chen *et al*, 2012**), and p53 binding in mESCs (**Li *et al*, 2012**). The resultant plots showed 20- and 10-fold enrichment for TRIM28 dependent H3K9 methylation at IAP and MusD loci respectively, in line with previous analyses (**Rowe et al, 2013**). In addition to TRIM28-dependent H3K9 methylation at IAP and MusD loci, I found that that TRIM28-dependent H3K9 methylation was enriched tenfold at Mus.g2, GLN and MMTV loci, and 20-fold at MLV loci. (**Figure 3.9**). All other ERV lineages analysed showed little or no evidence of TRIM28-dependent H3K9 methylation. The only lineage for which evidence could be found of p53 binding in mESCs was GLN, which showed ~tenfold enrichment of p53 binding relative to the genome background ((**Figure 3.19**). Finally, I found evidence of approximately five- and eight-fold enrichments respectively of NANOG and SMAD1 binding, and approximately 20-fold enrichment of SOX2 and ESRRB binding at Mus.g1 loci. (**Figure 3.9**).

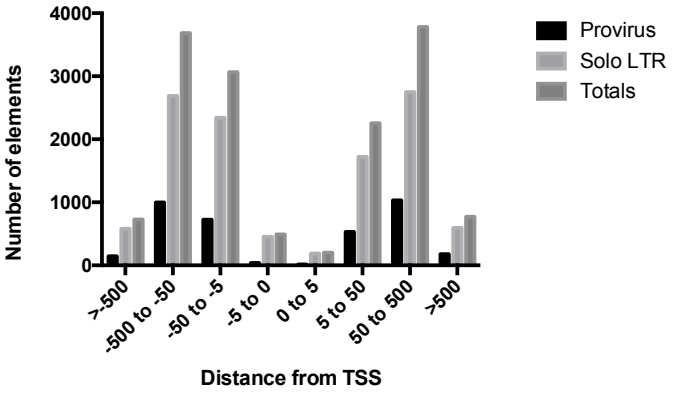

Figure 3.8: Enrichment of ERVs relative to gene transcription start sites. (TSS). Annotation tracks consisting of ERV proviruses and solo LTRs recovered in DIGS screens were uploaded to GREAT.

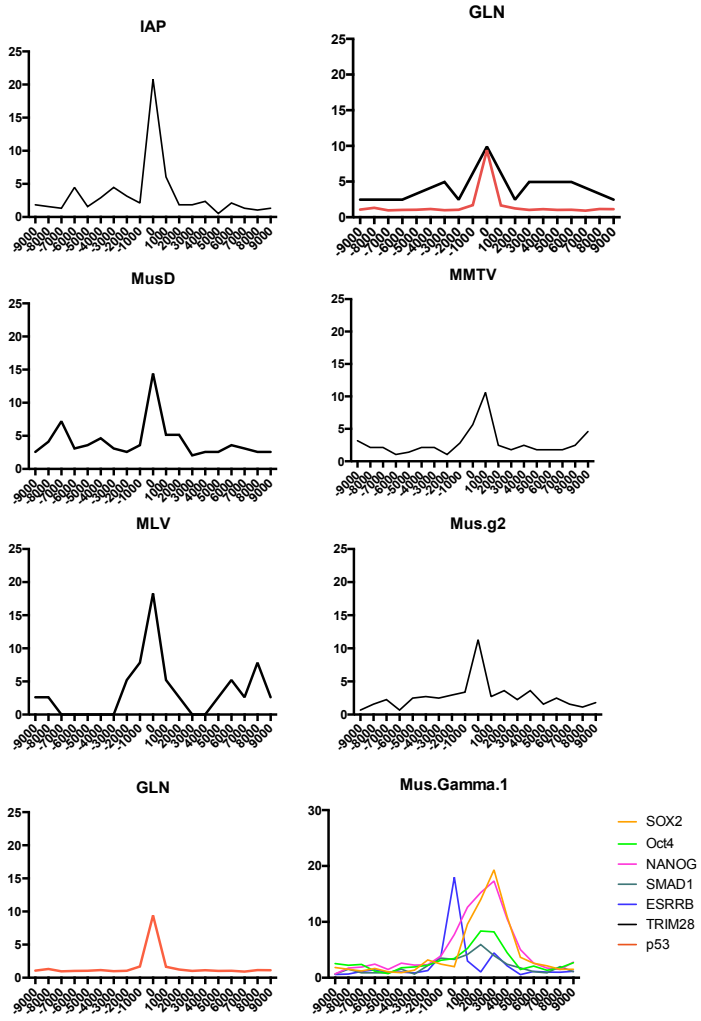

Figure 3.9: Enrichment of TF binding and histone methylation at ERV loci. Murine ERV LTR and Provirus annotation tracks were compared to ChIP-seq datasets of epigenetic changes and transcription factor binding in mESCs in CCG-ChIP-cor. Y axis indicates fold enrichment, X axis indicates position relative to ERV centre. Coloured plots indicate different factors as denoted by the key (bottom right). X axis indicates position of relative to the start of the locus, and Y axis indicates read enrichment relative to genome background.

Table 3.4 – IAP proviruses are significantly enriched proximal to gene groups involved in reproduction. GO Term name indicates GO (gene ontology) term for the gene group proximal to which IAP is enriched.

GO Term name	Binom Rank	Binom Raw P-Value	Binom FDR Q-Val	Binom Fold Enrichment	Binom Observed Region Hits	Binom Region Set Coverage	Hyper Rank	Hyper FDR Q-Val	Hyper Fold Enrichment	Hyper Observed Gene Hits	Hyper Total Genes	Hyper Ge Set Coverage
gamete generation	1	1.84E-42	1.86E-38	2.382744	299	0.1244796	1	6.45E-54	3.335832	208	737	0.1122504
sexual reproduction	2	2.15E-40	1.08E-36	2.30105	304	0.1265612	2	6.29E-49	3.078363	212	814	0.1144091
multicellular organismal reproductive process	3	7.66E-37	2.57E-33	2.147665	319	0.132806	3	4.65E-49	2.981801	222	880	0.1198057
single organism reproductive process	4	8.29E-37	2.09E-33	2.149975	318	0.1323897	5	1.09E-45	2.857949	221	914	0.1192661
multicellular organism reproduction	5	1.57E-36	3.16E-33	2.139232	319	0.132806	4	3.45E-48	2.944989	222	891	0.1198057

Table 3.5 – MusD proviruses are significantly enriched proximal to gene groups involved in reproduction. GO Term name indicates GO (gene ontology) term for the gene group proximal to which MusD is enriched.

GO Term name	Binom Rank	Binom Raw P-Value	Binom FDR Q-Val	Binom Fold Enrichment	Binom Observed Region Hits	Binom Region Set Coverage	Hyper Rank	Hyper FDR Q-Val	Hyper Fold Enrichment	Hyper Observed Gene Hits	Hyper Total Genes	Hyper Ge Set Coverage
gamete generation	1	6.30E-08	0.000634272	3.940927	21	0.2058824	1	2.29063E-06	6.071373	19	737	0.2043011
sexual reproduction	2	1.49E-07	0.000748376	3.743214	21	0.2058824	2	5.99978E-06	5.497054	19	814	0.2043011
single organism reproductive process	3	9.36E-07	0.003141419	3.343477	21	0.2058824	5	1.59426E-05	4.895626	19	914	0.2043011
multicellular organismal reproductive process	4	1.00136E-06	0.002520418	3.329415	21	0.2058824	3	1.43678E-05	5.084775	19	880	0.2043011
multicellular organism reproduction	5	1.06633E-06	0.002147156	3.316342	21	0.2058824	4	1.31879E-05	5.022	19	891	0.2043011
reproductive process	7	3.09655E-05	0.044537294	2.662587	21	0.2058824	6	0.000579437	3.850776	19	1162	0.2043011

Table 3.6 – Mus.g2 solo LTRs enriched proximal to protein coding genes. GO Term name indicates GO (gene ontology) term for the gene group proximal to which Mus.g2 is enriched

GO Term	Binom Rank	Binom Raw P-Value	Binom FDR Q-Val	Binom Fold Enrichment	Binom Observed Region Hits	Binom Region Set Coverage	Hyper Rank	Hyper FDR Q-Val	Hyper Fold Enrichment	Hyper Observed Gene Hits	Hyper Total Genes	Hyper Gene Set Coverage
Krueppel-associated box	1	3.66E-16	3.50E-12	6.175357	33	0.05400982	1	1.95748E-06	3.933549	30	360	0.06465517
Zinc finger C2H2-type/integrase DNA-binding domain	3	3.49E-08	0.000111225	2.676809	40	0.06546645	2	0.001935178	2.565358	35	644	0.07543103
Zinc finger, C2H2	4	4.06061E-06	0.009711954	2.209497	40	0.06546645	3	0.017048646	2.288214	35	722	0.07543103
Zinc finger, C2H2-like	6	1.64673E-05	0.026257062	2.077672	40	0.06546645	4	0.042667792	2.162422	35	764	0.07543103

3.2.9 – Dating ERV integrations

In ERV lineages where I could identify LTRs flanking proviruses, I used the LTR sequences to estimate the date of integration. I used the pairwise distance between LTRs, along with the mouse neutral substitution rate to estimate the date of integration. However, recombination or gene conversion that alters the LTR sequences can lead to inaccurate estimation of provirus dating (**Subramanian *et al*, 2011**). With this in mind, I used the phylogenies of ERV LTRs for 10 lineages (**Figure 3.7**) to identify the presence of recombinant sequences. Where sequences from the same provirus grouped separately to one another, they were discounted from the dating process. 655 eligible proviruses across 9 murine ERV lineages were identified. (**Appendix 3.2**).

635 (97%) of murine ERV loci sampled had a reported integration date of less than 500,000 years. 310 proviruses (47%) had an estimated integration date of 0, indicating integration within the last 0-50,000 years, occurring in the lineages MuERV-L, MMERV, MusD, GLN, MMERV, MMTV and MLV **(Figure 3.10**). The total mean integration date was 69908 years. The oldest locus in the lineages sampled was that of the Mus.g2 provirus, at ~1,200,000 years old. The lineages sampled with the most numerous loci (MLV, MuERV-L, GLN and MMERV) show signs of limited proliferation prior to bursts of expansion in recent history within the last ~660000 to ~230000 years. (**Figure 3.10**). By contrast, MusD and MuRRS show signs of relatively constant integration over million-year timescales, without the sharp increases in copy number occurring over ten thousand year timescales as shown for the lineages previously. Given that MURVY has replicated through segmental duplication on the Y chromosome, the shape of the MURVY data (two discrete groups) indicates two waves of duplication on the Y chromosome that have given rise to this lineage, approximately 50,000 and 160,000 years ago **(Figure 3.10**).

Where we could not identify flanking LTRs, we searched for orthologous integrations in other rodent species. (**see Methods**). We were able to identify orthologous integrations for: Mus.i1, Mus.s1, MuERV-Fc, and MYSERV (**Figure 3.11; Table 3.7**). The oldest integration we could find was an ortholog of the Mus.i1 integration in the Chinese hamster (*Cricetulus griseus*). (**Table 3.7;**

Figure 3.11) This gives the minimum estimated integration date for this murine ERV of 32.7 mya. Additionally, we found that the genomic region of this locus was syntenic (shared organisation) between the mouse and the rat, occurring in both species approximately 1kb 5' with respect to ZFP251 (**Figure 3.11**). For Mus.s1, I found an orthologous integration on chromosome 9 of the rat (*Rattus norvegicus*). (**Table 3.7**; **Figure 3.11**), giving the minimum integration date of this RT locus of approximately 20 mya (**Hedges *et al*, 2015**). I was able to find an ortholog of MuERV-Fc in the genome of the rat. (**Table 3.7; Figure 3.11**). Thus, this integration predates the divergence of genera *Mus* and *Rattus* approximately 20mya (**Hedges *et al*, 2015**). Finally, I discovered a MYSERV integration that was shared between the mouse and the wood mouse (*Apodemus sylvaticus*), making this lineage at least 14.5 mya (**Hedges *et al*, 2015**). (**Table 3.7: Figure 3.11**).

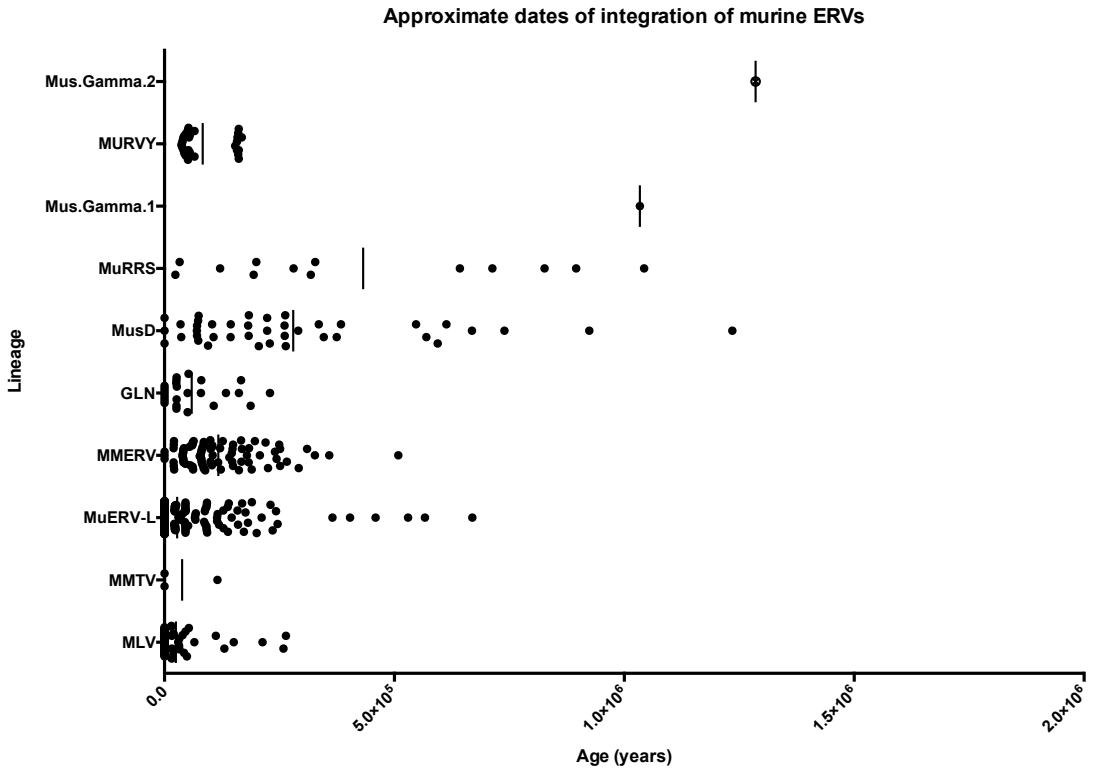

Figure 3.10 – Paired LTR estimation of murine ERV integration dates. Y axis indicates lineage, X axis indicates age in years. Each dot indicates a single murine ERV integration. The pairwise distance between two paired proviral LTRs was used in concert with the mouse background substitution rate to estimate the integration date of murine ERVs. Vertical lines indicate mean approximate integration date.

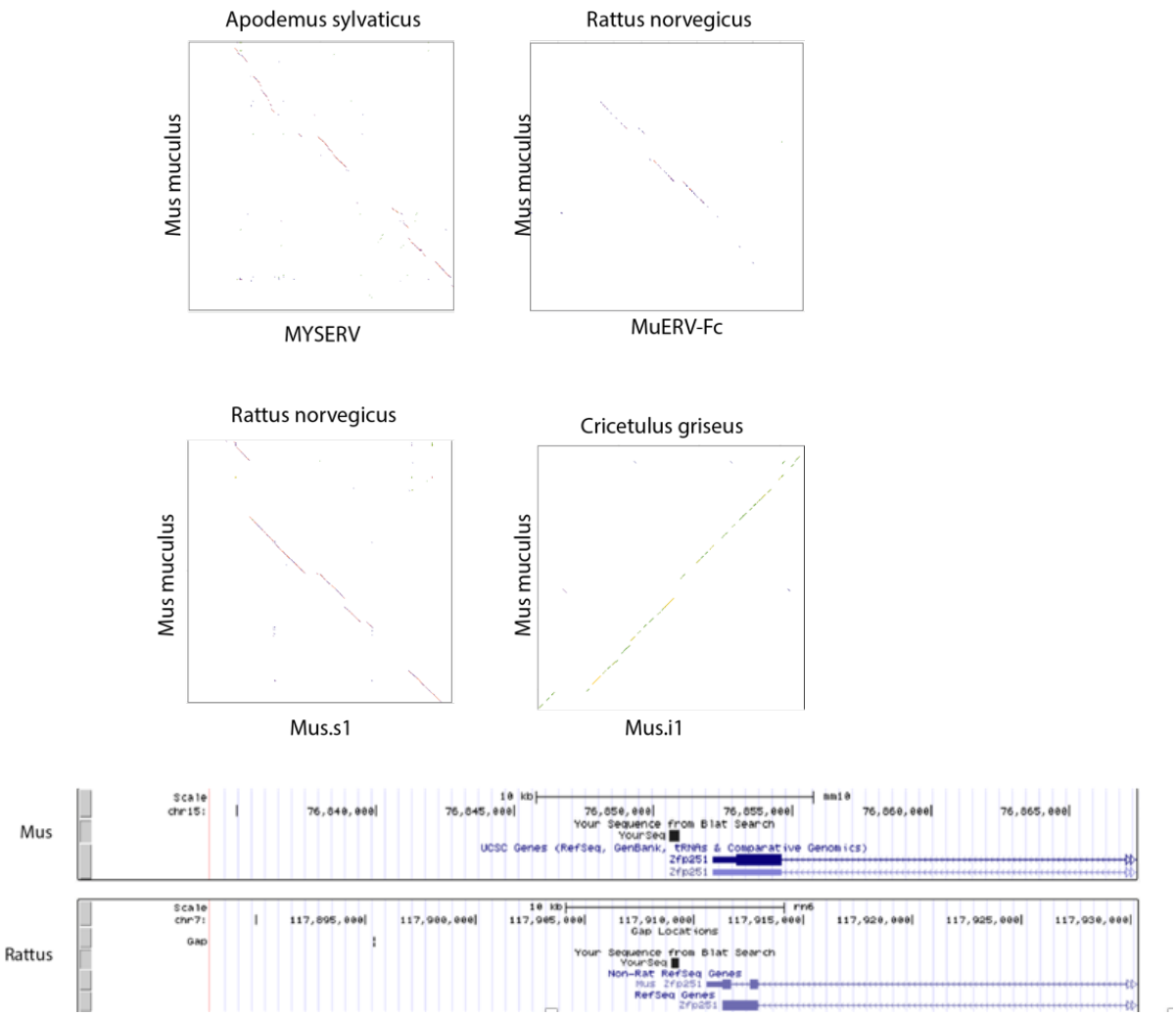

Figure 3.11 – **Orthologs of four murine ERV lineages.** Aligned genomic regions (20kb), containing orthologous ERV loci between the mouse and indicated species (top) for MYSERV (top left), MuERV-Fc (top right), Mus s1 (bottom left) and Mus.i1 (bottom right). The dotplot is marked where sequence similarity exceeds 80%. The sequence 'window' size for comparison was 20bp. The bottom (B) is a genome browser view of the genomic region containing Mus.i1 syntenic between the mouse and rat. YourSeq indicates Mus.i1 RT sequence as a BLAT search against the two genomes.

Table 3.7 – Orthologous murine ERV integrations, detailing the lineage and integration position of orthologous integrations

Lineage	Species	Assembly	Scaffold	Position
Mus.s1	Brown rat	rn6	chr6	64195658
Mus.i1	Chinese hamster	crigri1	KE377312	728267
Mus.z1	Brown rat	rn6	chr12	4696852
MYSERV	Wood mouse	ASM130590v1	LIPJ01000751.1	7314
MuERV-Fc	Brown rat	rn6	chr16	62193762

3.3 - Discussion

3.3.1 – Generation of a murine ERV dataset

Generations of inbreeding mean that C57BL/6J genome is representative of this mouse strain only. However, the overall ERV diversity is unlikely to be affected, instead manifesting itself in the form of varying copy number and polymorphism of ERV integrations, as has been noted for IAP, MMTV and MLV (**Ray *et al*, 2011; Kozak *et al*, 1987, Frankel *et al* 1990**). I generated a dataset consisting of alignments and metadata of murine ERVs. Generally, I found that this dataset compared well to other analyses that have been performed (**Table 3.1; Table 3.2**), in terms of ERV discovery and sequence retrieval. I make this dataset freely available in GLUE (**https://github.com/giffordlabcvr/Retroviridae-GLUE)** in the hope that it will assist in other analyses of murine ERVs.

Although comprehensive for many ERV lineages, I found that studying the highly polymorphic and numerous IAP lineage to be difficult, due to issues with aligning and curating data for such a large number of sequences. As such, the collection of sequences and data published in GLUE would be improved greatly through the addition of IAP provirus and LTR alignments.

3.3.2 – A transition in murine:ERV co-evolutionary history

I discovered five novel murine ERV lineages, and characterised in further detail the MuERV-Fc lineage identified by **Diehl *et al*, 2016**. I found that, without exception, the novel murine ERV lineages were highly degraded and low in copy number relative to the majority of murine ERVs lineages. The Mus.i1, Mus.s1, MuERV-Fc lineages represent ERV lineages that are apparently common to mammals, and in the case of ERV-I, vertebrates. (**Martin *et al*, 1997**; **Herniou *et al*, 1998; Diehl *et al*, 2016**). These lineages are demonstrably ancient, a fact underscored by the orthologs I found, making the integrations at least 20-30 million years old in mice (**Figure 3.12**). Based on these, and other published data (**Martin *et al*, 1997; Diehl *et al*, 2016**), I estimate that these ERV lineages are in fact the degraded, ancient remnants of the 'ancestral' ERV-I,

ERV-S and ERV-Fc lineages that are widespread in other animals and have been largely lost in the mouse, despite being maintained in other mammals. In lieu of ancestral mammal-common ERVs, the mouse appears to be instead populated by evolutionarily modern, transpositionally active gamma and betaretroviruses bearing high degrees of similarity to extant exogenous retroviruses. Furthermore, ERVs from these ancestral lineages (ERV-W and ERV-H, for example), play vital roles in mammalian physiology (**Mi *et al*, 2000; Chuong *et al*, 2016**). This poses a question: why and how has the mouse selectively lost ERV lineages from ancient, provisional genera whilst they are maintained, and play important physiological roles, in the human? (**Herniou *et al*, 1999; Hayward *et al*, 2015**).

The apparent loss of ancient ERV lineages in the mouse could be explained by genomic turnover of ERVs. Smaller mammals tend to have shorter lifespans than larger mammals. This creates a shorter generation time for smaller mammals which, in general, tends to speed up the rate of genome evolution. This would have the effect of loss of ancient ERVs occurring in smaller mammals at a greater rate than larger ones. The generation time hypothesis supports this, stipulating a higher level of genomic 'turnover' in rodents than in larger mammals (**Laird *et al*, 1969**).

It is conceivable that the higher level of genomic turnover would extend to transposable elements, leading to the loss of non-proliferating, inactive ERV lineages and the subsequent dominance of more modern sequences. Indeed, this concept has been explored by **Katzourakis *et al*, 2014;** who notice lower levels of active ERVs in larger mammals, and suggest, based on analysing the variation in ERV copy number in different mammals, that 68% of the variance in ERV age per genome can be explained by body size.

A comparative approach would therefore be an interesting future direction. Targeted genome screening of rodent genomes with 'ancestral ERV' polymerases to identify ERV diversity and distribution, and to see if this loss of ancestral ERVs is common in rodents. This approach would be well suited to DIGS – screening for a specific sequence diversity in a targeted group of species. Especially of interest would be the naked mole-rat (*Heterocephalus glaber*) – a rodent with a very long lifespan (~30 years) and higher reproductive age (~1-2 years) if our

hypothesis is correct, then one would expect the ERV diversity in the naked mole-rat to more closely resemble that of a human.

3.3.3 – Identification of ERVs potentially involved in physiological processes

We found that all active ERV lineages are repressed by TRIM28 in mESCs, with the exception of MuERV-L. This is consistent with the idea that TRIM28-based repression is a way to prevent inappropriate ERV activation during embryonic development, and of findings indicating MuERV-L is de-repressed in mESCs (**MacFarlan *et al*, 2012**). It has been suggested that ERV repression is a basis for transcriptional control of host genes during development (**Rowe *et al*, 2013; Gifford *et al*, 2013**). Accordingly, we found that MusD and IAP are enriched with respect to genes involved in gametogenesis and early embryonic development (**Table 3.4; Table 3.5**).

MMTV and MLV are repressed by TRIM28 dependent histone methylation in mESCs. These ERV lineages are very modern and possess exogenous counterparts. They are polymorphic between different mouse strains, with evidence indicative of exogenous replication followed by endogenisation. (**Wolf and Goff, 2009**). KRAB-ZFPs are highly diverse and evolving rapidly in mammals (**Emerson and Thomas, 2009**; **Schmitges, *et al*, 2016**). This functional diversity and rapid evolution mimics other antiviral gene families -PARP, for example (**Daugherty *et al*, 2016**). This raises the possibility that KRAB-ZFPs and ERVs represent a host:virus arms race.

Studies of the ERV:KRAB-ZFP conflict have found species specific modifications to ZFPs are used to repress novel transposon families as they emerge. In turn, this has led to the supposition that the genome is in an arms race with itself - as genomic transposons and ZFPs are in an evolutionary conflict (**Jacobs *et al*, 2014**). Our murine ERV dataset has indicators of this: Mus.g2 shows signatures of TRIM28-dependent histone methylation in mESCs. (**Figure 3.10**). In turn, Mus.g2 is enriched proximal to KRAB-ZFP genes. Additionally, I identified a Mus.i1 integration highly proximal to ZFP251 in mice and rats (**Figure 3.12**). (Highly) speculatively, a circular picture emerges, whereby ERV integrations proximal to ZFPs are repressed, leading to repression of ZFP gene

expression, leading to greater ERV success in the genome. Conversely, it could be that the host is utilising the Mus.g2 ERV sequences proximal to ZFPs to regulate ZFP expression in a physiologically useful context.

Mus.g1 shows signs of being bound by a diverse range of transcription factors in mESCs, including SOX2, OCT4 and NANOG. These transcription factors act as enhancers of transcription, and are all involved in cell differentiation regulation, being necessary for the maintenance of pluripotency in mammalian ESCs. Given that Mus.g1 does not encode any intact ORFs, (if it has a function), the possible functions include expression as an lncRNA, or regulation of nearby genes. Interestingly, HERV-H is an ERV that encodes an lncRNA expressed in response to OCT4, NANOG and SOX2 binding, and is implicated in the maintenance of pluripotency in human ESCs (**Göke and Ng, 2016; Friedli *et al*, 2014; Santoni *et al*, 2012).** HERV-H is an ancient ERV lineage, around 40 million years old in primates (**Mager and Freeman, 1995**). By contrast, Mus.g1 locus itself is little more than a million years old, highlighting the potential speed of functionalisation in core host processes.

The process of functional ERV turnover appears to be relatively common in mammals. **Ponting *et al*, 2011** reviewed studies of functional and non-functional genomic turnover in mammals, concluding that transposable elements make up a large proportion of transiently functional, noncoding genomic sequence, undergoing relatively rapid turnover. Interestingly, a comparative study by **Nellaker *et al*, 2012,** found that most TE integrations were introduced through the male germline. Although the data on the mechanics and consequences of this are scant, it is interesting to note that IAP and MusD sequences appear to be selectively enriched proximal to genes involved in spermatogenesis (**Table 3.4; Table 3.5**).

In vitro and *in vivo* study of the Mus.g1 locus will be interesting, and easily achievable – as shown by **Chuong *et al*, 2016** in their study of ERV regulation of nearby immune gene expression using a CRISPR-cas9 system to knock out ERV loci adjacent to genes. If Mus.g1 were important in pluripotency regulation, it would be an interesting example of evolutionary convergence, which has already been shown to occur for the Syncytin genes. Syncytin has been

exapted multiple times, from multiple different ERV lineages, in multiple mammalian lineages. (**Vernochet *et al*, 2015**). In this instance, it would be convergence on ERV regulation of ESC pluripotency – representing another example of how viruses have been essential drivers of mammalian evolution.

3.3.4 – Conclusions

In this chapter, I performed phylogenetic screening of the murine genome, identifying novel ERV lineages. Additionally, I generated a sequence resource for murine ERV study, using it to perform evolutionary and functional analyses of murine ERVs. I found that my data were largely in concordance with contemporary theories stipulating a rapid turnover of noncoding functional TE sequences, as well as increased levels of ERV turnover in smaller mammals. Additionally, I identified specific loci and lineages that may be involved in physiological processes, presenting numerous interesting aspects for future study of host:virus co-evolution.

I found that DIGS lent itself well to the study of murine ERVs. However, I found that my general approach of DIGS screening followed by phylogenetic and sequence analysis was most effective when applied to low-medium numbers of sequences (i.e. in the analysis of novel, low copy number murine ERV lineages, or orthologous ERV integrations) as opposed to high copy ERV lineages such as IAP, where it was very difficult to curate the high volume of sequence data. Additionally, this high volume of data made it difficult to adopt a comparative approach across different species. With refinement (i.e. the addition of high copy number ERV alignments) and development of methods to manage large numbers of degraded ERV sequences, I believe that the utility of the data gathered here would be advanced. Furthermore, a comparative analysis of ancient murine ERV lineages such as Mus.s1, Mus.i1, MYSERV and MuERV-Fc would be attractive due to their low copy number, the high availability of mammalian genomes, and the fact that they have remained relatively unstudied in recent years. As such this would be relatively straightforward, as I would not have to adapt my experimental framework overmuch to study this aspect of host:virus co-evolution in rodents.

Chapter 4: Paleovirological analyses of endogenous circoviruses

Text in this chapter has been adapted from Dennis *et al,* 2018 and Dennis *et al*, 2018(b), manuscripts written with the assistance of Dr Robert Gifford.

Figures 1, 3, 5, 7, 8, 9 are from Dennis *et al,* 2018 and Dennis, *et al*, 2018 (b)

Peter Flynn performed the PCR from which Figure 4.10 is derived, and co-wrote the corresponding text.

4.1 - Introduction

Circoviruses are members of the family Circoviridae, which is divided into two genera: Circovirus and Cyclovirus (**Rosario *et al*, 2017**) . These viruses possess a small, circular ssDNA genome of 1.7 – 2.2kb, and encode two ORFS; the replicase (Rep) and capsid (Cap), which are responsible for genome replication and particle formation respectively. The known natural hosts of circoviruses include bats, rodents, birds, invertebrates and domestic animals (**Liu *et al*, 2011; Baekbo *et al*, 2012; Tarján *et al*, 2014; Lorincz *et al*, 2011; Raidal *et al*, 2015**).

Viruses of genus *Circovirus* are known in veterinary medicine as the aetiologic agents behind postweaning multisystemic wasting syndrome (PWMS) in swine, and psittacine beak and feather disease (PBFD) in parrots. These fatal diseases of pigs and birds represent both a significant economic burden in swine farming (**Baekbo *et al*, 2012**), and an existential threat to many endangered psittacine species (**Raidal *et al*, 2015**). Circoviruses have also been implicated in a range of other diseases, including those of commercially farmed fish, and pigeons (**Lorincz *et al*, 2012; Stenzel and Konciki, 2017**). In addition to isolation of circoviruses from diseased hosts, genus *Circovirus* has been extensively studied *in vitro* through microscopy, sequence analysis and serology. By contrast, the genus *Cyclovirus* consists only of viruses that have been identified through sequencing methods. Because of this, their host associations are more uncertain. The discovery of cyclovirus nucleic acids in the CSF of patients with unexplained paraplegia in Malawi (**Smits et al, 2013**), and in the CSF of patients with acute CNS infections in Viet Nam (**Tan et al, 2013**), has led to the increasing prominence of cycloviruses as a potential cause of hitherto unexplained human diseases. These viruses have gained subsequently higher profiles, with the European Centre for Disease Control (ECDC) commissioning a risk assessment of cycloviruses (**Derrough *et al*, 2013**). In 2016, an article in The Conversation (**Rose, 2016**) discussed "Newly emerging viruses, such as cycloviruses […] causing neurological problems with children in Asia", and in The Scientist (**Palmer**, **2013**), leading with the with the headline of "New viruses attack Asia and Africa". Given the increasing profile of these viruses, it is important to gather more information on their host range and evolution.

The majority of sequences derived from *Circoviridae* have been isolated using metagenomics studies. Despite the capacity of metagenomics studies to generate a wealth of genome data associated with a given sample, the method has an important limitation with regards to virology: discovery of a sequence in an organism is insufficient evidence to prove that the virus infects that organism; seroconversion or replication data is required to prove that a virus infects a given host. By contrast, the discovery of an EVE in the WGS of that organism definitively proves an infection (past or present) by that virus of the host. By studying the past and present host range of viruses, inferences can be made that guide virus discovery and surveillance efforts, by indicating potential hosts for targeted virus sampling.

The study of ancient EVEs can give insights into host:virus co-evolution over deep timescales. Viruses have the potential for extremely rapid evolutionary rates. This is derived from rapid substitution, recombination, host switching and lineage extinction (**Charleston and Robertson, 2002; Holmes, 2003; McDonald *et al*, 2016**). The study of EVEs enables inferences to be made about more ancient virus sequences, as the host genomes in which EVEs are embedded invariably evolve more slowly than the original exogenous viruses. In addition, orthologs can be determined from EVEs and their flanking genomic sequences. Orthology represents a useful way to study ancient EVEs – relying on estimates of species evolution as inferred from a mixture of palaeontology, biogeography and molecular clock approaches (**Hedges *et al*, 2015**). The discovery of an EVE orthologous between multiple species sets a minimum estimated age for that virus integration, enabling calibration of the timescale of virus evolution.

Similarity-search based approaches have established the presence of circovirus EVEs (CVe) in animal germlines. **(Cui *et al*, 2014; Katzourakis and Gifford, 2010; Byeli *et al*, 2010; Liu *et al*, 2011**). The numbers of CVe in animal genomes are relatively low, indicating that CVe are quite rare. However, the full extent of CVe presence in animal genomes has yet to be established fully, and the rapidly growing body of publicly available WGS assemblies presents ample opportunity to do so. In this chapter, I screened a total of 676 animal

genomes, including 307 invertebrate species, with circoviral peptide sequences, with the aim of studying host:circovirus co-evolution.

4.2 - Results

4.2.1 - Identification and genomic characterisation of animal CVe

I used DIGS to screen animal genomes for CVe. 676 animal genomes were screened (**Appendix 1.1**). I employed a 'phylogenetic screening' approach used in other studies (**Tristem, 2000; Katzourakis and Gifford, 2010**), and **Chapter 3.,** with a reference library encompassing the known exogenous (**Table 4.2**) and endogenous diversity of Circoviridae. Using this approach, I identified 297 hits bearing significant similarity to circovirus and CRESS-DNA sequences. These sequences were derived from 75 species. (**Appendix 4.1; Table 4.1**).

The resulting hits were analysed to determine: which portion of the circovirus genome they were likely to be derived from, and how likely they were to be *bona fide* CVe, as opposed to exogenous viruses contaminating the genome assembly. For each of the 297 hits, I attempted to identify genomic flanking sequence. In the absence of genomic flanking regions, it is difficult to ascertain whether the sequence is truly an integrated EVE, as opposed to a viral sequence contaminating the host genome assembly. I used the presence of contig sequence that flanked regions of circovirus homology (>100bp) as evidence supporting the designation of a sequence as CVe. Furthermore, I used the presence of stop codons and frameshifts as supporting evidence that indicated lengthy residence in the host germline, and noted them in **Appendix 4.1**. Based on flanking genomic regions, at least ~94% (n=282) of the sequences found by DIGS are likely to be bona fide EVE sequences. Furthermore, 52% were degraded, containing at least one stop codon or frameshift. (**Appendix 4.1**). Exogenous (**Table 4.2**) and endogenous circovirus sequences were used to construct an MSA spanning the whole circovirus genome. This MSA was then used to identify the parts of the circovirus genome that had been incorporated as CVe. Most CVe represented only fragments of the full circovirus genome (**Figure 4.1**). 92% of the CVe were derived from whole or fragmented Rep peptides. I found four CVe that consisted of *cap* and *rep* sequences (**Table 4.1**), and a single CVe that consisted only of a *cap* gene (in the sloth) (**Table 4.1**).

I estimated the number of integration events that the CVe represented. Many species (or closely related species groups) genomes contained multiple CVe, but in some cases it was difficult to determine whether these sequences represented a single ancestral integration, or multiple distinct invasion events. The main causes of uncertainty were due to the sequence being present on a short contig, multiple CVe in a single organism that were located distally to one another, and were derived from regions of the circovirus genome that did not overlap one another. However, based on the relative rarity of CVe in animal genomes, (**Katzourakis and Gifford, 2010; Cui et al, 2014; Liu et al, 2011**) I assumed a single invasion event had taken place except where multiple integration events could be determined. I estimated that the 299 CVe identified here represented at least 35-39 distinct germline invasion events (**Appendix 4.1, Table 4.1, Figure 4.1**). Whether it is 35 or 39 depends on whether the CVe in ray-finned fish are are derived from a single invasion event, or seven distinct invasion events.(**see section 4.3.2**). The disparity between the number of CVe versus the number of invasion events was due to 30% (n=101) CVe in this study belonging to a group of highly duplicated carnivore CVe derived from a single invasion event.

To identify potentially expressed CVe, I used each CVe sequence in BLAST queries against the NCBI RefSeq RNA database for each host species. This revealed that 46 CVe sequences are predicted to be expressed (**Appendix 4.1, Table 4.3**). Where orthologs and tandem duplications are considered as single CVe, 13 CVe are predicted to be expressed. (**Table 4.3**).

Table 4.1: CVe and CVl detected in published vertebrate genome assemblies. Asterisks denote newly discovered circovirus-like sequences (CVl) and CVe. Crosses indicate instances with divergent sequences in the same species.

EVE name	Reference	Genes	# Hits	# Species
Agnatha				
CVl-Eptatretus*	This study	Rep	7	1
Actinoperygii				
CVe-Anguilla*	This study	Rep	1	1
CVe-Characiformes*	This study	Rep	1	1
CVe-Clupeiformes*	This study	Rep	1	1
CVe-Cypriniformes	Feher et al, 2013	Rep-Cap	18	2
CVe-Cyprinodontiformes*	This study	Rep	5	1
CVe-Perciformes*	This study	Rep	7	4
CVe-Salmoniformes*	This study	Rep	3	1
CVe-Mormyridiformes	This study	Rep	2	1
CVe-Cichliformes	This study	Rep	2	1
Amphibia				
CVe-Anura	Liu et al, 2011	Rep	2	2
Serpentes				
CVe-Serpentes	Gilbert et al, 2014	Rep-Cap	20	6
Aves				
CVe-Tinamou	Cui et al, 2014	Rep-Cap	2	1
CVe-Psittaciformes	Cui et al, 2014	Rep-Cap	4	3
CVe-Passeriformes	Cui et al, 2014	Rep	6	4
CVe-Egretta	Cui et al, 2014	Rep	1	1
CVe-Gallirallus*	This study	Rep	1	1
CVe-Picoides*	This study	Rep	1	1
Mammalia				
CVe-Chrysochloris*	This study	Rep, Cap	4	1
CVe-Carnivora	Katzourakis and Gifford, 2010	Rep	108	13
CVe-Mus.caroli*	This study	Rep	1	1
CVe-Heterocephalus *	This study	Rep	1	1
CVe-Phascolarctos*	This study	Rep	1	1
CVe-Sarcophilus *	This study	Rep	1	1
CVe-Monodelphis	This study	Rep	1	1
CVe-Galeopterus *	This study	Rep	2	1
CVe-Manis*	This study	Rep	1	1
CVe-Choloepus*	This study	Cap	2	1
Arthropoda				

CVe-Varroa†	Liu et al, 2011	Rep	19	1
CVe-Tropileilaps*†	This study	Rep	8	1
Cve-Pseudomyrmex*	This study	Rep	1	1
CVl.Galendromus*	This study	Rep	1	1
CVl.Metaseilius*	This study	Rep	1	1
CVl.Loxosceles*	This study	Rep	19	1
CVl.Oryctes*	This study	Rep	1	1
CVl.Ephydra_hians*†	This study	Rep	10	1
CVl.Ephydra_gracilis*†	This study	Rep	8	1
CVl.Parhyale*	This study	Rep	3	1
CVl.Caligus*	This study	Rep	4	1
CVl.Triops*	This study	Rep	1	1
Cnidaria				
CVl.Thelohanellus*	This study	Rep	1	1
Mollusca				
CVl.Modolius*	This study	Rep	4	1
CVl.Mytilus*	This study	Rep	4	1
CVl.Biomphalaria*	This study	Rep	1	1
CVl.Conus*	This study	Rep	3	1
CVl.Aplysia*	Liu et al, 2011	Rep	1	1
Platyhelminthes				
CVl.Taenia*	This study	Rep	3	1
Total			**300**	**75**

Table 4.2: NCBI reference sequences used in screening and initial alignment. Sequences were downloaded from GenBank using the accession number and included in MSAs and DIGS screens for CVe.

sequence-ID	name	full_name	clade
NC_001944	BFDV	Beak and feather disease virus	Avian-1
NC_003410	CaCV	Canary circovirus	Avian-1
NC_008033	SvCV	Starling circovirus	Avian-1
NC_008375	RaCV	Raven circovirus	Avian-1
NC_008521	GuCV	Gull circovirus	Avian-1
NC_008522	FiCV	Finch circovirus	Avian-1
NC_026945	ZfCV	Zebra finch circovirus	Avian-1
NC_007220	DuCV	Duck circovirus	Avian-2
NC_025247	SwCV	Swan circovirus	Avian-2
NC_015399	BarbCV	Barbel circovirus	Fish-1
NC_025246	SgCV	Wels catfish circovirus	Fish-2
NC_001792	PCV-1	Porcine circovirus 1	Mammal-1
NC_005148	PCV-2	Porcine circovirus 2	Mammal-1
NC_020904	CfCV	Canine circovirus 1	Mammal-1
NC_023885	MiCV	Mink circovirus	Mammal-1
NC_031753	PCV-3	Porcine circovirus 3	Mammal-2
NC_028045	TbCV	Mexican free-tailed bat circovirus	Mammal-2
NC_028045	TbCV	Mexican free-tailed bat circovirus	Mammal-2
NC_021707	CyCV-VN	Cyclovirus VN isolate hcf1	Cyclovirus-1
NC_020206	FWCasCYV	Florida woods roach-associated cyclovirus	Cyclovirus-3

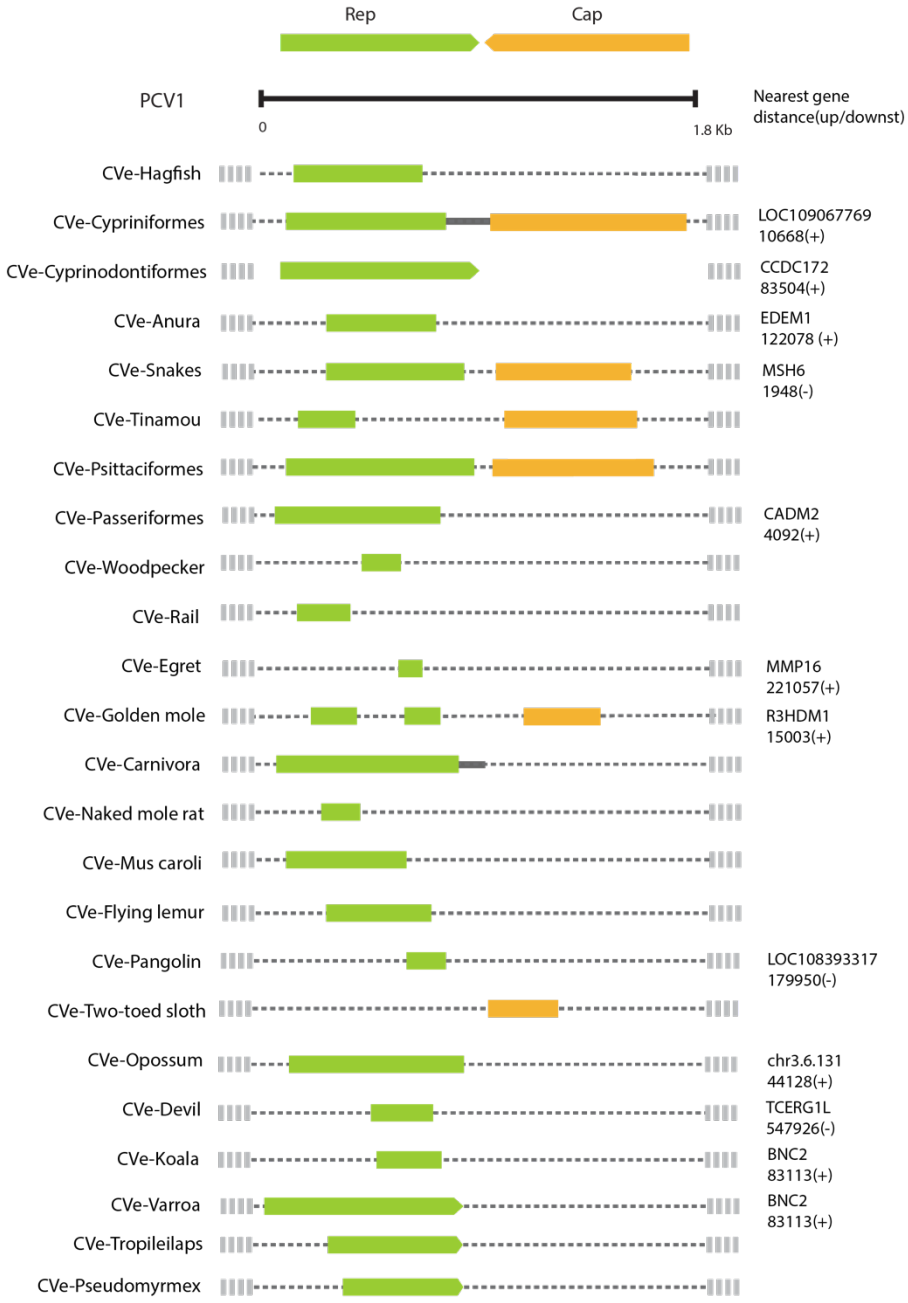

Figure 4.1: Genome structures of 24 endogenous circovirus (CVe) elements identified in animal genomes. Genome structures are shown relative to a porcine circovirus 1 (PCV1) reference genome (accession # NC_001792.2). CVe Rep and Cap coding sequences are represented schematically as green and yellow bars. A thickened grey line between the two ORFs indicates internal non-coding region of the circoviral genome. Dotted lines indicate regions of the viral genome that are not represented in CVe. The name of the nearest annotated gene, where one could be identified, is shown to the right of each element.

Table 4.3 – Expression predictions and accessions for CVe. sequenceID denotes sequence that was used in tBLASTn queries of the NCBI RefSeq RNA database. Significant expression predictions, with accession no and name, are listed on the right.

sequenceID	accession	expression_prediction_name
CVe.Carnivore.Mustela_putorius.4.1	XR_001180410.1	Mustela putorius furo uncharacterized LOC101671852
CVe.Carnivore.Ursus_maritimus.3	XR_571517.1	Ursus maritimus uncharacterized LOC103673978
CVe.Carnivore.Ursus_maritimus.4.1	XR_571516.1	Ursus maritimus uncharacterized LOC103673978
CVe.Cypriniformes.Cyprinus_carpio.3	XM_019084668.1	Cyprinus carpio uncharacterized LOC109067769
CVe.Cyprinodontiformes.Kryptolebias_marmoratus.1	XM_017439363.1	Kryptolebias marmoratus uncharacterized LOC108249773
CVe.Cyprinodontiformes.Kryptolebias_marmoratus.5	XM_017439647.1	Kryptolebias marmoratus 2-phosphoxylose phosphatase 1
CVe.Perciformes.Acanthochromis_polyacanthus	XM_022218352.1	Acanthochromis polyacanthus sequestosome 1
CVe.Pseudomyrmex_gracilis.1	**XM_020436189.1**	Pseudomyrmex gracilis uncharacterized LOC109858690
CVe.Salmoniformes.Salmo_salar.1	XM_014156888.1	Salmo salar TSC22 domain family protein 4-like
CVl.Varroa_destructor.11	XM_022795351.1	Varroa destructor uncharacterized LOC111246180
CVl.Varroa_destructor.14	XM_022796232.1	Varroa destructor uncharacterized LOC111246511
CVl.Varroa_destructor.18	XR_002679502.1	Varroa destructor uncharacterized LOC111246511
CVl.Varroa_destructor.3	XM_022854376.1	Varroa jacobsoni carbohydrate sulfotransferase 1-like

4.3.2 – Identification of orthologs and construction of a timeline of CVe evolution

I searched the flanking genomic regions of CVe for evidence of orthology. Where the flanking genomic regions displayed unambiguous (**see Materials and** Methods) homology, I designated them as orthologous. I was able to identify orthologous CVe insertions in five different species groups – Passeriformes, Carnivora, Serpentes, Cyprinidae, and *Varroa*. (**Figure 4.2**). In some cases sequence similarity and phylogenetic relationships (i.e. multiple CVe from the same species forming a monophyletic clade in a phylogeny) suggested orthology, but I was unable to definitively show this to be such using flanking sequences. Where this was the case, I designated them as potentially orthologous CVe. I found potentially orthologous CVe in Psittaciformes, Amphibia, Marsupialia, Perciformes and Actinoperygii. I used these data to create a timeline of circovirus invasion of vertebrate genomes (**Figure 4.3**). I could only find evidence of a single orthologous integration in invertebrate host groups (*Varroa*). Literature searches found no published divergence times between

Varroa jacobsoni and *Varroa destructor*. Because of this, I omitted invertebrate hosts from **Figure 4.3**.

When plotted onto a time-calibrated phylogeny, the minimum estimated integration dates of vertebrate CVe show that circoviruses have been infecting vertebrates since at least the end of the Mesozoic (**Figure 4.3**). The oldest ortholog, in *Serpentes*, dates CVe integration to 90mya. The remaining orthologs are clustered in the Paleogene (**Figure 4.3**). The potential orthologs suggest that *Circoviridae* may be even older than the *Serpentes* ortholog. Sequence similarity and phylogenetic analysis (**see Section 4.3.3)** indicate either that the CVe in fish represent an ancient ortholog predating the radiation of Actinoperygii, or co-divergence and endogenisation of exogenous fish-circoviruses (such as barbel circovirus – BarbCV) has led to this species group being represented as a monophyletic clade (**see Figure 4.5**).

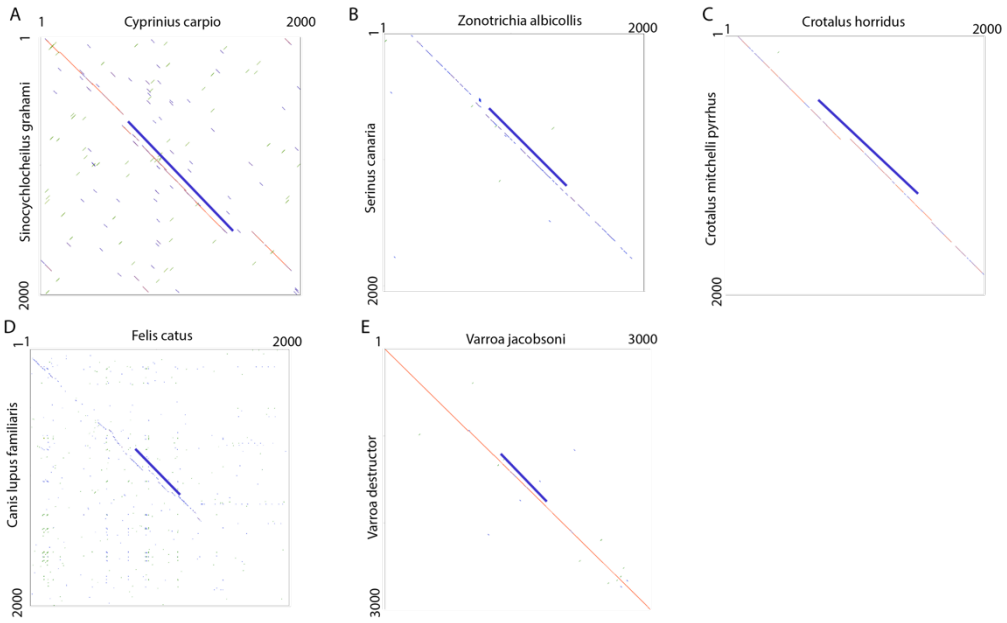

Figure 4.2: Orthologs of four CVe in fish (A), passerine birds (B), snakes (C), carnivores (D) and Varroa mites (E). The dotplots were generated by aligning 2kb (3 in case of Varroa) genomic extracts of CVe and genomic flanking regions. Threshold used for dot generation was 80%. Window size was 20bp. Thick blue lines indicate regions containing CVe.

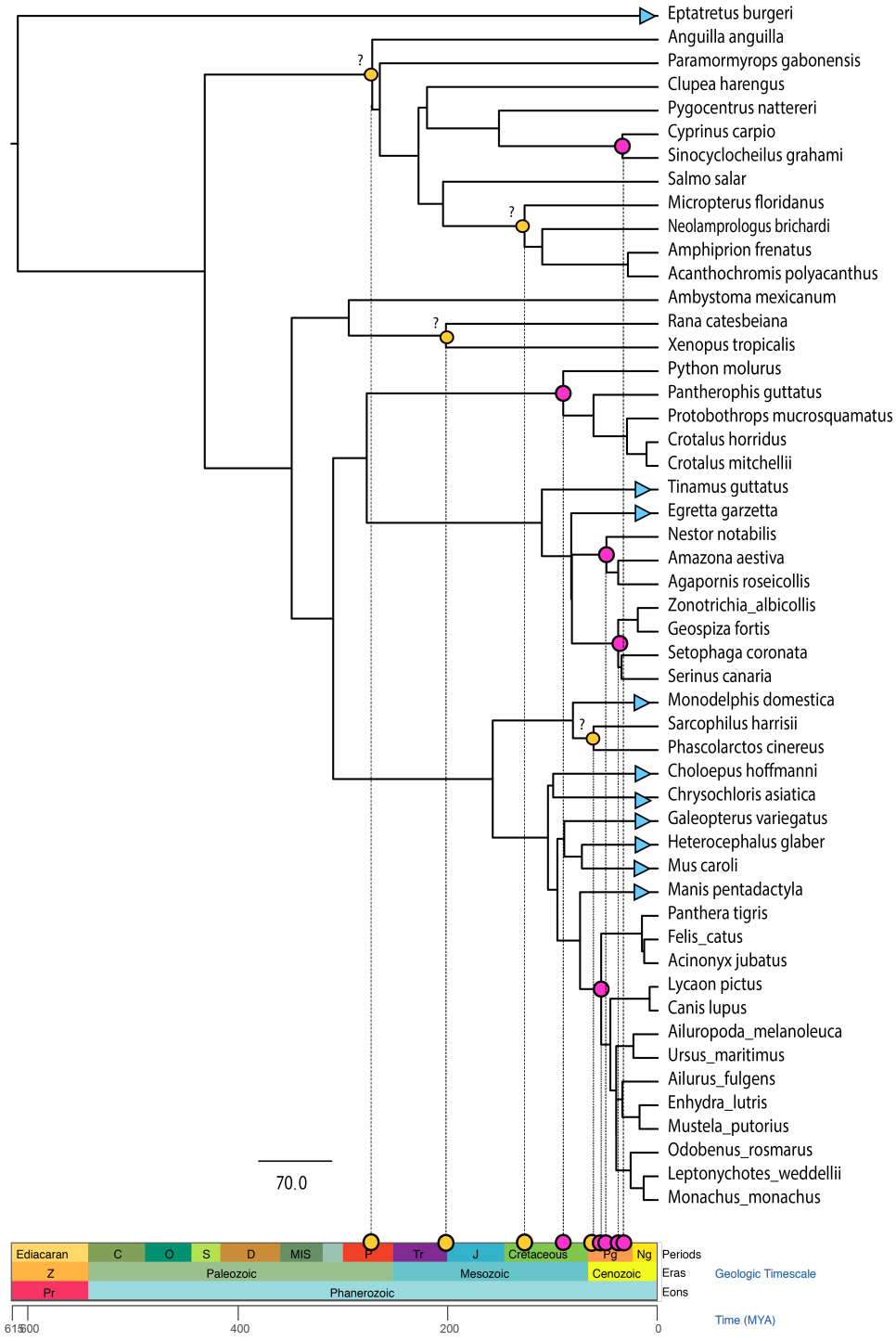

Figure 4.3: Evolutionary relationships of vertebrate species in which CVe have been identified, and timeline of CVe evolution. Pink circles indicate confirmed orthologs. Yellow circles indicate the presence of potential orthologs that have not been confirmed. Blue triangles indicate where CVe loci are present, but no information about their ages could be obtained. Phylogeny obtained from the TimeTree database (**Kumar et al, 2017**). *Kryptolebias marmoratus* was not present in the TimeTree database and was omitted from this figure.

4.3.3 - Phylogenetic analysis of Circoviridae Rep

We performed phylogenetic analysis of CVe. Because of the dominance of the Rep protein in our screening results, and the relative conservation of the Rep versus the Cap protein, the root Rep MSA was used in a preliminary phylogenetic analysis. The tree (**Figure 4.4**) disclosed three major clades, one of which was a group of sequences largely derived from marine invertebrates. I surmised that these sequences belonged to the CRESS-DNA group of viruses. This was based on the presence of the Avon-Heathcote Estuary associated virus (AHEaV) isolate (**Dayaram *et al*, 2015**), the basal position in the phylogeny, and a high degree of sequence divergence in this region of the tree (**Figure 4.4**). Bootstrap support for branching patterns in this region of the phylogeny were low, reflecting the highly diverse nature of CRESS-DNA viruses circulating in invertebrates (**Dayaram *et al*, 2015**), and relatively poor sampling of viruses in invertebrates. The only vertebrate sequence to group in this clade was a consensus sequence derived from the genome of the inshore hagfish (*Eptatretus burgeri)*. The CRESS sequences shown in **Figure 4.4,** are not considered to part of family Circoviridae (**Rosario *et al*, 2017**), and were therefore not considered to be CVe. Based on their basal grouping in phylogenies, and close matching to CRESS-DNA references, 65 sequences were designated as Circovirus like (CVl) CRESS. They were omitted from further analysis, removed from the MSA, and denoted in **Appendix 4.1**.

The MSA was then used to construct an ML phylogeny of Circoviridae Rep (**Figure 4.5**). The phylogeny showed two robustly supported clades emerging, corresponding to genera *Circovirus* and *Cyclovirus*. They contained a mixture of; (i) CVe from WGS assemblies; (ii) sequences obtained from virus isolates; (iii) sequences obtained from metagenomic samples (**Figure 4.5, Appendix 4.2**)

The first represented genus *Circovirus*. This grouping was based on the monophyletic grouping of well-described exogenous viruses, such as PCV-1/2, BaCV and BFDV, with CVe. (**Figure 4.5**). The second major clade corresponded to genus *Cyclovirus*, based on the monophyletic grouping of CVe and previously defined exogenous cycloviruses such as CyCV-VN, and Florida Woods cockroach

associated virus (**Figure 4.5**). In general, support at deeper nodes in the tree tended to be relatively low, with well supported subgroups emerging, reflecting the short and degraded nature of many CVe.

The *Circovirus* clade contained a mixture of CVe and exogenous sequences. Both were derived exclusively from vertebrates. Subgroups emerging within this clade showed well-supported definitions by host-class for three main groups, defined here as mammal 1, cyprinid 1 and avian 1. Each of these groups showed CVe grouping with exogenous sequences from the same host class (**Figure 4.5**). For example, beak and feather disease virus (BFDV) grouped robustly with a CVe that entered the germline of passeriform birds, while barbel circovirus grouped robustly with CVe from the genome of the golden line barbel, in a well-supported clade containing numerous CVe from ray-finned fish. The only sequence that contradicts this pattern is chimpanzee circovirus, which groups robustly within a cluster of avian viruses (**Figure 4.5**). However, this sequence was recovered in a metagenomics analysis of chimpanzee faeces, and a possible avian origin was noted following discovery (**Li *et al*, 2010**).

Figure 4.4: Phylogenetic relationships between Circoviridae and CRESS. Maximum likelihood phylogeny of aligned circovirus and CRESS amino acid Rep sequences. Genus/grouping indicated by labelled brackets. Scale bar indicates substitutions per site. Node labels indicate UFBOOT support (%).

The *Cyclovirus* clade contained three well-supported subclades, labelled cyclovirus 1-3 (**Figure 4.5**). Clade 1 consisted of viruses from vertebrate samples. Cyclovirus groups 2 and 3, however, contained intermingled sequences from vertebrate and invertebrate. In contrast to genus *Circovirus* (**Figure 4.5**) the clade structure did not reflect host associations in any way. Cyclovirus 2 contained invertebrate EVE sequences, derived from the genomes of the Varroa mite (*Varroa destructor*), and Asian bee mite (*Tropileilaps mercedesdae),* in addition to a cyclovirus derived from a bat. Cyclovirus 3 contained vertebrate and invertebrate cycloviruses derived exclusively from metagenomics analyses and CVe.Pseudomyrmex, derived from the genome of the elongate twig ant (*Pseudomyrmex gracilis*). Sequences in this clade did not follow a pattern according to host grouping. Branch lengths separating vertebrate from invertebrate viruses (and CVe) were relatively short, with the distance between

vertebrate and invertebrate cyclovirus sequences being far less than the distances separating vertebrate circoviruses, and CRESS sequences (**Figure 4.5**).

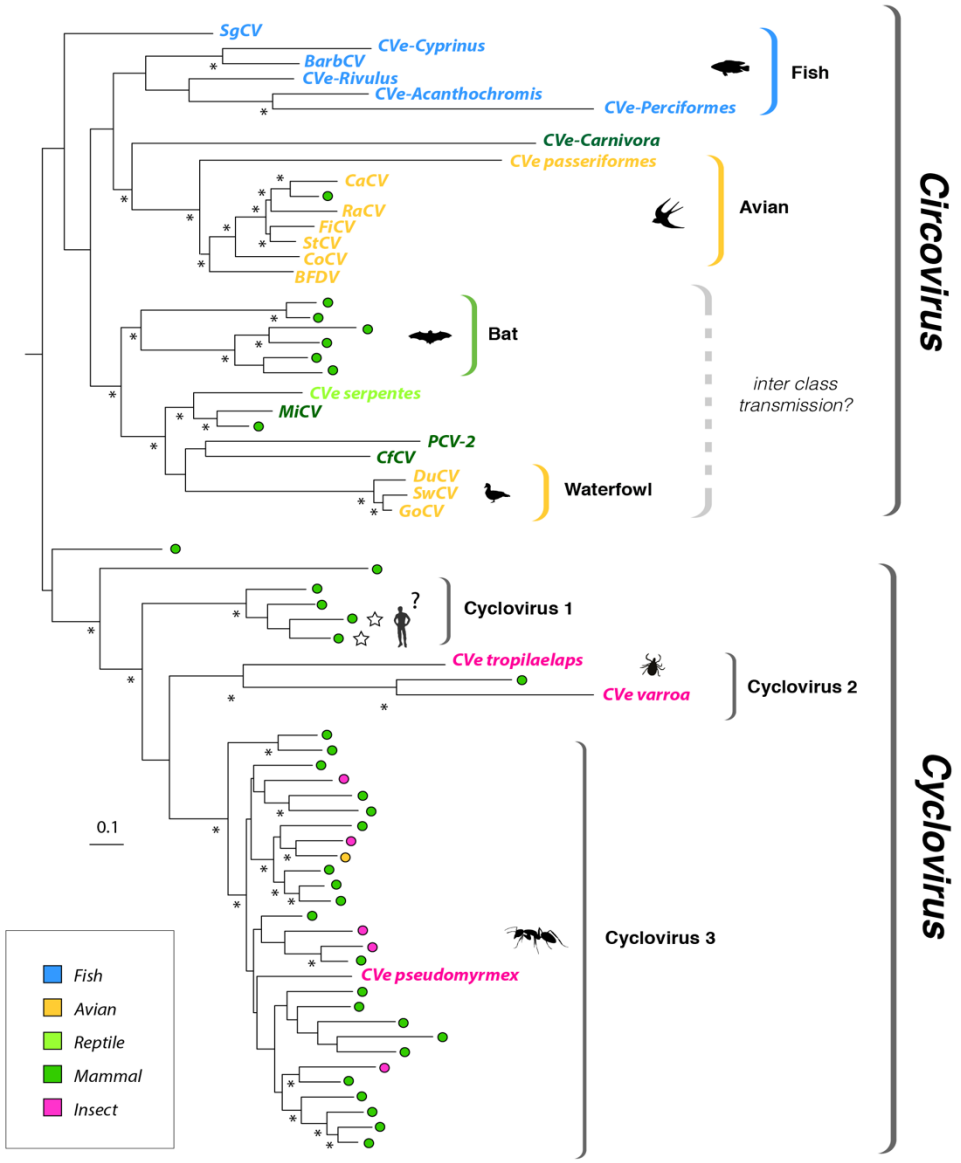

Figure 4.5: Maximum likelihood phylogeny of aligned amino acid Circoviridae Rep sequences. Sequences derived from metagenomic samples are indicated by colored circles. Taxa names are shown for sequences derived from viruses and CVe, and are coloured to indicate associations with host species groups, as shown in the key. Sequences derived from metagenomic sampling are indicated by circles coloured according to indicate sample associations with host species group. Pentagrams indicate viral taxa that have been linked to human disease. See **Appendices 4.1, 4.2 and 4.3** for accession numbers and details of taxa shown here. Asterisks indicate nodes with >70% bootstrap support. Scale bar indicates substitutions per site.

4.3.4 - Phylogenetic analysis of Circoviridae Cap

In general, the Cap protein was less conserved than the Rep protein, reflected by the higher level of substitutions per site (**Figure 4.6** – scale bar). Because of this, I excluded CRESS sequences from the alignment. The topology of the resulting phylogeny differed from that of the Rep phylogeny. In the midpoint-rooted Cap tree, Genus *Cyclovirus* remained monophyletic and was placed in a more derived position to Circovirus, grouping in the clade defined by fish circoviruses. The monophyletic clade comprising the remaining circovirus sequences grouped by host group (**Figure 4.6**) with mammal, snake, waterfowl, bat and bird clades forming in a in a similar manner to the Rep tree (**Figure 4.6**), but with weaker support. Comparatively robust (92%) support existed for the monophyletic clade containing the fish circoviruses and genus *Cyclovirus*. This topology suggests a different evolutionary history of Circoviridae Cap than that of Rep.

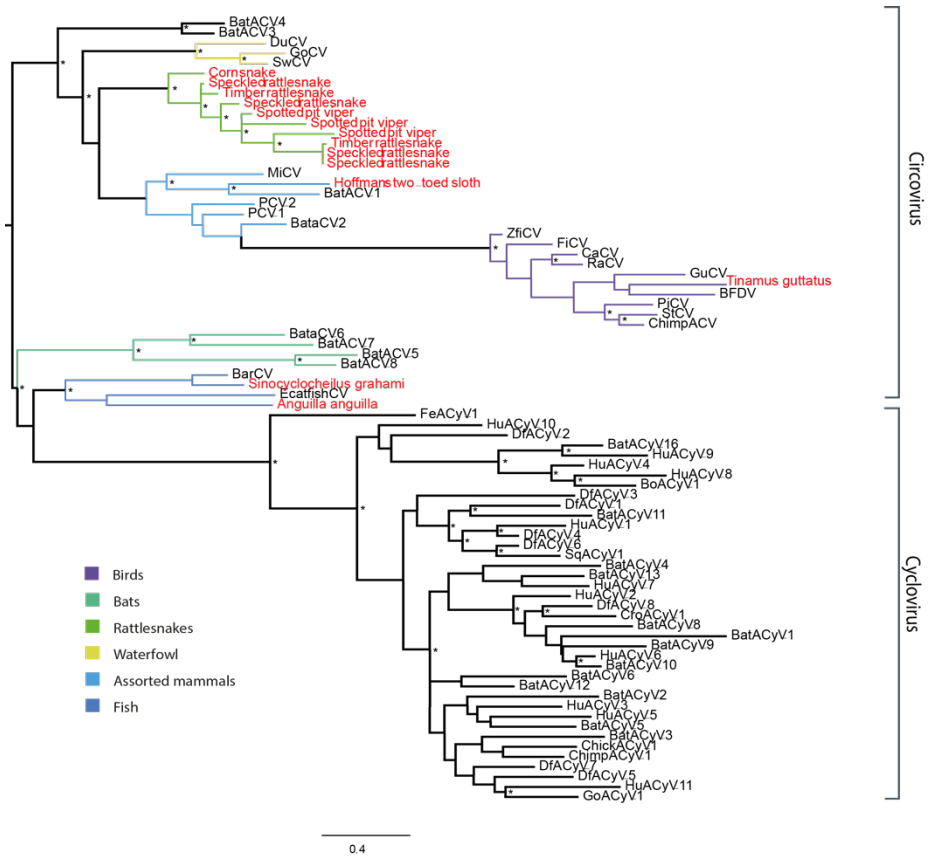

Figure 4.6: Maximum likelihood phylogeny of aligned amino acid Circoviridae Cap sequences. Genus indicated by labelled brackets. Scale bar indicates substitutions per site. Asterisks indicate UFBOOT support >95%. Red labels indicate CVe. Coloured branches indicate host grouping as denoted in the key.

4.3.5 - Species breakdown

4.3.5.1 - CVe in jawless fish

I found seven sequences similar to *rep* in the genome of the inshore hagfish (*Eptatretus burgeri*). These sequences are distinct from other circoviruses, grouping with CRESS sequences in the Rep phylogeny (**Figure 4.4**), and also showed relatively high genetic diversity relative to one another, forming three distinct clades (**Figure 4.7**), suggesting that these sequences are derived from three distinct germline invasions. This was supported by analysis of the putative Rep polypeptides encoded by these sequences. They contained several in-frame indels relative to one another, suggesting that the CVe in the hagfish are derived from three different germline invasions of three distinct, but relatively closely related viruses. These results should be taken with the caveat that no genomic flanking sequences were recovered for these sequences, making their status as CVe uncertain.

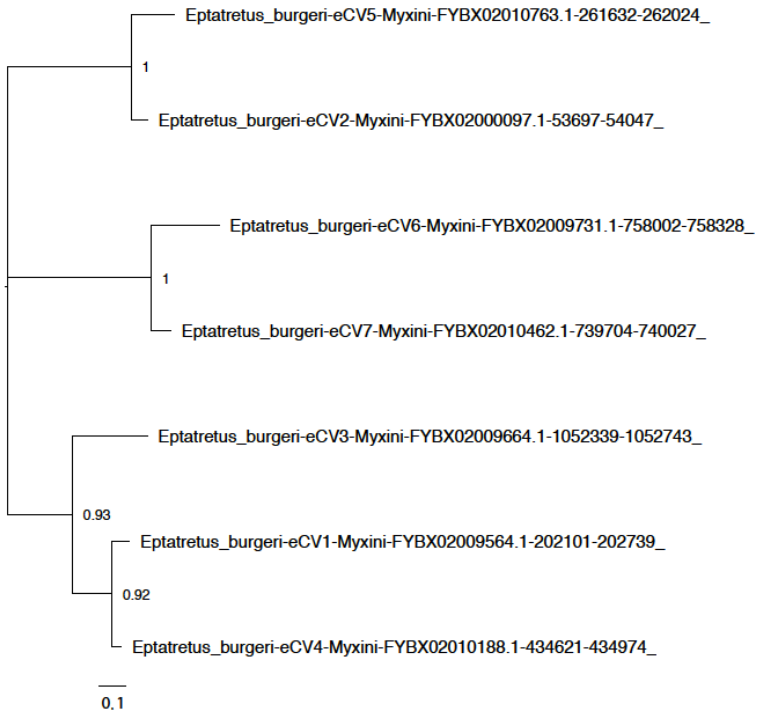

Figure 4.7: Maximum likelihood phylogeny of CVe Rep recovered from the genome of the inshore hagfish. (*Eptatretus burgeri*). Scale bar indicates substitutions per site. Node labels indicate % UFBOOT support.

4.3.5.2 - CVe in Actinoperygii

Endogenous and exogenous circoviruses have been reported in fish, with CVe being recovered in the genome of the Indian Rohu (*Labeo rohita)*, and circovirus nucleic acids and particles being recovered from fish tissues. **(Feher *et al*, 2013; Lorincz *et al*, 2012; Lorincz *et al*, 2011).** I identified additional CVe sequence in the genomes of ray-finned fishes (Class Actinopterygii) (**Table 4.1 and Table 4.3**). I found that CVe in the common carp (*Cyprinus carpio*) and golden-line barbel (*Sinocyclocheilus grahami*) genomes were orthologs with respect to each other, indicating they invaded the germline of cyprinid fish more than 39 million years ago (**Figure 4.2, Figure 4.3**). (**Ren *et al*, 2016**). These CVe consisted of multiple complete circovirus genomes arranged in tandem. They were highly similar (~70% nucleotide identity across 1654 nucleotides) to the barbell circovirus (BarbCV), grouping next to it in phylogenetic trees.

I also identified matches to *rep* in other species of ray-finned fish (**Table 4.1**). I could not determine how many integration events these CVe represented. All of these sequences group together in phylogenies (**Figure 4.5)**. A phylogeny of fish CVe shows that, generally, fish CVe can be sub-divided into based on host grouping. (**Figure 4.8**). Although the phylogeny is poorly supported, host group patterns can be loosely resolved. Perciform and cypriniform fish form the two largest host groups. An ortholog has been found in cyprinid fish, and the topology of the perciform fish clade is indicative of orthology, despite the grouping of a non-perciform sequence (CVe.Acanthochromis) within this clade. (**Figure 4.8**). These two groups cluster with single CVe from five other fish lineages in a manner that approximately follows that of the host species when rooted on the most basal host – the European eel (*Anguilla anguilla*) (**Figure 4.8)**. This indicates either a single ancestral integration event >200 million years ago (**Figure 4.3**)., or distinct invasion events. The observation that CVe elements in order Cypriniformes (golden-line barbel and carp) occur as full-length tandem genomes, whereas those in Perciformes (spiny chromis damselfish, tomato clownfish) are derived from more divergent fragments of rep, is suggestive of at least two separate invasion events. This is supported by the placement of the amphibian CVe.Xenopus and CVe.Rana in phylogenies, in which it splits the fish

CVe from one another, albeit with weak support (**Figure 4.8**). CVe in the mangrove rivulus (*Kryptolebias marmoratus*) and the common carp encoded complete intact rep genes (**Figure 4.1**) that are predicted to be expressed (**Table 4.3**), suggesting they may have been functionalized in some manner.

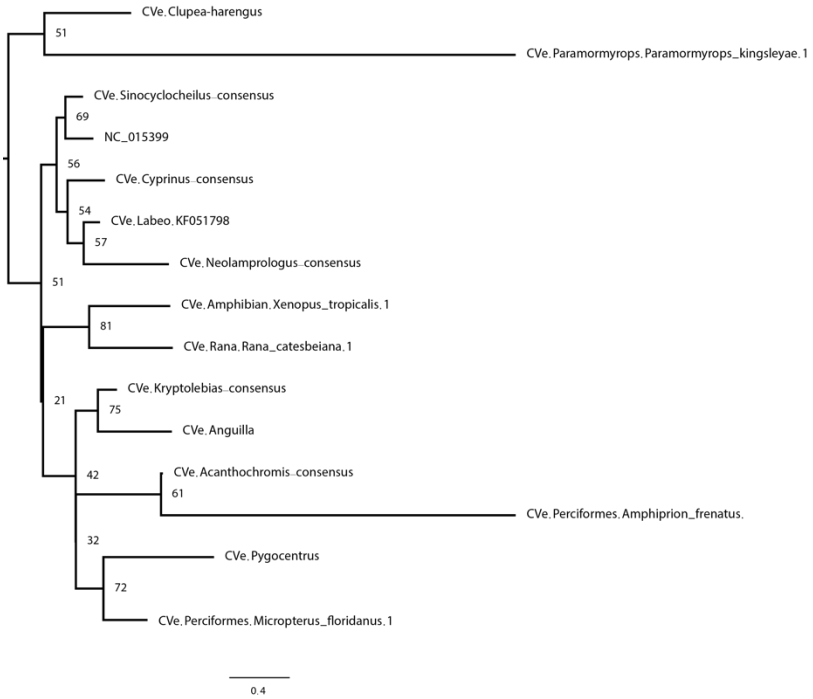

Figure 4.8: Phylogeny of aligned CVe Rep amino acid sequences recovered from fish and amphibian genomes. Node labels indicate % UFBOOT support. Scale bar indicates substitutions per site.

4.3.5.3 - CVe in amphibians

I identified a novel CVe in the genome of the American bullfrog (*Rana catesbiana*). This sequence overlaps slightly with a CVe previously identified in the genome of the Western clawed frog (*Xenopus tropicalis*) (**Liu *et al*, 2012**). The possibility that these CVe are orthologous implies a minimum estimated integration date of ~204 mya (**Figure 4.3**) (**Canatella *et al*, 2015; Roelants *et al*, 2011)**. However, I was unable to determine an orthology based on flanking genomic sequences, leading to their designation as potential orthologs only.

4.3.5.4 - CVe in reptiles

CVe orthologous in rattlesnakes (*Crotalus spp.*) have previously been identified (**Gilbert *et al*, 2015**). I identified CVe orthologous with respect to these in four additional snake species (**Table 4.1**). These orthologs indicate that this CVe is ~72-90 million years old – predating the radiation of sub-order *Serpentes*. (**Figure 4.3**). (**Chen *et al*, 2013; Wang *et al*, 2009**).

4.3.5.5 - CVe in birds

Similarity-search based analyses have found CVe in the genomes of the little egret (*Egretta garzetta*), white-throated tinamou (*Tinamus guttatus*), the kea (*Nestor notabilis*) and the medium ground-finch (*Geospiza fortis*), (**Cui *et al*, 2014**). I identified CVe in eight additional species (**Table 4.1**), some of which appeared to be orthologous with respect to previously reported CVe. I identified orthologs of the CVe previously identified in the medium ground finch in several other passeriform (perching/song bird) species (**Figure 4.3**). This demonstrates that this CVe integration predates the radiation of Passeroida approx. 38 mya (**Prum *et al*, 2015**; **Jetz *et al*, 2012**) (**Figure 4.2, Figure 4.3**). Furthermore, I identified potential orthologs of the CVe previously identified in the kea (**Cui *et al*, 2014**), in the Amazon (*Amazona aestiva)* and the rosy-faced lovebird (*Agapornis roseicollis*)**.** This suggests that these CVe are at least as old as the radiation of Psittaciformes, which occurred approx. 30-60 Mya (**Figure 4.3**). (**Prum *et al*, 2015**; **Jetz *et al*, 2012**).

In addition to the orthologous CVe in the genomes of Passerine and Psittaciform birds, I identified previously unreported CVe in the Japanese rail (*Gallirallus okinawae*: order Gruiformes) and downy woodpecker (*Picoides pubescens*: order Piciformes) (**Table 4.1**). Both these sequences were short and highly divergent, and as a result I could not determine their relationships to other CVe and circoviruses with confidence.

4.3.5.6 - CVe in mammals

The majority of CVe (n=100) identified in our screen were identified in carnivore genome assemblies. Phylogenetic and sequence analysis suggests that all of these CVe derive from 1-4 germline invasion events involving an ancient carnivore *rep* gene. This CVe appears to have undergone a striking intragenomic expansion in carnivore genomes: For example, the ferret genome contains at least 32 duplicated CVe sequences. **(Appendix 4.1)**.

To further study the evolutionary history of carnivore CVe, I generated an ML phylogeny of aligned carnivore Rep sequences. The phylogeny (**Figure 4.9**) indicated that at least four CVe insertions were present in the carnivore germline prior to the radiation of Carnivora, some of which have undergone expansion in carnivores. The phylogenetic relationships between carnivore CVe copies (**Figure 4.9**), particularly in the group CVe-Carnivora-4, indicate that these expansions have occurred independently in ursids (bears), pinnipeds (seals and walruses), and mustelids. It is possible that the reason why these CVe have proliferated is that they have become embedded into retroelements and copied along with these during intragenomic expansions. I investigated the regions flanking CVe-Carnivora and found that each CVe was embedded in a repeated 12kb region, that showed a high degree of homology to similar regions in the same and different species. I was unable to detect paired LTRs or retroviral genes, indicating that this sequence was not of ERV origin. The 15kb regions containing the repetitive sequence, and the carnivore CVe, were examined in DFAM (**Hubley *et al*, 2016**). This investigation showed that carnivore CVe were associated with long interspersed nuclear elements (LINE1). (**Appendix 4.3**) Three of the four expansion groups shown in Figure 4 were found to be associated with LINE-1 DNA, including the most basal. Given the age of this group of CVe in Carnivora to be at least 54 million years old (**Katzourakis and Gifford, 2010)**, it is likely that the fusion of CVe.Carnivora and LINE-1 predates the radiation of order Carnivora. The sequences appeared to be highly degraded, with the LINE1 ORFs disrupted by stop codons and indels. Curiously, DFAM was unable to locate one of the two canonical LINE1 ORFs (ORF1), instead only finding ORF2 and 3'/5' UTRs. (**Appendix 4.3**).

I identified a relatively intact CVe in the genome of the Ryukyu mouse (*Mus caroli*) that grouped with the exogenous canine circovirus (CaCV) recovered from dogs (**Decaro *et al*, 2014**). Additionally, I found degraded, divergent matches to *cap* and *rep* in the genome of the cape golden mole (*Chrysochloris asiatica*) on distinct contigs, meaning that I could not determine whether they derived from the same germline invasion event. CVe derived from cap were also identified in the genome of Hoffmann's two-toed sloth (*Choloepus hoffmanni*) (**Figure 4.1**).

CVe have previously been identified in marsupial genomes, having been found in the genome of the opossum (*Monodelphis domestica*). I identified CVe in the genomes of two australidelphian (Australian marsupial) species: the Tasmanian devil (*Sarcophilus harrisii*) and the koala (*Phascolarctos cinereus*). The Opossum CVe groups with mammalian CVe *rep* sequences derived from *Mus caroli*, PCV and CaCV. (**Figure 4.5**). However, the two novel CVe were short and highly degraded, and their position in phylogenies relative to other taxa was poorly supported. As such, the koala and devil sequences were omitted from **Figures 4.4 and 4.5**. I identified other short, divergent and degraded matches to Rep probes in the genome of the Sunda flying lemur (*Galeopterus variegatus*), Chinese pangolin (*Manis pentadactlya*), and the naked mole-rat (*Heterocephalus glaber*). (**Appendix 4.1, Figure 4.1**).

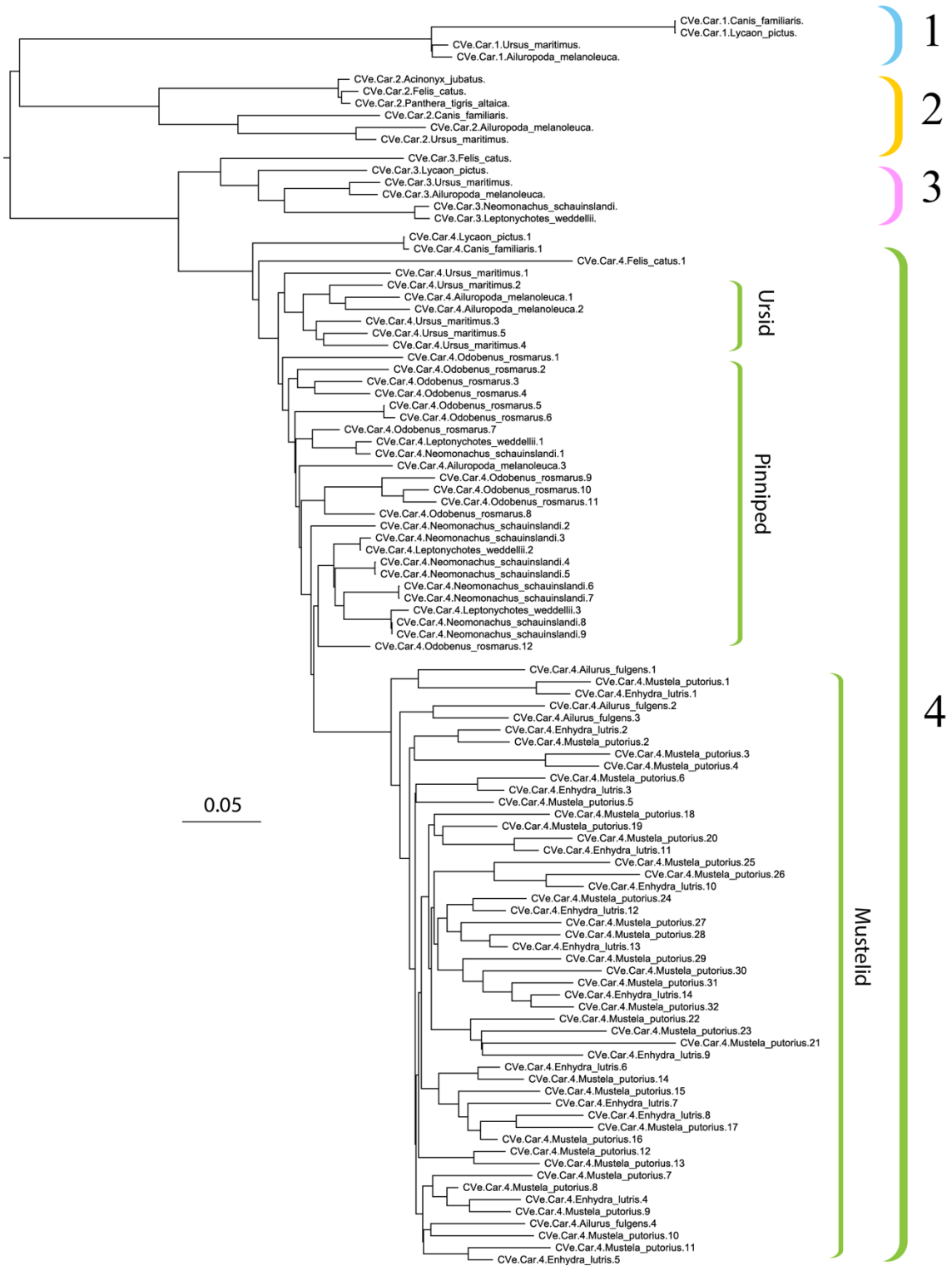

Figure 4.9: Phylogeny of CVe Rep amino acid sequences recovered from carnivore genome assemblies. At least four distinct CVe loci are present in the carnivore germline (clades I-IV) as indicated by the coloured brackets. Note the identifier (CVe.Car.#) indicating the expansion number. Scale bar indicates substitutions per site.

4.3.5.6 - CVe in mites

DNA sequences bearing a high similarity to cyclovirus *rep* genes were found in mite genomes. (**Appendix 4.1, Table 4.1**). CVe have previously been identified in the genome of the Varroa mite (**Liu *et al*, 2011**). I found 20 sequences matching cyclovirus Rep, and found evidence to support the designation of 18 of these as CVe (based on genomic flanking regions). Furthermore, an orthologous integration between *Varroa jacobsoni* and *Varroa destructor* was found using BLASTn searches (**Figure 4.2**). Unlike multicopy EVE integrations found in vertebrate species groups, Varroa CVe sequences were relatively divergent, and only 16 of them were incorporated into an alignment used to generate the consensus for CVe.Varroa. Four of these sequences were predicted to be expressed (**Table 4.3**).

Novel CVe were found in the genome of the Asian bee mite (*T. mercedesdae)*. Similarly to CVe.Varroa, the *Tropileilaps* CVe were relatively divergent, and of the eight recovered, only four were incorporated into the CVe.Tropileilaps consensus.

4.3.5.7 - CVe in ants

A Cyclovirus-like *rep* gene was also identified in the genome of the elongate twig ant (*Pseudomyrmex gracilis*). CVe-Pseudomyrmex was found to encode an intact *rep* ORF, which is predicted to be expressed (**Table 4.4**). CVe-Pseudomyrmex was no more distantly related to contemporary cycloviruses than many of them are to one another, including some that are associated with vertebrates (at least superficially) (**Figure 4.5**). This result was unexpected, and I sought confirm it through isolating CVe-Pseudomyrmex from the *P. gracilis* genome through PCR (performed by Peter Flynn).

Peter obtained genomic DNA from four species of ant belonging to genus *Pseudomyrmex (P. gracilis, P. elongatus, P. spinicola*, and *P. oculatus*). PCR was used to amplify a region encompassing part of the CVe, and part of the genomic flanking sequence. The PCR produced an amplicon of the expected size in all three DNA samples of *P.gracilis*: all other samples were negative (**Figure 4.10).**

The fact that CVe-Pseudomyrmex was not detected in other members of the genus suggests it invaded the *P. gracilis* germline after this species diverged from *P. elongatus*, *P. spinicola,* and *P. oculatus* in the mid-Miocene (**Gomez-Acevedo *et al*, 2010**). However, it was impossible to rule out that the failure to obtain an amplicon in these species was due to sequence divergence in the regions targeted by PCR primers.

4.3.5.8 - CVe in other invertebrates

I found 65 CRESS-DNA CVl sequences, derived from invertebrate genomes (**Table 4.1, Table 4.5**). The CVl were derived from a diverse range of hosts: Platyhelminthes, molluscs, arthropods and a cnidarian. CVl were highly divergent, with multicopy integrations from the same species group often bearing little similarity to one another. This was shown in **Figure 4.4,** with CRESS Rep sequences being separated from one another by long branches. This is a reflection both on the limited sampling of invertebrate taxa in genome sequencing efforts, and on the quality of many invertebrate genomes. Notably, a CVe was found in the genome of a Cnidarian – the parasitic myxosporean *Thelohanellus kiteaui*.

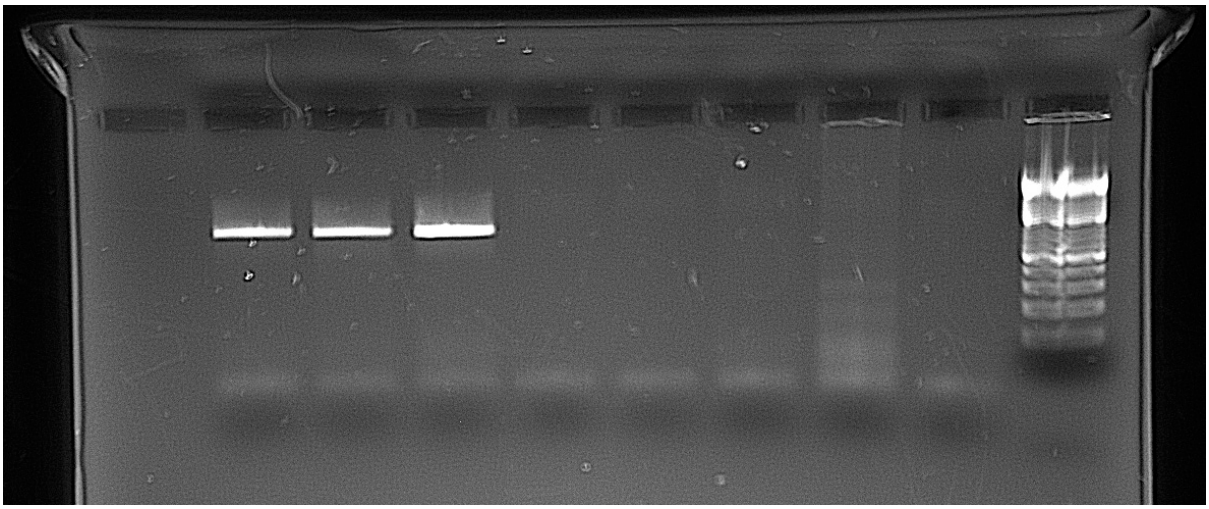

Figure 4.10. **PCR confirmation of CVe-*Pseudomyrmex* presence in three populations of *Pseudomyrmex gracilis*.** Columns: (1) negative control; (2) *Pseudomyrmex gracilis*, from the Florida Keys; (3) *P. gracilis*, from mainland Florida, USA; (4) *P. gracilis*, from Texas, USA; (5) *P. elongatus*, from the Florida Keys, USA; (6) *P. spinicola*, from Guanacaste Province, Costa Rica; (7) *P. oculatus*, from Cusco, Peru ; (8) *Cephalotes atratus,* from Cusco, Peru; (9) negative control; (10) ladder. The PCR was performed and figure prepared by Peter Flynn.

4.4 - Discussion

4.4.1 – Assessment of the methods

I found that DIGS was relatively well suited to searching animal genomes for CVe. I was able to screen a large number of animal genomes, and use MySQL to generate detailed summaries of CVe sequences broken down by host and virus taxonomy (**Table 4.1**). Downstream analysis of CVe was made more easy by the comparatively small size of the dataset (compared to **chapter 3**), and we were able to adopt a comparative approach whereby host:circovirus co-evolution was analysed across a wide range of species. Additionally, we were able to recover all CVe sequences and orthologs that have previously been described (**Table 4.1**), indicating that our approach for searching organism genomes for endogenous viral elements is robust.

4.4.2 - CVe provide retrospective information about circovirus evolution.

At least ten of the loci described here have been discovered previously (**Table 4.1**), and the majority of EVE sequences described here were orthologs (in the case of the cyprinid, bird and snake loci) or duplicated (in the case of the carnivore CVe) of these CVe. However, when these are discounted, this study represents an increase in known past circovirus host range of at least 39 species.

Novel orthologs were discovered in fish and birds, and previously characterised orthologs in snakes and carnivores were described in greater detail. This allowed us to establish the first minimum age estimates for some CVe loci, and to extend those of others **(Figure 4.2; Figure 4.3**). Thus, we were able to derive a more accurately calibrated timeline of evolution for the *Circovirus* genus, spanning multiple geological eras (**Figure 4.2**). The timeline indicated that circoviruses have been infecting a diverse range of vertebrate hosts over millions of years. The timeline (**Figure 4.3**), combined with the circovirus phylogeny (**Figure 4.5**) suggested a high degree of stability in circovirus host range, at least at high taxonomic levels. For example, fish circoviruses are at least 39 million years old and potentially up to 200 million years old (**Figure 4.3**). The Rep phylogeny shows this group of fish CVe and exogenous fish viruses to be monophyletic and exclusive to ray-finned fish

(**Figure 4.5**). As such, it appears that the circovirus-fish association is ancient, and that circoviruses have not been transmitted from fish to other species groups during their evolutionary history. This pattern is observed in other species groups – in snakes, birds and mammals, suggesting that circoviruses do not switch hosts readily. We observed an instance of potential inter-class transmission within one Circovirus sub-lineage that contains sequences obtained from anseriformes (waterfowl) and mammals (mink, bats, dogs and pigs) as well as CVe from snake genomes (see **Figure 4.5**). Within this clade, sequences from viruses of mink are robustly separated from those obtained from porcine and canine viruses by a CVe that invaded the serpentine germline >72 million years ago, based on its presence as an orthologous insertion in multiple snake species (**Gilbert et al, 2014**). This occurs as a single instance in the Rep phylogeny. When combined with the advanced age of the sequence groups involved, it suggests that if cross-class transmission occurs at all, it is relatively infrequent. However, many of the internal nodes in the Rep phylogeny lacked robust support, limiting our ability to extrapolate date calibrations determined by orthologous loci across the phylogeny. In addition, virus endogenisation is relatively rare, making the virus fossil record incomplete and obscuring other potential cross-species transmission events. Further sampling of circoviruses and CVe to allow the evolutionary relationships within the Circovirus genus to be determined with greater confidence, which will in turn allow more extensive calibrations to be made.

4.4.3 - Mapping the host range of circoviruses

The data presented in this chapter show a marked increase in the known past host range of circoviruses. Most of the new sequences discovered are orthologs or duplicates of one another (CVe.Carnivora, for example), but when these are discounted, 39 new species spanning 3 animal phyla are found to have been infected by circoviruses or circo-like viruses in their evolutionary past. Also of interest are the species groups *not* infected by circoviruses. Primates, for example, are well represented in genome sequence databases, yet evidence of primate infection by ciroviruses is conspicuously absent.

Knowledge of these host associations is useful, in that it can be used to guide contemporary virus discovery efforts, both in sampling exogenous viruses, and targeted genome sequencing of animal species for CVe.

The host makeup of the CRESS clade was, with the exception of CVl found in the genome of the inshore hagfish, confined to invertebrates, and largely those that occupy a marine habitat. The long branch lengths within this clade likely reflect divergences deep in evolutionary time, suggesting a wealth of diversity that is currently unknown. Circular ssDNA viruses have previously been associated with marine invertebrates, (**Rosario et al, 2012; Breitbart et al, 2015; Dunlap et al, 2013**) but until now, a conclusive host:virus interaction has not been proven beyond an association. The data presented definitively extend the host range of CRESS DNA to both marine and terrestrial invertebrates from three different phyla - Arthropoda, Cnidaria and Gastropoda. Further sampling of ssDNA viruses and invertebrate genomes would likely add valuable context to the nature and disease causing potential of this poorly studied, heterogenous group of viruses, as well as the possible origin of family Circoviridae.

Cycloviruses also have an uncertain host range. Prior to this analysis, cyclovirus sequences had only been identified through metagenomics analyses. In this study, we describe cyclovirus CVe in the elongate twig ant and two species of mite. As these are the only confirmed cyclovirus hosts, we propose that cycloviruses (at least clades 2 and 3) are invertebrate viruses that have contaminated metagenomics samples.

Arthropods are ubiquitous. For example, diverse mites live on animal skin and in house dust. (Making it interesting to note the presence of cyclovirus CVe in mites (**Table 4.3; Figure 4.5**)). As such, it is unsurprising that contamination would be commonplace. An alternative explanation is that transmission occurs readily between arthropods and vertebrates. When contamination is discounted, and the phylogeny is taken at face value, this appears to be the case, with ready host mixing between invertebrates and vertebrates taking place in cyclovirus clades 2 and 3 (**Figure 4.5**). This is unlikely, however, for three reasons:

Firstly, short branches between vertebrate and invertebrate cycloviruses give the impression of ready host mixing. This is incongruous with the phylogenetic evidence that circoviruses of the closely-related genus *Circovirus* are highly host-group specific. **(Figure 4.5**). Within this clade, host associations appear stable at high taxonomic levels (i.e. at the class level), with ancient CVe from particular host orders or classes seen grouping together with contemporary viruses, with possible evidence of infrequent cross-class transmission in occurring in *Circovirus* (**Figure 4.5**). In addition, the only sequence that seems to contradict this pattern of long-term host-group specificity is derived from a chimpanzee stool sample and likely reflects environmental contamination (**Li *et al*, 2010**). Also corroborating this this is the high distance between the invertebrate CRESS and the vertebrate circoviruses **(Figure 5**), which presents a contrast to the short distances between cycloviruses derived from vertebrate and invertebrate hosts (**Figure 4.5**). The apparent stability of circovirus host range demonstrated in the phylogeny (**Figure 4.5**) suggests that circoviruses do not readily cross barriers between distantly related hosts.

Secondly, there exists no evidence suggesting that circoviruses readily cross species barriers. Contemporary evidence exists suggesting that circoviruses do not travel between hosts of the same class. Since 2000, a large number of humans have been directly exposed to porcine circovirus 1 (PCV-1) through its presence as an contaminant in the live attenuated rotavirus vaccine (Rotarix) (**Victoria et al, 2010).** Despite this, PCV-1 is not thought to have infected any humans who received these vaccines, indicating that relatively strict barriers to cross-species transmission exist for these viruses.

Thirdly, humans exist in close proximity to animals that are routinely infected by circoviruses: birds and pigs. Despite the intimate role of these animals in human communities, no human circovirus infection has been reported. If cross-species transmission of circoviruses between distinct mammalian orders does not occur readily, then transmission between arthropod and vertebrate hosts appears unlikely.

The Varroa mite and elongate twig ant CVe are the first unambiguous evidence of host associations for cycloviruses. The confirmed presence of CVe-Pseudomyrmex in the ant genome (**Figure 4.10**), combined with the fact that it is predicted to be expressed, confidently indicates that CVe-Pseudomyrmex is an

endogenous cyclovirus. Since arthropods are the only confirmed cyclovirus hosts, contamination of vertebrate samples with arthropod viruses is the simplest explanation for the host associations described here.

Whereas the weight of evidence may favour cyclovirus clades 2 and 3 being exclusively arthropod viruses that frequently contaminate vertebrate samples, the status of the basal cyclovirus clade 1 is more equivocal. Cyclovirus 1 is comprised exclusively of sequences obtained from mammalian samples, and includes cycloviruses that have been posited as disease causing agents in humans (cyclovirus-VN and human cyclovirus VS5700009). Although the distance between this apparently mammalian lineage of cycloviruses and the invertebrate cycloviruses is still comparatively low compared to the distance between the CRESS and *Circovirus*, it is possible that these sequences represent a mammal or vertebrate-specific lineage of cycloviruses that is distinct from arthropod-infecting lineages. Notably, false-positive detection of human cyclovirus VS5700009 has been reported (**Chan *et al*, 2015**).

4.4.4 - Impact of CVe on host genome evolution

We found that 40% of CVe encoded intact ORFs, and that 13 were predicted to be expressed. The overwhelming majority of CVe consisted of Rep sequences. This may reflect a propensity of Rep to be exapted by the host, or some propensity of Rep sequences for insertion. This has been observed in general for EVEs derived from polymerase genes to encode and express intact ORFs. (**Katzourakis and Gifford, 2010; Arriagada and Gifford, 2014; Horie *et al*, 2016)**. An alternative explanation is that the higher degree of divergence in Cap proteins precludes it from discovery compared to the relatively conserved Rep.

We identified intact Rep ORFs that were predicted to be expressed as mRNA in fish and arthropods (**Table 4.3**). Particularly of interest is CVe-Pseudomyrmex, which we showed here to be a *bona fide* endogenous element (**Figure 4.10**). CVe-Pseudomyrmex encoded an intact *rep* gene product, and was predicted by genome annotation software (**see Methods**) to express mRNA (**Table 4.3**). The occurrence of an apparently fixed, intact, and expressed

circovirus *rep* gene in an ant genome provides further evidence that these genes have been co-opted or exapted by host species for as yet unknown functions. For example, functional genomic studies in insects indicate that endogenous viral element (EVE) sequences have been co-opted into RNA-based systems of antiviral immunity (**Whitfield *et al*, 2017**).

In addition, I found that the CVe.Carnivore sequence was predicted to be expressed in the ferret and polar bear (*Ursus maritimus*) as non-coding RNA (ncRNA), possibly reflecting its status as a retrotransposon **(Table 4.4)**. My discovery of expression predictions was dependent on genomes that have been subject to NCBI's annotation pipelines being present in NCBI's RefSeq RNA databases. As such, it is likely that my estimates here are conservative, and a potential for future study lies in the screening of organism transcriptome with CVe probes, to identify which CVe are expressed.

4.4.5 – Conclusions

In this chapter, I mined published animal genomes for endogenous circovirus and circovirus-like elements. This was the most comprehensive analysis of CVe to date, inasmuch as I searched all available animal genomes at the time of study, and retrieved the largest number of CVe of any previous studies of this type (**Liu *et al*, 2011; Katzourakis and Gifford, 2010; Belyi *et al*, 2010**). I analysed the CVe to gain insights into circovirus:host co-evolution by characterising their genome structures, assessing them for potential expression, making assessments of circovirus host range, calibrating the timeline of host:circovirus co-evolution, and performing phylogenetic analysis to assess the relationships of CVe to exogenous circoviruses.

I found that the approach I defined in **chapter 3** – phylogenetic screening of genome data, followed by detailed characterization, phylogenetic analysis and comparative analyses across multiple species, was well suited to the study of CVe, especially given that the relatively small dataset lent itself well to in-depth study (in contrast to my difficulty in leveraging large sequence volumes in **chapter 3**).

The ability of CVe to make inferences regarding host:circovirus co-evolution would be bolstered by the screening of more genomes. This would provide more CVe sequences for analysis. In turn, this would increase the potential for ortholog discovery, thus enabling more date calibrations to be made. In turn, the creation of more confident phylogenetic trees would allow more definitive statements to be made regarding circovirus:host co-evolution and range. In addition, the sampling of a greater diversity of invertebrate genomes would allow more insights into the poorly-defined CRESS-DNA viruses.

In this chapter, I show how the collection and study of CVe can be used to inform virus discovery efforts. The comprehensive screening of a large number of genome assemblies for EVE sequences can greatly extend the known host range of virus families, guiding sampling and discovery efforts. Secondly, where the phylogenetic relationships of virus sequences obtained using metagenomic approaches are examined in the relation to virus sequences for which a host association has been established, this provides a robust approach for checking the validity of metagenomics analyses: prior to the discovery and validation of CVe-Pseudomyrmex, the host association of cycloviruses was uncertain. EVE discovery provides robust conclusions where metagenomics lacks the ability to make firm statements on host range. Conversely, EVE discovery is limited by genome availability and quality, where metagenomics is limited only by the physical availability of samples. As such, the two approaches have the potential to make a powerful combination, greatly facilitating the discovery and classification of novel viruses and subsequent insights into host:virus co-evolution.

Chapter 5 – Searching for ISG expansion and loss in mammals

5.1 - Introduction

Antiviral genes can be used to gain insights into host:virus conflicts over evolutionary timescales. This approach is called indirect paleovirology (**Katzourakis, 2012**). In contrast to direct paleovirology – the study of endogenous viral elements- indirect paleovirology uses reconstruction of antiviral gene evolution to infer patterns of host-virus co-evolution. EVEs are relatively scant in animal genomes, and indirect paleovirology offers a method of studying host:virus co-evolution in the absence of EVEs, especially where the interface between host and virus has been studied. In addition to analysis of antiviral gene evolution, gene loci can be studied for dynamism: patterns of gene duplication, loss and recombination, as has been performed for the *mx*, *ifit*, and *parp* gene families, amongst others (**Braun *et al*, 2015; Mitchell *et al*, 2015; Daugherty *et al*, 2016; Daugherty *et al*, 2014**). Such analyses can be used to guide *in vitro* studies of species specificity and antiviral activity across multiple species.

The interferomes of various species have been studied using a single experimental method to allow the direct comparison of a snapshot of the interferome between multiple species. **Shaw *et al*,** (2017), used RNA-seq to characterise the interferomes of *Homo sapiens* (human), *Rattus norvegicus* (rat), *Bos taurus* (cow), *Ovis aries* (sheep), *Sus scrofa* (pig*), Equus caballus* (horse), *Canis lupus familiaris* (dog), *Myotis lucifugus* (little brown bat, microbat*), Pteropus vampyrus* (large flying fox, fruit bat), and *Gallus gallus* (chicken). In this study, we identified 90 mammalian ISGs – dubbed the 'core mammal' interferome that were consistently upregulated in response to interferon in nine mammalian species (**Shaw *et al*, 2017**). The 'core mammal' ISGs contained, amongst others, well known antiviral restriction factors, such as *mx1, ifit2* and *3, viperin* and *pml*, as well as proteins involved in multiple immune and cellular processes.

I used DIGS to investigate the evolution of core ISGs through identification of gene expansion and loss. I chose the core ISGs due to their increased

likelihood of (and in some cases known) involvement in the intrinsic immune response. (**Schoggins *et al*, 2010**). The transcriptomic analysis conducted by Shaw *et al,* (**2017**) was confined to nine species. By contrast, 111 mammalian genomes had been published at the time of study. As antiviral gene loci are prone to evolutionary innovation, including loss and expansion, the available genome data presented an opportunity to study this outwith established annotation pipelines. In the previous two chapters DIGS was used to interrogate a large number of animal genomes. DIGS can be configured to screen a large number of genome assemblies (**Dennis *et al*, 2018a; Dennis *et al*, 2018b)** and therefore provides a mechanism to screen mammalian genomes in search of lineage-specific patterns of ISG expansion and loss.

The aim of this chapter was to explore available mammalian WGS assemblies to find lineage specific patterns of ISG expansion and loss, and view these in the context of biological data. The availability of a wide range of mammalian genomes, coupled with the high potential for novel discoveries at antiviral gene loci (**Daugherty et al, 2016, Daugherty et al, 2015, Mitchell et al, 2015**), provides ample opportunity for a study of this kind. In addition to potential insights into host-virus co-evolution, this chapter presented an opportunity to assess the utility of DIGS in an analysis of host genes, as opposed to horizontally acquired viral sequences.

5.2 - Results

5.2.1 - Initial screening and results

I surveyed mammalian genomes for general, lineage specific patterns of gene expansion and deletion. I used DIGS to perform similarity searches of mammalian genomes with a panel of ISGs. The list of genes corresponded to the core mammalian ISGs we described in Shaw *et al*, (**2017**). (**Appendix 5.1**). For each gene, I extracted probe peptide sequences from ENSEMBL. The final list consisted of 90 core mammalian ISGs (**Appendix 5.1)**. I used a single reference library that consisted of all of the protein coding genes in the human genome, as this would then encompass existing human paralogs for each gene.

I screened 111 mammalian genomes (**Appendix 1.1**) with a library of 85 translated ISG peptide sequences **(Appendix 5.1**). The results were processed (see **Materials and Methods**) and displayed as a heatmap (**Figure 5.1**). When retrieving results from the database, we applied a bitscore cutoff to matching results to reduce the number of false-positive matches. The cutoff applied was the bitscore value which returned all of the exons for the probe gene from the human genome. For example, PARP9 has eleven exons. The highest bitscore that returned all of the eleven PARP9 exons from the human genome was 80. Therefore, only PARP9 hits with a bitscore of greater than 80 were retrieved from the database and counted. Data were normalised (see **Methods**) to account for the differences in exon counts between the genes screened.

I excluded a number of the 'core mammal' ISGs from the final results. The results for members of the *hla* (*a, b, c, d* and *e*), *apol* (*1, 2, 3* and *4*) and *tap* (*1/2/bp* and *bpl*) showed high levels of cross-matching and misclassification. I conjectured that this was because these gene groups are highly similar to one another, and occur in very gene-dense regions.

Overall, the heatmap showed a high degree of consistency for ISG counts across Mammalia (mean count of 1.17). (**Figure 5.1**). Generally, the heatmap did not show evidence of large-scale, lineage-specific gene expansion and deletion.

I found it difficult to establish with confidence whether the high counts for some genes (PKR, for example) were truly a result of functional gene duplication, pseudogenisation, or cross-matching of the probe with a similar but unrelated gene (for example, in protein families with modular domains i.e. KRAB-ZFPs). In previous chapters, I used methods that are relatively manually focussed following DIGS screens. Each ISG screen returned high numbers of hits matching the probe, even when a relatively high cutoff was applied, making it difficult to study in detail the results of the screen for each gene. As the aim of the study was to assess the more general landscape of ISG expansion and loss, it was unfeasible to apply a manually focussed approach to over one hundred large datasets. As such, I changed my focus to using this dataset to study gene deletions.

100 instances of count=0 for a gene in a species were shown by **Figure 5.1**. I documented these instances in **Table 5.1**. Where a gene possessed a count of zero, I followed it up by searching ENSEMBL (where the genome was listed), for evidence of that gene. This process was used in an attempt to confirm or refute the reported absence of a gene. Of the 100 instances of counts of zero, 64 appeared to be falsely reported, with annotations in ENSEMBL existing for these genes (**Table 5.1**). The remaining 36 represented possible gene deletions, with no sequence being reported in the database, and no annotation existing in ENSEMBL. I investigated these genes using BLAST and a genome browser (see **Materials and Methods**). *xaf1, ifit2* and *ifit3* were the genes that showed the most potential for multispecies deletions. *ifit2/3* were apparently missing in cetaceans (whales and dolphins, **Figure 5.1**), and *xaf1* was shown as missing in carnivores (**Figure 5.1**).

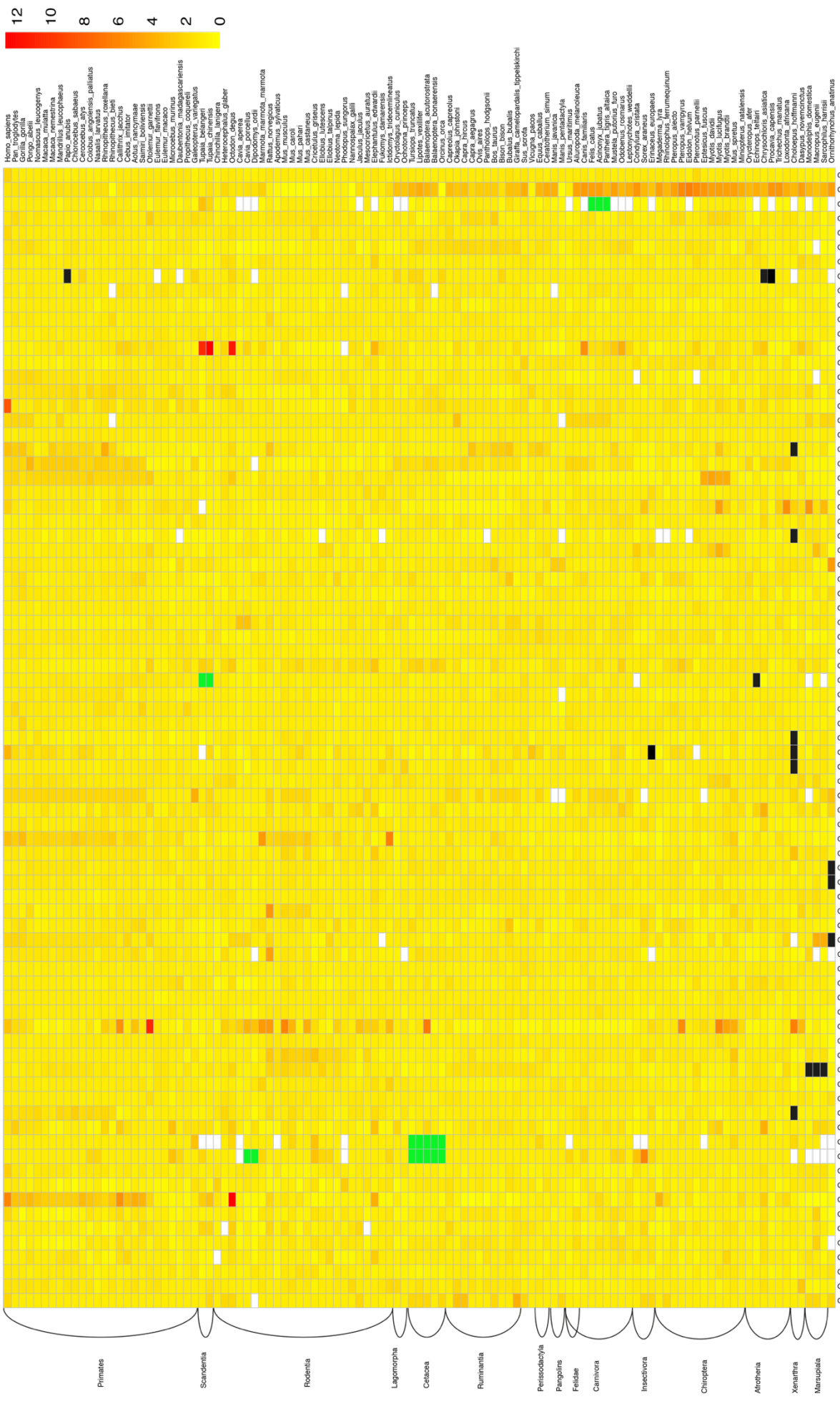
12
10
8
6
4
2
0
Homo_sapiens
Pan_troglodytes
Gorilla_gorilla
Pongo_abelii
Nomascus_leucogenys
Macaca_mulatta
Macaca_nemestrina
Mandrillus_leucophaeus
Papio_anubis
Chlorocebus_sabaeus
Colobus_angolensis_palliatus
Nasalis_larvatus
Rhinopithecus_roxellana
Rhinopithecus_bieti
Callithrix_jacchus
Cebus_imitator
Aotus_nancymaae
Saimiri_boliviensis
Otolemur_garnettii
Eulemur_flavifrons
Eulemur_macaco
Microcebus_murinus
Daubentonia_madagascariensis
Propithecus_coquereli
Galeopterus_variegatus
Tupaia_belangeri
Tupaia_chinensis
Chinchilla_lanigera
Heterocephalus_glaber
Octodon_degus
Cavia_aperea
Cavia_porcellus
Dipodomys_ordii
Marmota_marmota_marmota
Rattus_norvegicus
Apodemus_sylvaticus
Mus_musculus
Mus_spretus
Mus_pahari
Mus_castaneus
Cricetulus_griseus
Ellobius_lutescens
Ellobius_talpinus
Neotoma_lepida
Jaculus_jaculus
Mesocricetus_auratus
Elephantulus_edwardii
Elephantulus_rozeti
Ictidomys_tridecemlineatus
Oryctolagus_cuniculus
Ochotona_princeps
Sorex_araneus
Lipotes_vexillifer
Balaenoptera_acutorostrata
Balaenoptera_bonaerensis
Orcinus_orca
Capreolus_capreolus
Okapia_johnstoni
Capra_hircus
Capra_aegagrus
Ovis_aries
Panthalops_hodgsonii
Bos_taurus
Bison_bison_bison
Bubalus_bubalis
Giraffa_camelopardalis_tippelskirchi
Sus_scrofa
Vicugna_pacos
Equus_caballus
Ceratotherium_simum
Manis_javanica
Manis_pentadactyla
Ursus_maritimus
Ailuropoda_melanoleuca
Canis_lupus_familiaris
Felis_catus
Acinonyx_jubatus
Panthera_tigris_altaica
Mustela_putorius_furo
Odobenus_rosmarus
Leptonychotes_weddellii
Condylura_cristata
Sorex_araneus
Erinaceus_europaeus
Megaderma_lyra
Rhinolophus_ferrumequinum
Pteropus_alecto
Pteropus_vampyrus
Eidolon_helvum
Pteronotus_parnellii
Eptesicus_fuscus
Myotis_davidii
Myotis_lucifugus
Myotis_brandtii
Mus_spretus
Microtus_ochrogaster
Oryctolagus_afer
Echinops_telfairi
Chrysochloris_asiatica
Procavia_capensis
Trichechus_manatus
Loxodonta_africana
Choloepus_hoffmanni
Dasypus_novemcinctus
Monodelphis_domestica
Macropus_eugenii
Sarcophilus_harrisii
Ornithorhynchus_anatinus
CORE_ZNFX
CORE_ZCCH
CORE_XAF1
CORE_WARS
CORE_TDRD
CORE_SP11(
CORE_SLC2!
CORE_SERT/
CORE_FMR1
CORE_FAM4(
CORE_EHD4
CORE_DNAJ(
CORE_DNAJ/
CORE_CMTR
CORE_CMPK
CORE_CASP
CORE_C2
CORE_TNFS(
CORE_RICT(
CORE_LGAL!
CORE_IL15R
CORE_CD47
CORE_USP1(
CORE_TRIM/
CORE_TRAF(
CORE_SOCS
CORE_PARP
CORE_NMI
CORE_IFI35
CORE_CD27/
CORE_USP2(
CORE_TRIM/
CORE_STAT2
CORE_STAT1
CORE_RNF1
CORE_DDX5
CORE_MYD8
CORE_IFIH1
CORE_DHX5
CORE_IRF9
CORE_IRF7
CORE_IRF1
CORE_AZI2
CORE_UBE2(
CORE_UBA7
CORE_RNF3
CORE_RNF2
CORE_RNF1(
CORE_RBCK
CORE_PARP(
CORE_NUB1
CORE_N4BP
CORE_HERC
CORE_DTX3(
CORE_ZC3H(
CORE_SHISA
CORE_SAT1
CORE_RSAD
CORE_PML
CORE_EIF2A
CORE_PARP
CORE_OAS
CORE_MX1
CORE_MOV1
CORE_MOR(
CORE_ISG2(
CORE_ISG15
CORE_IFIT3
CORE_IFIT2
CORE_ADAR
CORE_RFX5
CORE_PSME
CORE_PSME
CORE_PSMB
CORE_PSMB
CORE_PSMB
CORE_NLRC
CORE_ERAP
CORE_B2M
Primates
Scandentia
Rodentia
Lagomorpha
Cetacea
Ruminantia
Perissodactyla
Pangolins
Felidae
Carnivora
Insectivora
Chiroptera
Afrotheria
Xenarthra
Marsupiala

Figure 5.1 (preceding page) – A landscape view of the heatmap showing normalised (**see Methods**) results of DIGS searches of 111 mammalian genomes for sequences disclosing homology to core mammalian ISGs. The X axis indicates gene common names (**listed in Appendix 5.1**). Text on the right is the latin binomial for species, with species groups indicated in brackets on the left. Heatmap colours keyed in the bar on the top right. Green indicates a confirmed deletion. White indicates spuriously reported counts of zero, and black indicates unconfirmed deletion.

Table 5.1: Reported core mammal ISG deletions in Figure 1. Genes were investigated in ENSEMBL, and their presence as annotations recorded in the 'ENSEMBL' column. Genes shown to be deleted in **Figure 1** without annotations in ENSEMBL were marked as 'FALSE' in the 'ENSEMBL' column and investigated further, the results of which were noted in the 'Reported' column.

Species	Gene name	Annotation in ENSEMBL	Reported
Ailuropodia melanoleuca	IFIT2	TRUE	
Apodemus sylvaticus	RFX5	TRUE	
Baleonoptera bonarensis	FMR1	TRUE	
Canis familiaris	XAF1	TRUE	
Cavia aperea	XAF1	TRUE	
Cavia porcellus	XAF1	TRUE	
Chinchilla lanigera	IFIT2	TRUE	
Choloepus hoffmani	DTX3L	TRUE	
Choloepus hoffmani	XAF1	TRUE	
Condylura cristata	CMPK2	TRUE	
Condylura cristata	IFIT2	TRUE	
Condylura cristata	RIG-I	TRUE	
Daubentonia madagascarensis	SERTAD1	TRUE	
Dipodomys ordii	B2M	TRUE	
Dipodomys ordii	IFIT2	TRUE	
Dipodomys ordii	ZC3HAV1	TRUE	
Echinops teifari	XAF1	TRUE	
Eidolon helvum	XAF1	TRUE	
Elephantulus edwardii	XAF1	TRUE	
Equus caballus	IRF1	TRUE	
Equus caballus	RFX5	TRUE	
Erinaceus europaeus	XAF1	TRUE	
Eulemur flavirons	RFX5	TRUE	
Eulemur macaco	RFX5	TRUE	
Heterocephalus glaber	PSMB10	TRUE	
Heterocephalus glaber	PSMB9	TRUE	
Jaculus jaculus	XAF1	TRUE	
Leptonychotes weddelli	XAF1	TRUE	
Macropus eugenii	PSME2	TRUE	

Macropus eugenii	RIG-I	TRUE	
Macropus eugenii	ZC3HAV1	TRUE	
Manis javanica	FMR1	TRUE	
Manis javanica	TNFSF10	TRUE	
Manis javanica	XAF1	TRUE	
Mesocricetus auratus	PSMB10	TRUE	
Mesocricetus auratus	PSMB9	TRUE	
Monodelphis domestica	PSME2	TRUE	
Monodelphis domestica	RIG-I	TRUE	
Monodelphis domestica	STAT2	TRUE	
Monodelphis domestica	XAF1	TRUE	
Mustela furo	XAF1	TRUE	
Nannospalax galili	ZCCHC2	TRUE	
Ochotona princeps	XAF1	TRUE	
Ochotona princeps	ZC3HAV1	TRUE	
Odobenus rosmarus	XAF1	TRUE	
Orcinus orca	ZCCHC2	TRUE	
Ornithorhyncus anatinus	PSME1	TRUE	
Ornithorhyncus anatinus	ZC3HAV1	TRUE	
Oryctolagus cuniculus	RFX5	TRUE	
Oryctolagus cuniculus	XAF1	TRUE	
Parnell's mustached bat	IRF7	TRUE	
Phodopus sungorus	FMR1	TRUE	
Phodopus sungorus	IFIT3	TRUE	
Phodopus sungorus	IFIT2	TRUE	
Propithecus coquereli	RFX5	TRUE	
Pteronotus parnelli	CMPK2	TRUE	
Rhinopithecus roxellana	TNFSF10	TRUE	
Sarcophils harrisi	PSME2	TRUE	
Sarcophils harrisi	RIG-I	TRUE	
Sorex aranaeus	IFIT2	TRUE	
Tupaia belangeri	IFIT3	TRUE	
Tupaia chinensis	IFIT3	TRUE	
Tupaia chinensis	TRIM21	TRUE	
Ursus maritimus	XAF1	TRUE	
Acinonyx jubatus	XAF1	FALSE	Confirmed
baleonoptera acurostrata	IFIT2	FALSE	Confirmed
baleonoptera acurostrata	IFIT3	FALSE	Confirmed
Baleonoptera bonarensis	IFIT2	FALSE	Confirmed
Baleonoptera bonarensis	IFIT3	FALSE	Confirmed
Cavia aperea	IFIT2	FALSE	Daugherty *et al,* 2016
Cavia porcellus	IFIT2	FALSE	Daugherty *et al,* 2016
Choloepus hoffmani	IRF1	FALSE	Contig missing
Choloepus hoffmani	IRF7	FALSE	Contig missing
Choloepus hoffmani	IRF9	FALSE	Contig missing

Choloepus hoffmani	ISG20	FALSE	Contig missing
Choloepus hoffmani	LGALS9	FALSE	Contig missing
Choloepus hoffmani	SOCS1	FALSE	Contig missing
Chrysochloris asiatica	SERTAD1	FALSE	False negative
Echinops teifari	RIG-I	FALSE	Contig missing
Erinaceus europaeus	ZC3HAV1	FALSE	Contig missing
Erinaceus europaeus	IRF7	FALSE	Contig missing
Felis catus	XAF1	FALSE	Confirmed
Lipotes vexillifer	IFIT2	FALSE	Confirmed
Lipotes vexillifer	IFIT3	FALSE	Confirmed
Macropus eugenii	Mx1	FALSE	Mitchell, 2015
Monodelphis domestica	Mx1	FALSE	Mitchell, 2015
Orcinus orca	IFIT2	FALSE	Confirmed
Orcinus orca	IFIT3	FALSE	Confirmed
Ornithorhyncus anatinus	DTX3L	FALSE	Contig missing
Ornithorhyncus anatinus	PARP9	FALSE	Contig missing
Ornithorhyncus anatinus	SERTAD1	FALSE	Contig missing
Panthera tigris altaica	XAF1	FALSE	Confirmed
Papio anubis	SERTAD1	FALSE	Probable deletion
Procavia capensis	SERTAD1	FALSE	Contig missing
Sarcophils harrisi	Mx1	FALSE	Mitchell, 2015
Tupaia belangeri	IRF7	FALSE	Contig missing
Tupaia belangeri	RIG-I	FALSE	Xu, 2016
Tupaia chinensis	RIG-I	FALSE	Xu, 2016
Tursiops truncatus	IFIT2	FALSE	Confirmed
Tursiops truncatus	IFIT3	FALSE	Confirmed

5.2.2 – IFIT

To investigate the *ifit* locus in cetaceans, I compared genomic regions containing *ifit2* and *ifit3* in the cow (*Bos taurus*) (**Figure 5.2**) to that of the bottlenose dolphin (*Tursiops truncatus)*, minke whale (*Balaenoptera acutorostrata),* baiji (*Lipotes vexillifer*), sperm whale (*Physeter catodon*) and the orca (*Orcinus orca*). I generated query sequences using the 5' flank, and first 80bp, of the cow *ifit2*, and used them in BLASTn searches of cetacean genomes (**Figure 5.2** – green bar). I found significant matches to the 5' flanking region, but no matches to the full *ifit3* sequence, **(Table 5.2)** in every cetacean species searched. I found matches to fragments of *ifit3* in the minke whale, baiji, and bottlenose dolphin, but aside from this, *ifit3* was deleted. This suggests that this gene has been totally lost in cetaceans.

In our genomic comparisons, I found evidence of *ifit2* sequences in cetaceans, despite them being shown as a count of zero on the heatmap (**Figure 5.1**). Upon investigation of the cetacean *ifit* locus using BLAST, I found that the *ifit2* sequence was present in cetaceans, albeit degraded. I performed confirmatory BLASTn searches using cow *ifit2* sequences, and found that cetacean *ifit2* was indeed present (**Table 5.3**). When these sequences were aligned to the cow *ifit2* as a reference, I found clear evidence of pseudogenisation. In 4/5 cases, the *ifit2* ORF was interrupted by a stop codon (**Figure 5.4**). The baiji sequence possessed a largely intact ORF, but was truncated by 62 residues in the C-terminal domain (**Figure 5.4**). When taken together, these findings show that *ifit3* has likely been lost in cetaceans, and that *ifit2* has undergone pseudogenisation. (**Figure 5.5**).

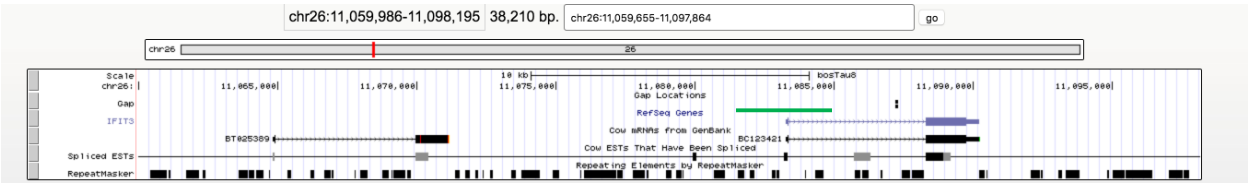

Figure 5.2: UCSC genome browser view of the *Bos taurus ifit2* and *ifit3* locus. Green line indicates region of sequence used as a BLAST query in cetaceans.

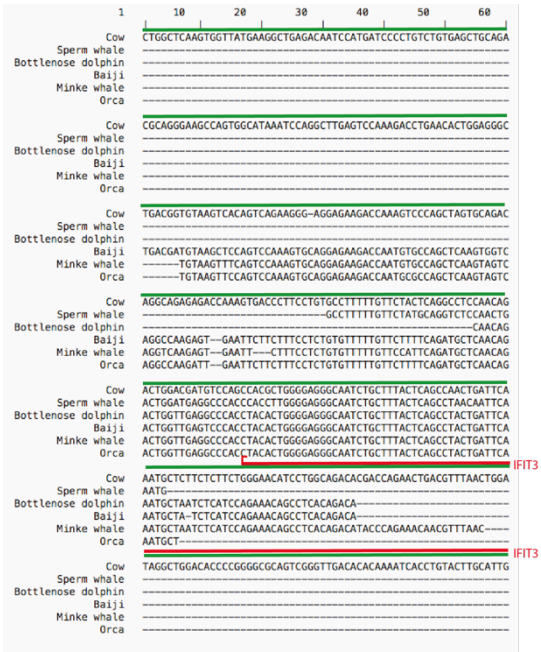

Figure 5.3: *ifit3* has been deleted in cetaceans. Text alignment view of the start of the *Bos taurus ifit3* CDS, and 5' flanking genomic region. Aligned to this region are BLAST hits from five cetacean species, retrieved using the orthologous cow region as a query (highlighted here, and in **Figure 5.2** in green), to identify the area where *ifit3* has been deleted, Red line indicates the *ifit3* CDS.

Table 5.2: Results for BLASTn queries of cetacean genomes with queries of cow genomic regions overlapping IFIT3 (indicated by the green line in **Figure 5.2** and **Figure 5.3**).

Species	Subject accession	% identity	length	mismatches	gap opens	q. start	q. end	s. start	s. end	evalue	bit score
Bhaji	NW_006773751.1	76.818	220	37	11	46	259	1657969	1658180	5.20E-21	111
Minke whale	NW_006729234.1	75.862	232	46	8	52	280	1172435	1172659	1.87E-20	110
Sperm whale	NW_019873573.1	86.316	95	11	2	135	228	31534129	31534222	3.13E-18	102
Orca	NW_004438446.1	76.796	181	37	5	52	230	13784843	13784666	1.46E-16	97.1
Bottlenose dolphin	NW_017844429.1	83.168	101	17	0	159	259	16138533	16138633	1.88E-15	93.5

Table 5.3: Results for BLASTn queries of cetacean genomes with queries of the cow IFIT2 sequence. Matching sequences were aligned to the cow IFIT2 sequence in **Figure 5.4**.

Species	Subject accession	% identity	length	mismatches	gap opens	q. start	q. end	s. start	s. end	evalue	bit score
Orca	NW_004438524.1	89.819	1385	137	4	1	1383	4421940	4423322	0	1773
Bottlenose	NW_017842131.1	89.17	1385	145	5	1	1383	8574877	8576258	0	1722
Sperm whale	NW_019873556.1	89.113	1387	140	11	1	1383	38546905	38545526	0	1714
Bhaji	NW_006785831.1	88.672	1386	152	5	1	1383	14738	16121	0	1685

Table 5.4: Results for BLASTn sequences of felid genomes with queries of the **cat** genomic region homologous to the **dog** *xaf1* containing genomic region. Matching sequences were aligned to the dog *xaf1* containing region in **Figure 5.5.**

Species	subject acc.ver	% identity	length	mismatches	gap opens	q. start	q. end	s. start	s. end	evalue	bit score
Cat	NC_018736.3	100	1080	0	0	1	1080	733139	734218	0	1948
Leopard	NW_017619906.1	97.407	1080	28	0	1	1080	13850281	13849202	0	1822
Cheetah	NW_015131202.1	95.833	1080	31	2	1	1080	540830	541895	0	1748
Amur tiger	NW_006712040.1	95.093	1080	29	1	1	1080	126126	125071	0	1727

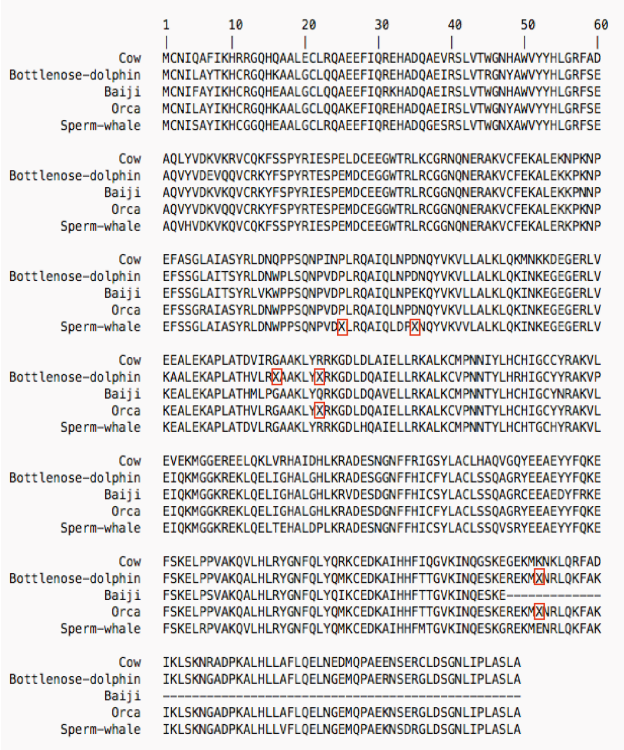

Figure 5.4: *ifit2* has been pseudogenised in cetaceans. Alignment view of translated in-frame cetacean pseudo- *ifit2*. Sequences were aligned to *Bos taurus ifit2* as a reference. Residues highlighted in red indicate stop codons.

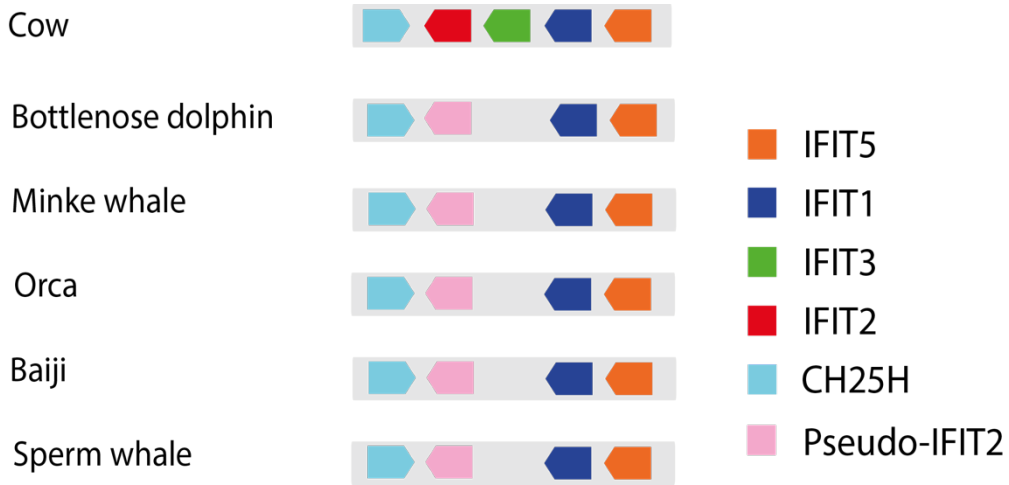

Figure 5.5: Genome browser schematic of the cetacean IFIT2 and IFIT3 loci. Shows a schematic of gene loss and pseudogenisation at IFIT2 and IFIT3 in cetaceans, using the syntenic cow region as a reference. Coloured arrows indicate gene and direction of transcription.

5.2.3 - XAF1

The heatmap showed counts of 0 for *xaf1* in most carnivores (**Figure 5.1**). Upon closer investigation in ENSEMBL, I found that the counts for *xaf1* in caniform (dog-like) carnivores was likely misrepresented, as annotations for *xaf1* were found. In felid carnivores (cats), I found no annotations in ENSEMBL. I investigated the genomic region containing *xaf1* in carnivores, by comparing the *xaf1* containing region in the domestic dog (*Canus lupus familiaris)* and the syntenic region in the cheetah (*Acinonyx jubatus*), cat (*Felis catus*), leopard (*Panthera pardus*), and Amur tiger (*Panthera tigris altaica*). In initial genome browser searches, and sequence alignments, I found that exons 4 and 5 of *xaf1* had likely been retained in cats, based on sequence conservation in this region shown in the UCSC genome browser(**Figure 5.6**; **Figure 5.8**). BLAST searches of the extracted dog sequences returned no matches in felids. I instead extracted this sequence from the cat genome and used it in BLASTn searches of felid genomes, returning matches in all of the species detailed above (**Table 5.4**). The matching sequences were used to generate an MSA of *xaf1* in carnivores (**Figure 5.5**). From the aligned flanking sequences, and the possible presence of exons 4 and 5, I suggest that *xaf1* has undergone at least a partial (if not a whole) deletion in felids (**Figure 5.8**).

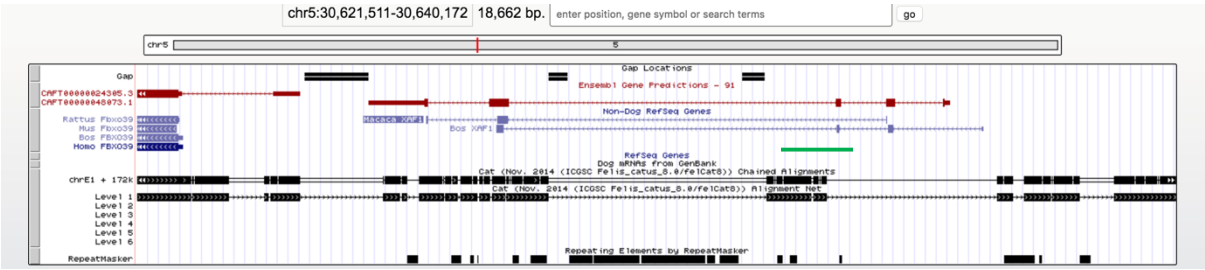

Figure 5.6: Screenshot of the UCSC genome browser centred on the dog *xaf1* containing region. Browser track chrE1 indicates aligned regions of the cat genome. Green bar indicates region used in BLAST queries of felid genomes.

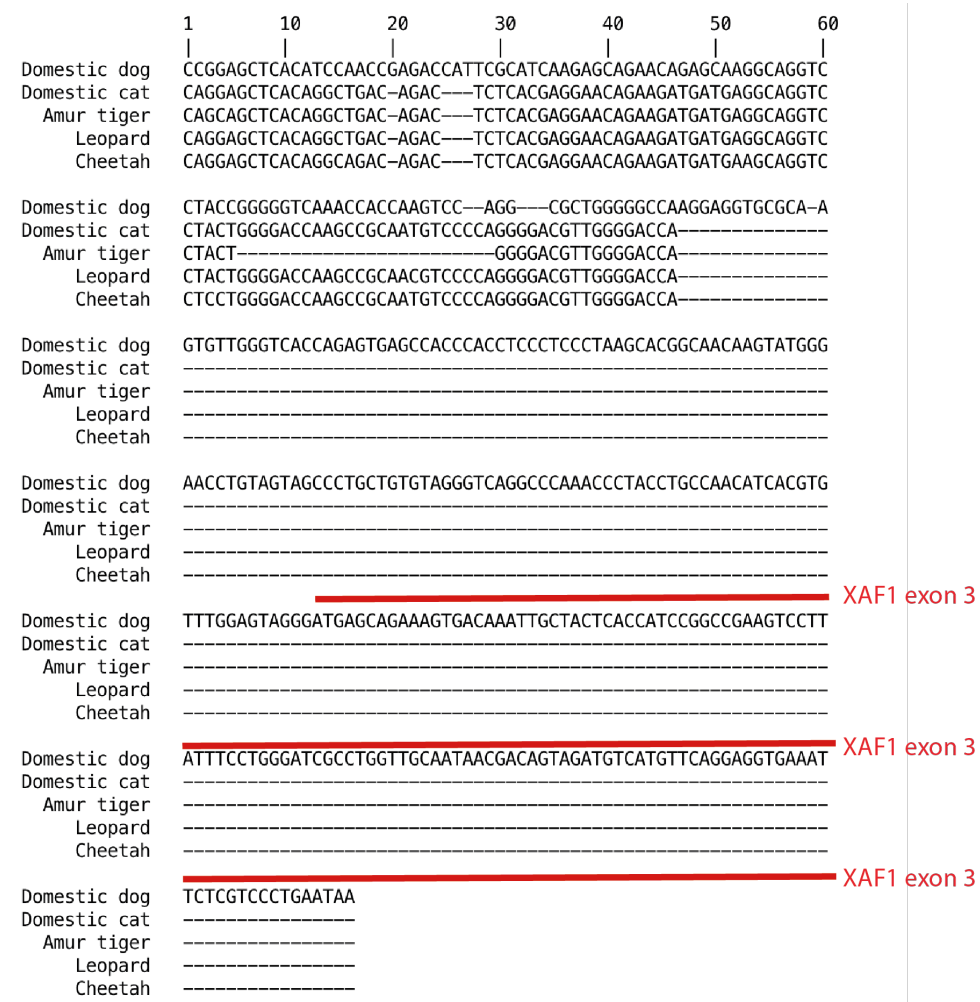

Figure 5.7 Alignment view of dog XAF1 intron and exon (indicated in red line) sequences. Aligned to this genomic region are matching felid genomic regions retrieved by BLAST searched of felid genomes with the probe sequence (indicated in **Figure 5.6**), showing the region where *xaf1* has been deleted.

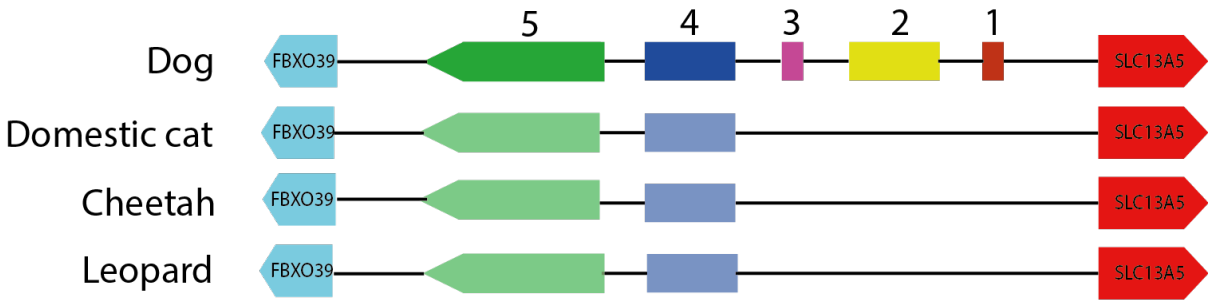

Figure 5.8 Genome browser schematic of the carnivore XAF1 locus, using the dog (top) as a reference. Numbered arrows indicate gene exons. Coloured labelled arrows indicate neighbouring genes.

5.2.4 - SERTAD1

The heatmap reported a count of zero for *sertad1* in the Cape golden mole (*Chrysochloris asiatica*). Searches in ENSEMBL in the rock hyrax (*Procavia* capensis), lesser hedgehog tenrec (*Echinops terifari*), Cape golden mole and elephant (*Loxodonta africana*) revealed that there were no annotations available for SERTAD1 in this species. This prompted a search for a potential deletion of *sertad1* in Afrotherians. I searched the region between the two genes flanking *sertad1* in mammals (*sertad3* and *prx*), in the UCSC genome browser, and found annotated sequences matching *sertad1* in all of the above species.(**Figure 5.9**).

5.2.5 - Other potentially deleted genes

We investigated the remaining 26 counts of zero (**Table 5.1**) using BLAST and a genome browser. I found seven deletions that had been previously characterised (**Table 5.1**). This included a deletion of *rig-I* in shrews (**Xu *et al*, 2016**), a deletion of *ifit2* in guinea pigs and marsupials (**Daugherty *et al*, 2016**) and a deletion of *mx1* in marsupials (**Mitchell *et al*, 2015**).

The olive baboon (*Papio anubis*), Platypus (*Ornythorhynchus anatinus),* the lesser hedgehog tenrec, the hedgehog (*Erinaceus europaeus)* and the sloth were the four remaining species with potential deletions (n=14 total). There were two problems in investigating these results. The first was the fractured nature of the genome assemblies – the generally low contig lengths made it difficult to study the genome organisation surrounding the locus of the gene of interest. Secondly, I experienced great difficulty in finding homologous flanking regions in the missing genes. For example, when we searched the sloth genome for evidence of *irf7* deletion, I used the sequence flanking the armadillo (*Dasypus novemcinctus*) *irf7* as a query. No significant results were found. As I was unable to find a shared genomic organisation, we were unable to extract these regions to compare them. I therefore considered the status of these genes in these species as speculative.

I marked the results of our investigation on the heatmap, with green squares showing confirmed gene deletions, black squares showing unconfirmed

gene deletions, and white squares indicating spuriously reported counts of zero. (**Figure 5.1**)

5.3 – Discussion

5.3.1 – Assessment of the methods

In this chapter, I encountered the same issue as I did in **chapter 3**: in some cases, screening with DIGS can produce a large amount of sequence data, and that in some cases issues can arise in processing and curating these datasets. Despite using relatively strict cutoffs, the screening produced a high level of artefacts, some of which were due to uneven genome quality, and others due to the rigorous cutoff applied. In view of this, I conceded that the heatmap was probably not representative of the true 'landscape' of the core mammalian interferome, and instead represented a more general view of ISG expansion and loss in mammals. This was a consequence of not being able to appropriately process the data. I had not accounted for the volume of data that would be produced for each hit. For each gene, numerous hits were produced, making it difficult to work out which of the hits corresponded to *bona fide* ISGs, and which were spurious cross-matches using the approach of relatively manual sequence alignment and analysis.

Using ENSEMBL, I found that the majority of reported counts of zero were artefactual (**Figure 5.1**; **Table 5.1**). This was probably due to the relatively strict cut-off applied to each gene when retrieving the sequence (see **Methods**), whereby more divergent results were filtered from the database. This is reflected in **Figure 5.1**, where the number of reported deletions increases with phylogenetic 'distance' from *Homo sapiens*. The high volume of data produced by the screening also made ISG expansions difficult to study with the methods we employed, making it difficult to establish with confidence whether a high count truly represented functional expansion of an ISG, or a spurious result caused by cross-matching to pseudogenes, transposon, or other genes. However, the manual approaches (study of relatively few loci using genome browsers, BLAST, phylogenetic analysis and sequence alignments) that were unsuited to the high volume of data produced in the ISG screens were well-suited to the study of gene deletions in a limited sample of species. This style of manually-focussed work is in the vein of Daugherty, *et al* (**2014**); Daugherty *et al*, (**2016)**

and Braun *et al,* (**2015**), in that it allows for robust, verifiable identification of gene deletion in three genes.

5.3.2 – Deletions in mammalian xaf1 and ifit

I found deletions in *xaf1* and the *ifit* gene family, which were supported by genomic and alignment data. I discovered that *xaf1* had been, at least partially, deleted in felids. Reviews of the literature surrounding *xaf1* find no evidence of direct antiviral activity, so it is hard to draw concrete inferences regarding the cause or consequence of *xaf1* loss in cats. More literature exists on the antiviral activity of *ifit2* and *ifit3*. Although conflicted, studies on the crystal structure and function of the IFIT protein family suggest that these proteins mediate their antiviral activity through binding either to viral dsRNA (**Yang *et al*, 2012**), or to other IFIT proteins (**Pichlmair *et al*, 2011**). IFIT2 appears to restrict replication in viruses lacking a 2'-O-methyl cap (**Daffis *et al*, 2010),** possibly by binding to, and stabilising the interaction between IFIT1 and viral RNA (**Fleith *et al*, 2018**). This raises the possibility that cetaceans have a) either evolved an alternative method for stabilising the IFIT1-viral RNA interaction, b) as proposed by **Braun *et al*, 2015** for *mx*, there is a selective disadvantage to encoding intact *ifit2* and *ifit3* or c) undergone a population bottleneck in which only the population containing the defective *ifit* genes survived. In any case, the loss of *ifit3,* and the inactivation of *ifit2* raise questions as to the impact of this gene loss on the cetacean response to virus infection, and to whether the IFIT family possess redundant antiviral specificities and properties, marking an opportunity for future study.

A key issue in using DIGS to study ISGs as shown here is to display the data in a manner that accurately represented an increased count or loss of ISGs. In this chapter, the data were likely contained within the database (the fault not lying with the BLAST searches used to screen the genomes), and my main issue was removing false positives and false negatives. Where gene losses were concerned, I identified a large number of artefactual results in my DIGS search for ISG loss in mammals, probably due to difficulties in applying an effective cut-off. A potential alternative is using ENSEMBL to study gene loss. Although in this chapter, we identify an instance where annotation is missing for a specific gene

(*sertad1*) in ENSEMBL, it is likely that interrogating ENSEMBL'S MySQL databases for information on mammalian ISGs would be a quicker and potentially more accurate way to identify potential gene losses. This approach would rely on existing annotations, and would not be clouded by uncertainty over which sequences corresponded to genes or not. This could then be followed up by studying each ISG locus in detail (as performed for *xaf1* and *ifit2/3)*. In addition to each query taking far less time than BLAST searches of over one hundred mammalian genomes, the query set could quickly be changed to consist of whichever query set of genes were of interest. For example, instead of studying a set of conserved genes, it would be interesting to study ISGs that are differentially expressed in a species-specific manner – easily achievable with the datasets described by Shaw *et al*, **2017**.

The difficulty in finding gene expansions was related to the volume of data produced. In addition to a different number of desired 'hits' for each gene (depending on the exons in each gene), each screen produced numerous hits that did not correspond to the gene, but were labelled as such in the database. Thus, it was difficult to determine whether a high result was due to cross-matching, or a functional gene duplication. A potential way to refine this approach (and make the counts reported in my screens more accurate) would be to screen only with peptides that are specific to conserved regions of antiviral genes (the A3G Z domain, for example). The advantage of this is twofold**:** Firstly, there would be less need to process the data so extensively – the use of a single peptide *in lieu* of a whole gene sequence circumvents the issues raised by differing exon numbers between genes. Secondly, when one sequence is used to screen a genome, the retrieved sequences can be more quickly and easily aligned and analysed, allowing easy classification. Classified sequences can then be added to the reference library and used to refine the screen until the results contained in the database accurately represent the number and classification of the gene in question in the genomes surveyed. This is a methodology more in keeping with the 'phylogenetic screening' approach successfully employed in previous studies (**Chapters 3, 4; Dennis *et al*, 2018a; Dennis *et al*, 2018b).**

5.3.3 – Conclusions

I searched mammalian genomes for signatures of ISG loss, expansion and evolution. DIGS was used to screen mammalian genome assemblies with peptides of 84 'core' mammalian ISGs identified by Shaw et al, (**2017**). Study of the resulting data disclosed instances of species-specific (and in two cases group-specific) gene losses. I found numerous artefacts and false positive results, and encountered difficulty in processing such a large dataset. However, I identified three instances of gene loss in felids and cetaceans, and set out detailed plans for future study whereby the large-scale mining of animal genomes with ISGs can be reconciled with the comparative phylogenetic approach employed in chapters 3 and 4, enabling more effective study of ISG loss and expansion.

Chapter 6 - General discussion

In this thesis, I performed three investigations into host:virus co-evolution. They were united by the use of DIGS to systematically mine WGS assemblies for sequences in heuristic investigations . In all three investigations, I was able to find novel results: inferring a history of ERV expansion, loss and functionalisation in the mouse, providing novel insights into circovirus host range, and identifying the loss of three genes in two mammalian groups. My approach involved assembling a library of peptides representing the diversity of the sequences under investigation (i.e. *Circovirus* peptides), and then collating a group of genomes corresponding to the diversity of organisms under investigations (i.e. genomes from kingdom Animalia). These were used in DIGS screens, the results of which were stored in MySQL. MySQL was used to interrogate and retrieve the resulting data, which were investigated using sequence alignments and phylogenetic analysis, along with generation of annotation tracks, identification of sequence features. I found that this general approach was successful in identifying novel results in all three projects, but the suitability varied depending on the application.

6.1 – Assessment of the methods

Generally, I found that DIGS was well-suited to mining genome databases in paleovirological studies. In my analysis of murine ERVs, I found that my approaches (DIGS screening, alignment and analysis, and annotation track generation) compared favourably to the published literature in terms of ERV copy number (**Table 3.2**), and was able to recapitulate findings of TRIM28-dependent histone methylation at ERV loci for MusD and IAP (**Rowe *et al*, 2013**), indicating that my approach was robust. I was able to use it to identify ancient, low copy number murine ERVs that have hitherto remained undiscovered. This was mirrored in **Chapter 4**, where I was able to identify numerous CVe integrations in animal genomes, leading to greater knowledge of circovirus host range and evolution. My paleovirological analyses were most powerful when DIGS was used to identify a low-medium (numbering a few hundred, for example) number of sequences across numerous species. This allowed for the creation of high quality alignments, often of degraded and ancient sequences (requiring

manual alignment) and in-depth study of host range, orthology, sequence features and evolution.

I found that my approach was less suited to analysing large sequence sets. This was shown in my difficulties studying IAP and in mining ISG sequences. The problem did not lie with DIGS, but was more to do with analysing the output. In **Chapter 3**, I experienced difficulties analysing the highly diverse and abundant IAP megafamily. Similarly in **Chapter 5**, the abundance of data produced in screening Mammalian genomes for ISGs was difficult to process using methods more suited to comparative approaches of relatively low numbers of sequences. As such, I contend that more conventional annotation-based approaches may be better when analysing extremely high copy number ERVs, or gene sequences. For ERVs, RetroTector, LTR-STRUC and LTR-Digest rely on the presence of relatively intact loci. As all of the high-copy number lineages in the mouse are relatively intact, the use to these software may be a more appropriate way to explore the evolution of ERV superfamilies. Similarly, studying genes may be more effective using the MySQL framework of ENSEMBL. However, this approach would come at the cost of genome availability, as the time taken between genome publication and ENSEMBL annotation is typically a few years.

In summary, I found that my approach (DIGS, combined with alignment and in-depth analysis across multiple species) was most powerful when adopting a comparative approach: screening a high diversity of animal genomes for a medium-low number of ERV, EVE or gene sequences, such as in **Chapter 4**. It was well-suited to the analysis of ancient and degraded sequences – a task that often necessitates careful manual analysis. In addition, DIGS can be applied to many contexts – exploring genome data for a wide variety of virus groups and non-virus sequences. The shortcomings of my approach were mostly related to processing large amounts of sequence data after retrieval from MySQL in DIGS. For **Chapters 3** and **5**, I experienced difficulty in analysing large sequence datasets, and cleaning noisy data (i.e. removing degraded ERV sequences from large alignments, removing false-positive and cross-matching ISG hits).

6.2 - Future directions

In the investigation of ISG evolution, it is possible to adapt my existing experimental framework to make screening for ISGs more effective. In **Chapter 5.3**, I suggested that an alternative approach may utilise the MySQL framework of ENSEMBL. However, the quality of these annotations are themselves subject to the quality of the genome, and ENSEMBL annotation may not reveal instances of gene loss and expansion (**Braun *et al*, 2014**). Furthermore, there is a gap between the publication of a WGS assembly and its annotation and storage in ENSEMBL. Therefore, there is scope for large scale mining of WGS assemblies for ISG sequences in a way that does not produce overwhelming volumes of data (studying fewer gene families using single, conserved peptides), and allows the investigator to 'stay ahead of the curve' by using genome data that has not been annotated.

In my analysis of murine ERVs, I produced a large dataset of aligned murine ERV sequences. These could be improved by the addition and categorization of proviruses and solo LTRs of the IAP megafamily. Additionally, analysis of murine ERVs in more detail: dating of solo LTR elements, mapping integration dates and functional ERV properties onto phylogenies and in-depth analysis of ERVs integrated proximal to genes, would allow greater insight of mouse:ERV co-evolution to be gained from these data.

Furthermore, it would be worth developing approaches that better leverage large sequence datasets produced by DIGS. The difficulty with the large ERV datasets was aligning and analysing a large number of degraded sequencing. In the ISG screens, the main difficulty was accurately determining what hits constituted ISGs in a noisy dataset. Developing methods to overcome these difficulties would allow (especially in the case of ERVs), more effective analysis of high copy number ERVs, and genes in species with poorly annotated genomes.

Having reflected on the findings and issues presented in this thesis, I propose two other future directions. First, moving the emphasis of study from high-copy number, modern ERV lineages and focussing on the ancient, low copy number ERVs that appear to be more closely related to HERVs (Mus.i1, Mus.s1,

MYSERV, MuERV-C). The relative low copy number of these ERVs in mice, combined with the diverse range of rodent genomes available, would make a comparative analysis of these lineages in rodents relatively straightforward – one could essentially use the same methods as I did for studying circovirus EVEs. Secondly, combining analysis of active, functionalised ERV lineages with functional genomics analyses in rodents would provide insight into the evolution of ERV functionalisation. Given that many physiological processes involving exapted ERV sequences in the mouse appear to involve recently acquired or active lineages (**MacFarlan *et al*, 2012; Rowe *et al*, 2013**), RetroTector, LTR-Digest, or ERVAP (**Zhu *et al*,** unpublished) may be more appropriate in dealing with the larger volume of intact ERV sequences.

My analysis of endogenous circoviral elements suggested a hitherto underappreciated stability in circovirus:host relationships. As such, it would be interesting to identify the blocks to cross-species transmission. Presently, no cell-culture system exists for circoviruses: if or once this has been developed, assaying circoviruses using ISG screens developed by **Schoggins *et al*, 2012** would allow the identification of anti-circoviral ISGs. Furthermore, targeted sampling of arthropod samples for cyclovirus CVe will clarify the specific nature of the relationships between cycloviruses and arthropods.

6.3 – Conclusions

This PhD project explored host:virus co-evolution by mining animal WGS assemblies for ERV, EVE and antiviral gene sequences. I used DIGS, a similarity-search based technique coupled with a relational database. I successfully mined animal genomes for insights into circovirus evolution, natural history and host range. My analysis of murine ERVs also produced interesting insights into ERV:mammal co-evolution and functionalisation, and I identified aspects of my approach that could be optimised. My indirect paleovirological study- attempting to survey ISG dynamism in mammals - uncovered three gene deletions in two species groups. However, I found that my approach in this instance was not well suited to the aim, and an overhaul of the experimental design may be required for further studies of this sort, as well as development of tools for analysing large sets of degraded ERV and gene sequences.

The increasing availability of biological data in public databases make exploration an attractive prospect. I found that I was able to leverage the available body of genome (and functional genomics) data to make insightful findings through heuristic investigations. As sampling of organism taxa for genome sequencing increases, the more paleovirological data will become available. This essentially makes genome screening iterative, with results updated whenever new organism sequence data becomes available. With increasing publication of new genome data – especially through long-read sequencing technology that has no bias against transposable elements (**Gordon *et al*, 2016; Whitfield *et al*, 2017**), studies using ERVs, EVEs and antiviral genes will be able to draw firmer conclusions on virus host range, evolution, and functionalisation. Overall, this PhD project highlights the benefits of mining genome data for insight into host:virus co-evolution, and provides tantalising glimpses of the insights that are to come with the ever increasing publication of sequence data.

Bibliography

Adams, D. J., Doran, A. G., Lilue, J., & Keane, T. M. (2015). The Mouse Genomes Project: a repository of inbred laboratory mouse strain genomes. *Mammalian Genome*, *26*(9–10), 403–412. http://doi.org/10.1007/s00335-015-9579-6

Adams, M. J., Lefkowitz, E. J., King, A. M. Q., Harrach, B., Harrison, R. L., Knowles, N. J., … Davison, A. J. (2017). *Changes to taxonomy and the International Code of Virus Classification and Nomenclature ratified by the International Committee on Taxonomy of Viruses (2017). Archives of Virology* (Vol. 162). http://doi.org/10.1007/s00705-017-3358-5

Adams, S. E., Rathjen, P. D., Stanway, C. a, Fulton, S. M., Malim, M. H., Wilson, W., … Kingsman, a J. (1988). Complete nucleotide sequence of a mouse VL30 retro-element. *Molecular and Cellular Biology*, *8*(8), 2989–98. http://doi.org/10.1128/MCB.8.8.2989

Aiewsakun, P., & Katzourakis, A. (2015). Endogenous viruses: Connecting recent and ancient viral evolution. *Virology*, *479–480*, 26–37. http://doi.org/10.1016/j.virol.2015.02.011

Alarcon, P., Rushton, J., & Wieland, B. (2013). Cost of post-weaning multi-systemic wasting syndrome and porcine circovirus type-2 subclinical infection in England - An economic disease model. *Preventive Veterinary Medicine*, *110*(2), 88–102. http://doi.org/10.1016/j.prevetmed.2013.02.010

Altschul, S. F., Madden, T. L., Schäffer, A. A., Zhang, J., Zhang, Z., Miller, W., & Lipman, D. J. (1997). Gapped BLAST and PSI-BLAST: A new generation of protein database search programs. *Nucleic Acids Research*. http://doi.org/10.1093/nar/25.17.3389

Ambrosini, G., Dreos, R., Kumar, S., & Bucher, P. (2016). The ChIP-Seq tools and web server: A resource for analyzing ChIP-seq and other types of genomic data. *BMC Genomics*, *17*(1), 1–15. http://doi.org/10.1186/s12864-016-3288-8

Arjan-Odedra, S., Swanson, C. M., Sherer, N. M., Wolinsky, S. M., & Malim, M. H. (2012). Endogenous MOV10 inhibits the retrotransposition of endogenous retroelements but not the replication of exogenous retroviruses. *Retrovirology*, *9*, 1–13. http://doi.org/10.1186/1742-4690-9-53

Arriagada, G., & Gifford, R. J. (2014). Parvovirus-Derived Endogenous Viral Elements in Two South American Rodent Genomes. *Journal of Virology*, *88*(20), 12158–12162. http://doi.org/10.1128/JVI.01173-14

Baekbo, P., Kristensen, C. S., & Larsen, L. E. (2012). Porcine Circovirus Diseases: A review of PMWS. *Transboundary and Emerging Diseases*, *59*(SUPPL. 1), 60–67. http://doi.org/10.1111/j.1865-1682.2011.01288.x

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*(11), 5784–98. http://doi.org/10.1128/JVI.78.11.5784-5798.2004

Ballandras-Colas, A., Naraharisetty, H., Li, X., Serrao, E., & Engelman, A. (2013). Biochemical Characterization of Novel Retroviral Integrase Proteins. *PLoS ONE*, *8*(10). http://doi.org/10.1371/journal.pone.0076638

Bamunusinghe, D., Liu, Q., Lu, X., Oler, A., & Kozak, C. a. (2013). Endogenous Gammaretrovirus Acquisition in Mus musculus Subspecies Carrying Functional Variants of the XPR1 Virus Receptor. *Journal of Virology*, *87*(17), 9845–9855. http://doi.org/10.1128/JVI.01264-13

Bamunusinghe, D., Naghashfar, Z., Buckler-White, A., Plishka, R., Baliji, S., Liu, Q., … Kozak, C. a. (2016). Sequence diversity, intersubgroup relationships, and origins of the mouse leukemia gammaretroviruses of laboratory and wild mice. *Journal of Virology*, (February), JVI.03186-15. http://doi.org/10.1128/JVI.03186-15

Baust, C., Gagnier, L., Baillie, G. J., Harris, M. J., Juriloff, D. M., & Mager, D. L. (2003). Structure and expression of mobile ETnII retroelements and their coding-competent MusD relatives in the mouse. *Journal of Virology*, *77*(21), 11448–58. http://doi.org/10.1128/JVI.77.21.11448

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*(17), 9437–9442. http://doi.org/10.1128/JVI.02216-06

Belyi, V. A., Levine, A. J., & Skalka, A. M. (2010). Sequences from Ancestral Single-Stranded DNA Viruses in Vertebrate Genomes: the Parvoviridae and Circoviridae Are More than 40 to 50 Million Years Old. *Journal of Virology*, *84*(23), 12458–12462. http://doi.org/10.1128/JVI.01789-10

Bénit, L., De Parseval, N., Casella, J. F., Callebaut, I., Cordonnier, a, & Heidmann, T. (1997). Cloning of a new murine endogenous retrovirus, MuERV-L, with strong similarity to the human HERV-L element and with a gag coding sequence closely related to the Fv1 restriction gene. *Journal of Virology*, *71*(7), 5652–7. Retrieved from /pmc/articles/PMC191811/?report=abstract

Bénit, L., Lallemand, J. B., Casella, J. F., Philippe, H., & Heidmann, T. (1999). ERV-L elements: a family of endogenous retrovirus-like elements active throughout the evolution of mammals. *Journal of Virology*, *73*(4), 3301–8. Retrieved from http://www.ncbi.nlm.nih.gov/pubmed/10074184%5Cnhttp://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=PMC104094

Bénit, L., Dessen, P., & Heidmann, T. (2001). Identification , Phylogeny , and Evolution of Retroviral Elements Based on Their Envelope Genes Identification , Phylogeny , and Evolution of Retroviral Elements Based on Their Envelope Genes, *75*(23), 11709–11719. http://doi.org/10.1128/JVI.75.23.11709

Bézier, A., Herbinière, J., Lanzrein, B., & Drezen, J. M. (2009). Polydnavirus hidden face: The genes producing virus particles of parasitic wasps. *Journal of Invertebrate Pathology*, *101*(3), 194–203. http://doi.org/10.1016/j.jip.2009.04.006

Bill, C. A., & Summers, J. (2004). Genomic DNA double-strand breaks are targets for hepadnaviral DNA integration. *Proceedings of the National Academy of Sciences*, *101*(30), 11135–11140. http://doi.org/10.1073/pnas.0403925101

Blanco-melo, D., Gifford, R. J., & Bieniasz, P. D. (2018). Reconstruction of a replication-competent ancestral murine endogenous retrovirus-L. *Retrovirology*, *15*, 1–17. http://doi.org/10.1186/s12977-018-0416-3

Blinkova, O., Rosario, K., Li, L., Kapoor, A., Slikas, B., Bernardin, F., … Delwart, E. (2009). Frequent detection of highly diverse variants of Cardiovirus, Cosavirus, Bocavirus, and Circovirus in sewage samples collected in the United States. *Journal of Clinical Microbiology*, *47*(11), 3507–3513. http://doi.org/10.1128/JCM.01062-09

Boursot, P., Auffray, J. C., Britton-Davidian, J., & Bonhomme, F. (1993). The Evolution of House Mice. *Annual Review of Ecology and Systematics*, *24*(1), 119–152. http://doi.org/10.1146/annurev.es.24.110193.001003

Brattås, P. L., Jönsson, M. E., Fasching, L., Nelander Wahlestedt, J., Shahsavani, M., Falk, R., … Jakobsson, J. (2017). TRIM28 Controls a Gene Regulatory Network Based on Endogenous Retroviruses in Human Neural Progenitor Cells. *Cell Reports*, *18*(1), 1–11. http://doi.org/10.1016/j.celrep.2016.12.010

Braun, B. a., Marcovitz, A., Camp, J. G., Jia, R., & Bejerano, G. (2015). Mx1 and Mx2 key antiviral proteins are surprisingly lost in toothed whales. *Proceedings of the National Academy of Sciences*, *112*(26), 8036–8040. http://doi.org/10.1073/pnas.1501844112

Breitbart, M., Benner, B. E., Jernigan, P. E., Rosario, K., Birsa, L. M., Harbeitner, R. C., … Nejstgaard, J. C. (2015). Discovery, prevalence, and persistence of novel circular single-stranded DNA viruses in the ctenophores mnemiopsis leidyi and Beroe ovata. *Frontiers in Microbiology*, *6*(DEC), 1–10. http://doi.org/10.3389/fmicb.2015.01427

Bromham, L., Clark, F., & McKee, J. (2001). Discovery of a Novel Murine Type C Retrovirus by Data Mining. *Journal of Virology*, *75*(6). http://doi.org/10.1128/JVI.75.6.3053

Brûlet, P., Kaghad, M., Xu, Y. S., Croissant, O., & Jacob, F. (1983). Early differential tissue expression of transposon-like repetitive DNA sequences of the mouse. *Proceedings of the National Academy of Sciences of the United States of America*, *80*(18), 5641–5. http://doi.org/10.1073/pnas.80.18.5641

Camacho, C., Coulouris, G., Avagyan, V., Ma, N., Papadopoulos, J., Bealer, K., & Madden, T. L. (2009). BLAST+: Architecture and applications. *BMC Bioinformatics*, *10*, 1–9. http://doi.org/10.1186/1471-2105-10-421

Campos-Sánchez, R., Cremona, M. A., Pini, A., Chiaromonte, F., & Makova, K. D. (2016). Integration and Fixation Preferences of Human and Mouse Endogenous Retroviruses Uncovered with Functional Data Analysis. *PLoS Computational Biology*, *12*(6), 1–41. http://doi.org/10.1371/journal.pcbi.1004956

Cannatella, D. (2015). Xenopus in Space and Time: Fossils, Node Calibrations, Tip-Dating, and Paleobiogeography. *Cytogenetic and Genome Research*, *145*(3–4), 283–301. http://doi.org/10.1159/000438910

Cantrell, M. a, Ederer, M. M., Erickson, I. K., Swier, V. J., Baker, R. J., & Wichman, H. a. (2005). MysTR: an endogenous retrovirus family in mammals that is undergoing recent amplifications to unprecedented copy numbers. *Journal of Virology*, *79*(23), 14698–14707. http://doi.org/10.1128/JVI.79.23.14698-14707.2005

Carlson, J., Suchman, E., & Buchatsky, L. (2006). Densoviruses for Control and Genetic Manipulation of Mosquitoes. *Advances in Virus Research*, *68*(06), 361–392. http://doi.org/10.1016/S0065-3527(06)68010-X

Casper, J., Zweig, A. S., Villarreal, C., Tyner, C., Speir, M. L., Rosenbloom, K. R., … Kent, W. J. (2018). The UCSC Genome Browser database: 2018 update. *Nucleic Acids Research*, *46*(D1), D762–D769. http://doi.org/10.1093/nar/gkx1020

Chan, M. C. W., Kwok, S. W. Y., & Chan, P. K. S. (2015). False-positive PCR detection of cyclovirus Malawi strain VS5700009 in human cerebrospinal fluid. *Journal of Clinical Virology*, *68*, 76–78. http://doi.org/10.1016/j.jcv.2015.05.007

Chance, M. R., Sagi, I., Wirt, M. D., Frisbie, S. M., Scheuring, E., Chen, E., … South, T. L. (1992). Extended x-ray absorption fine structure studies of a retrovirus: equine infectious anemia virus cysteine arrays are coordinated to zinc. *Proc.Natl.Acad.Sci.U.S.A.*, *89*(21), 10041–10045.

Charleston, M. A., & Robertson, D. L. (2002). Preferential host switching by primate lentiviruses can account for phylogenetic similarity with the primate phylogeny. *Systematic Biology*, *51*(3), 528–535. http://doi.org/10.1080/10635150290069940

Chen, C., Morris, Q., & Mitchell, J. A. (2012). Enhancer identification in mouse embryonic stem cells using integrative modeling of chromatin and genomic features. *BMC Genomics*, *13*(1), 152. http://doi.org/10.1186/1471-2164-13-152

Chen, X., Huang, S., Guo, P., Colli, G. R., Nieto Montes de Oca, A., Vitt, L. J., … Burbrink, F. T. (2013). Understanding the formation of ancient intertropical disjunct distributions using Asian and Neotropical hinged-teeth snakes (Sibynophis and Scaphiodontophis: Serpentes: Colubridae). *Molecular Phylogenetics and Evolution*, *66*(1), 254–261. http://doi.org/10.1016/j.ympev.2012.09.032

Cheslock, S. R., Poon, D. T. K., Fu, W., Rhodes, T. D., Henderson, L. E., Nagashima, K., … Hu, W.-S. (2003). Charged assembly helix motif in murine leukemia virus capsid: an important region for virus assembly and particle size determination. *Journal of Virology*, *77*(12), 7058–7066. http://doi.org/10.1128/JVI.77.12.7058-7066.2003

Cheung, A. K. (2012). Porcine circovirus: Transcription and DNA replication. *Virus Research*, *164*(1–2), 46–53. http://doi.org/10.1016/j.virusres.2011.10.012

Cheung, A. K., & Greenlee, J. J. (2011). Identification of an amino acid domain encoded by the capsid gene of porcine circovirus type 2 that modulates intracellular viral protein distribution during replication. *Virus Research*, *155*(1), 358–362. http://doi.org/10.1016/j.virusres.2010.09.021

Chuong, E. B., Elde, N. C., & Feschotte, C. (2016). Regulatory evolution of innate immunity through co-option of endogenous retroviruses. *Science*, *351*(6277), 1083–1087. http://doi.org/10.1126/science.aad5497

Chuong, E. B., Rumi, M. a K., Soares, M. J., & Baker, J. C. (2013). Endogenous retroviruses function as species-specific enhancer elements in the placenta. *Nature Genetics*, *45*(3), 325–329. http://doi.org/10.1038/ng.2553

Claramunt, S., & Cracraft, J. (2015). A new time tree reveals Earth historys imprint on the evolution of modern birds. *Science Advances*, *1*(11), e1501005–e1501005. http://doi.org/10.1126/sciadv.1501005

Coffin, J. M., Hughes, S. H., & Varmus, H. E. (1997). *Retroviruses*. *Retroviruses*. Retrieved from http://www.ncbi.nlm.nih.gov/pubmed/21433340

Cordonnier, a, Casella, J. F., & Heidmann, T. (1995). Isolation of novel human endogenous retrovirus-like elements with foamy virus-related pol sequence. *Journal of Virology*, *69*(9), 5890–7. Retrieved from http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=189468&tool=pmcentrez&rendertype=abstract

Costas, J. (2003). Molecular characterization of the recent intragenomic spread of the murine endogenous retrovirus MuERV-L. *Journal of Molecular Evolution*, *56*(2), 181–186. http://doi.org/10.1007/s00239-002-2392-3

Cotmore, S. F., Agbandje-McKenna, M., Chiorini, J. a., Mukha, D. V., Pintel, D. J., Qiu, J., … Davison, A. J. (2014). The family Parvoviridae. *Archives of Virology*, *159*(5), 1239–1247. http://doi.org/10.1007/s00705-013-1914-1

Crochu, S., Cook, S., Attoui, H., Charrel, R. N., De Chesse, R., Belhouchet, M., … de Lamballerie, X. (2004). Sequences of flavivirus-related RNA viruses persist in DNA form integrated in the genome of Aedes spp. mosquitoes. *Journal of General Virology*, *85*(7), 1971–1980. http://doi.org/10.1099/vir.0.79850-0

Cui, J., & Holmes, E. C. (2012). Endogenous Hepadnaviruses in the Genome of the Budgerigar (Melopsittacus undulatus) and the Evolution of Avian

Hepadnaviruses. *Journal of Virology*, *86*(14), 7688–7691. http://doi.org/10.1128/JVI.00769-12

Cui, J., Zhao, W., Huang, Z., Jarvis, E., Gilbert, M., Walker, P., … Zhang, G. (2014). Low frequency of paleoviral infiltration across the avian phylogeny. *Genome Biology*, *15*(12), 539. http://doi.org/10.1186/s13059-014-0539-3

Dalton, A. J., Potter, M., & Merwin, R. M. (1961). Some ultrastructural characteristics of a series of primary and transplanted plasma-cell tumors of the mouse. *Journal of the National Cancer Institute*, *26*(5), 1221–1267. http://doi.org/10.1093/jnci/26.5.1221

Daugherty, M. D., Schaller, A. M., Geballe, A. P., & Malik, H. S. (2016). Evolution-guided functional analyses reveal diverse antiviral specificities encoded by ifit1 genes in mammals. *ELife*, *5*(MAY2016), 1–22. http://doi.org/10.7554/eLife.14228

Daugherty, M. D., Young, J. M., Kerns, J. a., & Malik, H. S. (2014). Rapid Evolution of PARP Genes Suggests a Broad Role for ADP-Ribosylation in Host-Virus Conflicts. *PLoS Genetics*, *10*(5). http://doi.org/10.1371/journal.pgen.1004403

Dayaram, A., Goldstien, S., Argüello-Astorga, G. R., Zawar-Reza, P., Gomez, C., Harding, J. S., & Varsani, A. (2015). Diverse small circular DNA viruses circulating amongst estuarine molluscs. *Infection, Genetics and Evolution*, *31*(February 2015), 284–295. http://doi.org/10.1016/j.meegid.2015.02.010

Decaro, N., Martella, V., Desario, C., Lanave, G., Circella, E., Cavalli, A., … Buonavoglia, C. (2014). Genomic characterization of a circovirus associated with fatal hemorrhagic enteritis in dog, Italy. *PLoS ONE*, *9*(8). http://doi.org/10.1371/journal.pone.0105909

Delwart, E. L. (2007). Viral metagenomics. *Reviews in Medical Virology*, *17*(2), 115–131. http://doi.org/10.1002/rmv.532

Delwart, E., & Li, L. (2012). Rapidly expanding genetic diversity and host range of the Circoviridae viral family and other Rep encoding small circular ssDNA genomes. *Virus Research*, *164*(1–2), 114–121. http://doi.org/10.1016/j.immuni.2010.12.017.Two-stage

Demirov, D. G., & Freed, E. O. (2004). Retrovirus budding. *Virus Research*, *106*(2 SPEC.ISS.), 87–102. http://doi.org/10.1016/j.virusres.2004.08.007

Denner, J., Specke, V., Schwendemann, J., & Tacke, S. J. (2001). Porcine endogenous retroviruses (PERVs): adaptation to human cells and attempts to infect small animals and non-human primates. *Annals of Transplantation : Quarterly of the Polish Transplantation Society*.

Denner, J., & Mankertz, A. (2017). Porcine circoviruses and xenotransplantation. *Viruses*, *9*(4), 1–13. http://doi.org/10.3390/v9040083

Dennis, T. P. W., de Souza, W. M., Marsile-Medun, S., Singer, J. B., Wilson, S. J., & Gifford, R. J. (2018). The evolution, distribution and diversity of endogenous circoviral elements in vertebrate genomes. *Virus Research*, (March). http://doi.org/10.1016/j.virusres.2018.03.014

Dennis, T. P.W., Flynn, P. J., de Souza, W. M., Singer, J. B., Moreau, C. S., Wilson, S. J., & Gifford, R. J. (2018). Insights into circovirus host range from the genomic fossil record. *Journal of Virology*, *92*(16). http://doi.org/doi:10.1128/JVI.00145-18

Derrough, T., Jansa, J., Plachouras, D., & Sudre, B. (2013). RAPID RISK ASSESSMENT Cyclovirus in cerebrospinal fluid of patients with central nervous system infection ECDC internal response team Consulted experts, (July), 1–5.

Dewannieux, M., Dupressoir, A., Harper, F., Pierron, G., & Heidmann, T. (2004). Identification of autonomous IAP LTR retrotransposons mobile in mammalian cells. *Nature Genetics*, *36*(5), 534–539. http://doi.org/10.1038/ng1353

Dewannieux, M., Harper, F., Richaud, A., Letzelter, C., Ribet, D., Pierron, G., & Heidmann, T. (2006). Identification of an infectious progenitor for the multiple-copy HERV-K human endogenous retroelements. *Genome Research*, *16*(12), 1548–1556. http://doi.org/10.1101/gr.5565706

Diehl, W. E., Patel, N., Halm, K., & Johnson, W. E. (2016). Tracking interspecies transmission and long-term evolution of an ancient retrovirus using the genomes of modern mammals. *ELife*, *5*, 1–25. http://doi.org/10.7554/eLife.12704

Dill, J. a., Camus, A. C., Leary, J. H., Di Giallonardo, F., Holmes, E. C., & Ng, T. F. F. (2016). Distinct Viral Lineages from Fish and Amphibians Reveal the Complex Evolutionary History of Hepadnaviruses. *Journal of Virology*, *90*(17), 7920–7933. http://doi.org/10.1128/JVI.00832-16

Drummond, A. J., Pybus, O. G., & Rambaut, A. (2003). Inference of evolutionary rates from molecular sequences. *Advances in Parasitology*, (54), 331–358.

Dunlap, D. S., Ng, T. F. F., Rosario, K., Barbosa, J. G., Greco, A. M., Breitbart, M., & Hewson, I. (2013). Molecular and microscopic evidence of viruses in marine copepods. *Proceedings of the National Academy of Sciences of the United States of America*, *110*(4), 1375–80. http://doi.org/10.1073/pnas.1216595110

Dunn, B. M., Goodenow, M. M., Gustchina, A., & Wlodawer, A. (2002). Retroviral proteases. *Genome Biology*, *3*(4), Reviews3006.1-7. http://doi.org/10.1186/gb-2002-3-4-reviews3006

Edgar, R. (2002). Gene Expression Omnibus: NCBI gene expression and hybridization array data repository. *Nucleic Acids Research*, *30*(1), 207–210. http://doi.org/10.1093/nar/30.1.207

Edgar, R. C. (2004). MUSCLE: a multiple sequence alignment method with reduced time and space complexity. *BMC Bioinformatics*, *5*, 113. http://doi.org/10.1186/1471-2105-5-113

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

Fehér, E., Székely, C., Lorincz, M., Cech, G., Tuboly, T., Singh, H. S., … Farkas, S. L. (2013). Integrated circoviral rep-like sequences in the genome of cyprinid fish. *Virus Genes*, *47*(2), 374–377. http://doi.org/10.1007/s11262-013-0928-9

Fonslow, B. R., Stein, B. D., Webb, K. J., Xu, T., Choi, J., Kyu, S., & Iii, J. R. Y. (2013). NIH Public Access, *10*(1), 54–56. http://doi.org/10.1038/nmeth.2250.Digestion

François, S., Filloux, D., Roumagnac, P., Bigot, D., Gayral, P., Martin, D. P., … Ogliastro, M. (2016). Discovery of parvovirus-related sequences in an unexpected broad range of animals. *Scientific Reports*, *6*(1), 30880. http://doi.org/10.1038/srep30880

Frankel, W. N., Stoye, J. P., Taylor, B. a, & Coffin, J. M. (1989). Genetic analysis of endogenous xenotropic murine leukemia viruses: association with two common mouse mutations and the viral restriction locus Fv-1. *Journal of Virology*, *63*(4), 1763–1774.

Freed, E. O. (2002). MINIREVIEW, *76*(10), 4679–4687. http://doi.org/10.1128/JVI.76.10.4679

Friedli, M., Turelli, P., Kapopoulou, A., Rauwel, B., Castro-d, N., Rowe, H., … Trono, D. (2014). Retroelements During Reprogramming To Pluripotency, *I*(1), 1251–1259. http://doi.org/10.1101/gr.172809.114.

Fujino, K., Horie, M., Honda, T., Merriman, D. K., & Tomonaga, K. (2014). Inhibition of Borna disease virus replication by an endogenous bornavirus-like element in the ground squirrel genome. *Proceedings of the National Academy of Sciences*, *111*(36), 13175–13180. http://doi.org/10.1073/pnas.1407046111

Fujisawa, R., McAtee, F. J., Zirbel, J. H., & Portis, J. L. (1997). Characterization of glycosylated Gag expressed by a neurovirulent murine leukemia virus: identification of differences in processing in vitro and in vivo. *Journal of Virology*, *71*(7), 5355–60. Retrieved from http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=191773&tool=pmcentrez&rendertype=abstract

Galloway, D. A., Laimins, L. A., Division, B., & Hutchinson, F. (2016). HHS Public Access, *20*(2), 87–92. http://doi.org/10.1016/j.coviro.2015.09.001.Human

Garigliany, M.-M., Hagen, R. M., Frickmann, H., May, J., Schwarz, N. G., Perse, a., … Schmidt-Chanasit, J. (2014). Cyclovirus CyCV-VN species distribution is

not limited to Vietnam and extends to Africa. *Scientific Reports*, *4*, 1–7. http://doi.org/10.1038/srep07552

Geuking, M. B., Weber, J., Dewannieux, M., Gorelik, E., Heidmann, T., Hengartner, H., … Hangartner, L. (2009). Recombination of retrotransposon and exogenous RNA virus results in nonretroviral cDNA integration. *Science*, *323*(5912), 393–396. http://doi.org/10.1126/science.1167375

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. *Proc Natl Acad Sci U S A*, *105*(51), 20362–20367. http://doi.org/10.1073/pnas.0807873105\r0807873105 [pii]

Gifford, R. J. (2012). Viral evolution in deep time: Lentiviruses and mammals. *Trends in Genetics*, *28*(2), 89–100. http://doi.org/10.1016/j.tig.2011.11.003

Gifford, R., & Tristem, M. (2003). The evolution, distribution and diversity of endogenous retroviruses. *Virus Genes*, *26*(3), 291–315. http://doi.org/Doi 10.1023/A:1024455415443

Gifford, W. D., Pfaff, S. L., & MacFarlan, T. S. (2013). Transposable elements as genetic regulatory substrates in early development. *Trends in Cell Biology*, *23*(5), 218–226. http://doi.org/10.1016/j.tcb.2013.01.001

Gilbert, C., Meik, J. M., Dashevsky, D., Card, D. C., Castoe, T. a, & Schaack, S. (2014). Endogenous hepadnaviruses, bornaviruses and circoviruses in snakes. *Proceedings. Biological Sciences / The Royal Society*, *281*(1791), 20141122. http://doi.org/10.1098/rspb.2014.1122

Gilbert, C., & Feschotte, C. (2010). Genomic fossils calibrate the long-term evolution of hepadnaviruses. *PLoS Biology*, *8*(9). http://doi.org/10.1371/journal.pbio.1000495

Göke, J., & Ng, H. H. (2016). CTRL+INSERT: retrotransposons and their contribution to regulation and innovation of the transcriptome. *EMBO Reports*, *17*(8), 1131–1144. http://doi.org/10.15252/embr.201642743

Gómez-Acevedo, S., Rico-Arce, L., Delgado-Salinas, A., Magallón, S., & Eguiarte, L. E. (2010). Neotropical mutualism between Acacia and Pseudomyrmex: Phylogeny and divergence times. *Molecular Phylogenetics and Evolution*, *56*(1), 393–408. http://doi.org/10.1016/j.ympev.2010.03.018

Gordon, D., Huddleston, J., Chaisson, M. J. P., Hill, C. M., Kronenberg, Z. N., Munson, K. M., … Wilson, R. K. (2016). Long Read Sequence Assembly of the Gorilla Genome, *0344*. http://doi.org/10.1126/science.aae0344

Göttlinger, H. G., Dorfman, T., Sodroski, J. G., & Haseltine, W. a. (1991). Effect of mutations affecting the p6 gag protein on human immunodeficiency virus particle release. *Proceedings of the National Academy of Sciences of the United States of America*, *88*(8), 3195–3199. http://doi.org/10.1073/pnas.88.8.3195

Grandi, N., Cadeddu, M., Blomberg, J., & Tramontano, E. (2016). Contribution of type W human endogenous retroviruses to the human genome: characterization of HERV-W proviral insertions and processed pseudogenes. *Retrovirology*, *13*(1), 67. http://doi.org/10.1186/s12977-016-0301-x

Green, R., Moosey, M., & Bittner, J. (1946). Serial Transmission of the iMilk Agent of Mouse Mammary Carcinoma*. *PRQC. SOC. EXP. BIOL. AND MED*, *61*(115), 362–363.

Gross, L. (1953). A Filterable Agent, Recovered from Ak Leukemic Extracts, Causing Salivary Gland Carcinomas in C3H Mice. *Experimental Biology and Medicine*, *83*(2), 414–421.

Guerrero, R. B., & Roberts, L. R. (2005). The role of hepatitis B virus integrations in the pathogenesis of human hepatocellular carcinoma. *Journal of Hepatology*. http://doi.org/10.1016/j.jhep.2005.02.005

Hahn, C. M., Iwanowicz, L. R., Cornman, R. S., Conway, C. M., Winton, J. R., & Blazer, V. S. (2015). Characterization of a Novel Hepadnavirus in the White Sucker (Catostomus commersonii) from the Great Lakes Region of the United, *89*(23), 11801–11811. http://doi.org/10.1128/JVI.01278-15.Editor

Hayward, A., Cornwallis, C. K., & Jern, P. (2015). Pan-vertebrate comparative genomics unmasks retrovirus macroevolution. *Proceedings of the National Academy of Sciences*, *112*(2), 464–469. http://doi.org/10.1073/pnas.1414980112

Hayward, A., Grabherr, M., & Jern, P. (2013). Broad-scale phylogenomics provides insights into retrovirus-host evolution. *Proceedings of the National Academy of Sciences of the United States of America*, *110*(50), 20146–51. http://doi.org/10.1073/pnas.1315419110

Hedges, S. B., Marin, J., Suleski, M., Paymer, M., & Kumar, S. (2015). Tree of life reveals clock-like speciation and diversification. *Molecular Biology and Evolution*, *32*(4), 835–845. http://doi.org/10.1093/molbev/msv037

Henzy, J. E., & Johnson, W. E. (2013). Pushing the endogenous envelope. *Philosophical Transactions of the Royal Society B: Biological Sciences*, *368*(1626), 20120506–20120506. http://doi.org/10.1098/rstb.2012.0506

Herniou, E., Martin, J., Miller, K., Cook, J., Wilkinson, M., & Tristem, M. (1998). Retroviral diversity and distribution in vertebrates. *Journal of Virology*, *72*(7), 5955–5966.

Holmes, E. C. (2003). Molecular Clocks and the Puzzle of RNA Virus Origins. *Journal of Virology*, *77*(7), 3893–3897. http://doi.org/10.1128/JVI.77.7.3893-3897.2003

Horie, M., Kobayashi, Y., Honda, T., Fujino, K., Akasaka, T., Kohl, C., … Tomonaga, K. (2016). An RNA-dependent RNA polymerase gene in bat genomes derived from an ancient negative-strand RNA virus. *Scientific Reports*, *6*(April), 1–9. http://doi.org/10.1038/srep25873

Hsu, H., Lin, T., Wu, H., Lin, L., Chung, C., Chiou, M., & Lin, C. (2016). High detection rate of dog circovirus in diarrheal dogs. *BMC Veterinary Research*, 8–13. http://doi.org/10.1186/s12917-016-0722-8

Hubley, R., Finn, R. D., Clements, J., Eddy, S. R., Jones, T. a., Bao, W., … Wheeler, T. J. (2016). The Dfam database of repetitive DNA families. *Nucleic Acids Research*, *44*(D1), D81–D89. http://doi.org/10.1093/nar/gkv1272

Huder, J. B., Böni, J., Hatt, J.-M., Soldati, G., Lutz, H., & Schüpbach, J. (2002). Identification and characterization of two closely related unclassifiable endogenous retroviruses in pythons (Python molurus and Python curtus). *Journal of Virology*, *76*(15), 7607–15. http://doi.org/10.1128/JVI.76.15.7607-7615.2002

Hué, S., Gray, E. R., Gall, A., Katzourakis, A., Tan, C. P., Houldcroft, C. J., … Towers, G. J. (2010). Disease-associated XMRV sequences are consistent with laboratory contamination. *Retrovirology*, *7*(1), 111. http://doi.org/10.1186/1742-4690-7-111

Hutchison, K. W., & Eicher, E. M. (1989). An Amplified Endogenous Retroviral Sequence on the Murine-Y Chromosome Related to Murine Leukemia Viruses and Viruslike 30s Sequences. *Journal of Virology*, *63*(9), 4043–4046.

Hutchison, K. W., & Eicher, E. M. (1989). An Amplified Endogenous Retroviral Sequence on the Murine-Y Chromosome Related to Murine Leukemia Viruses and Viruslike 30s Sequences. *Journal of Virology*, *63*(9), 4043–4046.

Itin, A., & Keshet, E. L. I. (1986). A Novel Retroviruslike Family in Mouse DNA, *59*(2), 301–307.

Jern, P., Sperber, G. O., & Blomberg, J. (2005). Use of Endogenous Retroviral Sequences (ERVs) and structural markers for retroviral phylogenetic inference and taxonomy. *Retrovirology*, *2*, 1–12. http://doi.org/10.1186/1742-4690-2-50

Jern, P., Stoye, J. P., & Coffin, J. M. (2007). Role of APOBEC3 in genetic diversity among endogenous murine leukemia viruses. *PLoS Genetics*, *3*(10), 2014–2022. http://doi.org/10.1371/journal.pgen.0030183

Jetz, W., Thomas, G. H., Joy, J. B., Hartmann, K., & Mooers, a. O. (2012). The global diversity of birds in space and time. *Nature*, *491*(7424), 444–448. http://doi.org/10.1038/nature11631

Johnson, W. E. (2015). Endogenous Retroviruses in the Genomics Era. *Annual Review of Virology*, *2*(1), 135–159. http://doi.org/10.1146/annurev-virology-100114-054945

Jolicoeur, P., & Baltimore, D. (1976). Effect of Fv-1 gene product on proviral DNA formation and integration in cells infected with murine leukemia viruses. *Proc Natl Acad Sci U S A*, *73*(7), 2236–40. http://doi.org/10.2307/65744

Jurka, J. (2000). Repbase Update: a database and an electronic journal of repetitive elements. *Trends in Genetics*, *16*(9), 418–20.

Kalyaanamoorthy, S., Minh, B. Q., Wong, T. K. F., von Haeseler, a, & Jermiin, L. S. (2017). ModelFinder: Fast Model Selection for Accurate Phylogenetic Estimates. *Nature Methods*, *in press*(April), 1–5. http://doi.org/10.1038/nmeth.4285

Kassiotis, G. (2014). Europe PMC Funders Group Endogenous retroviruses and the development of cancer, *192*(4), 1343–1349. http://doi.org/10.4049/jimmunol.1302972.Endogenous

Katzourakis, A. (2013). Paleovirology: inferring viral evolution from host genome sequence data. *Philosophical Transactions of the Royal Society of London. Series B, Biological Sciences*, *368*(1626), 20120493. http://doi.org/10.1098/rstb.2012.0493

Katzourakis, A., & Gifford, R. J. (2010). Endogenous viral elements in animal genomes. *PLoS Genetics*, *6*(11). http://doi.org/10.1371/journal.pgen.1001191

Katzourakis, A., Rambaut, A., & Pybus, O. G. (2005). The evolutionary dynamics of endogenous retroviruses. *TRENDS in Microbiology*, *13*(10). http://doi.org/10.1016/j.tim.2005.08.004

Kearse, M., Moir, R., Wilson, A., Stones-Havas, S., Cheung, M., Sturrock, S., … Drummond, A. (2012). Geneious Basic: An integrated and extendable desktop software platform for the organization and analysis of sequence data. *Bioinformatics*, *28*(12), 1647–1649. http://doi.org/10.1093/bioinformatics/bts199

Keckesova, Z., Ylinen, L. M. J., Towers, G. J., Gifford, R. J., & Katzourakis, A. (2009). Identification of a RELIK orthologue in the European hare (Lepus europaeus) reveals a minimum age of 12 million years for the lagomorph lentiviruses. *Virology*, *384*(1), 7–11. http://doi.org/10.1016/j.virol.2008.10.045

Keshet, E., & Itin, A. (1982). Patterns of genomic distribution and sequence heterogeneity of a murine “retrovirus-like” multigene family. *J.Virol.*, *43*(1), 50–58. Retrieved from http://www.ncbi.nlm.nih.gov/pubmed/6287016

Khan, E., Mack, J. P. G., Katz, R. A., Kulkosky, J., & Skalka, A. M. (1991). Retroviral integrase domains: DNA binding and the recognition of LTR sequences. *Nucleic Acids Research*, *19*(6), 1358. http://doi.org/10.1093/nar/19.6.1358

Kobayashi, Y., Horie, M., Nakano, A., Murata, K., Itou, T., & Suzuki, Y. (2016). Exaptation of Bornavirus-Like Nucleoprotein Elements in Afrotherians. *PLoS Pathogens*, *12*(8), 1–21. http://doi.org/10.1371/journal.ppat.1005785

Kotin, R. M., Linden, R. M., & Berns, K. I. (1992). Characterization of a preferred site on human chromosome 19q for integration of adeno-associated virus

DNA by non-homologous recombination. *The EMBO Journal*, *11*(13), 5071–8. http://doi.org/10.1182/blood-2008-10-181479

Kozak, C., Peters, G., Pauley, R., Morris, V., Michalides, R., Dudley, J., … Callahan, R. (1987). A standardized nomenclature for endogenous mouse mammary tumor viruses. *J.Virol.*, *61*(5), 1651–1654.

Kozak, C. A. (2013). Evolution of different antiviral strategies in wild mouse populations exposed to different gammaretroviruses. *Current Opinion in Virology*, *3*(6), 657–663. http://doi.org/10.1016/j.coviro.2013.08.001

Kozak, C. a. (2015). Origins of the endogenous and infectious laboratory mouse gammaretroviruses. *Viruses*, *7*(1), 1–26. http://doi.org/10.3390/v7010001

Krupovic, M., & Forterre, P. (2015). Single-stranded DNA viruses employ a variety of mechanisms for integration into host genomes. *Annals of the New York Academy of Sciences*, *1341*(1), 41–53. http://doi.org/10.1111/nyas.12675

Krupovic, M., & Koonin, E. V. (2017). Multiple origins of viral capsid proteins from cellular ancestors. *Proceedings of the National Academy of Sciences*, *114*(12), E2401–E2410. http://doi.org/10.1073/pnas.1621061114

Kumar, S., & Subramanian, S. (2002). Mutation rates in mammalian genomes. *Proceedings of the National Academy of Sciences of the United States of America*, *99*(2), 803–8. http://doi.org/10.1073/pnas.022629899

Kunarso, G., Chia, N.-Y., Jeyakani, J., Hwang, C., Lu, X., Chan, Y.-S., … Bourque, G. (2010). Transposable elements have rewired the core regulatory network of human embryonic stem cells. *Nature Genetics*, *42*(7), 631–634. http://doi.org/10.1038/ng.600

Kunitake, T., Kitamura, T., Guo, J., Taguchi, F., Kawabe, K., & Yogo, Y. (1995). Parent-to-child transmission is relatively common in the spread of the human polyomavirus JC virus. *Journal of Clinical Microbiology*, *33*(6), 1448–1451. http://doi.org/10.1128/JVI.01169-10

Lamprecht, B., Walter, K., Kreher, S., Kumar, R., Hummel, M., Lenze, D., … Mathas, S. (2010). Derepression of an endogenous long terminal repeat activates the CSF1R proto-oncogene in human lymphoma. *Nature Medicine*, *16*(5), 571–579. http://doi.org/10.1038/nm.2129

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

Lavialle, C., Cornelis, G., Dupressoir, A., Esnault, C., Heidmann, O., Vernochet, C., & Heidmann, T. (2013). Paleovirology of “syncytins”, retroviral env genes exapted for a role in placentation. *Philosophical Transactions of the Royal Society B: Biological Sciences*, *368*(1626), 20120507–20120507. http://doi.org/10.1098/rstb.2012.0507

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 of London. Series B, Biological Sciences*, *368*(August), 20120503. http://doi.org/10.1098/rstb.2012.0503

Lee, K. H., You, R. N., Greenhalgh, D. G., & Cho, K. (2012). Identification of a group of Mus dunni endogenous virus-like endogenous retroviruses from the C57BL/6J mouse genome: Proviral genomes, strain distribution, expression characteristics, and genomic integration profile. *Chromosome Research*, *20*(7), 859–874. http://doi.org/10.1007/s10577-012-9322-z

Lee, Y. N., Malim, M. H., & Bieniasz, P. D. (2008). Hypermutation of an Ancient Human Retrovirus by APOBEC3G. *Journal of Virology*, *82*(17), 8762–8770. http://doi.org/10.1128/JVI.00751-08

Lerner, D. L., Wagaman, P. C., Phillips, T. R., Prospero-Garcia, O., Henriksen, S. J., Fox, H. S., … Elder, J. H. (1995). Increased mutation frequency of feline immunodeficiency virus lacking functional deoxyuridine-triphosphatase. *Proceedings of the National Academy of Sciences of the United States of America*, *92*(16), 7480–7484. http://doi.org/10.1073/pnas.92.16.7480

Lewinski, M. K., Yamashita, M., Emerman, M., Ciuffi, A., Marshall, H., Crawford, G., … Bushman, F. D. (2006). Retroviral DNA integration: Viral and cellular determinants of target-site selection. *PLoS Pathogens*, *2*(6), 0611–0622. http://doi.org/10.1371/journal.ppat.0020060

Li, C. X., Shi, M., Tian, J. H., Lin, X. D., Kang, Y. J., Chen, L. J., … Zhang, Y. Z. (2015). Unprecedented genomic diversity of RNA viruses in arthropods reveals the ancestry of negative-sense RNA viruses. *ELife*, *4*, 1–26. http://doi.org/10.7554/eLife.05378

Li, J., Akagi, K., Hu, Y., Trivett, A. L., Hlynialuk, C. J. W., Swing, D. A., … Symer, D. E. (2012). Mouse endogenous retroviruses can trigger premature transcriptional termination at a distance. *Genome Research*, *22*(5), 870–884. http://doi.org/10.1101/gr.130740.111

Li, L., Kapoor, a., Slikas, B., Bamidele, O. S., Wang, C., Shaukat, S., … Delwart, E. (2010). Multiple Diverse Circoviruses Infect Farm Animals and Are Commonly Found in Human and Chimpanzee Feces. *Journal of Virology*, *84*(4), 1674–1682. http://doi.org/10.1128/JVI.02109-09

Li, M., He, Y., Dubois, W., Wu, X., Shi, J., & Huang, J. (2012). Distinct Regulatory Mechanisms and Functions for p53-Activated and p53-Repressed DNA Damage Response Genes in Embryonic Stem Cells. *Molecular Cell*, *46*(1), 30–42. http://doi.org/10.1016/j.molcel.2012.01.020

Li, W., Lee, M. H., Henderson, L., Tyagi, R., Bachani, M., Steiner, J., … Nath, A. (2015). Human endogenous retrovirus-K contributes to motor neuron disease. *Science Translational Medicine*, *7*(307). http://doi.org/10.1126/scitranslmed.aac8201

Li, X., Mak, J., Arts, E. J., Gu, Z., Kleiman, L., Wainberg, M. A., & Parniak, M. A. (1994). Effects of alterations of primer-binding site sequences on human immunodeficiency virus type 1 replication. *Journal of Virology*, *68*(10), 6198–6206.

Lian, H., Liu, Y., Li, N., Wang, Y., Zhang, S., & Hu, R. (2014). Novel circovirus from mink, China. *Emerging Infectious Diseases*, *20*(9), 1548–1550. http://doi.org/10.3201/eid2009.140015

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 of the United States of America*, *72*(6), 2315–9. http://doi.org/10.1073/pnas.72.6.2315

Liu, H., Fu, Y., Xie, J., Cheng, J., Ghabrial, S. a., Li, G., … Jiang, D. (2011). Widespread Endogenization of Densoviruses and Parvoviruses in Animal and Human Genomes. *Journal of Virology*, *85*(19), 9863–9876. http://doi.org/10.1128/JVI.00828-11

Liu, H., Fu, Y., Li, B., Yu, X., Xie, J., Cheng, J., … Jiang, D. (2011). Widespread Horizontal Gene Transfer from Circular Single-stranded DNA Viruses to Eukaryotic Genomes. *BMC Evolutionary Biology*, *11*(1), 276. http://doi.org/10.1186/1471-2148-11-276

Llorens, C., Fares, M. a, & Moya, A. (2008). Relationships of Gag-pol Diversity between Ty3/Gypsy and Retroviridae LTR retroelements and the three kings hypothesis. *BMC Evolutionary Biology*, *8*(1), 276. http://doi.org/10.1186/1471-2148-8-276

Lorincz, M., Csagola, a., Farkas, S. L., Szekely, C., & Tuboly, T. (2011). First detection and analysis of a fish circovirus. *Journal of General Virology*, *92*(8), 1817–1821. http://doi.org/10.1099/vir.0.031344-0

Lorincz, M., Dán, Á., Láng, M., Csaba, G., Tóth, Á. G., Székely, C., … Tuboly, T. (2012). Novel circovirus in European catfish (Silurus glanis). *Archives of Virology*, *157*(6), 1173–1176. http://doi.org/10.1007/s00705-012-1291-1

Louis, A., Muffato, M., & Crollius, H. R. (2013). Genomicus: Five genome browsers for comparative genomics in eukaryota. *Nucleic Acids Research*, *41*(D1), 700–705. http://doi.org/10.1093/nar/gks1156

Lowe, C. B., Bejerano, G., & Haussler, D. (2007). Thousands of human mobile element fragments undergo strong purifying selection near developmental genes. *Proceedings of the National Academy of Sciences*, *104*(19), 8005–8010. http://doi.org/10.1073/pnas.0611223104

Macera, L., Focosi, D., Vatteroni, M. L., Manzin, A., Antonelli, G., Pistello, M., & Maggi, F. (2016). Cyclovirus Vietnam DNA in immunodeficient patients. *Journal of Clinical Virology*, *81*, 12–15. http://doi.org/10.1016/j.jcv.2016.05.011

Macfarlan, T. S., Gifford, W. D., Driscoll, S., Lettieri, K., Rowe, H. M., Bonanomi, D., … Pfaff, S. L. (2012). Embryonic stem cell potency fluctuates with endogenous retrovirus activity. *Nature*, *487*(7405), 57–63. http://doi.org/10.1038/nature11244

Mager, D. L., & Freeman, J. D. (1995). HERV-H Endogenous Retroviruses: Presence in the New World Branch but Amplification in the Old World Primate Lineage. *Virology*, *213*(2), 395–404. http://doi.org/10.1006/viro.1995.0012

Mager, D. L., & Stoye, J. P. (2015). Mammalian Endogenous Retroviruses. *Microbiology Spectrum*, *3*(1), 1–20. http://doi.org/10.1128/microbiolspec.MDNA3-0009-2014

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*(19), 7385–7390. http://doi.org/10.1073/pnas.1200913109

Magiorkinis, G., Belshaw, R., & Katzourakis, A. (2013). “There and back again”: revisiting the pathophysiological roles of human endogenous retroviruses in the post-genomic era. *Philosophical Transactions of the Royal Society of London. Series B, Biological Sciences*, *368*(1626), 20120504. http://doi.org/10.1098/rstb.2012.0504

Maksakova, I. a., Romanish, M. T., Gagnier, L., Dunn, C. a., Van De Lagemaat, L. N., & Mager, D. L. (2006). Retroviral elements and their hosts: Insertional mutagenesis in the mouse germ line. *PLoS Genetics*, *2*(1), 1–10. http://doi.org/10.1371/journal.pgen.0020002

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*(6), 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*(9), 1307–1318. http://doi.org/10.1101/gr.145000

Malim, M. H., & Bieniasz, P. D. (2012). HIV restriction factors and mechanisms of evasion. *Cold Spring Harbor Perspectives in Medicine*, *2*(5), 1–16. http://doi.org/10.1101/cshperspect.a006940

Markopoulos, G., Noutsopoulos, D., Mantziou, S., Gerogiannis, D., Thrasyvoulou, S., Vartholomatos, G., … Tzavaras, T. (2016). Genomic analysis of mouse VL30 retrotransposons. *Mobile DNA*, *7*(1), 1–14. http://doi.org/10.1186/s13100-016-0066-8

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*(1), 437–443.

McDonald, S. M., Nelson, M. I., Turner, P. E., & Patton, J. T. (2016). Reassortment in segmented RNA viruses: mechanisms and outcomes. *Nature Reviews Microbiology*, *14*(7), 448–460. http://doi.org/10.1038/nrmicro.2016.46

Mclean, C. Y., Bristor, D., Hiller, M., Clarke, S. L., Schaar, B. T., Wenger, A. M., & Bejerano, G. (2016). Regions, *28*(5), 495–501. http://doi.org/10.1038/nbt.1630.GREAT

McNatt, M. W., Zang, T., Hatziioannou, T., Bartlett, M., Fofana, I. Ben, Johnson, W. E., … Bieniasz, P. D. (2009). Species-specific activity of HIV-1 Vpu and positive selection of tetherin transmembrane domain variants. *PLoS Pathogens*, *5*(2), 9–12. http://doi.org/10.1371/journal.ppat.1000300

Medsker, B., Forno, E., Simhan, H., Juan, C., & Sciences, R. (2016). HHS Public Access, *70*(12), 773–779. http://doi.org/10.1097/OGX.0000000000000256.Prenatal

Mi, S., Lee, X., Li, X.-P., Veldman, G. M., Finnerty, H., Racie, L., … McCoy, J. M. (2000). Syncytin is a captive retroviral envelope protein involved in human placental morphogenesis. *Nature*, *403*(6771), 785–789. http://doi.org/10.1038/35001608

Minh, B. Q., Nguyen, M. A. T., & Von Haeseler, A. (2013). Ultrafast approximation for phylogenetic bootstrap. *Molecular Biology and Evolution*, *30*(5), 1188–1195. http://doi.org/10.1093/molbev/mst024

Miranda, C., & Thompson, G. (2016). Canine parvovirus: The worldwide occurrence of antigenic variants. *Journal of General Virology*. http://doi.org/10.1099/jgv.0.000540

Mitchell, P. S., Young, J. M., Emerman, M., & Malik, H. S. (2015). Evolutionary Analyses Suggest a Function of MxB Immunity Proteins Beyond Lentivirus Restriction. *PLoS Pathogens*, *11*(12), 1–21. http://doi.org/10.1371/journal.ppat.1005304

Moreau, C. S. (2014). A practical guide to DNA extraction , PCR , and gene-based DNA sequencing in insects. *Halteres*, *5*, 32–42.

Morissette, G., & Flamand, L. (2010). Herpesviruses and Chromosomal Integration. *Journal of Virology*, *84*(23), 12100–12109. http://doi.org/10.1128/JVI.01169-10

Münk, C., Willemsen, A., & Bravo, I. G. (2012). An ancient history of gene duplications, fusions and losses in the evolution of APOBEC3 mutators in mammals. *BMC Evolutionary Biology*, *12*(1). http://doi.org/10.1186/1471-2148-12-71

Nasir, A., & Caetano-Anollés, G. (2015). A phylogenomic data-driven exploration of viral origins and evolution. *Science Advances*, *1*(8). http://doi.org/10.1126/sciadv.1500527

Nellåker, C., Keane, T. M., Yalcin, B., Wong, K., Agam, A., Belgard, T. G., … Ponting, C. P. (2012). The genomic landscape shaped by selection on transposable elements across 18 mouse strains. *Genome Biology*, *13*(6), R45. http://doi.org/10.1186/gb-2012-13-6-r45

Nguyen, L. T., Schmidt, H. a., Von Haeseler, A., & Minh, B. Q. (2015). IQ-TREE: A fast and effective stochastic algorithm for estimating maximum-likelihood phylogenies. *Molecular Biology and Evolution*, *32*(1), 268–274. http://doi.org/10.1093/molbev/msu300

Oroszlan s. (1990). Retroviral proteinases. *Curr.Top. Microbiol .Immunol*, *146*, 2578–185.

Padilla-Rodriguez, M., Rosario, K., & Breitbart, M. (2013). Novel cyclovirus discovered in the Florida woods cockroach Eurycotis floridana (Walker). *Archives of Virology*, *158*(6), 1389–1392. http://doi.org/10.1007/s00705-013-1606-x

Palinski, R. M., Mitra, N., & Hause, B. M. (2016). Discovery of a novel Parvovirinae virus, porcine parvovirus 7, by metagenomic sequencing of porcine rectal swabs. *Virus Genes*, *52*(4), 564–567. http://doi.org/10.1007/s11262-016-1322-1

Palmer, C. (2013). New Viruses Attack Asia and Africa. *The Scientist*. Retrieved from http://www.the-scientist.com/?articles.view/articleNo/36137/title/New-Viruses-Attack-Asia-and-Africa/

Parvoviruses, H. (2017). crossm, *30*(1), 43–113.

Patel, M. R., Emerman, M., & Malik, H. S. (2011). Paleovirology - Ghosts and gifts of viruses past. *Current Opinion in Virology*, *1*(4), 304–309. http://doi.org/10.1016/j.coviro.2011.06.007

Phan, T., Mori, D., Deng, X., Rajidrajith, S., Ranawaka, U., Ng, F., … Delwart, E. (2015). Small viral genomes in unexplained cases of human encephalitis, diarrhea, and in untreated sewage. *Virology*, (482), 98–104. http://doi.org/10.14574/ojrnhc.v14i1.276.Using

Phan, T. G., Luchsinger, V., Avendaño, L. F., Deng, X., & Delwart, E. (2014). Cyclovirus in nasopharyngeal aspirates of Chilean children with respiratory infections. *Journal of General Virology*, *95*(PART 4), 922–927. http://doi.org/10.1099/vir.0.061143-0

Pichlmair, A., Lassnig, C., Eberle, C.-A., Górna, M. W., Baumann, C. L., Burkard, T. R., … Superti-Furga, G. (2011). IFIT1 is an antiviral protein that recognizes 5′-triphosphate RNA. *Nature Immunology*, *12*(7), 624–630. http://doi.org/10.1038/ni.2048

Prakash, A., Jeffryes, M., Bateman, A., & Finn, R. D. (2017). The HMMER Web Server for Protein Sequence Similarity Search. *Current Protocols in Bioinformatics*, 3.15.1-3.15.23. http://doi.org/10.1002/cpbi.40

Price, M. N., Dehal, P. S., & Arkin, A. P. (2010). FastTree 2 - Approximately maximum-likelihood trees for large alignments. *PLoS ONE*, *5*(3). http://doi.org/10.1371/journal.pone.0009490

Prum, R. O., Berv, J. S., Dornburg, A., Field, D. J., Townsend, J. P., Lemmon, E. M., & Lemmon, A. R. (2015). A comprehensive phylogeny of birds (Aves) using targeted next-generation DNA sequencing. *Nature*, *526*(7574), 569–573. http://doi.org/10.1038/nature15697

Qin, C., Wang, Z., Shang, J., Bekkari, K., Liu, R., Pacchione, S., … Storer, R. D. (2010). Intracisternal a particle genes: Distribution in the mouse genome, active subtypes, and potential roles as species-specific mediators of susceptibility to cancer. *Molecular Carcinogenesis*, *49*(1), 54–67. http://doi.org/10.1002/mc.20576

Qiu, J., Söderlund-Venermo, M., & Young, N. S. (2017). Human parvoviruses. *Clinical Microbiology Reviews*. http://doi.org/10.1128/CMR.00040-16

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

Raidal, S. R., Sarker, S., & Peters, a. (2015). Review of psittacine beak and feather disease and its effect on Australian endangered species. *Australian Veterinary Journal*, *93*(12), 466–470. http://doi.org/10.1111/avj.12388

Ranwez, V., Harispe, S., Delsuc, F., & Douzery, E. J. P. (2011). MACSE: Multiple alignment of coding SEquences accounting for frameshifts and stop codons. *PLoS ONE*, *6*(9). http://doi.org/10.1371/journal.pone.0022594

Ray, A., Rahbari, R., & Badge, R. M. (2011). IAP Display: A simple method to identify mouse strain specific IAP insertions. *Molecular Biotechnology*, *47*(3), 243–252. http://doi.org/10.1007/s12033-010-9338-6

Ren, Q., & Mayden, R. L. (2016). Molecular phylogeny and biogeography of African diploid barbs, 'Barbus', and allies in Africa and Asia (Teleostei: Cypriniformes). *Zoologica Scripta*, *45*(6), 642–649. http://doi.org/10.1111/zsc.12177

Reuter, G., Boros, Á., Delwart, E., & Pankovics, P. (2014). Novel circular single-stranded DNA virus from turkey faeces. *Archives of Virology*, *159*(8), 2161–2164. http://doi.org/10.1007/s00705-014-2025-3

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*(4), 1888–1898. http://doi.org/10.1128/JVI.02051-06

Ribet, D., Harper, F., Dupressoir, A., Dewannieux, M., Pierron, G., & Heidmann, T. (2008). An infectious progenitor for the murine IAP retrotransposon: Emergence of an intracellular genetic parasite from an ancient retrovirus. *Genome Research*, *18*(4), 597–609. http://doi.org/10.1101/gr.073486.107

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*(9), 4413–4419. http://doi.org/10.1128/JVI.02141-07

Rice, P., Longden, I., Bleasby, A. (2000). EMBOSS: The European Molecular Biology Open Software Suite. *Trends in Genetics*, *16*(6), 276–277.

Ristevski, S., Purcell, D. F., Marshall, J., Campagna, D., Nouri, S., Fenton, S. P., … Kannourakis, G. (1999). Novel endogenous type D retroviral particles expressed at high levels in a SCID mouse thymic lymphoma. *J Virol*, *73*(0022-538X (Print)), 4662–4669. Retrieved from http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=112507&tool=pmcentrez&rendertype=abstract

Roelants, K., Haas, a., & Bossuyt, F. (2011). Anuran radiations and the evolution of tadpole morphospace. *Proceedings of the National Academy of Sciences*, *108*(21), 8731–8736. http://doi.org/10.1073/pnas.1100633108

Rosa, A., Chande, A., Ziglio, S., Sanctis, V. De, Goh, S. L., Mccauley, S. M., … Pizzato, M. (2016). Incorporation, *526*(7572), 212–217. http://doi.org/10.1038/nature15399.HIV-1

Rosario, K., Breitbart, M., Harrach, B., Segalés, J., Delwart, E., Biagini, P., & Varsani, A. (2017). Revisiting the taxonomy of the family Circoviridae: establishment of the genus Cyclovirus and removal of the genus Gyrovirus. *Archives of Virology*, *162*(5), 1447–1463. http://doi.org/10.1007/s00705-017-3247-y

Rosario, K., Breitbart, M., Harrach, B., Segalés, J., Delwart, E., Biagini, P., & Varsani, A. (2017). Revisiting the taxonomy of the family Circoviridae: establishment of the genus Cyclovirus and removal of the genus Gyrovirus. *Archives of Virology*. http://doi.org/10.1007/s00705-017-3247-y

Rosario, K., Duffy, S., & Breitbart, M. (2009). Diverse circovirus-like genome architectures revealed by environmental metagenomics. *Journal of General Virology*, *90*(10), 2418–2424. http://doi.org/10.1099/vir.0.012955-0

Rosario, K., Duffy, S., & Breitbart, M. (2012). A field guide to eukaryotic circular single-stranded DNA viruses: Insights gained from metagenomics. *Archives of Virology*. http://doi.org/10.1007/s00705-012-1391-y

Rose, J. (2016). Brazil's sewage woes reflect the growing global water quality crisis. *The Conversation*. Retrieved from https://theconversation.com/brazils-sewage-woes-reflect-the-growing-global-water-quality-crisis-63172

Ross, S. R. (2010). Mouse mammary tumor virus molecular biology and oncogenesis. *Viruses*. http://doi.org/10.3390/v2092000

Rosseel, T., Pardon, B., De Clercq, K., Ozhelvaci, O., & Van Borm, S. (2014). False-positive results in metagenomic virus discovery: A strong case for

follow-up diagnosis. *Transboundary and Emerging Diseases*, *61*(4), 293–299. http://doi.org/10.1111/tbed.12251

Rous. (1911). A Sarcoma of the Fowl Transmissible by an Agent Separable from the Tumor Cells. *The American Journal of the Medical Sciences*, *142*(2), 312. http://doi.org/10.1097/00000441-191108000-00079

Rowe, H. M., Jakobsson, J., Mesnard, D., Rougemont, J., Reynard, S., Aktas, T., … Trono, D. (2010). KAP1 controls endogenous retroviruses in embryonic stem cells. *Nature*, *463*(7278), 237–240. http://doi.org/10.1038/nature08674

Rowe, H. M., Kapopoulou, A., Corsinotti, A., Fasching, L., Macfarlan, T. S., Tarabay, Y., … Trono, D. (2013). TRIM28 repression of retrotransposon-based enhancers is necessary to preserve transcriptional dynamics in embryonic stem cells. *Genome Research*, *23*(3), 452–461. http://doi.org/10.1101/gr.147678.112

Rowe, H. M., & Trono, D. (2011). Dynamic control of endogenous retroviruses during development. *Virology*, *411*(2), 273–287. http://doi.org/10.1016/j.virol.2010.12.007

Sanchez, A., Trappier, S. G., Mahy, B. W., Peters, C. J., Nichol, S. T., Mohan, G. S., … Diseases, I. (2012). Ackn owled gmen ts. *Science*.

Santoni, F. A., Guerra, J., & Luban, J. (2012). HERV-H RNA is abundant in human embryonic stem cells and a precise marker for pluripotency. *Retrovirology*, *9*(1), 1. http://doi.org/10.1186/1742-4690-9-111

Sawyer, S. L., Wu, L. I., Emerman, M., & Malik, H. S. (2005). Positive selection of primate TRIM5 identifies a critical species-specific retroviral restriction domain. *Proceedings of the National Academy of Sciences*, *102*(8), 2832–2837. http://doi.org/10.1073/pnas.0409853102

Sawyer, S. L., Emerman, M., & Malik, H. S. (2007). Discordant evolution of the adjacent antiretroviral genes TRIM22 and TRIM5 in mammals. *PLoS Pathogens*, *3*(12), 1918–1929. http://doi.org/10.1371/journal.ppat.0030197

Schmid, C. D., & Bucher, P. (2010). MER41 repeat sequences contain inducible STAT1 binding sites. *PLoS ONE*, *5*(7). http://doi.org/10.1371/journal.pone.0011425

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*(10), 3461–3470. http://doi.org/10.1093/nar/13.10.3461

Schoggins, J. W., Wilson, S. J., Panis, M., Murphy, M. Y., Jones, C. T., Bieniasz, P., & Rice, C. M. (2011). A diverse range of gene products are effectors of the type i interferon antiviral response. *Nature*, *472*(7344), 481–485. http://doi.org/10.1038/nature09907

Schröder, A. R. W., Shinn, P., Chen, H., Berry, C., Ecker, J. R., Bushman, F., … Jaenisch, R. (2002). HIV-1 integration in the human genome favors active genes and local hotspots. *Cell*, *110*(4), 521–9. http://doi.org/10.1016/s0092-8674(02)00864-4

Shaw, A. E., Hughes, J., Gu, Q., Behdenna, A., Singer, J. B., Dennis, T., … Palmarini, M. (2017). Fundamental properties of the mammalian innate immune system revealed by multispecies comparison of type I interferon responses. *PLoS Biology*, *15*(12), 1–23. http://doi.org/10.1371/journal.pbio.2004086

Shi, M., Lin, X.-D., Tian, J.-H., Chen, L.-J., Chen, X., Li, C.-X., … Zhang, Y.-Z. (2016). Redefining the invertebrate RNA virosphere. *Nature*, *540*(7634), 1–12. http://doi.org/10.1038/nature20167

Simmonds, P., Adams, M. J., Benkő, M., Breitbart, M., Brister, J. R., Carstens, E. B., … Zerbini, F. M. (2017). Consensus statement: Virus taxonomy in the age of metagenomics. *Nature Reviews Microbiology*, *15*(3), 161–168. http://doi.org/10.1038/nrmicro.2016.177

Singer, J. B., Thomson, E. C., McLauchlan, J., Hughes, J., & Gifford, R. J. (2018). GLUE: A flexible software system for virus sequence data. *BioRxiv*. http://doi.org/10.1101/269274

Smedley, D., Haider, S., Durinck, S., Pandini, L., Provero, P., Allen, J., … Kasprzyk, A. (2015). The BioMart community portal: An innovative alternative to large, centralized data repositories. *Nucleic Acids Research*, *43*(W1), W589–W598. http://doi.org/10.1093/nar/gkv350

Smit, AFA, Hubley, R. & G. (n.d.). RepeatMasker Open-4.0.

Smit, A. F. A. (1993). Identification of a new, abundant superfamily of mammalian LTR-transposons. *Nucleic Acids Research*, *21*(8), 1863–1872. http://doi.org/10.1093/nar/21.8.1863

Smits, S. L., Zijlstra, E. E., van Hellemond, J. J., Schapendonk, C. M. E., Bodewes, R., Sch??rch, A. C., … Osterhaus, A. D. M. E. (2013). Novel cyclovirus in human cerebrospinal fluid, Malawi, 2010-2011. *Emerging Infectious Diseases*, *19*(9), 1511–1513. http://doi.org/10.3201/eid1909.130404

Sonigo, P., Wain-Hobson, S., Bougueleret, L., Tiollais, P., Jacob, F., & Brulet, P. (1987). Nucleotide sequence and evolution of ETn elements. *Proceedings of the National Academy of Sciences*, *84*(11), 3768–3771. http://doi.org/10.1073/pnas.84.11.3768

Souza, W. M. De, Romeiro, M. F., Fumagalli, M. J., Modha, S., Araujo, J. De, Queiroz, L. H., … Gifford, R. J. (2017). Chapparvoviruses occur in at least three vertebrate classes and have a broad biogeographic distribution. *Journal of General Virology*, *98*(2), 225–229. http://doi.org/10.1099/jgv.0.000671

Sperber, G. O., Airola, T., Jern, P., & Blomberg, J. (2007). Automated recognition of retroviral sequences in genomic data - RetroTector©. *Nucleic Acids Research*, *35*(15), 4964–4976. http://doi.org/10.1093/nar/gkm515

Stanaway, J. D., Flaxman, A. D., Naghavi, M., Fitzmaurice, C., Vos, T., Abubakar, I., … Cooke, G. S. (2016). The global burden of viral hepatitis from 1990 to 2013: findings from the Global Burden of Disease Study 2013. *Lancet (London, England)*, *388*(10049), 1081–8. http://doi.org/10.1016/S0140-6736(16)30579-7

Steinbiss, S., Willhoeft, U., Gremme, G., & Kurtz, S. (2009). Fine-grained annotation and classification of de novo predicted LTR retrotransposons. *Nucleic Acids Research*, *37*(21), 7002–7013. http://doi.org/10.1093/nar/gkp759

Stenzel, T., & Koncicki, A. (2017). The epidemiology, molecular characterization and clinical pathology of circovirus infections in pigeons – current knowledge, *2176*(May). http://doi.org/10.1080/01652176.2017.1325972

Stewart, M. E., Perry, R., & Raidal, S. R. (2006). Identification of a novel circovirus in Australian ravens (Corvus coronoides) with feather disease. *Avian Pathology*, *35*(2). http://doi.org/10.1080/03079450600597345

Stocking, C., & Kozak, C. a. (2008). Endogenous retroviruses: Murine endogenous retroviruses. *Cellular and Molecular Life Sciences*, *65*(21), 3383–3398. http://doi.org/10.1007/s00018-008-8497-0

Stoye, J. P., & Coffin, J. M. (1988). Polymorphism of murine endogenous proviruses revealed by using virus class-specific oligonucleotide probes. *Journal of Virology*, *62*(1), 168–75. Retrieved from http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=250515&tool=pmcentrez&rendertype=abstract

Stoye, J. P., Moroni, C., & Coffin, J. M. (1991). Virological events leading to spontaneous AKR thymomas. *Journal of Virology*, *65*(3), 1273–85. Retrieved from http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=239902&tool=pmcentrez&rendertype=abstract

Stoye, J. P., & Coffin, J. M. (1987). The Four Classes of Endogenous Murine Leukemia Virus: Structural Relationships and Potential for Recombination. *JOURNAL OF VIROLOGY. Sept*, *61*(9), 2659–2669.

Subramanian, R. P., Wildschutte, J. H., Russo, C., & Coffin, J. M. (2011). Identification, characterization, and comparative genomic distribution of the HERV-K (HML-2) group of human endogenous retroviruses. *Retrovirology*, *8*(1), 90. http://doi.org/10.1186/1742-4690-8-90

Sundaram, V., Cheng, Y., Ma, Z., Sundaram, V., Cheng, Y., Ma, Z., … Wang, T. (2014). innovation of gene regulatory networks Widespread contribution of transposable elements to the innovation of gene regulatory networks, 1963–1976. http://doi.org/10.1101/gr.168872.113

Swofford, D. L. (2002). Phylogenetic Analysis Using Parsimony. *Options*, *42*(2), 294–307. http://doi.org/10.1007/BF02198856

Tan, L. Van, Doorn, H. R. Van, Trung, D., Hong, T., Phuong, T., & Vries, M. De. (2013). Identification of a New Cyclovirus in Cerebrospinal Fluid of Patients with Acute Central Nervus System Infections. *MBio*, *4*(3), e00231-13. http://doi.org/10.1128/mBio.00231-13.Invited

Tarján, Z. L., Pénzes, J. J., Tóth, R. P., & Benkő, M. (2014). First detection of circovirus-like sequences in amphibians and novel putative circoviruses in fishes. *Acta Veterinaria Hungarica*, *62*(1), 134–44. http://doi.org/10.1556/AVet.2013.061

Tarlinton, R. E., Meers, J., & Young, P. R. (2006). Retroviral invasion of the koala genome. *Nature*, *442*(7098), 79–81. http://doi.org/10.1038/nature04841

Taylor, D. J., Dittmar, K., Ballinger, M. J., & Bruenn, J. A. (2011). Evolutionary maintenance of filovirus-like genes in bat genomes. *BMC Evolutionary Biology*, *11*(1), 336. http://doi.org/10.1186/1471-2148-11-336

Temin, H. M. (1976). The DNA provirus hypothesis. *Science (New York, N.Y.)*, *192*(4244), 1075–1080. http://doi.org/10.1126/science.58444

Temin, H. M., & Rubin, H. (1958). Characteristics of an assay for Rous sarcoma virus and Rous sarcoma cells in tissue culture. *Virology*, *6*(3), 669–688. http://doi.org/10.1016/0042-6822(58)90114-4

Todd, D., Scott, A. N. J., Fringuelli, E., Shivraprasad, H. L., Gavier-Widen, D., & Smyth, J. A. (2007). Molecular characterization of novel circoviruses from finch and gull. *Avian Pathology*, *36*(1), 75–81. http://doi.org/10.1080/03079450601113654

Todd, D., Weston, J. H., Soike, D., & Smyth, J. A. (2001). Genome sequence determinations and analyses of novel circoviruses from goose and pigeon. *Virology*, *286*(2), 354–362. http://doi.org/10.1006/viro.2001.0985

Todd, D., Weston, J., Ball, N. W., Borghmans, B. J., Smyth, J. A., Gelmini, L., & Lavazza, A. (2001). Nucleotide sequence-based identification of a novel circovirus of canaries. *Avian Pathology*, *30*(4), 321–325. http://doi.org/10.1080/03079450120066322

Tristem, M. (2000). Identification and characterization of novel human endogenous retrovirus families by phylogenetic screening of the human genome mapping project database. *Journal of Virology*, *74*(8), 3715–30. http://doi.org/10.1128/JVI.74.8.3715-3730.2000

Vernochet, C., Redelsperger, F., Harper, F., Souquere, S., Catzeflis, F., Pierron, G., … Dupressoir, A. (2014). The Captured Retroviral Envelope syncytin-A and syncytin-B Genes Are Conserved in the Spalacidae Together with Hemotrichorial Placentation1. *Biology of Reproduction*, *91*(6), 1–16. http://doi.org/10.1095/biolreprod.114.124818

Victoria, J. G., Wang, C., Jones, M. S., Jaing, C., McLoughlin, K., Gardner, S., & Delwart, E. L. (2010). Viral Nucleic Acids in Live-Attenuated Vaccines: Detection of Minority Variants and an Adventitious Virus. *Journal of Virology*, *84*(12), 6033–6040. http://doi.org/10.1128/JVI.02690-09

Vogt, P. K. (2012). Retroviral oncogenes: A historical primer. *Nature Reviews Cancer*. http://doi.org/10.1038/nrc3320

Walsh, C. P., Chaillet, J. R., & Bestor, T. H. (1998). Transcription of IAP endogenous retroviruses is constrained by cytosine methylation [4]. *Nature Genetics*, *20*(2), 116–117. http://doi.org/10.1038/2413

Waterston, R. H., Lindblad-Toh, K., Birney, E., Rogers, J., Abril, J. F., Agarwal, P., … Mouse Genome Sequencing, C. (2002). Initial sequencing and comparative analysis of the mouse genome. *Nature*, *420*(6915), 520–562. http://doi.org/10.1038/nature01262

Whitcomb, J. M., Ortiz-Conde, B. a, & Hughes, S. H. (1995). Replication of avian leukosis viruses with mutations at the primer binding site: use of alternative tRNAs as primers. *Journal of Virology*, *69*(10), 6228–6238.

Whitfield, Z. J., Dolan, P. T., Kunitomi, M., Tassetto, M., Seetin, M. G., Oh, S., … Andino, R. (2017). The Diversity, Structure, and Function of Heritable Adaptive Immunity Sequences in the Aedes aegypti Genome. *Current Biology*, *27*(22), 3511–3519.e7. http://doi.org/10.1016/j.cub.2017.09.067

Wichman, H. a, Potter, S. S., & Pine, D. S. (1985). Mys, a family of mammalian transposable elements isolated by phylogenetic screening. *Nature*, *317*(6032), 77–81. Retrieved from http://eutils.ncbi.nlm.nih.gov/entrez/eutils/elink.fcgi?dbfrom=pubmed&id=2993916&retmode=ref&cmd=prlinks%5Cnpapers2://publication/uuid/5964753B-3921-4034-8366-325F798774FD%5Cnhttp://www.ncbi.nlm.nih.gov/pubmed/2993916

Widegren, B., Kjellman, C., Aminoff, S., Sahlford, L. G., & Sjogren, H. O. (1996). The structure and phylogeny of a new family of human endogenous retroviruses. *J Gen Virol*, *77 (Pt 8)*, 1631–1641. http://doi.org/10.1099/0022-1317-77-8-1631

Wills, J. W., Cameron, C. E., Wilson, C. B., Xiang, Y., Bennett, R. P., & Leis, J. (1994). An assembly domain of the Rous sarcoma virus Gag protein required late in budding. *Journal of Virology*, *68*(10), 6605–18. Retrieved from http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=237081&tool=pmcentrez&rendertype=abstract

Wolf, D., & Goff, S. P. (2009). Embryonic stem cells use ZFP809 to silence retroviral DNAs. Nature 458:1201–1204. http://doi.org/10.1038/nature07844.Embryonic

Wolf, D., & Goff, S. P. (2007). TRIM28 Mediates Primer Binding Site-Targeted Silencing of Murine Leukemia Virus in Embryonic Cells. *Cell*, *131*(1), 46–57. http://doi.org/10.1016/j.cell.2007.07.026

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*(9), 7459–7466.

Wu, X., Li, Y., Crise, B., Burgess, S. M., & Lj, Y. (2015). Human Targets Start Regions Are Favored in the Genome for MLV, *300*(5626), 1749–1751. http://doi.org/10.1126/science.1083413

Xiong, Y., & Eickbush, T. H. (1990). Origin and evolution of retroelements based upon their reverse transcriptase sequences. *The EMBO Journal*, *9*(10), 3353–62. Retrieved from http://www.ncbi.nlm.nih.gov/pubmed/1698615

Xu, L., Yu, D., Fan, Y., Peng, L., Wu, Y., & Yao, Y. (2016). Loss of RIG-I leads to a functional replacement with MDA5 in the Chinese tree shrew. *Proceedings of the National Academy of Sciences of the United States of America*, *113*(39), 1–6. http://doi.org/10.1073/pnas.1604939113

Yang, S., Liu, Z., Wang, Y., Li, W., Fu, X., Lin, Y., … Zhang, W. (2016). A novel rodent Chapparvovirus in feces of wild rats. *Virology Journal*, *13*(1), 133. http://doi.org/10.1186/s12985-016-0589-0

Young, N. L., & Bieniasz, P. D. (2007). Reconstitution of an infectious human endogenous retrovirus. *PLoS Pathogens*, *3*(1), 0119–0130. http://doi.org/10.1371/journal.ppat.0030010

Zhang, Y., Maksakova, I. A., Gagnier, L., Van De Lagemaat, L. N., & Mager, D. L. (2008). Genome-wide assessments reveal extremely high levels of polymorphism of two active families of mouse endogenous retroviral elements. *PLoS Genetics*, *4*(2). http://doi.org/10.1371/journal.pgen.1000007

Zhao, Y., Jacobs, C. P. D., Wang, L., & Hardies, S. C. (1999). MuERVC: A new family of murine retrovirus-related repetitive sequences and its relationship to previously known families. *Mammalian Genome*, *10*(5), 477–481. http://doi.org/10.1007/s003359901026

Zheng, R., Jenkins, T. M., & Craigie, R. (1996). Zinc folds the N-terminal domain of HIV-1 integrase, promotes multimerization, and enhances catalytic activity. *Proceedings of the National Academy of Sciences of the United States of America*, *93*(24), 13659–64. http://doi.org/10.1073/pnas.93.24.13659

Zhou, Y., & Holmes, E. C. (2007). Bayesian estimates of the evolutionary rate and age of hepatitis B virus. *Journal of Molecular Evolution*, *65*(2), 197–205. http://doi.org/10.1007/s00239-007-0054-1

Zhu, H., Dennis, T., Hughes, J., & Gifford, R. J. (2018). Database-integrated genome screening (DIGS): exploring genomes heuristically using sequence similarity search tools and a relational database. *BioRXIV*.

Appendices

Appendix 2.1: Taxonomy and details of WGS of organisms analysed in this thesis.

Appendix 3.1: Details and accession numbers of sequences used in MSAs of Retroviridae RT

Appendix 3.2: Proviral LTRs recovered, filtered and dated in pairwise-LTR based dating of murine ERVs

Appendix 4.1: Details of CVe recovered in screening of animal WGS assemblies. Sequence-ID indicates bespoke sequence identifier. WGS_v is assembly version. Bp indicates length, desc indicates where the Cve was first identified.

Appendix 4.2: Sequences used in ML phylogenies of Circoviridae Rep (Figure 4.4, Figure 4.5 and others)

Appendix 4.3: DFAM output of Carnivore Cve sequences

Appendix 5.1: ENSEMBL ID, common name, and full name of ISGs used in screens in chapter 5

Due to length considerations, (~100 pages) appendices are found at:

https://drive.google.com/file/d/1fXWL_QYL6kNncXz7slPSWOfYeJiEpCd3/view?usp=sharing