A phylogenomic study characterising the co-option and evolutionary history of endogenous viral elements in bats (order Chiroptera)

Ilya Levantis

School of Biological and Chemical Sciences

Queen Mary University London

Mile End Road, London, E1 4NS

Submitted in partial fulfillment of the requirements of the

Degree of Doctor of Philosophy, January 2021

Author's Declaration

I, Ilya Levantis, confirm that the research included within this thesis is my own work or that where it has been carried out in collaboration with, or supported by others, that this is duly acknowledged below and my contribution indicated. Previously published material is also acknowledged below. I attest that I have exercised reasonable care to ensure that the work is original, and does not to the best of my knowledge break any UK law, infringe any third party's copyright or other Intellectual Property Right, or contain any confidential material. I accept that the College has the right to use plagiarism detection software to check the electronic version of the thesis. I confirm that this thesis has not been previously submitted for the award of a degree by this or any other university. The copyright of this thesis rests with the author and no quotation from it or information derived from it may be published without the prior written consent of the author.

Signature: I.Levantis

Date: 27/01/2021

Details of collaborations and data acquisition:

- A total of 47 raw RNA-seq read libraries were generated by members of the RossiterLab research group (Kalina Davies, Josh Potter), and Liliana Dávalos, which were used in Chapter 5.
- Four unpublished genome assemblies were provided by members of the RossiterLab research group (Georgia Tsagkogeorga, and Josh Potter) which were used in Chapters 4 and 5.
- Bat spleen samples were collected from Copenhagen Zoo by Stephen Rossiter, Josh Potter, and Rosie Drinkwater.
- RNA isolation, library preparation, and sequencing was undertaken by researchers at Novogene.

Acknowledgements

First and foremost, I would like to thank my primary supervisor Steve Rossiter for the invaluable advice and guidance over the past four years that have contributed to making this such an enjoyable project to work on. My thanks also to my secondary supervisor, Aris Katzourakis, for making this exciting collaborative project a reality, and for being a rich seam of knowledge in the tricks of the paleovirology trade. A special thanks is also due to Nadine Mogford for being the most incredible organiser of all things related to the LIDo DTP.

My gratitude also extends to all the members of my research group and SBCS for the invaluable discussions and their input, as well as all the great times that were had – you lot have been pretty fantastic! In particular, Josh Potter, Kalina Davies, and Georgia Tsagkogeorga have all contributed so much in laying the foundations for my research through the extensive amount of fieldwork collecting samples, and labwork preparing sequencing libraries. Thank you to Roddy Pracana for teaching me so much about bioinformatics and being a frequent source for consultation, especially in the early days. To Sandra Álvarez, Joe Williamson, James Gilbert, Rosie Drinkwater, and Sabhrina Gita Aninta, you have been such great people with whom to bounce ideas around, and your wisdom in statistics and phylogenetics has helped shaped this thesis into something I am proud of.

I must also thank all the members of London Biohackspace and the wider DIYBio community who provided the many fresh ideas and projects that kept me excited about biology over the years, and without whom I never would have pursued this PhD. Chief among them are Edo Gianni, Tom Hodder, Sam Thompson, and Bethan Wolfenden, who were incredible crewmates and friends for our expedition into unexplored waters.

Finally, I would like to thank everyone I met at SBCS, through the LIDo DTP, and all my existing friends that have made the pubs, trips, and other activities so enjoyable. Thanks for keepin' it real through lockdown and thesis writing!

Above all, thanks Jemma.

I am thankful to the Biotechnology and Biological Sciences Research Council and The London Interdisciplinary Biosciences Consortium for providing the funding and the opportunity to carry out this work. I also send my support and gratitude to Alexandra Elbakyan for fighting the good fight in academia and bringing about real change for the better.

Abstract

Endogenous viral elements (EVEs) are the genetic fragments of viruses found within the genomes of eukaryotes. These elements represent 'fossils' of viruses that once infected the ancestors of the host, millions of years in the past, and thus contain a wealth of information about the deep history of viruses. In this thesis, I develop a novel, automated approach to sensitively detect EVEs and accurately recover their genetic sequence, even in the presence of multiple frame-disrupting mutations. I then use this approach to carry out the most phylogenetically comprehensive, systematic census of EVEs in bat genomes to date, providing new data for the study of viruses associated with bats. Bats (order Chiroptera) are a clade of extreme species richness. Constituting around one fifth of all mammal species, they harbour a similarly high diversity of viruses – including many that pose public health risks as zoonoses, making their associations with viruses and their potential as zoonotic reservoirs important areas of research. The EVE census I carried out enabled me to compare the distribution of endogenous retrovirus (ERV) lineages across bats, infer rates of cross-species retrovirus transmission between bats over the course of evolutionary time, and ultimately show significant differences in this long-term transmission rate between different phylogenetic levels and different clades of bats. In addition, I use bat EVEs to provide new evidence characterising the ancient biogeography, host-range, and macro-evolutionary patterns of parvoviruses, filoviruses, and bornaviruses. Finally, I combine the EVE data I generated with a large set of transcriptomic data to examine the broad patterns of ERV co-option occurring in bats, and to locate examples of specific EVE loci that are potentially co-opted for roles in the bat immune system. Together, these analyses show the diverse insights about both hosts and viruses that large comparative studies of EVEs can provide.

Contents

Glossary **12**

1 General introduction **20**

1.1 Bats are the products of a unique evolutionary radiation 20

1.2 The unique bat immune system 22

1.3 Do bats have a special relationship with viruses? 26

1.3.1 Inter- and intra-host virus dynamics in bats 26

1.3.2 Are bats a special reservoir for zoonotic viruses? 27

1.4 Endogenous viral elements (EVEs) 29

1.4.1 EVEs are found across eukaryotic genomes 29

1.4.2 EVEs are a viral fossil record 30

1.5 EVEs are a source of genomic novelty 34

1.5.1 ERVs can enhance genomic plasticity 34

1.5.2 EVEs can be co-opted as *cis*-regulatory elements 35

1.5.3 EVE co-option for generation of non-coding transcripts 35

1.5.4 EVE co-option for generation of coding transcripts 37

1.5.5 EVE-derived immunity genes 37

1.6 Outline of work 39

2 Improving upon current EVE mining methods 40

Abstract 41

2.1 Introduction 41

2.2 Methods 44

2.2.1 Selection of genome assemblies 44

2.2.2 Restricting search regions to putative ERV regions 44

2.2.3 RT locus identification using four search methods 46

2.2.4 Sensitivity analysis 47

2.2.5 Predicted sequence completeness analysis 47

2.3 Results 48

2.3.1 Differences in number and length of identified RT sequence between search methods 52

2.3.2 Differences in detected RT domain abundance between genome assemblies 53

2.3.3 DVORFS: An easy-to-use wrapper and post-processor for EVE mining with GeneWise 55

2.4 Discussion 56

2.4.1 Frameshift-aware protein pHMM to DNA alignment search improves EVE detection and sequence recovery 56

2.4.2 Genome assembly strategies have a large impact on assembled TE content 58

2.4.3 The need for a high sensitivity search tool for EVE identification . 60

3 ERV distribution and retrovirus transmission within the Chiroptera 62

Abstract 63

3.1 Introduction 63

3.1.1 Bats are a panglobal clade of extreme species richness 63

3.1.2 Bats are zoonotic reservoirs for viruses of public health concern . 64

3.1.3 ERVs are a record of retrovirus infections 65

3.2 Methods 67

3.2.1 RT alignment and retrovirus reference phylogeny creation 67

3.2.2 Phylogenetic placement of RT sequences and Edge PCA comparison 69

3.2.3 Inference of horizontal ERV transfers 70

3.2.4 Statistical modelling of ERV transfer rates 71

3.3 Results 72

3.3.1 Differences in ERV composition among bats species 72

3.3.2 Rates of cross-species retrovirus transmission in bats over evolutionary time 76

3.4 Discussion 83

3.4.1 An unbiased analysis of ERV activity across Chiroptera 83

3.4.2 Numerous expansions and extinctions of bat ERV lineages 84

3.4.3 RT sequence gene reconciliation reconstructs the retrovirus sharing history of Chiroptera 86

3.4.4 Phylogenetic relatedness impacts the rate of retrovirus transmission 87

3.4.5 Increased rate of cross-species retrovirus transmission within noctilionoid bats 88

4 Non-retroviral EVEs in bat genomes **90**

Abstract 91

4.1 Introduction 91

4.1.1 Bats harbour a diversity of viruses 91

4.1.2 Non-retrovirus EVEs are rare in vertebrate genomes 92

4.1.3 Bornaviruses and bornavirus-derived elements 92

4.1.4 Filoviruses and filovirus-derived elements 94

4.1.5 Parvoviruses and parvovirus-derived elements 96

4.1.6 Non-retroviral EVEs in bat genomes 97

4.2 Methods 98

4.2.1 EVE mining 98

4.2.2 Orthology clustering 99

4.2.3 Sequence alignment and phylogenetic tree construction 100

4.2.4 Selection analysis 101

4.2.5 Structural protein modelling 101

4.3 Results 102

4.3.1 Bornavirus-derived elements in bat genomes 103

4.3.2 chEBL elements with conserved ORFs 107

4.3.3 Filovirus-derived elements in bat genomes 109

4.3.4 Parvovirus-derived elements in bat genomes 114

4.4 Discussion 117

4.4.1 Ancient cultervirus and carbovirus infections in bats 117

4.4.2 Phylogenetic distribution of bornavirus elements among bats . . . 118

4.4.3 Phylogenetic distribution of filovirus elements among bats 120

4.4.4 Phylogenetic distribution of parvovirus elements among bats . . . 121

4.4.5 Putatively co-opted EBL and EFL elements in bats 122

5 Patterns of EVE expression and co-option across the Chiroptera 124

Abstract . . . 125

5.1 Introduction . . . 125

5.1.1 Bats are a speciose clade with unique physiological traits . . . 125

5.1.2 ERVs are important sources of genetic novelty . . . 126

5.1.3 ERV loci play an important role in antiviral immunity . . . 129

5.2 Methods . . . 130

5.2.1 Spleen transcriptome sequencing . . . 130

5.2.2 EVE annotation . . . 131

5.2.3 EVE orthology analysis . . . 131

5.2.4 Transcript prediction and abundance estimation . . . 132

5.2.5 Differential expression analysis . . . 133

5.2.6 Statistical comparison of expressed and unexpressed ERVs . . . 133

5.3 Results . . . 133

5.3.1 Transcribed non-retroviral EVEs . . . 134

5.3.2 Differences in degradation between transcribed and untranscribed ERV loci . . . 134

5.3.3 Immune response RNA-seq datasets show significant differential expression of ERV loci . . . 137

5.3.4 Searching for co-opted EVEs using evolutionary conservation . . . 141

5.4 Discussion . . . 145

5.4.1 Transcribed ERV loci are often less degraded than untranscribed ERV loci . . . 145

5.4.2 Transcriptomic immune responses in bats include the up- and downregulation of ERV-derived transcripts . . . 147

5.4.3 Candidate EVE loci for further investigation 148

6 General Discussion **151**

References **156**

Appendices

A Chapter 2 supplementary material **172**

A.1 Supplementary tables 172

A.2 Supplementary methods 172

A.2.1 Genome assemblies used in custom ERV library construction . . . 172

B Chapter 3 supplementary material **174**

B.1 Supplementary methods 174

B.1.1 Genome assemblies used in Chapter 3 174

C Chapter 4 supplementary material **175**

C.1 Supplementary tables 175

C.2 Supplementary Figures 186

C.3 Supplementary methods 187

C.3.1 Genome assemblies used in Chapter 5 187

D Chapter 5 supplementary material **188**

D.1 Supplementary tables 188

D.2 Supplementary figures 193

D.3 Supplementary methods 196

D.3.1 Genome assemblies used in Chapter 5 196

Glossary

AIC Akaike information criterion. An estimator of the relative quality of a statistical model where better quality models have a lower AIC. AIC considers both the simplicity of the model (number of parameters) and the goodness of fit of the model.

BEB posterior probabilities Bayes empirical Bayes probabilities in a PAML analysis give the posterior probabilities that each codon is from a particular class of site (e.g. posterior probability of a site evolving under positive selection).

Backbone alignment An MSA of a set of preselected sequences which provides the basis for further sequences to be added to the MSA following the patterns of homology already present in the backbone alignment.

Bayesian phylogenetic tree A phylogenetic tree generated by searching for trees with the highest posterior probability that the tree is correct for the given sequence alignment, substitution model, and a prior probability distribution.

Bonferroni correction A method to correct for FWER when performing multiple hypothesis tests by adjusting the P value significance threshold proportional to the number of tests carried out.

Boreoeutheria Clade of placental mammals with estmiated divergence date of 83 Mya consisting of the Laurasiatheria and the Euarchontoglires (rodents, lagomorphs, treeshrews, colugos, primates).

Contig N50 The length of the shortest contig (in base pairs) at 50% of the total length of a genome assembly.

EDI EVE-derived immunity. The result of EVEs being co-opted by the host for immune functions.

ERV Endogenous retrovirus. A retrovirus genome that is integrated into the host cell genome, forming the provirus stage of the retrovirus life-cycle. ERVs can multiply within a genome through intracellular retrotransposition categorising them as class I TEs as well as EVEs.

EVE Endogenous viral element. EVEs are virus-derived sequences found in cellular genomes.

Edge PCA Edge principal component analysis. The sum of LWRs across the branches of a phylogenetic tree resulting from a set of phylogenetic placements can be analysed using edge PCA. Each branch is weighted according to the distribution of LWRs on either side of it such that branches of the tree across which there is a high level of between-sample heterogeneity receive a large weight. Standard PCA is then used to identify principal components across the branches which can then be used to identify clusters or trends in the samples based on differences in their phylogenetic composition.

Emballonuroidea Yangochiropteran superfamily of bats containing Emballonuridae and Nycteridae families.

Eocene Geological epoch spanning from 56 to 34 Mya. The beginning of the Eocene is marked by a short period of intense global warming. During the Eocene Australia split from Antarctica and the northern supercontinent of Laurasia began to fragment. The end of the Eocene is marked by major floral and fuanal turnover.

FWER Family wise error rate. The probability of false positives (type I errors) when multiple hypothesis tests are performed.

GAM Generalised additive model. A statistical model that is an extension of a GLM, where the response variable is related to a sum of predictor variables each of which is first transformed by a fitted smooth function.

GLM Generalised linear model. A generalised form of linear regression that relates a response variable to a linear combination of a set of numerical or categorical predictor variables and their interactions. The relationship of the linear combination to the response variable can be one of a number of different link functions and the response variable can be modelled as one of a number of different probability distributions.

GeneWise 3:33L alignment model A hidden markov model for alignment of protein profiles (e.g. a protein pHMM) against DNA sequence. This model allows for codon matches, codon insertions, codon deletions, frameshifting nucleotide insertions, and frameshifting nucleotide deletions, but does not model intron splice sites.

Genome annotation Information marking the locations of specific regions of interest in a genome and information about those regions, for example the genomic positions of a specific gene's exons or promoter sequence.

HERV Human endogenous retrovirus.

Helitron element A family of class II TEs that replicate by a rolling circle replication mechanism.

IFN Interferon. IFNs are a group of antiviral cytokines categorised into three types (type I IFNs, type II IFNs, and type III IFNs) based on the receptors through which they signal.

IFNα Interferon-alpha. A type I IFN that forms a key part of the innate antiviral response signalling. It is regulated by IRF3/IRF7 and binds to cell membrane receptors found on most cell types, leading to strong and ubiquitous antiviral responses through the JAK-STAT pathway.

IFNκ Interferon-kappa. A type I IFN.

IFNλ Interferon-lambda. A type III IFN which is distantly related to type I IFNs but is also involved in innate antiviral immune response and likewise signals through the JAK-STAT pathway.

IFNω Interferon-omega. A type I IFN.

IL-10 Interleukin 10. An anti-inflammatory cytokine. IL-10 can inhibit activity of NF-κB (nuclear factor kappa-light-chain-enhancer of activated B cells), and is involved in regulation of the JAK-STAT signalling pathway.

IL-1β Interleukin 1 beta. A pro-inflammatory cytokine - expression mediated by the NLRP3 inflammasome.

ILS Incomplete lineage sorting. A phenomenon where polymorphism within an ancestral population followed by speciation events leads to a gene tree that is discordant with the species tree.

IRF3 Interferon regulatory factor 3. Acts as a central regulator of innate antiviral responses. When activated, IRF3 induces the expression of type I IFNs.

IRF7 Interferon regulatory factor 7. Like IRF3, IRF7 is a transcription factor that plays an important role in the expression of type I IFNs during innate antiviral responses.

ISG IFN-stimulated genes. A broad group of genes that are upregulated in response to IFN signalling which inhibit virus replication through a variety of different mechanisms.

JAK-STAT signalling pathway A signalling pathway that involves cytokine-binding cell-surface receptors, Janus kinases (JAKs), and signal transducer and activator of transcription proteins (STATs). The JAK-STAT pathway plays a key role in cytokine signalling, transducing extracellular cytokine signals to effect transcriptional changes in the cell nucleus.

LINE Long interspersed element. A family of autonomous non-LTR retrotransposons which contain an ORF coding for an RNA binding protein and an ORF coding for a protein containing an endonuclease and reverse transcriptase domain.

LRT Likelihood ratio test. A statistical test to determine whether a statistical model has a significantly better fit than a model with contains a subset of the focal model's parameters using a ratio of the two models' likelihoods.

LTR Long terminal repeat. LTRs are non-coding sequences that are found as sequence-identical pairs flanking the internal, gene containing region of ERVs and other retrotransposons. Recombination between the two LTRs of a retroelement leads to the formation of a solo-LTR.

LWR Likelihood weight ratio. Statistic calculated as part of phylogenetic placement which is the maximum likelihood value for placement of a sequence on a branch of the reference tree normalised so that all placement location maximum likelihood values sum to one.

Laurasiatheria Clade of placental mammals with estimated divergence date of 76 Mya consisting of the Laurasiatheria and the Eulipotyphla (hedgehogs, shrews, moles, solenodons).

MDA5 Melanoma Differentiation-Associated protein 5. A cytosolic pathogen RNA sensing PRR which activates antiviral and inflammatory cytokine pathways.

MSA Multiple sequence alignment. A set of sequences aligned with each other in a matrix - rows correspond to sequences, columns correspond to homologous sites along the set of sequences.

Mann-Whitney U test A statistical test that is non-parametric and allows the comparison of two groups without the assumption that values are normally distributed.

Maximum likelihood phylogenetic tree A phylogenetic tree generated by searching for trees with the highest likelihood for the given sequence alignment and substitution model.

Maximum likelihood tree bootstrapping A method of providing support values or confidence values for splits in a ML phylogenetic tree. The columns in the sequence alignment are subsetted into multiple pseudoreplicate alignments each of which is then used to calculate an ML tree. Bootstrap support values are the proportion of these pseudoreplicate trees that have the same split.

Mya Million years ago.

NAHR Non-allelic homologous recombination. When two different regions of a chromosome contain highly similar sequences, homologous recombination can occur between these regions, resulting in large-scale genomic deletions or duplications.

NLRP3 inflammasome A protein complex composed of NLRP3 (NOD-, LRR- and pyrin domain-containing protein 3) and other proteins. When activated leads to release of pro-inflammatory cytokines, including IL-1β.

Noctilionoidea Yangochiropteran superfamily of bats containing Thyropteridae, Furipteridae, Noctilionidae, Mormoopidae, Phyllostomidae, Myzopodidae, and Mystacinidae families.

ORF Open reading frame.

PAMP Pathogen associated molecular pattern. Molecular motifs typically associated with pathogens which are recognised by PRRs and exploited as a signal for immune activation by host cells.

PCA Principal component analysis. A method for reducing the dimensionality of a dataset by defining new uncorrelated variables that are composed of weightings of the original variables which maximise the variance of the dataset.

PRR Pattern recognition receptor. Receptors expressed by host cells enabling the sensing of PAMPs or molecules associated with cell damage.

PYHIN genes Pyrin and hematopoietic interferon-inducible nuclear (HIN) domain-containing protein gene family. The PYHIN family of genes codes for cytosolic DNA-sensing proteins that activate inflammasome and interferon pathways.

Phylogenetic placement A method for estimating the phylogenetic position of a given sequence on a pre-existing phylogenetic tree (i.e. a reference tree). For the given sequence, a likelihood weight ratio (LWR) is calculated for each branch on the phylogenetic tree that represents the relative likelihood that the sequence would adjoin the phylogenetic tree at that branch.

Phylogenetic tree reconciliation A method which fits a secondary phylogenetic tree (e.g. a gene tree or co-evolving species tree) to a species tree by modelling speciation, duplication, transfer, and loss events in the secondary tree. This provides an estimated evolutionary history of the secondary tree in the context of the species tree.

Polytomy A split or node in a phylogenetic tree with more than two descendant branches.

Pteropodidae Pteropodidae (Old World fruit bats) are a family within Yinpterochiroptera that form a sister clade to the Rhinolophoidea superfamily. Pteropodoidea is sometimes used to refer to the superfamily that contains only the Pteropodidae and is therefore functionally the same taxonomic group as the Pteropodidae. Pteropodid species cover ranges across the Old World continents.

RIG-I Retinoic acid-inducible gene-I. A cytosolic pathogen RNA sensing PRR which activates antiviral and inflammatory cytokine pathways.

Rhinolophoidea A superfamily within Yinpterochiroptera that form a sister clade to the Pteropodidae. The Rhinolophoidea consists of the Craseonycteridae, Hipposideridae, Megadermatidae, Rhinolophidae, Rhinonycteridae, and Rhinopomatidae families.

SINE Short interspersed element. A family of non-autononomous non-LTR retrotransposons typically 100-700 bp in length. SINEs are derived from retrotranscribed ribosomal RNAs and tRNAs and rely on the replication machinery produced by LINEs for retrotransposition.

STING Stimulator of interferon genes. A cytosolic DNA-sensing protein that activates type I interferons.

Scrotifera Clade of placental mammals within the Laurasiatheria with estmiated divergence date of 73 Mya. Scrotifera consists of the Chiroptera (bats), Carnivora , Pholidota (pangolins), Perissodactyla (odd-toed ungulate) and Cetartiodactyla (even-toed ungulate) orders.

Search query In alignment-based sequence search methods the query is compared against a database of sequences to find sequences or subsequences within the database that are similar to the query.

Sequencing coverage (read depth) The mean number of sequencing reads aligned to each base pair in a genome assembly.

Software pipeline A collection of software used to process data in defined successive steps.

Software wrapper A piece of software with the purpose of calling/running specific other software in order to abstract away some the implementation details of running that software.

Substitution matrix A matrix describing the relative rate at which each amino acid or nucleotide residue is substituted for each other residue.

Substitution model A model of the rate of nucleotide or amino acid substitutions at a given site and the distribution of substitutions across the entire sequence. Substitution models are used to optimise topology and calculate branch lengths when inferring phylogenetic trees from MSAs.

TE Transposable element. A selfish genetic element that can replicate within a genome. Retrotransposons or retroelements (class I TEs) replicate through retrotransposition. DNA transposons (class II TEs) can change genomic location through cut-and-paste or rolling-circle transposition which sometimes results in duplication.

TLR8 Toll-like receptor 8. Endosomal TLR which recognises viral nucleic acids.

TLR9 Toll-like receptor 9. Endosomal TLR which recognises viral nucleic acids.

TLR Toll-like receptor. A family of transmembrane receptors which trigger the expression of proinflammatory genes in response to binding to PAMPs. Most mammals have 10 different TLRs, each recognizing different ligands.

TNFα Tumour necrosis factor alpha. Inflammatory cytokine involved in both pro-inflammatory and anti-inflammatory pathways.

UTR Untranslated region.

Vespertilionoidea Yangochiropteran superfamily with a cosmopolitan distribution. Contains the Vespertilionidae and Molossidae families.

Yangochiroptera Suborder of Chiroptera consisting of the Rhinolophoidea, Emballonuroidea, Noctilionoidea, and Vespertilionoidea superfamilies.

Yinpterochiroptera Suborder of Chiroptera consisting of the Pteropodidae family, and Rhinolophoidea superfamily.

c-Rel The c-Rel protein is a member of the NF-κB (nuclear factor kappa-light-chain-enhancer of activated B cells) family of transcription factors.

lncRNA Long non-coding RNA.

pHMM Profile-hidden markov model. A probabilistic model that describes the variation present at each position along a set of aligned sequences.

piRNA Piwi-interacting RNA. Short RNAs that form a complex with Piwi protein to act as an RNAi silencing mechanism.

1

General introduction

Contents

1.1 Bats are the products of a unique evolutionary radiation 20
1.2 The unique bat immune system 22
1.3 Do bats have a special relationship with viruses? 26
1.3.1 Inter- and intra-host virus dynamics in bats 26
1.3.2 Are bats a special reservoir for zoonotic viruses? 27
1.4 Endogenous viral elements (EVEs) 29
1.4.1 EVEs are found across eukaryotic genomes 29
1.4.2 EVEs are a viral fossil record 30
1.5 EVEs are a source of genomic novelty 34
1.5.1 ERVs can enhance genomic plasticity 34
1.5.2 EVEs can be co-opted as *cis*-regulatory elements 35
1.5.3 EVE co-option for generation of non-coding transcripts 35
1.5.4 EVE co-option for generation of coding transcripts 37
1.5.5 EVE-derived immunity genes 37
1.6 Outline of work 39

1.1 Bats are the products of a unique evolutionary radiation

Comprising around one fifth of all mammal species, bats number over 1300 named species across over 200 genera, and 21 families, making the Chiroptera the second-most speciose mammal order after the Rodentia (Burgin et al. 2018). The order is split into two suborders with an estimated early divergence date of ~64 Mya: Yinpterochiroptera, consisting of the Pteropodidae family and Rhinolophoidea superfamily, and Yangochi-

roptera, consisting of the Rhinolophoidea, Emballonuroidea, Noctilionoidea, and Vespertilionoidea superfamilies (Teeling et al. 2005; Jebb et al. 2020). Yinpterochiroptera likely originated in Asia and is found exclusively in the Old World. In contrast, Yangochiroptera has a global distribution with a more uncertain location of origin (Teeling et al. 2005; Jones and Teeling 2006). Overall, twice as many bat species are found in the Old World compared with the New World, although bats achieve their greatest alpha species richness in South America (Racey 2015). Across the 21 families, species richness also varies greatly among families, from one and two species in the Craseonycteridae and Noctilionidae respectively, to over 200 species in the Phyllostomidae and nearly 500 species in the Vespertilionidae (Burgin et al. 2018).

Bats likely diverged approximately 65 Mya from the other orders of the Scrotifera clade, within the superorder Laurasiatheria (X. Zhou et al. 2012; Jebb et al. 2020). After this split, bats evolved the capability of powered flight, followed by echolocation (Jones and Teeling 2006; Simmons et al. 2008), which together allowed bats to occupy and eventually dominate nocturnal aerial feeding niches that are relatively free of competitors and predators, diversifying as they spread across the continents (J.J. Shi and Rabosky 2015). This evolutionary radiation has resulted in the diversity of bat species with the range of feeding strategies, life histories, and physiologies observed in modern bats.

Although ~70% of bat species are insectivorous, some species have transitioned to carnivory, preying on small land vertebrates and, in some cases, specialising on small fishes. Other species have evolved different dietary specialisations and feeding strategies altogether, including frugivory, nectarivory, sanguivory (vampire bats are the only mammals to feed exclusively on blood), and omnivory (Racey 2015; Teeling et al. 2018). Remarkably, obligate frugivory and nectarivory have evolved independently in at least two separate bat lineages – the yinpterochiropteran Pteropodidae family and the yangochiropteran Phyllostomidae family (K. Wang et al. 2020).

When not foraging, bats exploit a range of roosting locations. They are the only group of vertebrates to have successfully made use of caves as daytime shelters, benefiting from temperatures that allow individuals to reduce the energy cost of homeothermy (Kunz 1982). Cave roosts are frequently inhabited by extremely large numbers of individuals, with up to tens of millions of individuals sometimes aggregating in a single cave in the case of *Tadarida brasiliensis* (McCracken 2003). Furthermore, caves are often co-habited by multiple species of bat (Kunz 1982). Aside from cave-roosting bats, other species are found to roost in tree cavities, within foliage such as unfurling leaves, or roost by hanging from branches (Racey 2015). Interestingly, some bats (especially vespertilionids), have become some of the most synanthropic of all vertebrates, having frequently

taken up – and benefited from – roosting within roof spaces of houses, churches and other buildings (Racey 2015).

As with birds, the ability of bats to fly has contributed to their patterns of dispersal, as well as their ability to migrate over mid to long distances, although less than 7% of bats are suspected to migrate (Racey 2015). Indeed, most migratory bat species are considered to be regional migrants, moving around 100-500 km each way – this is seen in some temperate species of bat which migrate to and from hibernacula (Krauel and McCracken 2013). Additionally, a small number of temperate species migrate >2000 km in order to overwinter in milder climates and avoid hibernation (Krauel and McCracken 2013). In tropical or subtropical areas, migrating bats appear to follow transient opportunities for feeding on fruit and nectar (Krauel and McCracken 2013).

In addition to a range of behavioural changes, the evolution of flight within the Chiroptera led to a number of bat-specific physiological traits, likely related to the high metabolic demands of powered flight. These traits include extreme longevity for their body size, high levels of metabolic activity (Healy et al. 2014), low rates of cancer (Olds et al. 2015), small genome sizes (Kapusta et al. 2017), and an apparently unique immune system (see Section 1.2).

1.2 The unique bat immune system

Bats appear to have an idiosyncratic physiological response to viral infections, and although the details of bat immune systems are only just beginning to be elucidated, specific differences from typical mammal immune systems have been identified.

Investigations focussing on bat immune responses to RNA viruses suggest that bats have much the same machinery for sensing RNA virus infection as do other mammals. Pattern recognition receptors (PRRs) are a conserved feature of animal cells, which detect pathogen associated molecular patterns (PAMPs) derived from viruses and other pathogens. Toll-like receptors (TLRs), an important class of PRRs, have begun to be investigated in bats and appear to perform the same exogenous RNA sensing functions as those found in human cells (Schad and Voigt 2016). However, TLR8 does appears to differ structurally between bat species and appears to be under greater positive selection in bats compared with other mammals (Schad and Voigt 2016; Escalera-Zamudio et al. 2015). Cytosolic RNA sensing PRRs such as retinoic acid-inducible gene-I (RIG-I) and melanoma differentiation-associated gene 5 (MDA5) are well-characterised in

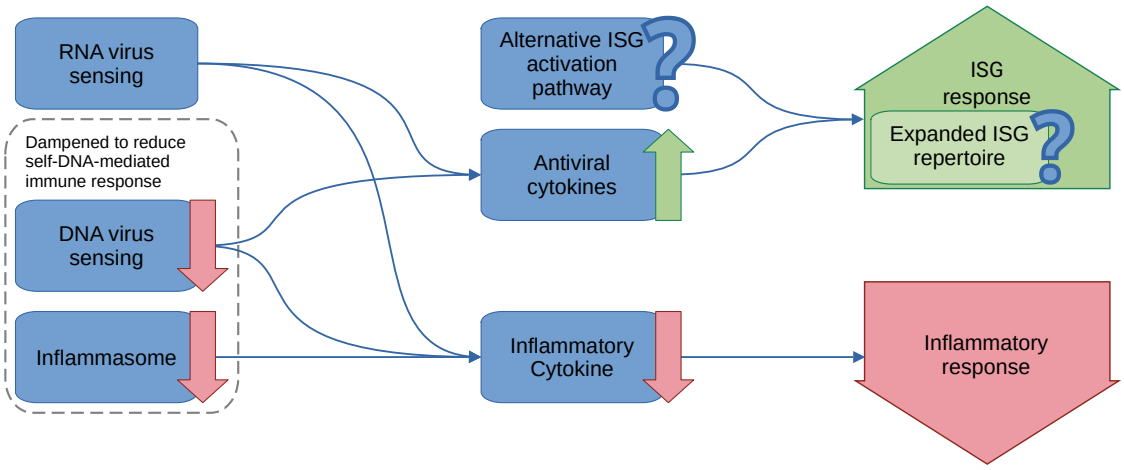


Figure 1.1: Overview of innate immunity pathways in bats showing relative differences in strength of responses and pathway components compared with other mammals. Antiviral cytokine expression and interferon stimulated gene (ISG) response is stronger in bats compared to other mammals. Additionally, indirect evidence suggests a possible bat-specific alternative pathway of ISG activation and expanded ISG repertoire not found in other mammals. On the other hand, DNA sensing, and inflammatory pathways are dampened in bats, leading to a reduced inflammatory response compared with other mammals.

human cells and are also found to be present and with conserved functionality in bat cells (Banerjee et al. 2017).

Although the initial sensing of viral RNA in bat cells appears to be similarly robust to that of human cells, the downstream inflammatory responses appear to be strongly dampened in bats. In *Eptesicus fuscus* cells, expression of the pro-inflammatory cytokine TNFα is inhibited by the protein c-Rel binding to the *TNF* promoter sequences (Banerjee et al. 2017). In parallel, activation of the NLRP3 inflammasome is also dampened in both yin- and yangochiropteran bats, thus reducing the production of another key inflammatory cytokine, IL-1β, in response to viral infection (Ahn et al. 2019). Furthermore, a study using *Myotis myotis* macrophages showed sustained high-level transcription of the anti-inflammatory cytokine IL-10 (Kacprzyk et al. 2017). Together this evidence suggests that bats have adapted to strongly limit their inflammatory response to viral infection, and in this way limit the often lethal pathogenesis associated with strong and chronic inflammation (Banerjee, Baker, et al. 2020).

In comparison to their limited inflammatory responses to RNA virus invasion, current evidence suggests that bat cells exhibit strong and rapidly-induced antiviral cytokine responses. Activation of PRRs leads to the downstream activation of interferon regulatory factor 3 (IRF3) or IRF7, which activate interferons (IFNs), an important group of antiviral cytokines. These in turn activate the expression of IFN-stimulated genes (ISGs) in neighbouring cells via the JAK-STAT signalling pathway. ISGs are a broad group

of genes that inhibit virus replication through different mechanisms (Schoggins et al. 2011; Shaw et al. 2017). Across bats, IRF3 appears to be particularly adapted to trigger an enhanced antiviral cytokine response compared to humans and other mammals (Banerjee, X. Zhang, et al. 2020). Other studies using *P. alecto* cells have also identified broad constitutive expression of IRF7 and downstream constitutive expression of IFNα and associated ISGs (P. Zhou et al. 2014; P. Zhou et al. 2016). However, constitutive expression of IFNα was not observed in *R. aegyptiacus* cells (Pavlovich et al. 2018). Broader ISG transcript expression dynamics have also been studied in *P. alecto* cells, and when treated with IFNα, a universal, rapid induction and subsequent rapid decline in levels of all ISG transcripts that were studied was observed (Cruz-Rivera et al. 2018). Further functional studies showing that IFNα and other type I IFNs (IFNκ and IFNω), as well as type III IFNs (IFNλs) are involved in RNA virus response pathways have also been carried out in *E. fuscus*, *E. serotinus*, *P. vampyrus*, *R. aegyptiacus*, and *P. alecto* cells (Virtue et al. 2011; P. Zhou et al. 2011; He et al. 2014; Pavlovich et al. 2018). These corroborate the importance of the antiviral cytokine based immune response in bats. Furthermore, expansion of ISG families (J.A. Hayward et al. 2018) and atypical ISG expression (compared to human cells) has been observed in *P. alecto* cells (Cruz-Rivera et al. 2018) and *P. vampyrus* cells (Glennon et al. 2015) showing that bats have unique sets of ISGs in addition to those studied in humans. Of particular note is the fact that some of these ISGs required infection by Newcastle disease virus rather than IFNα treatment alone in order for expression to occur. This suggests that bats may have evolved unique sensors and/or additional signalling pathways independent of IFNs to stimulate the expression of ISGs during viral invasion. It is also notable that major variation in the number and subtypes of IFN loci, the different IFN expression patterns, and differences in ISG repertoires between different bat species show that adaptations in antiviral immunity are species-specific and differ significantly across Chiroptera.

Compared to the numerous pieces of evidence pointing to the robust, or even enhanced, RNA virus sensing capabilities of bat cells, our understanding of DNA virus sensing in bat cells is limited. Nevertheless, some studies show key evidence suggesting reduced or dampened DNA virus sensing in bat cells. Firstly, it was discovered that the PYHIN family of immune-related, cytosolic DNA-sensing genes are entirely absent across chiropteran genomes, having likely been deleted independently during the evolution of separate bat lineages (Ahn et al. 2016). Secondly, it has been found that Stimulator of IFN genes (STING), an essential adaptor protein involved in multiple DNA sensing pathways, has a reduced ability to induce IFN expression in bat cells and this allows herpes simplex virus (HSV) to replicate more rapidly in *P. alecto* cells (Xie et al. 2018).

Whether DNA sensing pathways in bats are globally dampened is still speculative, given that bats are host to several DNA viruses that do not cause overt pathogenesis in their hosts (Hayman 2016). However, it is reasonable to speculate that bats have a need for a dampened DNA sensing pathway as a side-effect of the evolution of flight.

During flight, high metabolic rates and elevated body temperatures (sometimes over 41°C; Hock 1951) produce reactive oxygen species that may cause DNA damage and release of DNA into the cytoplasm (Barzilai et al. 2002). Bats show evidence of positive selection in various genes related to DNA repair (G. Zhang et al. 2013). This selection pressure could therefore also be the driving factor for loss of cytosolic DNA sensors such as the PYHIN gene family and reduction of IFN activation ability by STING in order to reduce self-DNA-mediated immunopathology. It is possible that the adaptations observed in TLR9 and autophagy pathways in bats evolved to compensate for this dampened cytosolic DNA response (Escalera-Zamudio et al. 2015; Laing et al. 2019).

Compared with the innate immune system of bats, our understanding of bat adaptive immunity is currently very poor. Although serological studies have shown bats produce antigen-specific antibody responses, the strength and longevity of this response appears to vary significantly between studies (Banerjee, Baker, et al. 2020). Furthermore, the function of antibodies during viral infection in bats is unknown. Studies investigating neutralising antibodies in response to viral infections in bats show conflicting patterns of seroconversion compared with incidence of detectable viral shedding or virus replication within individuals (Banerjee, Baker, et al. 2020). These data suggest that antibodies arising in response to virus infection in bats could control viruses via a mechanism independent of neutralisation. Genomic evidence also points towards an increased importance of the naive immunoglobulin repertoire for response to infections in bats compared to other mammals (Bratsch et al. 2011).

Overall, the emerging picture of bat immunity is one of a diversity of adaptations that have occurred within different bat lineages. However, evidence from a number of species point to a general downregulation of inflammatory responses and a dampened cytosolic DNA sensing pathway. These changes may be compensatory adaptations to mitigate immunopathology caused as a side effect of the physiological stresses of flight. In turn, these changes have repercussions in terms of host-virus interactions in bats.

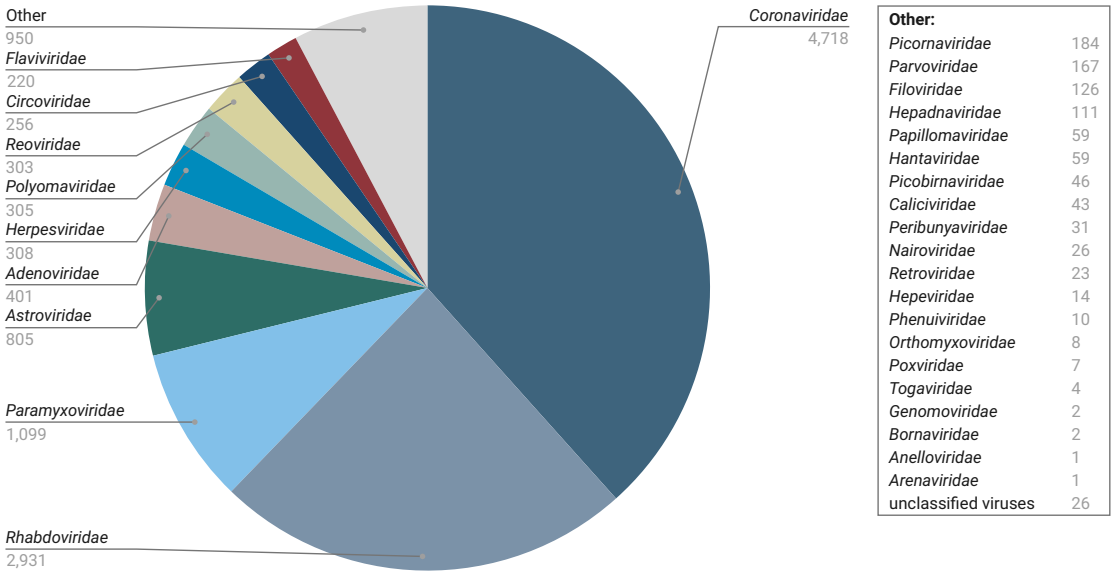


Figure 1.2: Overview of currently described diversity of bat viruses based on numbers of publicly available sequences of bat-associated viruses for each viral family found in bats. *Coronaviridae* and *Rhabdoviridae* comprise well over half of known bat virus diversity, although this may in part be due to sampling bias towards these viral groups of human concern. Data from DBatVir (L. Chen et al. 2014), accessed on 01/12/2020.

1.3 Do bats have a special relationship with viruses?

A particularly high diversity of viruses have been found circulating among bat populations in comparison to other orders of mammals (Luis et al. 2013; Hayman 2016) with over 12,000 bat-associated species of virus discovered and catalogued, representing 30 virus families (Fig. 1.2) (L. Chen et al. 2014). Of these, coronaviruses are widely detected in bats from most regions of the world, and represent a large portion of diversity among bat viruses as well as the largest diversity of coronaviruses among mammals (Wong et al. 2019). Similarly, bat astroviruses are also hugely diverse and can be found in over 30% of individual bats screened (Young and Olival 2016). Other groups of viruses with high diversity found in bats include the *Lyssavirus* genus within the *Rhabdoviridae* family, which contains *Rabies virus*, and the *Paramyxoviridae,* which contains *Hendra virus* and *Nipah virus*. Across this viral diversity, a broad range of pathologies are induced during infections of bats that often have significant effects on bat populations (Mühldorfer et al. 2011).

1.3.1 Inter- and intra-host virus dynamics in bats

Bat physiology and bat immune systems appear to be atypical amongst mammals, indicating the potential for unique dynamics between bats and viruses. However, both

the inter-host dynamics and the underlying within-host dynamics of viral infections in bats is poorly understood, with a number of different hypotheses supported by conflicting evidence (Plowright et al. 2016). Fortunately, due to their underlying nature, better understanding of within-host dynamics would also be predictive of the inter-host dynamics of viral spread at the population level.

One hypothesis posits that bats generally experience acute viral infections followed by long-term immunity (Plowright et al. 2011). If this is the case, then the main drivers of epidemic cycles in bats would be expected to be oscillations of herd immunity levels, population size, and connectivity between populations (Plowright et al. 2016). The short infectious period shown by bats inoculated in captivity with Hendra virus and Nipah virus suggest this dynamic may occur in some cases (Halpin et al. 2011). However, other observations are inconsistent with this dynamic, such as a lack of expected association between flying fox population size and Hendra virus prevalence in the Australian subtropics (Giles et al. 2018; Letko et al. 2020).

An alternative hypothesis states that bats exhibit persistent or recurrent viral infections (Plowright et al. 2016). In this case, the immune competence of individuals would control viral shedding and thus virus transmission and it follows that factors causing stress would likely be drivers of shedding peaks and bursts of viral transmission (Plowright et al. 2016). This hypothesis has been supported by the general assumption, and common paradigm, that reservoir hosts carry persistent infections, as well as the suggestion that bats have immune systems with dampened viral responses (see Section 1.2). However, more recently, specific instances of shedding and inferred transmission events have given support to this idea (Plowright et al. 2016). In one study, seven months after the inoculation of a group of 36 *R. aegyptiacus* bats with Marburg virus, a number of naive individuals became infected, suggesting that the virus persisted within the experimental group for months before being transmitted to the naive bats (Schuh et al. 2017).

1.3.2 Are bats a special reservoir for zoonotic viruses?

Approximately 75% of emerging infectious diseases for humans are zoonoses (L.-F. Wang and Anderson 2019), and over the past few decades, numerous zoonotic viruses have surfaced with serious public health consequences. Of these, bat-borne viruses occupy a greater proportion than viruses from any other mammalian order (L.-F. Wang and Anderson 2019). For example, Marburg virus, Hendra virus, Nipah virus, and Sosuga virus have been directly isolated from bats (Towner et al. 2009; Halpin et al. 2011;

Amman et al. 2015). In addition, there is strong evidence suggesting that while other mammal hosts may have acted as intermediate reservoirs before human infection, Ebola viruses, severe acute respiratory syndrome coronavirus (SARS-CoV), SARS-CoV-2 and Middle East respiratory coronavirus (MERS-CoV), also originated in bats (Olival and Hayman 2014; Latinne et al. 2020; Memish et al. 2013; Leroy et al. 2005).

Based on a dataset of mammal-virus associations which covers >10% of mammal diversity and catalogues >500 viruses (Olival et al. 2017), it appears that, along with rodents and primates, bats account for much of viral sharing over broad phylogenetic distances (Carlson et al. 2019). Statistical modelling based on this dataset using generalised additive models (GAMs), suggests a major factor determining zoonotic risk from a species is the total viral richness present for that species, supporting the idea that the number of zoonotic viruses scales with the total pool of viruses associated with a species (Olival et al. 2017). In addition, this modelling suggests that a significantly higher proportion of bat viruses are zoonotic compared with viruses associated with any other mammalian order. These results add weight to the idea that the unique physiological and ecological traits of bats make them a 'special reservoir' for zoonotic viruses, and leads to elevated numbers of zoonoses originating from this group (Luis et al. 2013). The results of this modelling are also consistent with previously observed dynamics in which the phylogenetic proximity of reservoir host to recipient host increases the likelihood of viral transmission (Cleaveland et al. 2001; Luis et al. 2015; Streicker et al. 2010), and also seem to support human–animal contact – such as that resulting from hunting, increased human population density, and increased urbanisation – as important in defining per-species zoonotic risk.

Interestingly, based on a more stringent dataset cataloguing the associations of >400 viruses with robust evidence of persistence within 11 orders of mammals and birds, a more recent analysis used similar methods as those by Olival et al. (2017) but differed in some of its conclusions (Mollentze and Streicker 2020). Although Mollentze and Streicker (2020) agree with the analysis by Olival et al. (2017) in finding that the number of zoonotic viruses scales with the total pool of viruses associated with a taxonomic group, they did not find Chiroptera (or any other mammalian or avian order) to host a significantly higher proportion of zoonotic viruses than other taxonomic orders, casting doubt on the 'special reservoir' hypothesis. Instead, their findings suggest that the high observed number of zoonotic diseases originating in bats is proportional to the higher total number of viruses harboured by bats, in turn due to the higher number of bat species. This is consistent with the findings of a study using a similar modelling approach but based on a different database of host-virus associations across several mammalian taxa, which found high levels of viral sharing within Chiroptera but low levels

of viral sharing between bats and species from other mammalian orders (Albery et al. 2020). Additionally, the results from Mollentze and Streicker (2020) also suggest that the greater species richness of a host group, such as bats, does not facilitate zoonotic transmission by selecting for a broader host range of the host group-associated viruses.

Although the physiology and ecology of bats may not lead to an enhanced zoonotic capability of their associated viruses, distinct traits, including particular features of bat immunity, are still likely to have important influences on viral communities in bats, with repercussions for recipient hosts of zoonotic viruses. One such trait is the elevated antiviral cytokine-based immune response that even appears to be constitutive in some species of bats (see Section 1.2 for details) (P. Zhou et al. 2014; P. Zhou et al. 2016). Models based on virus infectivity assays in bat cell lines suggest that the heightened innate immune response of bat cells limits pathogen-induced cellular morbidity which, in turn, leads to accelerated within-host propagation rates of bat viruses (Brook et al. 2020). Thus, viruses evolved under this cellular regime may be likely to cause increased virulence in hosts which do not possess the same elevated innate immune responses present in bats.

Altogether, this suggests that the disproportionate impact of zoonotic diseases originating in bats may be an outcome of a higher average virulence of bat zoonoses rather than the generation of a disproportionate number of zoonoses. Regardless of proportionality, the absolute number of zoonoses generated in bats is still second only to rodents, and therefore any general understanding of bat-virus interactions that applies across the Chiroptera applies to a substantial fraction of all zoonotic threats and is of extreme value from a public and veterinary health standpoint.

1.4 Endogenous viral elements (EVEs)

1.4.1 EVEs are found across eukaryotic genomes

Virus-like sequences, or endogenous viral elements (EVEs), have been discovered in the genomes of cellular organisms, showing that genetic material can be transferred from a virus into a host cell genome. When such a transfer of genetic material occurs within a germline cell, the resulting EVE can be passed vertically to the host's offspring and may eventually become fixed within the host population. The vast majority of EVEs are retrovirus-derived EVEs, typically contributing 5-15% of the total DNA content of animal genomes (Smit et al. 2013). These endogenous retroviruses (ERVs) are formed

during a retrovirus infection, where the virus integrates a DNA copy of its genome into the host cell chromosome as part of its replication cycle. ERVs can also multiply within a genome through the efficient intracellular mechanism of retrotransposition (Stocking and Kozak 2008), leading to extraordinarily high copy numbers (Magiorkinis et al. 2012). ERVs are therefore also classified under the umbrella of autonomous transposable elements (TEs).

In comparison to the ubiquity of ERVs within host genomes, non-retrovirus-derived EVEs are relatively rare, typically numbering fewer than twenty or thirty loci within a genome (Katzourakis and Gifford 2010; Aiewsakun and Katzourakis 2015). Whereas every retroviral infection has the potential to create a vertically transmitted ERV due to the nature of retrovirus replication, and further duplications are then possible through retrotransposition, integrations of non-retrovirus genetic sequences are low probability events. One mechanism for the generation of non-retroviral EVEs involves mRNA generated by the virus being captured by the retrotranscribing machinery of retroelements in the cell, such as LINEs and ERVs. Reverse transcription generates DNA from this RNA which may then be concatenated with the fragments of the retroelement and subsequently incorporated into the host cell genome by mechanisms of retroelement integration (Belyi et al. 2010b). A second mechanism that can occur in the case of DNA virus integration is direct homologous recombination of the viral genomic DNA and the host cell chromosomal DNA (Belyi et al. 2010a). Ultimately, over the course of millions of years, these low probability events have led to EVEs representing every Baltimore Classification group to be represented within host eukaryotic genomes (Aiewsakun and Katzourakis 2015). Within animals specifically, EVEs have been found that represent *Bornaviridae*, *Bunyaviridae*, *Filoviridae*, *Orthomyxoviridae*, *Reoviridae*, and *Rhabdoviridae* families (Aiewsakun and Katzourakis 2015).

1.4.2 EVEs are a viral fossil record

Studies which have sequenced samples of viruses across multiple timepoints have shown that many exogenous viruses evolve at a rate of around 10^{-3} substitutions per site per year (s/n/y) (Sanjuán 2012). Although these studies examined closely related viruses, it is also commonplace in the literature to characterise diversity and evolutionary history of viruses by aligning viral genes or genomes sampled from across viral genera or families and building phylogenies from these alignments. The high evolutionary rate of viruses would therefore suggest that phylogenetic trees, and the alignments on which they are based, are only likely to represent recently diverged

EVE dating method	Description
Orthologue dating	If orthologues of an EVE can be found in multiple host species, the divergence date of the host species can be considered a minimum age for the EVE.
Paralogue dating	In the case of two or more paralogous copies of an EVE, an alignment can be used to count the number of mutations between each copy. Assuming that the sequences have experienced a neutral selection regime, the time since duplication can be inferred using the count of mutations and the host neutral mutation rate. This provides a minimum age for the EVE.
LTR dating	ERVs are flanked on both sides by LTR regions which can be used to date the ERV integration date. During retroviral integration, the LTRs are identical in sequence. As with paralogue dating, the number of genetic differences between the two sequences is directly proportional to their age (assuming neutral selection) and thus the date of the ERV insertion.
Stop codon dating	The number of premature stop codons that have accrued within an EVE sequence should be proportional to the age of the EVE since it can be assumed that at integration time no premature stop codons were present in the sequence. Thus a count of accumulated stop codons within an EVE sequence can be used to estimate the age of the EVE. In practice, the number of stop codons can be hard to estimate due to the presence of frameshift mutations.

Table 1.1: Overview of EVE dating methods.

viruses, with a high potential for spurious phylogenetic relationships at deeper levels. Older homology and sequence relationships would be lost behind a phylogenetic horizon, beyond which an excessive number of substitutions would render sequences unalignable or saturation of substitutions would make it impossible to determine the true phylogenetic relationships between sequences. Interestingly, viral clade groupings produced by such phylogenies *do* appear to mirror phenotypic similarities well (Nishimura et al. 2017; M. Shi et al. 2018; Amarasinghe et al. 2019), therefore suggesting that such phylogenies are reasonably accurate. Interestingly, EVEs provide evidence showing that these deep phylogenetic relationships are not only accurate, but are rooted millions of years in the past, indicating that the previously calculated evolutionary rates may be inaccurate over long time scales.

EVE sequences are, in most cases, of neutral fitness value and therefore mutations occur at the neutral substitution rate of the host. Since cellular organisms have neutral mutation rates of around 10^{-9}s/n/y (Kumar and Subramanian 2002), EVE sequences mutate several orders of magnitude more slowly over time compared to their exogenous counterparts, and thus provide relatively unaltered snapshot of a virus genome at the time of insertion millions of years ago. It is this property that allows EVEs to be considered viral 'fossils' allowing comparison of modern exogenous viral sequences with sequences closely related to ancestral exogenous viral sequences from millions of years ago. Specifically, EVEs can be dated through a number of methods (see Table 1.1) and therefore can be used to time-calibrate virus phylogenies in much the same way that geological fossils – which provide time information based on geological dating – can

be used to time-calibrate animal and plant phylogenies. By including dated EVEs in virus phylogenies, it has been established that viral clades are generally in the order of millions of years old rather than estimates based on short-term measured evolutionary rates which may predict ages in the order of tens of thousands of years old (M. Shi et al. 2018). One such example is the discovery of orthologous hepadnavirus EVEs in turtle genomes, which dates the minimum age of the *Hepadnaviridae* to the Triassic period >207 Mya (Suh et al. 2014).

Strikingly, when EVE sequences of a known age are compared with modern exogenous viruses of the same family, many fewer substitutions are present than would be expected with an evolutionary rate of 10^{-3}s/n/y, and in fact, this comparison suggests an evolutionary rate closer to the cellular evolutionary rate of 10^{-9}s/n/y (Aiewsakun and Katzourakis 2016; Simmonds et al. 2019a). This fits with other findings that the determined evolutionary rate of a virus is dependent on the time-scale over which it is measured; the longer the timescale, the slower the evolutionary rate (Aiewsakun and Katzourakis 2016).

To explain this, Simmonds et al. (2019a) posit a niche-filling model in which viruses evolve to fit a niche which is defined by their host (but see Holmes and Duchêne 2019, and Simmonds et al. 2019b). Being dependent on the host biology, this niche changes according to the host's evolution and thus at the same rate as the host's evolutionary rate. Therefore, under this model, the long-term evolutionary rate of the virus is determined by changes in the host-determined niche and short-term evolutionary rate is reflective of evolutionary local exploration of that niche. However, although highly conserved viral genes (such as negative-strand RNA viral polymerases) fit this pattern, and are conserved enough to allow alignment across entire orders or families of viruses, other genes are family-specific and their origins cannot be traced back as far despite their apparent slow long-term evolutionary rate. Notably, some of these family-specific genes are structurally and functionally homologous with genes specific to other families. For example, the VP40 gene present in filoviruses, and the M gene in distantly related bornaviruses, share no significant sequence homology but appear to serve similar functions in immune suppression and genome replication, and also show striking structural homology (Neumann et al. 2009). This might suggest a common origin of these genes, followed by rapid divergence at a rate much higher than the rate of change of the host-defined niche, before each of the two divergent genes settled into slower long-term evolutionary rates. Under this scenario, it would seem that the niche in which the viruses reside sometimes *does* allow viruses to harness their ability to rapidly adapt for long stretches of time rather than just in the short-term.

An alternate explanation for functionally similar but genetically distant genes, is that such genes could be a case of a long-distance journey through genotype-space which corresponds to a looping journey through phenotype-space in which phenotypic start and endpoints are close together. This would therefore still be a case of local exploration of the host-defined niche and be consistent with the niche-filling model proposed by Simmonds et al. (2019a). However owing to a lack of a framework to measure viral evolutionary rate where sequences are unalignable, it is not possible to quantitatively characterise this type of evolutionary trajectory and thus only qualitative speculations can be made.

Beyond revealing this curious phenomenon, EVEs have also helped to expand our knowledge of host ranges for families of viruses (Aiewsakun and Katzourakis 2015). Since EVEs are a record of a past viral infection, detecting an EVE derived from a particular clade of viruses shows that the host species has a previous (and likely ongoing) relationship with that group of viruses. For example, the identification of endogenous filovirus elements in South American and Australian marsupial genomes that are phylogenetically basal to extant filoviruses suggests that extant filoviruses (which circulate amongst mammals in the Old World) may have originated in the New World (D.J. Taylor et al. 2010). This result, points to a broader geographical and phylogenetic range of filoviruses than was previously deduced. This approach is therefore particularly useful, because although host ranges of extant viruses have been extensively surveyed amongst humans, livestock, and crop species, the host ranges of viruses beyond these limited groups is poorly characterised. Using such data to determine broader patterns of virus host ranges is therefore subject to the heavy information bias of agriculturally and medically important viruses. EVE based surveys not only reveal new virus-host relationships where no exogenous virus has been previously associated with a host (M. Shi et al. 2018), but also produce less biased virus-host association data because the diversity of genome sequences available for analysis is much broader than exogenous virus surveillance data (Aiewsakun and Katzourakis 2015). Despite this, the rarity of EVE endogenisation events make EVEs an extremely sparse record of viral infections within a host, and, furthermore, there are stark differences in the likelihood of endogenisation for different viral clades (Kryukov et al. 2018).

In addition to simply revealing broader host ranges, EVEs can help to elucidate the long-term evolutionary patterns of viral lineages, helping to characterise the propensity of a virus to undergo host-switching, as well as identifying ancestral host populations. For

example, analysis of a specific lineage of gammaretroviruses found throughout mammals showed wide dissemination between ~33 and ~15 Mya which involved frequent jumps between mammal species and continents as well as viral recombination (Diehl et al. 2016). In contrast, ERVs have revealed the phylogeny of foamy viruses to be remarkably congruent with that of their eutherian mammal hosts, pointing to an extremely stable co-speciation dynamic occurring over the past 100 million years (Switzer et al. 2005; Katzourakis et al. 2014).

Despite the rarity of endogenisation events, EVEs are an extremely valuable resource for the understanding of virus evolution and host-range. With the onset of large sequencing consortia aiming to produce genome sequences for thousands of species spanning broad phylogenetic ranges (Koepfli et al. 2015; Lewin et al. 2018), a wealth of EVE data will likely become available in the near future. As a result, EVEs will become an ever more valuable means of filling in gaps in our knowledge of virus-host associations, and will ultimately lead to a better understanding of long-term viral evolutionary dynamics, with associated potential implications for the development of viral control and prevention strategies.

1.5 EVEs are a source of genomic novelty

Most EVEs are non-functional, either being the product of a fragmented piece of viral genetic material or being strictly silenced and undergoing pseudogenisation (Feschotte and Gilbert 2012). However in many cases, ERVs and non-retroviral EVEs can play a beneficial role in broadly shaping host genome biology, or can contribute useful coding or non-coding sequence that is co-opted by the host.

1.5.1 ERVs can enhance genomic plasticity

The ability of active ERVs to duplicate through retrotransposition leads to high copy numbers within genomes and brings these elements under the umbrella of transposable elements (TEs) or interspersed repetitive elements. Interspersed elements are known to facilitate large-scale genomic rearrangements, duplications, and deletions through nonallelic homologous recombination (NAHR) events (Konkel and Batzer 2010). Highly similar transposable element copies within a genome (such as a multiple near-identical copies of an ERV) provide an opportunity for the pairing of different regions of sister chromatids during homologous recombination, which in

turn results in translocation, duplication, or loss of genomic segments (Konkel and Batzer 2010). An increased number of transposable elements, such as ERVs, therefore leads to what is often termed genomic instability (regarded as deleterious), but could also be considered as increased genomic plasticity. The effects of increased genomic plasticity can prove to be useful substrates for evolution, enabling the duplication of large numbers of genes (in turn enabling subfunctionalisation), or the removal of expendable regions of the genome. The latter case may be particularly important for species where genome size is constrained by high metabolic demands, such as bats and birds (Kapusta et al. 2017). In such circumstances, evidence points to short genome size being maintained by an 'accordion model' of genome evolution, wherein an equilibrium exists between genome growth and TE-mediated NAHR events that result in the beneficial loss of DNA (Kapusta et al. 2017).

1.5.2 EVEs can be co-opted as *cis*-regulatory elements

EVEs, especially ERVs, are rich in *cis*-regulatory sequences which originally functioned to recruit cellular transcriptional machinery for replication of the original virus. Once integrated into the host genome, these sequences have the potential to control the activation of adjacent cellular genes, and therefore have the potential to be co-opted as transcription factor binding sites, enhancers, or promoters (Chuong et al. 2017). For example, transcription factor binding sites present in the long terminal repeat (LTR) regions of human ERVs (HERVs) control the transcription of a substantial number of target genes for human tumour suppressor protein p53 (T. Wang et al. 2007). Additionally, enhancers present in lineage-specific ERVs have independently contributed to the transcriptional networks underlying interferon (IFN) response across diverse mammalian genomes (Chuong et al. 2016), while HERVs appear to function as active promoters in the regulation networks involved in pluripotency control in human cells (J. Wang et al. 2014).

1.5.3 EVE co-option for generation of non-coding transcripts

When EVEs, ERVs, and other TEs insert into the host cell genome, it can result in the creation of new exonic sequence where previously none existed, a process termed 'exonisation'. Exonisation is the result of the splice site-like structures in the element creating new splice patterns when an element insertion occurs in the intronic region of an existing gene, which leads to the creation of a novel exon or exons being incorporated

into the mature transcript (Sela et al. 2010; Schmitz and Brosius 2011). Such a transcript may now include some of the sequence from the EVE or TE itself and thus the result is a chimeric transcript. However, due to the potential for insertions to disrupt reading frames, selection has heavily biased their distribution towards occurring in non-coding transcripts or, in rare cases, in the UTR regions of protein coding transcripts (Kapusta et al. 2013). In addition, the transcription start sites present in TEs or viral elements are frequently generators of de-novo long non-coding RNAs (lncRNAs), in some cases generating transcripts which consist mostly of the element sequence and in other cases generating a transcript derived from adjacent intergenic or intronic DNA (Kapusta et al. 2013).

These two mechanisms together likely account for the fact that TE sequence fragments – including ERVs – were found to be present in ~75%, ~68%, and ~66% of lncRNAs in humans, mouse, and zebrafish genomes, respectively, suggesting that this pattern may hold across all vertebrates (Kapusta et al. 2013). Interestingly, within humans, ERVs appear to be overrepresented compared to other TE families, suggesting HERVs may be particularly adapted to recruiting transcriptional machinery (Kelley and Rinn 2012; Kapusta et al. 2013). The functionality of lncRNA repertoires is unfortunately not nearly as well characterised as that of coding genes, owing to the volatility of their evolution (Ponjavic et al. 2007; Kapusta et al. 2013; Johnsson et al. 2014; Necsulea et al. 2014; Craig et al. 2018); however, some specific examples of ERV-derived lncRNAs have been characterised. One example is lncRNA_ES3, generated by a transcription start site and splice structure of a HERV locus, found to be a *trans*-acting regulator of pluripotency in human embryonic stem cells (Ng et al. 2012). A second example, again found in human cells, is lnc-EPAV, a *trans*-acting positive regulator of virus induced cytokine response in human cells (B. Zhou et al. 2019).

A further type of non-coding RNAs that appears to be generated by ERVs is microRNA (miRNA), which appears to be ubiquitously involved in regulation of a broad diversity of biological processes (Roberts et al. 2014). Beyond miRNAs and lncRNAs – which most commonly appear to play a regulatory role – EVEs have also been found to contribute to another form of non-coding RNA, called Piwi-interacting RNA (piRNA). piRNA clusters are repetitive sequence clusters in the genome that are transcribed to form piRNA precursor molecules, which are then processed to form piRNAs. piRNAs form a complex with Piwi protein which then acts as an RNAi silencing mechanism for RNA molecules complementary to the piRNA (Ophinni et al. 2019). Interestingly, non-retroviral EVEs have been found to be enriched within piRNA clusters in some mammal species, which has been suggested as evidence that this system acts to capture and integrate viral genetic sequence during infections and subsequently result in piRNA-guided RNAi, in turn

mediating adaptive immunity (Ophinni et al. 2019). In particular, bornavirus-derived EVEs have been found to be independently enriched within piRNA clusters in both rodents and primates, and give rise to piRNAs that are antisense to the bornaviral nucleoprotein mRNA (Parrish et al. 2015).

1.5.4 EVE co-option for generation of coding transcripts

EVEs are derived either from viral mRNAs coding for viral proteins or entire viral genomes that are extremely ORF dense. Thus it is no surprise that this coding potential within EVEs has also been targeted by selection, resulting in viral protein coding genes being expressed for the benefit of the host. The most striking example of this is the well studied syncytins, derived from retroviral env genes (Mi et al. 2000). Syncytins are glycoproteins present in the species of a number of mammalian orders, as well as in some viviparous lizard species, and carry out cell-cell fusion functions during placental development (Mi et al. 2000; Cornelis et al. 2017). Fascinatingly, syncytins are thought to be the result of 17 independent co-option events throughout the evolution of vertebrates, with each case derived from a different lineage of retrovirus (Broecker and Moelling 2019). Due to their ubiquity in vertebrate genomes, retroviruses have thus provided natural selection many opportunities to co-opt retroviral coding genes. Correspondingly, ERV-derived coding transcripts have regularly been discovered in sequenced genomes (Frank and Feschotte 2017; Broecker and Moelling 2019). In fact, a recent systematic phylogenomic study searched a broad sampling of vertebrate genomes and identified 177 putatively co-opted retroviral *gag*- and *env*-derived genes across these genomes (J. Wang and Han 2020). Although non-retroviral EVEs are much rarer than ERVs, bornavirus-, filovirus- and parvovirus-derived genes with conserved coding regions and evidence of expression have been identified in vertebrate genomes and often appear to possess immune-related functions (Fujino et al. 2014; Kondoh et al. 2017; Lau et al. 2017).

1.5.5 EVE-derived immunity genes

Examples of each of the mechanisms of direct EVE co-option mentioned above are overwhelmingly found to result in immune related functionality in the host, whether the co-option results in new regulatory sequences, generates new non-coding RNAs, or produces functional proteins. Co-option of EVEs resulting in immune functions is a phenomenon termed EVE-derived immunity (EDI) (Aswad and Katzourakis 2012;

Broecker and Moelling 2019). This apparent predisposition is likely a result of the fact that viruses are already adapted to exploiting and interfacing with host immune responses, thus, viral genomes are rich in features that are preadapted and can readily 'plug in' to host immune pathways. This is well demonstrated by the broad finding that the promoter/enhancer sequences in ERVs are a major component of the interferon response transcriptional network in humans (Chuong et al. 2016; M. Wang et al. 2020). Furthermore, broad differential expression of ERV loci has been observed in response to infections by numerous virus species including HIV, Epstein-Barr virus (EBV), hepatitis B virus (HBV), human T-lymphotropic virus (HTLV), and Dengue virus (DENV) (Contreras-Galindo et al. 2007; Hsiao et al. 2009; Toufaily et al. 2011; M. Wang et al. 2020), which also tallies with findings that lncRNAs (a large proportion of which are ERV-derived) are involved with human interferon signalling pathways (Peng et al. 2010).

The nucleotide homology of EVEs to their exogenous counterparts is also something that has been exploited by hosts (Aswad and Katzourakis 2012; Roberts et al. 2014). An example of this is EVEs included in piRNA clusters which potentially aid host immunity by recognising infecting viral RNA through complementarity as potentially seen with some bornavirus-derived EVEs in mammals (Parrish et al. 2015).

Finally, at the protein coding level, a number of examples of EVE co-option for immune function have been well characterised (Broecker and Moelling 2019). One form of immune-related role taken up by EVEs is the direct conferral of innate immunity against infecting viruses by blocking entry of exogenous viruses into cells, an example of which is the retroviral in origin *Fv4* gene found in mice (G.M. Taylor et al. 2001). When *Fv4 is* expressed, the translated Env protein competes with exogenous murine leukemia virus (MuLV) in binding to the cell surface receptor used for viral entry and thus inhibits cell entry, thereby conferring immunity against this virus (G.M. Taylor et al. 2001). An expressed counterfeit viral protein encoded by an EVE can also hamper the intracellular stages of viral life cycle. For example, another ERV-derived gene in mice, *Fv1,* produces a Gag-like protein that binds to, and blocks, retroviral capsid cores shortly after cell entry (Best et al. 1996; Boso et al. 2018). Interestingly, *Fv1* and its homologues in other rodents, appear to confer immunity to a range of retroviruses unrelated to MuLV (to which it is most closely related), suggesting that *Fv1* acts upon conserved structural patterns rather than specific amino acid motifs (Yap et al. 2014). Another example of a protein coding EDI gene is *itEBLN,* a protein-coding bornavirus nucleocapsid gene co-opted in the ground squirrel which appears to aid immunity against bornaviruses through binding to the bornaviral ribonucleoprotein complex and causing suppression of both inter- and intracellular stages of the viral life cycle (Geib et al. 2003; Fujino et al. 2014;

Kim et al. 2020). One more mode of immune related function that can be performed by a co-opted coding viral gene is that of immune regulation. A potential example of this is a filovirus VP35 gene co-opted in *Myotis* bats which has retained its immune suppressing function that evolved in the exogenous virus and appears to have been repurposed by the bats as a regulator of interferon (IFN) signalling (Edwards et al. 2018).

In summary, EVEs can represent an important source of genetic novelty for their hosts. Indeed, EVEs have provided the opportunity for hosts to exploit every aspect of viral genomes during their co-option, and most frequently have done so for immune-related function.

1.6 Outline of work

In this project I carry out the first large-scale comparative survey of EVEs found in the growing number of sequenced bat genomes and attempt to glean a maximal amount of information about the viruses that are represented by them. By developing a new method for *in silico* EVE detection (Chapter 2), I attempt to thoroughly and systematically characterise the EVEs present in bat genomes that are as yet unexplored in their EVE content, and improve upon the previous EVE-mining efforts in other bat genomes. Based on this, I am able to examine the phylogenetic distribution of ERVs across bats and extract information about the rates of cross-species retrovirus transmission between bat lineages across an evolutionary timescale. This allows me to compare the rates of cross-species transmission occurring within different bat clades and to test the effects of phylogenetic relatedness (Chapter 3). I am also able to identify new information about the evolutionary history and long-term association with bats, of filoviruses, parvoviruses, and the recently discovered *Cultervirus* and *Carbovirus* genera of bornaviruses (Chapter 4). Additionally, I provide evidence that the *Tetraparvovirus* genus is more ancient than previously understood (Chapter 4). Finally, with the aid of transcriptomic data, I identify putative EVE co-option events in bats, and examine their potential relationship to the unique immune systems found in bats (Chapters 4 and 5).

2

Improving upon current EVE mining methods

Contents

Abstract **41**
2.1 Introduction **41**
2.2 Methods **44**
2.2.1 Selection of genome assemblies 44
2.2.2 Restricting search regions to putative ERV regions 44
2.2.3 RT locus identification using four search methods 46
2.2.4 Sensitivity analysis 47
2.2.5 Predicted sequence completeness analysis 47
2.3 Results **48**
2.3.1 Differences in number and length of identified RT sequence between search methods 52
2.3.2 Differences in detected RT domain abundance between genome assemblies 53
2.3.3 DVORFS: An easy-to-use wrapper and post-processor for EVE mining with GeneWise 55
2.4 Discussion **56**
2.4.1 Frameshift-aware protein pHMM to DNA alignment search improves EVE detection and sequence recovery 56
2.4.2 Genome assembly strategies have a large impact on assembled TE content 58
2.4.3 The need for a high sensitivity search tool for EVE identification 60

Abstract

Endogenous viral elements (EVEs) can be found integrated within the genomes of nearly all eukaryotic organisms, having resulted from the insertion of genetic material from a virus into the genome of the host organism. The majority of EVEs occur as pseudogenes and thus experience the slow neutral mutation rates of eukaryotic organisms. EVEs therefore provide a snapshot of the viruses infecting hosts millions of years in the past and thus provide an invaluable resource for the study of the evolutionary history of viruses. However, identifying EVEs and recovering their genetic sequence is bioinformatically challenging due to the frequent presence of multiple frameshifting mutations and their often distant homology to known viral sequences. In this study I benchmark four *in silico* approaches for the detection of EVEs and compare their performance in the detection and recovery of endogenous retroviral reverse transcriptase sequences in a panel of ten bat genome assemblies. I show that a search strategy based on frameshift-aware alignment of protein profile-hidden markov models is well suited for EVE mining due to its superior recovery of fragmented viral sequence. Based on this, I develop an easy-to-use tool – Disrupted Viral ORF Search (DVORFS) – that implements this novel EVE mining strategy, improving upon currently used methods. The benchmark results also highlight the major impact of genome assembly strategy on the transposable element (TE) content present in genome assemblies. This suggests that future studies comparing the TE content across different genome assemblies must take into account the assembly strategies of the genomes being compared for a valid comparison to be made.

2.1 Introduction

A broad diversity of virus-derived sequences, termed endogenous viral elements (EVEs), can be found scattered across eukaryotic genomes, with examples having been discovered that together represent every class of virus in the Baltimore classification system (Aiewsakun and Katzourakis 2015). EVEs are a product of the process of endogenisation, in which virus-derived DNA integrates into host germline chromosomes, leading to vertical transmission from parent to offspring, and, in some cases, eventual fixation in a population (Katzourakis and Gifford 2010). Endogenous retroviruses (ERVs) are often prolific within the genomes of their vertebrate hosts because they enter and proliferate within host chromosomes as an obligate step during their life cycle (Stocking and Kozak 2008). In vertebrate genomes, the vast majority of EVEs are retrovirus-derived, and typically around 5-15% of the genome consists of retrovirus derived sequence (Smit et al. 2013). In comparison, non-retroviral EVEs are rarely found in numbers that exceed twenty or thirty loci within a given genome because their integration occurs accidentally and at a low rate (Katzourakis and Gifford 2010).

Most retroviral and non-retroviral EVEs occur as inactive pseudogenes. Inactivation can arise from becoming fragmented during the process of endogenisation, subsequent disruption of the reading frame through the accrual of mutations, or being silenced through methylation based cellular machinery (Groh and Schotta 2017). As a result of

deactivation, most EVEs mutate under a neutral selection regime at the neutral substitution rate of their host (~10^{-9}s/n/y; Kumar and Subramanian 2002), which is around three orders of magnitude lower than the substitution rate of the exogenous virus from which the EVE is derived (~10^{-3}s/n/y; Sanjuán 2012). Thus, EVEs represent 'fossilised' viral sequences with high identity to the ancient viruses that were endogenised. The study of EVEs, termed paleovirology, is therefore a means to elicit information about the ancient biology (M. Chen and Cui 2019), host distribution (Metegnier et al. 2015; Pénzes et al. 2019), and evolution of viral groups (Aiewsakun and Katzourakis 2015; Diehl et al. 2016; M. Shi et al. 2018) using these 'fossilised' sequences.

Despite their abundance, conducting a thorough census of EVEs within a genome assembly is bioinformatically challenging. Firstly, being derived from viruses, the sequence diversity among EVEs will reflect the vast diversity present within viral groups. Unfortunately, due to a paucity of sampling, viral sequence databases fail to capture the vast sequence diversity present even in extant exogenous viruses, and therefore represent an even smaller proportion of the diversity in extinct lineages of viruses that could be found as EVEs (Aiewsakun and Simmonds 2018). Thus, when using a homology-based search method, successful detection of an EVE will frequently require the detection of a sequence with remote homology from query sequences in a reference database. Although detection of ERVs is commonly accomplished using RepeatMasker (Smit et al. 2013) – which relies on nucleotide against nucleotide search – most approaches for identifying non-retroviral EVEs have utilised translated BLAST searches with viral protein sequences as queries against a host genome (Camacho et al. 2009; Katzourakis and Gifford 2010; Kondo et al. 2015; Metegnier et al. 2015; Skirmuntt and Katzourakis 2019). Yet because coding sequences are more conserved at the amino acid level than the nucleotide level, the sensitivity of a translated search is higher than a nucleotide against nucleotide search because it enables the identification of EVEs within a larger genetic distance of a reference viral sequence database.

Sensitivity to remote homologues may be increased further by using profile-based search methods, such as position-specific scoring matrix (PSSM) search in PSI-BLAST (Altschul et al. 1997) or profile-hidden markov model (profile-HMMs or pHMMs) search implemented in HMMER (Eddy 2009). A profile-HMM is generated from a multiple sequence alignment (MSA) of a representative set of homologous sequences to produce a probabilistic model that describes the diversity and conservation present within the MSA and, therefore, the broader sequence family. The extra information captured in this model compared to individual sequences therefore provides higher sensitivity when searching for remote homologues (Park et al. 1998) and thus makes

pHMM search an especially useful tool for detecting and aligning viral sequences. For example, HMMER has been used with protein pHMM queries to improve the sensitivity of: detecting virus coding sequences in metagenomic reads (Skewes-Cox et al. 2014); detecting EVEs with ORFs greater than 80 amino acids in length (Nakagawa and Takahashi 2016); and identification of previously undetected tobamovirus-derived genes in fly genomes (Kirsip and Abroi 2019).

Although previous studies have used translated BLAST searches to identify pseudogenes (Harrison et al. 2002; Lerat and Ochman 2004), methods that rely on protein sequence queries to identify pseudogenes lack sensitivity for loci with numerous frameshifts. This is because most such tools rely on searching each translated frame of a nucleotide sequence separately, which causes acute problems when numerous frameshifts are present at a locus. For regions of pseudogenised coding sequence that contain multiple frameshifts, homology within each particular translation frame will only be present in very short interspersed stretches. For both HMMER and translated BLAST, each sub-alignment or 'high scoring pair' (HSP) can only occur within a single frame, and proximal and consecutive HSPs in different frames are joined only after they have been identified, meaning that each HSP must individually meet a threshold of detection. In the case of detecting EVEs, where numerous frameshifts and insertions frequently occur, both methods lack sensitivity because the homology at an EVE locus is often too 'spread out' across the different reading frames for the EVE to be detected

Maximising the sensitivity of a search method for identifying EVEs thus requires an approach that combines searching a nucleotide sequence in amino acid-space using a protein pHMM query while simultaneously accounting for frameshifts and insertions. The only currently available tool that fulfils these requirements is GeneWise from the *Wise2* package (Birney et al. 2004). GeneWise implements a number of complex, hidden markov model-based, alignment models which can account for frameshift mutations in coding sequences and should therefore provide improved sensitivity for detecting EVEs.

EVE mining using GeneWise often requires preprocessing of genome assemblies to reduce excessive computation times, conversion of modern pHMM formats to be compatible with GeneWise, specific parameters to enable effective detection of EVEs, and post processing of hits to create more accurate EVE annotations. Here I describe the development of a new tool, named DVORFS (Disrupted Viral ORF Search), that simplifies these preprocessing steps and integrates a post-processing method in conjunction with the GeneWise 3:33L alignment model for effective detection of EVEs in genomic data. In this study I show the benefits of using this method for the detection of EVEs by comparing the results of a search for endogenous retrovirus reverse transcriptase (RT) domains

in 10 bat genome assemblies using GeneWise, tBLASTn (Camacho et al. 2009), HMMER, and RetroTector — a tool specialised for detection of ERV coding domains (Sperber et al. 2007). Additionally, I use the results to show the significant impact which different genome assembly strategies (such as the use of different sequencing technologies or genome assembly software) have on the resulting ERV content of a genome assembly.

2.2 Methods

2.2.1 Selection of genome assemblies

To assess EVE search performance of a number of EVE mining methods across varying types of genome assembly, a panel of ten genome assemblies (listed in Table 2.1) were selected from across the order Chiroptera, representing three families of bats (Rhinolophidae, Pteropodidae, Phyllostomidae) that span both the Yinpterichiroptera and Yangochiroptera suborders. To allow more direct comparison of assembly methods, genome assemblies were chosen from the same or closely related species within the three bat families. Genome assemblies were further selected to represent a diversity of assembly strategies to include assemblies produced with a range of short-read coverage depths, different short read assembly software, and assemblies produced using long reads. Genome assemblies within each bat family were selected to represent at least three of four categories of genome assembly strategy: short read, low coverage, assembled by SOAPdenovo (R. Li et al. 2010); short read, high coverage, assembled by SOAPdenovo; short-read, medium coverage, assembled by DISCOVAR *de novo* (BROAD Institute 2015); and assemblies that variously utilised long reads.

2.2.2 Restricting search regions to putative ERV regions

In order to reduce spurious hits to non-retrovirus-derived RT domains, genome assemblies were first screened using RepeatMasker loaded with a previously constructed custom repeat library to identify putative ERV regions within which tBLASTn, HMMER, RetroTector, and GeneWise would be used to search for RT domains. The custom repeat library was created with the aim of covering as much bat ERV diversity as possible. First, ERV sequences were extracted from genome assemblies from 15 bat species (Appendix A.2.1) using RetroTector (version 1.0.1, using the `SweepDNA` then `SweepScripts` commands with default parameters and default motif library). LTR regions for all ERV sequences discovered were then removed to leave only the internal

Species (assembly name)	Accession	Contig N50 (bp)	Sequencing Technology	Cov.	Assembly Method
Rhinolophidae:					
R. ferrumequinum (ASM46549v1)	GCA_000465495.1	11,659	Illumina HiSeq	17x	CLC NGS Cell; SOAPdenovo
R. ferrumequinum (RhiFer_v1_BIUU)	GCA_007922735.1	127,195	Illumina HiSeq	40x	DISCOVAR de novo
R. ferrumequinum (mRhiFer1_v1.p)	GCA_004115265.1	31,900,497	PacBio Sequel; 10X Genomics linked reads; Bionano Genomics 2 enzyme; Phase Genomics HiC	53x	DAmar; Arrow; FreeBayes; Longranger; Bionano Solve; Salsa2
Pteropodidae:					
E. helvum (ASM46528v1)	GCA_000465285.1	12,668	Illumina HiSeq	18x	CLC NGS Cell; SOAPdenovo
P. alecto (ASM32557v1)	GCA_000325575.1	31,841	Illumina HiSeq	110x	SOAPdenovo
R. aegyptiacus (RouAeg_v1_BIUU)	GCA_004024865.1	100,260	Illumina HiSeq	36x	DISCOVAR de novo
E. spelaea (Espe.v1)	GCA_003508835.1	8,002,591	PacBio Sequel	80x	Falcon
Phyllostomidae:					
D. rotundus (ASM294091v2)	GCA_002940915.2	80,250	Illumina Hiseq	94x	SOAPdenovo; HiRise
T. saurophila (TonSau_v1_BIUU)	GCA_004024845.1	141,649	Illumina HiSeq	46x	DISCOVAR de novo
P. discolor (mPhyDis1_v1.p)	GCA_004126475.1	6,892,556	PacBio Sequel; 10X Genomics linked reads; Bionano Genomics 2 enzyme; Arima Genomics HiC	66x	DAmar; Arrow; FreeBayes; Longranger; Bionano Solve; Salsa2

Table 2.1: Genome assemblies used in this study to benchmark EVE search methods.

regions. Nested LINE and SINE elements within these ERV sequences were identified using RepeatMasker (version 4.0.7, parameters: `-species mammalia -e ncbi`) loaded with RepBase repeat library release 20170127 (Bao et al. 2015) and removed using a custom script. Next, redundant sequences from this ERV library were removed based on the '80-80' rule, which considers two repeat sequences to be of the same family if more than 80% of their length can be aligned with over 80% identity (Wicker et al. 2007). This was carried out using CD-HIT (version 4.7, cd-hit-est with parameters: `-n 5 -c 0.8 -G 0 -aS 0.8`). Finally, this library was merged with RepBase repeat library release 20170127 and ERV sequences obtained by Zhuo et al. (2013), again removing redundant sequences according to the '80-80' rule using CD-HIT (version 4.7, cd-hit-est-2d with parameters: `-n 5 -c 0.8 -G 0 -aS 0.8`). Each of the ten genomes in the panel were annotated using RepeatMasker (version 4.0.7, with the parameter `-e ncbi`) loaded with this custom repeat library. Any regions annotated as ERV loci which were within 200 nucleotides of each other were then merged into single regions and this set of regions was then defined as putative ERV regions to be searched for RT loci.

2.2.3 RT locus identification using four search methods

To evaluate and compare the EVE mining performance of different methods, the putative ERV regions identified in the previous stage were searched using four approaches, three of which are based on general homology search tools (HMMER, GeneWise, tBLASTn). In order to ensure equivalent information was provided to each of these tools, the same pHMMs were used as queries with both HMM based tools (HMMER and GeneWise) and the sequences that comprised the seed alignment for these pHMMs were used as query sequences with tBLASTn. The four methods used were: **1)** the DNA sequences of the ERV regions were searched directly using GeneWise (`genewisedb` from Wise2, version 2.4.1) employing the 3:33L model using all RT pHMMs from GypsyDB (Llorens et al. 2011) as queries with the parameters `-alg 333 -aalg 333L -init local -gap 6 -ext 0 -subs 1e-2 -indel 1e-4 -cut 15`; **2)** the DNA sequences of the ERV regions were translated into all six frames using EMBOSS transeq (version 6.5.7.0), the resulting amino acid sequences were searched using HMMER (`hmmsearch` from HMMER3, version 3.1b2) with default parameters and the same query pHMMs as those used with GeneWise, the translated hit locations were then converted back to genomic locations; **3)** the DNA sequences of the ERV regions were searched directly using tBLASTn from the BLAST+ package (version 2.5.0+) with default parameters, using a query library consisting of all the sequences used to construct the pHMMs used in the previous methods; **4)** the DNA sequences of the ERV regions were searched directly using RetroTector (version 1.0.1) using the `SweepDNA` then `SweepScripts` commands with default parameters and default motif library, the predicted Pol proteins containing at least one RT motif were considered as hits to RT loci.

For each search method, the same algorithm was employed in a post-processing step to eliminate overlapping hits and to combine multiple fragmented hits in the same coding region into combined single hits. First, where hits from the same query overlapped, only the highest scoring hit was retained. Then, hits from the same query, that were within 1000bp of each other and were in the correct orientation and order with respect to the query, were merged into a combined hit. After merging, overlapping hits from different queries were filtered to retain only the highest scoring hit. Finally, the hits were filtered to remove those with fewer than a total of 30 codons aligned to the query.

2.2.4 Sensitivity analysis

In order to quantify the sensitivity of each of the four RT identification methods, the loci detected by each method were compared to the combined set of loci detected by all methods combined. First, the hits produced in the previous steps for each of the ten genome assemblies were combined into a single set for each assembly and any overlapping hits were merged into a single combined hit. This set of hits was used as the definitive set of putative RT loci against which each of the methods could be compared and the number of these definitive RT loci which overlapped with hits produced by each search method were counted (Table A.1 and Fig. 2.1). This approach ensured that search methods which produced multiple short hits within a single locus did not receive a higher count than methods which successfully identified the whole locus in a single hit. An upset plot was generated to visualise the numbers of putative RT loci identified by each combination of methods (Fig. 2.3). Plots were generated using the seaborn and UpSetPlot packages in Python.

2.2.5 Predicted sequence completeness analysis

I compared the ability of each method to reconstruct the amino acid sequence of the detected RT loci in the presence of ORF disruptions. This was achieved using an approach based on alignment of the predicted RT amino acid sequences against a reference alignment. Despite RT being the most conserved domain found in ERVs and exogenous retroviruses, the high diversity present between different retrovirus genera causes many automatic alignment methods to produce inaccurate alignments of this domain when input sequences originate from distant retroviral genera. The approach used here attempts to minimise this error by using a manually curated backbone alignment (spanning the diversity of all retrovirus genera) to which new sequences are aligned with a custom substitution matrix in two stages. First, predicted RT amino acid sequences from each search method were split into two categories: 'complete sequences' derived from hits with greater than 180 codons matching the query HMM or sequence; and 'fragmented sequences' with fewer than 180 matching codons (RetroTector-derived sequences were considered 'complete' or 'fragmented' if they contained >3 or ≤3 RT motifs respectively). Where a predicted sequence was derived from multiple fragmentary hits that were merged during the search post-processing step, the missing sections in relation to the query pHMM or sequence were represented by lengths of ambiguous residues equal to the length of the missing sections.

A manually curated alignment of RT sequences from 83 exogenous retroviruses and well-characterised endogenous retroviruses, as well as retroelement outgroups (Gypsy, Micropia, Ty3, and Del elements), was used as a backbone amino acid alignment. First, an amino acid substitution matrix for further alignment steps was derived from this manual alignment using ReplacementMatrix (Dang et al. 2011). Next, for each method, and for each assembly, the 'complete' RT amino acid sequences were added to the initial alignment using the E-INS-i alignment method in MAFFT (version 7.467) using the derived substitution matrix and with the `--add` parameter (Katoh and Standley 2013). Next, the 'fragmented' RT amino acid sequences were added to these alignments using MAFFT with the same substitution matrix and the parameters `--addfragments --multipair`.

In order to count the proportion of the RT domain that was recovered in each hit, the alignment was trimmed to include only the 255 columns from the original alignment. The number of non-ambiguous residues for each predicted sequence was then counted to give the number of 'alignable amino acids' (AAAs) which can be used as an estimate of the proportion of the domain that was recovered in the predicted sequence. Importantly, this approach normalises the varying numbers of insertions found in each locus. The distributions of AAAs were plotted for each search method and each genome (Fig. 2.2) using the seaborn python package.

2.3 Results

A panel of ten bat genome assemblies from three bat families (Rhinolophidae, Pteropodidae, Phyllostomidae) were selected for benchmarking the ERV RT identification capabilities of four sequence search methods (GeneWise, HMMER, tBLASTn, RetroTector). In order to assess performance across varying types of genome assembly, genome assemblies within each bat family were selected to represent at least three of four categories of assembly strategy: short-read, low coverage, assembled by SOAPdenovo; short-read, high coverage, assembled by SOAPdenovo; short-read, medium coverage, assembled by DISCOVAR *de novo*; and assemblies that variously utilised long reads (see Table 2.1 for details). Putative ERV regions within each assembly were identified using RepeatMasker loaded with an expanded bat ERV library (see Section 2.2.2) and each method was used to independently search within these regions. Hits from all methods were then merged to create combined putative RT loci. Across the ten assemblies, a total of 16,434 putative RT loci were identified (mean 1,819, 1,629, 1,487 for Rhinolophidae,

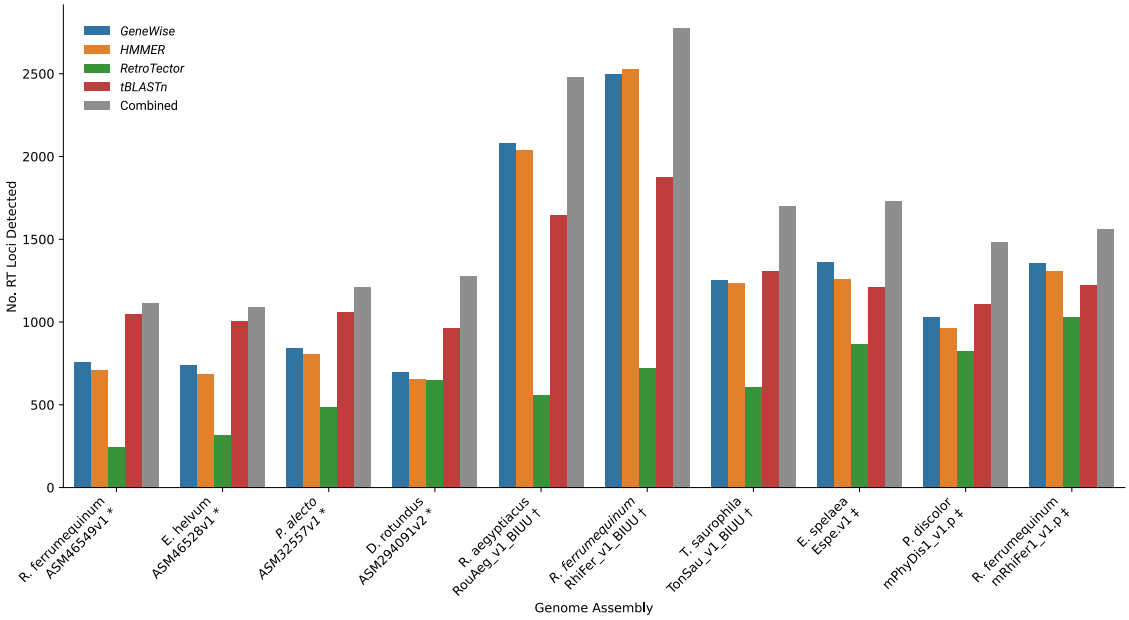


Figure 2.1: Counts of number of putative RT loci detected by each search method compared with the total number of putative RT loci (grey), across genome assemblies. The number of putative RT loci is determined by combining all hits from all search methods and merging any overlapping hits to create a set of combined positions. X-axis ordered by assembly strategy and ascending contig N50 value. * SOAPdenovo-assembled genomes. † DISCOVAR *de novo*-assembled genomes. ‡ Long-read-based genome assemblies.

Pteropodidae, and Phyllostomidae, respectively), see Fig. 2.1 and Table A.1. To assess the completeness of the predicted amino acid sequences, the predicted amino acid sequences for each hit by each method were aligned using MAFFT to an existing amino acid alignment of retrovirus RT sequences with 255 columns. To normalise for varying numbers of insertions between loci, columns not present in the original alignment were removed, and the number of residues remaining were counted to give the number of 'alignable amino acids' (AAAs) , which ranged from 0 to a maximum of 255 (the length of the original RT alignment) for each hit. The distributions of AAAs are plotted for each assembly in Fig. 2.2.

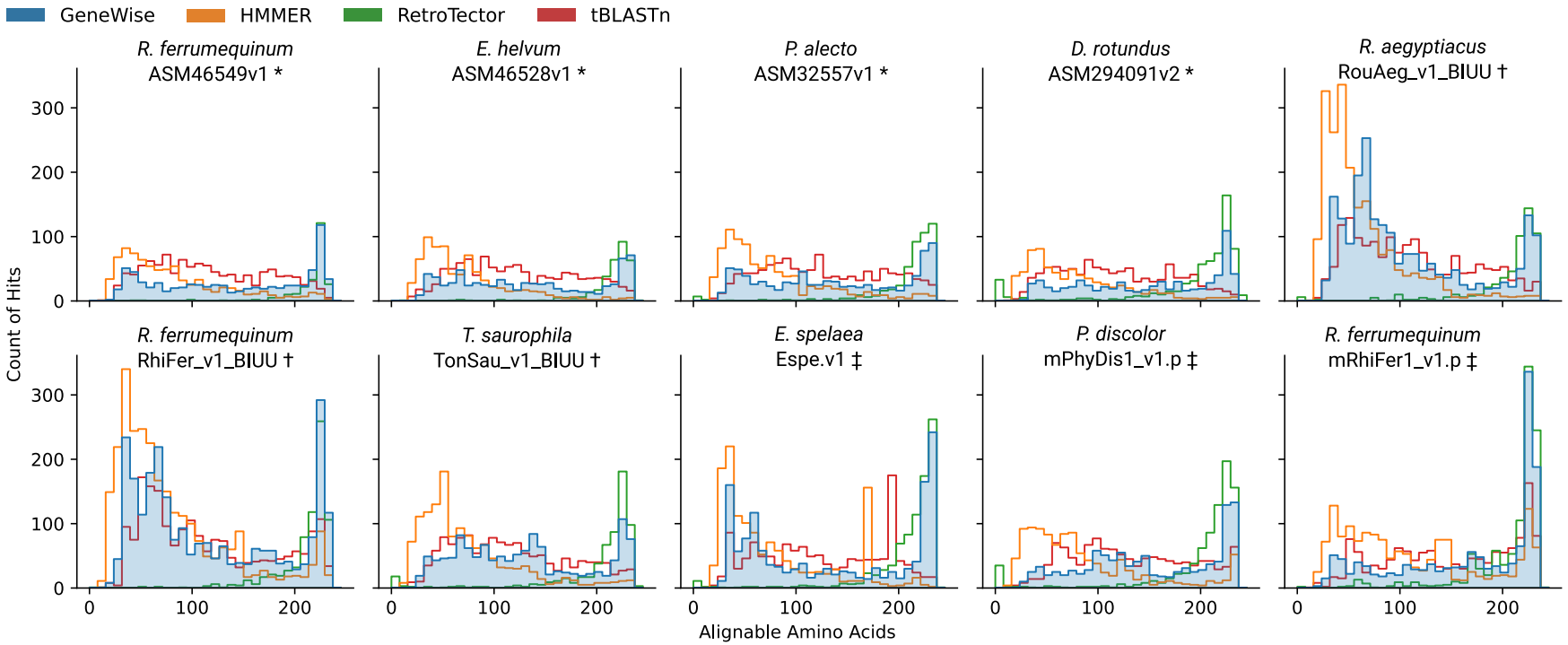


Figure 2.2: Distributions of 'alignable amino acids' (AAAs) of RT domain hits for each search method, across genome assemblies. AAA is calculated as the number of amino acids in the predicted protein sequence of a hit that align to the original 255 columns of the manually curated reference RT alignment. Number of AAAs is therefore a measure of the completeness of a recovered RT sequence, with hits possessing close to 255 AAAs considered to represent complete sequences. Plots are ordered by assembly strategy and ascending contig N50 value. * SOAPdenovo-assembled genomes. † DISCOVAR *de novo*-assembled genomes. ‡ Long-read-based genome assemblies.

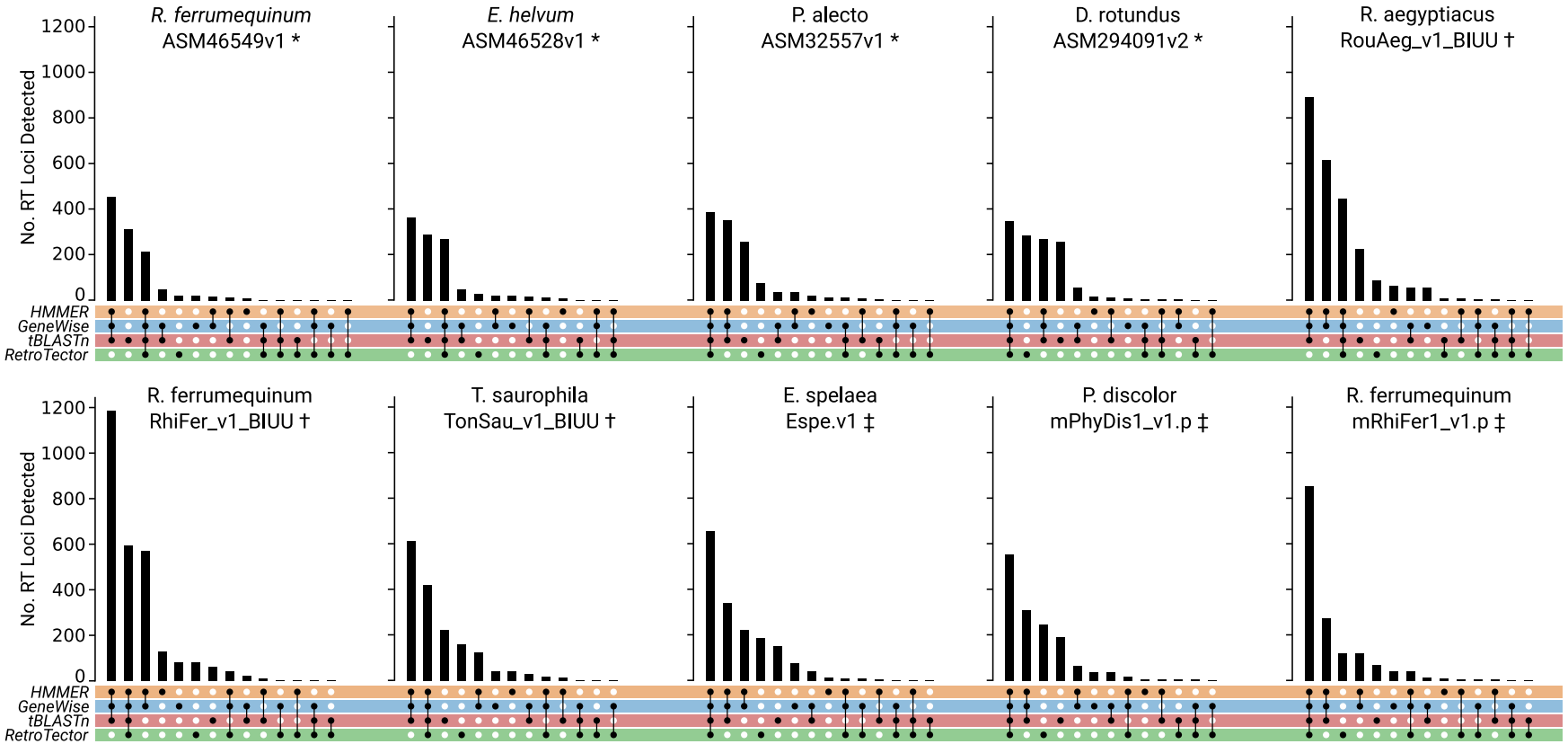


Figure 2.3: Upset plot to show the number of putative RT loci detected by each grouping of methods, across genome assemblies. The membership of each group (along the x-axis) is displayed with filled circles (included), empty circles (not included). Ordering of groups is by descending number of loci detected by each group. Plots are ordered by assembly strategy and ascending contig N50 value. * SOAPdenovo-assembled genomes. † DISCOVAR *de novo*-assembled genomes. ‡ Long-read-based genome assemblies.

2.3.1 Differences in number and length of identified RT sequence between search methods

To assess the ability of GeneWise to recover EVEs from genome assemblies, I compared the hits and predicted protein sequences generated by four methods (GeneWise, tBLASTn, HMMER, and RetroTector) when searching for retrovirus-derived RT domains in 10 bat genome assemblies. To avoid spurious hits to LINE RT domains and to enable direct comparisons between all methods within reasonable time limits, searches were restricted to putative ERV regions identified by RepeatMasker in each genome assembly. GeneWise, tBLASTn, and HMMER are general search sequence methods and can be used to search for any endogenous viral domains, however, RetroTector is specialised for ERV domain discovery. For each of the general search methods, equivalent queries were used in order to provide a fair comparison (see Section 2.2.3). ERV RT domains were chosen as a means to benchmark these tools due to their high abundance within vertebrate genomes and the wide range of states of degradation of the encoding open reading frames (ORFs), thus providing an extensive dataset to robustly assess performance. It also enabled me to compare against the specialised tool, RetroTector, which is specifically tuned to recover long sections of ERV coding domains and successfully predict the encoded protein sequence.

Across the regions of each genome that were scanned, hits from all four tools were combined and overlapping hits were merged to give a combined total set of putative ERV RT loci. This was used to assess the sensitivity of each tool for detecting the presence of endogenised RTs, each tool was considered to have detected a particular locus if it produced hits that overlapped with the putative locus.

RetroTector detected the fewest putative loci in each genome (Fig. 2.1 and Table A.1), however a higher contig N50 value appears to correspond to a higher number of loci detected, suggesting that RetroTector requires longer stretches of coding sequence, undisrupted by a more fragmented genome assembly, to detect a hit. This is confirmed by the distribution of Retrotector's hit lengths (visualised in Fig. 2.2), with the vast majority of hits consisting of >200 AAAs and almost none below 150 AAAs. Both HMMER and GeneWise use protein pHMMs as search queries, and so, perhaps unsurprisingly, detected very similar numbers of loci in each genome assembly. Although within each assembly, GeneWise detected more loci that HMMER was unable to detect than *vice versa* (see Fig. 2.3). Interestingly, the detection sensitivity of tBLASTn relative to other tools varied by genome assembly. Specifically, in short-read, SOAPdenovo assemblies tBLASTn detected substantially more putative RT loci than any of the other tools. However in

genomes assembled from long-read data, tBLASTn detects a very similar number of loci to HMMER and GeneWise, and in DISCOVAR *de novo* assemblies, tBLASTn appears to detect fewer loci than HMMER and GeneWise.

Although HMMER, GeneWise, and to some extent, tBLASTn were able to detect similar numbers of RT loci in each genome, these tools differ substantially in terms of the length of hits they recover. Across SOAPdenovo assemblies, and long-read assemblies, tBLASTn detected a relatively uniform number of RT fragments of every length, however in DISCOVAR *de novo* assemblies, tBLASTn detected an increased number of shorter fragments. In mRhiFer1_v1.p and Espe.v1 assemblies tBLASTn detected an increased number of longer fragments – in line with HMMER and GeneWise. Across all assemblies, HMMER produced large numbers of short hits and only relatively few long hits, meaning that only short fragments of coding sequence could be reconstructed using this tool, despite its sensitivity. In each assembly, RetroTector produced only longer hits, with almost no hits containing fewer than 150 AAAs. The number of long hits produced by RetroTector was consistently much higher than the number produced by HMMER or tBLASTn, indicating the strength of RetroTector in reconstructing long sections of ERV coding domains. Across all assemblies GeneWise was able to reconstruct a comparable number of long hits to Retrotector, and in addition, GeneWise was able to detect and reconstruct the sequence of shorter fragments that RetroTector lacked the sensitivity to detect at all. This suggests that GeneWise is particularly successful at detecting and reconstructing the full extent of both short and long fragments of pseudogenised RT coding sequence.

2.3.2 Differences in detected RT domain abundance between genome assemblies

Although not the main purpose of this study, conducting a thorough search of ERV RT domains across this dataset highlights the major impact of different assembly strategies on the content of the genome assembly. Specifically, the major differences that are reflected in the results of an *in silico* ERV census. ERVs are transposable elements which are often found in regions with many tandem-duplicated copies (Ahmed and Liang 2012). These repetitive genomic regions, which contain a high proportion of the total ERV content of a genome, often prove particularly difficult to assemble with short-read sequencing data. The low coverage, short-read genome assemblies (ASM46549v1 and ASM46528v1, assembled with SOAPdenovo in both cases) have contig N50 values below 13Kbp, implying they are highly fragmented assemblies, and unsurprisingly, contain low amounts of detectable ERVs. In these cases, the assembler likely struggles to

assemble a high proportion of the repetitive regions which contain many of the ERVs. Furthermore, the lower read coverage may lead to multiple copies of ERVs being collapsed into single contigs as there are too few reads to distinguish between the different copies given the low amount of variation between them.

The genome assemblies ASM32557v1 and ASM294091v2 were both assembled with SOAPdenovo using short-read datasets but with around 100x coverage (more than five times the read depth used in the low coverage short read assemblies) and consequently have several fold higher contig N50 values (~32Kbp and ~80Kbp for ASM32557v1 and ASM294091v2 respectively). Interestingly, these two assemblies only show a 13% increase in mean number of RT loci detected compared with the many fold increase in contig N50 values. This suggests that although the non-ERV enriched genomic sections may be much less fragmented in these assemblies, the assembler still has issues with assembling ERV containing regions of the genome despite the higher read coverage.

All three assemblies produced using PacBio Sequel long-read sequencing technology have contig N50 values which are 1-2 orders of magnitude greater than those produced using short reads, indicating vastly less fragmented assemblies. This corresponds to an increase of up to 59% in the total number of RT loci detected compared with genomes of closely related species assembled with short reads by SOAPdenovo. Additionally, the number of highly complete RT sequences that were recovered from these assemblies was markedly higher than those produced by SOAPdenovo or DISCOVAR *de novo* (Fig. 2.2). This suggests that the long-read technology enables, not only assembly of more ERV-derived nucleotides in total, but also enables much more successful assembly of the ERVs in their entirety.

The panel of genome assemblies also contains three assemblies (RouAeg_v1_BIUU, RhiFer_v1_BIUU, and TonSau_v1_BIUU) produced using the assembly tool DISCOVAR *de novo* with short reads with coverages between 36x and 46x. Impressively, the contig N50 values for these assemblies ranged from 100Kbp to 142Kbp despite only using short-read libraries with middling read coverage. Additionally, up to 2.5 times the number of RT loci were detected in these assemblies than in SOAPdenovo assemblies for closely related species of bats (for example 1214 putative RT loci in ASM32557v1 compared with 2482 in RouAeg_v1_BIUU despite a coverage of 110x and 36x respectively). This suggests that DISCOVAR *de novo* is much more capable of assembling ERV-enriched regions than SOAPdenovo, even given fewer reads. Interestingly, more RT loci were also detected in each of these assemblies than in those produced using long-read sequencing technology. This could indicate that DISCOVAR *de novo* is also more

effective at assembling ERV regions than the long-read assembly strategies or it could be a result of DISCOVAR *de novo* producing an excess of artifactual contigs containing ERV sequence. Notably, in RouAeg_v1_BIUU and RhiFer_v1_BIUU, a large proportion of the increased number of RT loci compared with other assemblies were loci which only spanned shorter regions of the RT domain (Fig. 2.2), implying that the additional ERV regions present in the DISCOVAR *de novo* assemblies are highly fragmented.

2.3.3 DVORFS: An easy-to-use wrapper and post-processor for EVE mining with GeneWise

In order to provide an easy-to-use means of applying GeneWise for the sensitive detection and reconstruction of EVE sequences, I developed an integrated GeneWise wrapper and post-processing tool: Disrupted Viral ORF Search (DVORFS), which can be accessed at https://github.com/ilevantis/dvorfs. DVORFS is implemented in Python 3 and makes use of: GeneWise (Birney et al. 2004) as the core tool for searching for EVEs in a genome; BEDTools (Quinlan and Hall 2010) for manipulating genomic coordinates and extracting specific regions of fasta files; EMBOSS transeq (Rice et al. 2000) for efficiently translating DNA sequence for searching with HMMER; and HMMER (Eddy 2011) for carrying out rapid preliminary searches of large genomes. DVORFS is intended to be used with selected pHMMs from databases such as PFAM (El-Gebali et al. 2019), GyDB (Llorens et al. 2011), and SUPERFAMILY (Gough et al. 2001) or from pHMMs derived from user curated protein sequence alignments.

DVORFS carries out a search for viral domains (from a given set of viral protein pHMMs) in a given set of DNA sequences in three stages: **1)** 'presearch' stage – the DNA sequences are windowed and translated, then searched with HMMER using lenient thresholds, then translated hit coordinates are converted back to genomic coordinates; **2)** search stage – hit locations from the 'presearch', as well as their surrounding genomic regions, are windowed and searched using the 3:33L model in GeneWise; **3)** post-process stage – the coordinates and predicted coding sequences of hits from the search stage are combined where appropriate and filtered to meet desired threshold values. Optionally, the first stage can be skipped and a predefined set of regions can be supplied to set the regions within which the search stage will be restricted. The output of DVORFS is a set of hits and the corresponding predicted amino acid sequence of the EVE domains as well as a preliminary codon alignment for each query domain based on the query pHMM. During the process of experimenting with the use of GeneWise as a tool for mining EVEs, a number of bugs were discovered in Wise2, for which DVORFS contains workarounds to enable ease of use for EVE mining.

2.4 Discussion

2.4.1 Frameshift-aware protein pHMM to DNA alignment search improves EVE detection and sequence recovery

The software GeneWise can directly perform protein pHMM to DNA alignments whilst accounting for frameshifts in the DNA sequence. Since most ERVs and other EVEs present in a genome consist of coding sequences that are highly degraded, the HMM-based, frameshift-aware alignment model implemented in GeneWise is theoretically, particularly well-suited to the detection and coding sequence prediction of EVEs. In this study, I compare GeneWise against HMMER, tBLASTn, and RetroTector and show that GeneWise is, in practice, more successful in detecting and reconstructing the full extent of both short and long fragments of pseudogenised RT coding sequence across ten bat genome assemblies. Thus, this novel application of GeneWise, implemented in DVORFS, provides an improved method for *in silico* EVE detection and characterisation over previously used EVE mining methods.

Across the ten genome assemblies analysed, each of the tools showed different characteristics in terms of sensitivity to detecting the presence of ERV RT domains and ability to reconstruct the coding sequence of the disrupted locus. Both HMMER and tBLASTn were broadly able to detect a similar number of RT loci in each genome (Fig. 2.1), however, the amount of coding sequence recovered for many of the RT loci was relatively short compared to that recovered by GeneWise and RetroTector (Fig. 2.2). The differences between the quantity of coding sequence recovered by these two groups is likely a reflection of these tools' abilities to deal with frameshifts in the ERV RT loci. HMMER (with post-processing) and tBLASTn both align protein queries to each frame of a DNA sequence individually before stitching sequential HSPs found in different frames into a single hit. In comparison, RetroTector uses a 'fragment threading' sequence alignment approach which is based on short motifs and is robust to numerous frameshifts. The GeneWise alignment model is also able to account for frameshifting insertions or deletions because it can identify codons that match the protein query in any frame. Furthermore, a frameshift-aware alignment method is able to continue extending an initial seed alignment across frameshifts, thus recovering longer sections of a disrupted coding sequence despite its existence across multiple frames. In comparison, frameshifted sections of sequence may be entirely missed by the single-frame alignment algorithms in HMMER and tBLASTn as the section of sequence in a particular frame may have too little homology to the query in isolation in order for an initial hit to occur. As a result,

GeneWise and RetroTector are particularly successful in comparison to HMMER and tBLASTn at recovering the full extent of longer ERV RT loci.

Although the 'fragment threading' approach of RetroTector is robust to frameshifts, it requires the presence of two or more specific motifs to be conserved at the locus for detection to occur. This means that only loci which are long enough to contain these specific motifs will be detected. This condition, in turn, explains why RetroTector is unable to detect shorter RT fragments and therefore detects the fewest number of putative RT loci over all. In comparison, GeneWise does not rely on specific motifs within the query but can find homology anywhere along the query HMM, enabling the detection of shorter coding sequence fragments from any region of the original ORF. As a consequence, GeneWise is able to detect shorter fragments with sensitivity comparable to, or better than, HMMER and tBLASTn, both of which are similarly able to detect homology anywhere along the query.

Overall, the benchmarking demonstrates that GeneWise is able to detect short, more fragmented, sequences with similar sensitivity to tBLASTn and HMMER, while also being able to more completely recover highly degraded coding sequence, like RetroTector, but is not limited to detecting only ERV coding sequences. These capabilities make GeneWise particularly well-suited to detecting any EVEs or transposable elements (TEs) that contain coding sequence within host genome assemblies. In particular, GeneWise should prove useful for recovering older, more degraded, pseudogenised coding sequences in which numerous frameshifts have accumulated. Therefore, compared with other protein-based TE detection methods, a GeneWise-based search enables a more accurate census of TEs to be conducted due to the increased chance of detecting the more ancient and degraded elements in a genome. When used for EVE detection, the ability to reconstruct longer sections of EVE coding sequence is particularly useful as it provides more biological and phylogenetic information about the original virus that was endogenised. Furthermore, the utilisation of protein pHMM-based search provides an increase in sensitivity for detecting remote homologues of the sequences represented in the query profile as shown by previous studies (Park et al. 1998; Skewes-Cox et al. 2014; Kirsip and Abroi 2019). This is of particular need when searching for non-retroviral EVEs due to the sparse sampling within viral sequence databases, and the large genetic distances between reference exogenous viruses and the ancient viral lineages found endogenised as EVEs. This search method therefore has the potential to greatly improve the breadth of viral lineages that can be discovered, as well as the information that can be recovered about these lineages, in studies of paleoviruses.

A shortcoming of GeneWise is the extreme computational intensity of its alignment algorithms (such as the 3:33L algorithm tested here), compared with the much more efficient HMMER and BLAST algorithms. Unfortunately, using GeneWise to search an entire eukaryotic genome with a substantial number of query pHMMs can therefore take many thousands of CPU hours. In order to reduce the computational time required, it is reasonable to use a more efficient, though less sensitive, search algorithm with lenient cutoff thresholds to identify genomic regions within which GeneWise should search. In this benchmarking study the searches for RT domains was restricted to putative ERV regions identified by RepeatMasker (based on nucleotide BLAST search). However, when searching for non-retroviral EVEs it would be reasonable to use the highly optimised HMMER software (Eddy 2011) to find regions with tentative hits to the query pHMMs or a BLAST search with a relevant query database. The tool developed here (DVORFS), for EVE detection with GeneWise, uses HMMER by default to implement this approach in order to analyse genomes within a reasonable timeframe without sacrificing too much sensitivity. DVORFS therefore provides an easy to use tool which implements the novel EVE/TE mining method described and benchmarked in this study.

2.4.2 Genome assembly strategies have a large impact on assembled TE content

Using a panel of ten genome assemblies, this study provides the first systematic comparison of different genome assembly strategies with respect to ERV content, clearly showing, for the first time, that assembly strategy has a major effect on the amount of detectable ERV content within a genome assembly. The ten assemblies were selected to encompass regularly used genome assembly strategies, including Illumina short-read-based assemblies with varying coverage using two different assembly programs, as well as PacBio long-read-based assemblies. To allow like-for-like comparisons between assemblies, genome assemblies of closely related, or the same, species were chosen from three bat families.

An important factor in the assembly of ERVs and other TEs during genome assembly is the length of sequencing reads used for the assembly. Specifically, assembly algorithms have difficulty with genomic regions containing repeat sequences that are longer than the inputted reads. Unsurprisingly therefore, the results here show that long-read-based assemblies contain a substantially higher total count of ERVs than SOAPdenovo short-read assemblies (Fig. 2.1), and furthermore, contain a much higher proportion of full length ERVs than short-read assemblies produced by either SOAPdenovo or DISCOVAR *de novo* (Fig. 2.2). This reflects the fact that, unlike short-read Illumina sequencing

libraries, long-read sequencing libraries contain reads long enough to span the entirety of single repetitive elements, including ERVs, and even multiple tandem repeats. This in turn enables the assembly of higher numbers of solitary ERVs as well as regions of ERV tandem repeats that account for a substantial proportion of the ERVs in humans and likely other eukaryotic genomes (Ahmed and Liang 2012). The successful assembly of TE regions also leads to an overall markedly less fragmented assembly which is précised by the 1-2 order of magnitude increase in contig N50 values that long-read-based assemblies achieve over short-read-based assemblies.

For short-read assemblies, the results of this study clearly show that the most important factor determining the quantity of assembled ERVs appears to be the assembly program used. Assemblies produced using SOAPdenovo had very similar quantities of ERVs regardless of whether they were produced with read coverage around 20x or 100x despite having significantly higher contig N50 values in the latter case. Interestingly, the three short-read assemblies produced using DISCOVAR *de novo* contain much higher numbers of complete RT sequences than the SOAPdenovo assemblies and have the highest over all counts of detected RTs even compared with the corresponding long-read assemblies. The three assemblies of the *R. ferrumequinum* genome are particularly illustrative of this pattern. Both the DISCOVAR *de novo* assembly and the long-read assembly have high numbers of complete RT sequences compared with the SOAPdenovo assembly, however the DISCOVAR *de novo* assembly has a somewhat increased number of detected medium length RT fragments compared with the SOAPdenovo and long-read assemblies. Most strikingly, however, is the vastly increased number of short RT fragments detected in the DISCOVAR *de novo* assembly compared with either of the other two assemblies. This suggests that DISCOVAR *de novo* assemblies contain many short fragments of ERVs in addition to the successfully fully assembled ERVs. It is unclear whether these fragmented ERVs are artefacts produced by the DISCOVAR *de novo* assembly method or whether they represent true regions of the genome which are not assembled by the other assembly strategies. Over all, these results show that DISCOVAR *de novo* appears to be vastly more successful at assembling ERV-containing regions than SOAPdenovo given short-read sequencing libraries, this corresponds with data from other studies (R.R. Love et al. 2016; Scheben et al. 2020).

In conclusion, these results show clearly, for the first time, that the measured ERV content of a genome assembly is highly dependent on the strategy used to create the assembly. This likely arises due to the varying success with which different strategies are able to assemble the repetitive regions of a genome and therefore the effect likely extends to the measurement of all TE content within a genome assembly. Importantly,

this highlights the need to ensure like-for-like comparisons are used when TE content of genomes are compared between species, and raises the need for reappraisal of studies comparing TE content in genome assemblies produced using non-uniform methods (for example Kapusta et al. 2017 and Pasquesi et al. 2018). In future studies where measuring the total TE abundance or abundance of the most prevalent TEs within a genome is the most important factor in the analysis, an assembly-free approach such as RepeatExplorer (Novák et al. 2010) should be used as it will likely more accurately identify and quantify the high copy number TEs comprising the majority of total TE content. Where between-species comparisons of low copy number TEs are of interest, assembly-based identification strategies are likely necessary, but the bias of varying assembly strategies must be taken into account. Here it should be noted that contig N50 may be a bad heuristic by which to solely assess the proportion of TE sequence one expects to be successfully incorporated into an assembly; instead, the specific assembly strategy should be considered.

2.4.3 The need for a high sensitivity search tool for EVE identification

EVEs are particularly challenging to detect and extract from genomes due to the tremendous sequence diversity of viral genes coupled with the high frequency of frameshifts present in these often pseudogenised sequences. These dual problems cause most homology-based search tools such as BLAST and HMMER to lack sensitivity for initial detection of an EVE or lack the ability to identify the full extent of its coding region. GeneWise is well suited to overcome both of these problems because it is frameshift-aware and uses protein pHMM to DNA alignment. In this study I show empirically that it is consequently well-suited to the task of EVE detection and EVE coding sequence reconstruction and its application to EVE mining provides an improvement over current methods. Unfortunately, GeneWise is now an unmaintained piece of software, last updated in 2007, which can be finicky to work with due to occasional bugs and incompatibility with more modern data formats. To alleviate these issues I created an easy-to-use tool, DVORFS, which simplifies the process of EVE mining with GeneWise. DVORFS provides a wrapper for GeneWise that includes workarounds for a number of bugs, carries out an automatic 'pre-search' stage which reduces the computational requirements, and carries out a post-processing stage which refines obtained hits into more complete EVE predictions. Ultimately, DVORFS enables straightforward use of this novel EVE mining approach.

This study was conducted using pHMMs from a pre-existing database to conduct single-pass scans of the genome assemblies, however an iterative approach could be employed

to increase remote homologue sensitivity further. In such a workflow, newly identified EVE sequences from the first round of search would be aligned with reference viral sequences, and this alignment could then be used to produce an improved pHMM with which to search the genome again. These steps could be iterated until no new EVEs are discovered. Another strategy, which has been shown to improve remote homologue sensitivity for EVE detection , would be to use pHMMs that incorporate structural information, for example from the SUPERFAMILY database (Gough et al. 2001).

This study highlights the utility of a frameshift-aware protein pHMM to DNA aligner for the purposes of characterising EVEs, and I provide a wrapper to more easily use GeneWise for this purpose. However, the reliance of DVORFS on an unmaintained piece of software is not ideal, and additionally, faster algorithms for searching using pHMMs (Eddy 2011) have been developed since the development of GeneWise. Future work in applying faster pHMM-based algorithms to the protein pHMM to DNA alignment would likely provide a large improvement over DVORFS. Recently, the tool MachineBoss (Silvestre-Ryan et al. 2020) has been published which could provide a simplified method to prototype frameshift-aware search and alignment models tailored particularly to EVEs and perhaps provide even better EVE characterisation than the GeneWise alignment algorithm.

3

ERV distribution and retrovirus transmission within the Chiroptera

Contents

Abstract **63**
3.1 Introduction **63**
3.1.1 Bats are a panglobal clade of extreme species richness 63
3.1.2 Bats are zoonotic reservoirs for viruses of public health concern 64
3.1.3 ERVs are a record of retrovirus infections 65
3.2 Methods **67**
3.2.1 RT alignment and retrovirus reference phylogeny creation . . 67
3.2.2 Phylogenetic placement of RT sequences and Edge PCA comparison 69
3.2.3 Inference of horizontal ERV transfers 70
3.2.4 Statistical modelling of ERV transfer rates 71
3.3 Results **72**
3.3.1 Differences in ERV composition among bats species 72
3.3.2 Rates of cross-species retrovirus transmission in bats over evolutionary time 76
3.4 Discussion **83**
3.4.1 An unbiased analysis of ERV activity across Chiroptera 83
3.4.2 Numerous expansions and extinctions of bat ERV lineages . . 84
3.4.3 RT sequence gene reconciliation reconstructs the retrovirus sharing history of Chiroptera 86
3.4.4 Phylogenetic relatedness impacts the rate of retrovirus transmission 87
3.4.5 Increased rate of cross-species retrovirus transmission within noctilionoid bats 88

Abstract

Endogenous retroviruses (ERVs) are ubiquitous across the genomes of modern vertebrates, having become fixed in ancestral populations and accumulated over millions of years. Through their ability to duplicate as transposable elements they have come to comprise substantial portions of host genomes. In this chapter, I mine the genomes of 19 bat species across 11 families for endogenous retroviral reverse transcriptase (RT) loci, and use the recovered sequences to make an unbiased comparison of the ERV composition across Chiroptera. This comparison shows a broadly consistent abundance of ERVs across bats, which is dominated in each case by betaretroviral and gammaretroviral ERVs. In contrast to the broad trend, *Megaderma lyra* and *Lasiurus borealis* show evidence of major ERV lineage expansions while other species show substantially reduced total ERV abundance and/or extinctions of particular classes of ERVs. As the result of past retroviral infections, ERVs also provide a historical record of retrovirus infections. Notably, bats are host to a broad diversity of viruses, a number of which pose serious veterinary and public health threats as zoonotic viruses. Thus, elucidating historical trends of cross-species viral transmission is of particular public health and veterinary interest, as well as evolutionary research interest. Using a phylogenetic reconciliation approach I infer historical cross-species retrovirus transmission events occurring amongst bats based on the ERVs recovered in the 19 bat genomes. This broad-scale, systematic analysis recapitulates previous findings that phylogenetic relatedness has an important effect on cross-species viral transmission rates and shows that retrovirus transmission between noctilionoid bat species has occurred at a historically higher rate than has occurred in bat superfamilies with more dispersed ranges.

3.1 Introduction

3.1.1 Bats are a panglobal clade of extreme species richness

Bats number over 1300 extant species across more than 200 genera (Racey 2015), forming a clade (order Chiroptera) of extreme species richness within mammals. The order Chiroptera is split into two suborders with an estimated divergence date of ~64 Mya: Yinpterochiroptera consisting of the Pteropodidae family, and Rhinolophoidea superfamily; and Yangochiroptera consisting of the Rhinolophoidea, Emballonuroidea, Noctilionoidea, and Vespertilionoidea superfamilies (Teeling et al. 2005; Jebb et al. 2020). Yinpterochiroptera likely originated in Asia and are found exclusively in the Old World. In contrast, yangochiropteran species have a panglobal distribution with a more uncertain location of origin (Teeling et al. 2005; Jones and Teeling 2006). Within the Yangochiroptera, with a few exceptions, most members of the Noctilionoidea superfamily are endemic to South America, and the clade is particularly ecologically diverse, having undergone an evolutionary radiation after reaching the Neotropics in the Eocene (Gunnell et al. 2014; Rojas et al. 2016). In contrast, the Vespertilionoidea superfamily is a clade containing the extremely cosmopolitan Vespertilionidae and Molossidae

families, which likely originated in the northern hemisphere in the early Eocene, before dispersing across the globe to occupy ecological niches in every continent except Antarctica (Teeling et al. 2005; Lack and Van Den Bussche 2010; Ruedi et al. 2013; J.J. Shi and Rabosky 2015).

The frequent and rapid range expansions of bats across landmasses during the Eocene were likely to be facilitated by their unique (amongst mammals) ability for powered flight (J.J. Shi and Rabosky 2015). However the evolution of powered flight led to high metabolic demands, which turn appears to be related to the origin of several derived traits in bats. These traits include a small genome size (Kapusta et al. 2017), and a unique immune system adapted to dampen self-mediated immune activation and inflammation (Ahn et al. 2016; Xie et al. 2018; Ahn et al. 2019).

3.1.2 Bats are zoonotic reservoirs for viruses of public health concern

The extremely high species and ecological diversity of bats is mirrored by a particularly high diversity of bat viruses (Luis et al. 2013; Hayman 2016). To date, over 12,000 bat-associated species of virus have been discovered, representing 30 virus families (L. Chen et al. 2014). Of these, a number pose serious veterinary and public health threats as zoonotic viruses, implying that bats are a particularly important zoonotic reservoir. For example, bats are implicated as reservoirs directly responsible for transmission of Marburg, Nipah, and Hendra viruses to humans (Towner et al. 2009; Halpin et al. 2011), as well as being the likely origin of Ebola virus, and SARS-CoV-2 (Olival and Hayman 2014; Latinne et al. 2020).

The peculiarities of bat physiology and bat immune systems would suggest that there are likely to be unique dynamics present between bats and viruses, and that these may be related to the high number of zoonotic spillovers recorded for bat-associated viruses. Indeed, previous studies have suggested that bats give rise to a higher abundance of zoonoses in comparison to other orders of mammal, and this may be a product of their unique physiology and ecology, and perhaps particularly due to their flight and tendency to form large social groups (Turmelle and Olival 2010; Luis et al. 2013; Olival et al. 2017). However, a more recent study based on an expanded and more robust dataset suggested that bats as a group were not any more likely to transmit viruses across large phylogenetic distances than any other order of mammals once clade size was taken into account (Mollentze and Streicker 2020).

Nonetheless, the dynamics of viral infections amongst and between bat populations (i.e. within species or across smaller phylogenetic distances) are poorly understood, and a range of different hypotheses explaining these inter-host dynamics are each supported by conflicting evidence (Plowright et al. 2016; Letko et al. 2020). Some studies have explored viral sharing networks between bat species using a literature-search approach, and have suggested that the most important factors determining viral sharing between bat species are gregariousness, extent of range overlap (Luis et al. 2015), and potential for co-roosting (Willoughby et al. 2017). A recent study, which used direct sampling of bat-associated alpha- and betacoronaviruses with phylogenetic reconstructions to characterise host-switching events, showed that cross-species transmission rates were not homogeneous across bat families, and that the cave-roosting *Rhinolophus* genus was particularly central to alphacoronavirus transmission (Latinne et al. 2020). This study also highlights the striking differences in cross-species transmission between the two genera of virus examined, and recapitulates the finding by Mollentze and Streicker (2020) that virus-specific traits are one of the most important factors determining the rate of viral host-switching, rather than host-related factors.

Regardless, the species richness of Chiroptera means that any findings relating to bat–virus interactions and inter-host transmission which generalise across the clade, will be applicable to a large proportion of potential zoonoses.

3.1.3 ERVs are a record of retrovirus infections

Retroviruses are single stranded positive-sense RNA viruses that infect vertebrates. The family *Retroviridae* comprises the monogeneric subfamily *Spumavirinae*, and the *Orthoretrovirinae*. Within the *Orthoretrovirinae* there are three classes, defined by major divergences in their reverse transcriptase genes: Class I, containing *Epsilon-* and *Gammaretrovirus*; Class II, containing *Alpha-*, *Beta-*, *Lenti-*, and *Deltaretrovirus*; and Class III, containing the ERVL-like retroviruses (Llorens et al. 2008; A. Hayward et al. 2015; Gifford et al. 2018). During infections, retroviruses integrate a DNA copy of their genome (~10 Kbp in length) into the host cell chromosome as part of their replication cycle, creating an endogenous provirus, termed an endogenous retrovirus (ERV). This provirus then continues the replication cycle by producing RNA copies of the viral genome that are either packaged into virions, or are integrated at another location in the genome through retrotransposition (Stocking and Kozak 2008). In the latter case, the efficient intracellular nature of replication can lead to many more copies of the original provirus spreading throughout a host's genome (Magiorkinis et al. 2012). When

an integration occurs in a germline cell, the ERV may also be transmitted vertically to the host's offspring, and has the opportunity to eventually become fixed in the host population. Thus, through repeated insertions and vertical transmission, ERVs can accumulate within host genomes to the extent that they typically comprise 5% to 15% of the total genomic DNA of most mammals (Smit et al. 2013). Interestingly, although active flight imposes metabolic limits on bats, and has likely led to a constrained genome size, ERV activity within bat genomes does not appear to be reduced (Kapusta et al. 2017).

Due to the highly deleterious effects of active ERVs, host organisms have evolved mechanisms to strictly silence ERVs (Groh and Schotta 2017; Hurst and Magiorkinis 2017). This causes most ERV insertions that do not disrupt existing genes to be of negligible evolutionary impact, and thus these loci mutate under a neutral selection regime at the host's neutral mutation rate. The neutral mutation rate for cellular organisms is estimated of around 10^{-9}substitutions per nucleotide per year (Kumar and Subramanian 2002), which is several orders of magnitude slower than the mutation rates found in exogenous viruses (Aiewsakun and Katzourakis 2016). As a result, ERVs are well preserved copies — or 'molecular fossils' — of the exogenous retroviruses that infected the host millions of years in the past, so providing sequence information about ancient viruses that is lost from extant exogenous viruses due to their rapid mutation rates. Furthermore, the vertical inheritance of ERVs within hosts means that orthologous ERVs found in multiple host species can provide a robust means of time-calibrating retrovirus phylogenies based on the divergence times of the host species (Aiewsakun and Katzourakis 2015).

This wealth of information means that ERVs have been crucial in determining the deeper phylogeny and co-evolutionary history of retroviruses, which has revealed a likely marine origin of the clade in amphibians and fish >450 Mya (Aiewsakun and Katzourakis 2017), followed by diversification that involved frequent host switching amongst vertebrate hosts (A. Hayward et al. 2015; Xu et al. 2018). ERVs have also provided evidence that bats, along with rodents, have been particularly important in shaping retrovirus diversity through evolutionary history, including the possibility that bats may be the origin of the *Gammaretrovirus* genus (Cui et al. 2012; Cui et al. 2015). ERVs have also enabled more detailed analyses of the co-evolutionary history and specific cross-species transmission events for specific retrovirus lineages, for example: revealing the intercontinental and cross-species spread of the gammaretroviral ERV-Fc lineage occurring ~33-15 Mya in mammals (Diehl et al. 2016); and the exogenous and endogenous history of a specific gammaretrovirus lineage, transmitted between vesper bats, felid cats, and pangolins, before being independently endogenised in each of these host lineages ~13–25 Mya (Zhuo and Feschotte 2015).

Previous studies have characterised the distribution of ERVs from different retroviral lineages within some bat species, finding the ERV landscape to be dominated by beta- and gammaretroviruses in both yinpterochiropterans and yangochiropterans (Zhuo et al. 2013; A. Hayward et al. 2013; J.A. Hayward et al. 2013; A. Hayward et al. 2015; Jebb et al. 2020), and have identified the rare occurrence of endogenous deltaretroviruses in genomes across Chiroptera (Hron et al. 2019). In this study I carry out the most extensive survey of bat ERVs to date, characterising their phylogenetic distribution across 19 genomes spanning 11 families of bats and characterising the expansions and extinctions of ERV lineages within different bat clades. I then use the identified ERV sequences to determine historical patterns of cross-species retrovirus transmission within Chiroptera, enabling me to test for differences in the rates of cross-species transmission at different phylogenetic levels and between different bat clades.

3.2 Methods

3.2.1 RT alignment and retrovirus reference phylogeny creation

I obtained data from 19 publicly available bat genomes (Appendix B.1.1), which were selected based on the uniform sequencing and assembly method used to produce them. This was necessary as previous analysis showed biases in retroelement composition of genome assemblies depending on assembly method (see Chapter 2). A pipeline based on the DVORFS tool was used to search these 19 genomes for reverse transcriptase domains (RT domains) from endogenous retroviruses with a frameshift-aware strategy (see Chapter 2). RT domain protein HMM profiles from GyDB (Llorens et al. 2011) were used as queries with bit-cutoff of 20.0 and length-cutoff of 30 amino acids. The recovered sequences were split into two categories: 'complete' RT sequences with 180 or more amino acids; and 'fragmented' RT sequences comprising fewer than 180 amino acids.

A manually curated codon-alignment of RT sequences from 83 exogenous retroviruses and well-characterised endogenous retroviruses, as well as retroelement outgroups (Gypsy, Micropia, Ty3, and Del elements), was used as a starting point for providing a high quality alignment of retrovirus RT sequences. First, a custom amino acid substitution matrix for further alignment steps was derived from this manual alignment using ReplacementMatrix (Dang et al. 2011). Next, the 'complete' RT amino acid sequences were added to the initial alignment using the E-INS-i alignment method in MAFFT (version 7.467) using the custom substitution matrix and with the `--add` parameter (Katoh and Standley 2013). Finally, the 'fragmented' RT amino acid

sequences were added to this alignment using MAFFT with the same substitution matrix and the parameters `--addfragments --multipair`. The resulting alignment was then trimmed to contain only columns which were present in the original manually curated alignment and columns that contain less than 10% ambiguous or gap positions.

For downstream analysis, a reference phylogeny of retrovirus RT sequences is required. In order to construct an accurate reference phylogeny with good representation of the evolutionary diversity encompassed by the RT sequences recovered from the bat genomes, the following steps were carried out. First, the trimmed alignment generated by the previous steps was thinned to only retain sequences with over 180 non-ambiguous amino acids. The best fitting amino acid substitution model for this alignment was determined to be 'JTT+F+R6' using ModelFinder (Kalyaanamoorthy et al. 2017) and this model was used for constructing all further phylogenetic trees. An approximate maximum likelihood tree was then built using IQ-TREE (version 1.6.12) with the `-fast` parameter (L.-T. Nguyen et al. 2015).

Next, a phylogenetic thinning algorithm was applied to this tree in order to remove lower quality RT sequences (sequences with more ambiguous or gap positions) that were only a short phylogenetic distance from better quality sequences. Briefly, clusters of RT sequences were identified by searching the phylogenetic tree for subtrees in which the number of descendent branches at a genetic distance of 0.15 substitutions per site from the root node of the cluster was over 50% of the number of tips in the cluster. From each cluster, the sequence with the most non-ambiguous positions was selected as representative, when sequences are tied on this metric, the sequence (i.e. tip) with a distance to the subtree root node closest to the median distance of tips to the subtree root node is selected. All sequences not selected as representative within a cluster are discarded unless they are from the original manually curated alignment. The reduced alignment produced by this algorithm is used to create a maximum likelihood tree using IQ-TREE with 1000 ultrafast bootstrap replicates and nearest neighbour interchange (Hoang et al. 2018). The phylogenetic thinning algorithm followed by maximum likelihood tree generation strategy was iterated once more, until the number of sequences was small enough to be computationally tractable for more computationally intense but more accurate tree building strategies (below 200 sequences).

Using the final thinned alignment, a maximum likelihood tree evaluated with 1000 non-parametric bootstrap replicates was constructed using IQ-TREE and a greedy Bayesian tree was additionally constructed using BAli-Phy (version 3.5, parameters: `-S jtt+f+Rates.free[n=6] -I none`) (Suchard and Redelings 2006). For the

Bayesian tree, MCMC was run for enough iterations such that the effective sample size (ESS) for all estimated parameters was greater than 200 (Nascimento et al. 2017). Both trees gave the same topology of basal nodes, however, the Bayesian tree was able to better reproduce the topology found by previous literature of the *Alpha-* and *Betaretrovirus* clade and was thus selected for use as the retrovirus reference tree.

3.2.2 Phylogenetic placement of RT sequences and Edge PCA comparison

To assess accurately the differences in ERV composition among the focal bat genomes, both the longer extracted RT sequences and the shorter more fragmented RT sequences were examined. The phylogenetic placement tool EPA-ng (version 0.3.6, Barbera et al. 2019) was used to place the 39,980 RT sequences on the reference tree based on the alignment. This strategy is often used in metabarcoding studies for determining the phylogenetic distribution of organisms within a sample using short read data from amplicons, but applies equally well to the analogous situation of determining the phylogenetic distribution of ERVs based on RT sequences extracted from a genome. Phylogenetic placement assigns each 'query sequence' a likelihood weight ratio (LWR) for each branch on the reference tree. This is the relative likelihood that the query sequence would adjoin the reference tree at that branch. Because this strategy incorporates evolutionary history (based on the reference tree) and placement uncertainty into its output, it is a more accurate method for determining phylogenetic composition than BLAST based assignment using reference sequences when dealing with distantly related sequences (Berger et al. 2011). It is thus a particularly good fit for determining ERV composition of a genome as retroviruses have high sequence diversity and copies within a genome are often highly fragmented sequences. After phylogenetic placement, any RT sequences with a total LWR >50% within the outgroup clade were removed.

The distributions of phylogenetic placements were visualised using Gappa (version 0.6.0) (Czech et al. 2020) to count the weight of placements within retroviral clades, to create placement heatmaps showing ERV abundances over the reference retrovirus tree, and to perform edge principal component analysis (edge PCA). Abundance of ERVs across the reference retrovirus tree can be characterised by the sums of placement LWRs on each branch (termed edge masses) and can be visualised with a heatmap. Edge PCA uses the imbalance of edge masses either side of each branch to find branches of the tree across which there is a high level of between-sample heterogeneity, then weights each branch accordingly, before applying classical PCA to identify principal components. These principal components can then be visualised on the original reference tree by the weight of each branch's contribution to the principal component.(Iv and Evans 2013; Czech and Stamatakis 2019).

3.2.3 Inference of horizontal ERV transfers

To infer the occurrence of horizontal ERV transfers and therefore the occurrence of cross-species retroviral transmission, the RT sequences mined from the 19 bat genomes were analysed under a DTL gene reconciliation framework. First, the RT sequences were split into smaller phylogenetic subgroups based on their phylogenetic placement in the previous step. Sequences were included within a subgroup if their accumulated placement LWR within that clade was greater than 0.95 and were further filtered to remove sequences shorter than 200 amino acids.

To produce RT gene trees for reconciliation, the sequences for each group were re-aligned as nucleotide sequences, thereby increasing the accuracy and resolution of the alignments and resulting phylogenetic trees. This was carried out using the E-INS-i alignment method in MAFFT along with a manually selected outgroup for each viral grouping. A maximum-likelihood RT gene tree was then built for each separate group alignment with IQ-TREE using ultrafast bootstrapping with 1000 replicates and the in-built ModelFinder to determine the best fitting nucleotide model for each alignment. Branches on these RT gene trees with a support of less than 75% were collapsed to polytomies. Monophyletic groups of RT sequences from the same bat genome, or polytomies containing only groups of RT sequences from monophyletic bat genomes, were filtered to leave only the sequence with median distance to the group root node. The retained sequences were then used to repeat this tree building and collapsing process and this was iterated until no further RT sequences were removed from the tree.

This procedure was used to reduce the RT subgroup sequence trees to be of a computationally tractable size for tree reconciliation by removing only sequences that are likely to be the result of duplications within a bat genome and therefore retain regions of the tree showing speciation and potential horizontal transfers of RT sequences. By only retaining one RT sequence per bat genome within each polytomy, this process is pessimistic with respect to including parts of the RT tree representing horizontal transfer events, and thus the reconciliation graph produced from such input trees likely provides an underestimate of the number of horizontal transfer events.

The final, reduced, subgroup RT trees were then reconciled with a fully dated species tree of the 19 bat species. The divergence dates and topology for the Noctilionoidea superfamily were obtained from Potter (2019) and divergence dates and topology for the rest of the phylogeny were obtained from TimeTree (Kumar et al. 2017). For each subgroup RT tree, ecceTERA (Jacox et al. 2016) was used to perform a maximum

parsimony reconciliation with the species tree (branches with less than 75% bootstrap support were collapsed to polytomies; the default event costs were used i.e. 2:3:1 for duplication, transfer and loss respectively; strategy S5 from To et al. (2015) was used for exploring Pareto optimal reconciliations; horizontal gene transfers outside the given species tree were not allowed; the symmetric median reconciliation was used for downstream analysis).

A gene tree-species tree reconciliation traces the evolution of a gene tree within the context of a species tree by considering speciation as well as gene duplication, transfer and loss events (DTL). ecceTERA uses a maximum parsimony approach to explore possible reconciliations in which each type of event (duplication, transfer, or loss) is given a cost and reconciliations with the minimum possible total cost are considered part of the set of maximum parsimony reconciliations (MPRs), which are often numerous. From the set of MPRs for a given set of event costs and the surrounding possible reconciliations with slightly adjusted event costs, ecceTERA takes the median reconciliation in order to select an accurate estimate of events. The support for each event in the median reconciliation is defined as the frequency of the particular event within the set of reconciliations considered (T.-H. Nguyen et al. 2013; To et al. 2015).

3.2.4 Statistical modelling of ERV transfer rates

To estimate rates of cross-species retrovirus transmission, the inferred ERV transmission events were used to fit a Poisson GLM. Horizontal gene transfer events (HGTs) with over 90% support were counted in the RT tree–bat tree reconciliations produced for each viral grouping and the counts summarised by branch-pair. HGT counts for each branch-pair were categorised by the superfamily of the donor branch and whether the transfers took place between branches within the same superfamily (intra-superfamily HGT) or between branches from different superfamilies (inter-superfamily HGT). For each branch-pair, the length of time the branches coexisted in millions of years was also calculated.

A Poisson GLM, with a log link function, was then constructed and fitted, using HGT_{count} (count of estimated ERV transfers between a given branch-pair) as the response variable. Following Reitan and Nielsen (2016), *overlap* (branch coexistence time in millions of years) was used as an exposure term so that the modelled property, λ, would represent HGT_{rate} (the rate of ERV transfers per million years for each branch pair). Three explanatory variables were used: *superfamily* (superfamily of the donor branch), *intra*

vs inter (intra- vs inter-superfamily HGT), and *divergence* (branch-pair divergence age). A maximal model with all three variables and all two-way and three-way interactions between them was fitted, and terms were then successively removed and the model refitted until AIC stopped decreasing, resulting in a model that included the terms *superfamily*, *intra vs inter*, *divergence*, and the interaction terms *superfamily* × *intra vs inter*, and *superfamily* × *divergence* (compared to the next best model ΔAIC = -1.98, df = 567). Sequential likelihood ratio tests (LRTs) were also conducted which showed all terms significantly improved the model.

To separately test the differences in ERV transfer rate occurring within the Phyllostomidae and Vespertilionidae families a second model was constructed. For this model, the data were restricted to branch-pairs within the Phyllostomidae and within the Vespertilionidae. A Poisson GLM with a log link function was then constructed using the same response variable and exposure terms as the previous model but with different explanatory variables: *family* (family of the donor branch), and *divergence* (branch-pair divergence age) as well as the interaction between them. The model was fitted (df = 46) and sequential likelihood ratio tests (LRTs) were conducted which showed all terms significantly improved the model. All models were constructed and fitted, and LRTs were carried out, using R (version 3.6.2, R Core Team 2019). Post hoc significance testing of group comparisons was then carried out using the emmeans R package (version 1.5.0, Lenth 2020).

3.3 Results

The DVORFS tool identified 39,980 retrovirus RT sequences or sequence fragments within the 19 bat genomes analysed. Of these, 9,331 sequences matched >180 out of 255 amino acids in the query profile HMM and were used in the construction of a reference RT phylogeny (Section 3.2.1). All 39,980 RT sequences and fragments were placed on this reference phylogeny and any that were placed with over 50% likelihood weight ratio in the outgroup clade were discarded (see Section 3.2.2), leaving 39,968 retrovirus RT sequences distributed across 19 genomes for further analysis.

3.3.1 Differences in ERV composition among bats species

Almost all retrovirus RT sequences identified were placed within ERVL-like, *Gammaretrovirus*, *Betaretrovirus,* or closely related clades, with a very small number

of exceptions that had placements of varying certainty on more basal branches of the reference phylogeny (Fig. 3.1 and Fig. 3.2). Specifically, 109 placements occurred with >50% of their total LWR in the basal branches of class I ERVs or in the *Epsilonretrovirus* clade, two placements occurred with >50% of their total LWR in the basal branches of class II ERVs, one placement occurred with 53% of its total LWR in the *Alpharetrovirus* clade, and one placement occurred with 68% of its total LWR in the *Spumaretrovirus* clade.

Twelve of the nineteen bat species analysed had a total count of between 1200 and 2200 retrovirus RTs, however *L. borealis* and *M. lyra* had much higher counts of 3875 and 6077 retrovirus RT loci, respectively, whereas *P. pipistrellus* and *M. schreibersii* and had much lower counts of only 762 and 516 retrovirus RT loci, respectively. The surprisingly low quantity of retrovirus RT loci detected in *M. schreibersii* may be a reflection of the smaller genome size found in this genus (Smith et al. 2013; Gregory 2020). However, the weak correlation and small effect size of assembly size (in Mbp) on RT locus count (R^2=0.389, θ=2.68) and the lack of correlation of cytometric estimates of genome size and RT locus count suggests this is not an important factor across the breadth of the species in this study. Additionally, no correlation or pattern between RT locus count and sequencing coverage or scaffold N50 was observed, suggesting that the total count differences observed are not likely to be caused by artefacts in specific genome assemblies.

The large total number of RT loci in the *M. lyra* genome is mainly due to a much higher number of RT loci from ERVL-like retroviruses than are found in any other bat genome (1987 edge mass within the ERVL-like clade compared to the next highest 863 *in R. ferrumequinum*) as well as a much higher number of gammaretroviruses than the other yinpterochiropteran genomes included in this study (2088 edge mass within the gammaretrovirus clade compared to the next highest 720 in *H. galeritus*). Specifically, there is a highly expanded number of RT loci from three particular *Gammaretrovirus* lineages: GALV/KORV related; HERV-E/PERV-E related; meg_lyr-1314 related (Fig. 3.2). Additionally, the *M. lyra* genome contains a high abundance of RT loci from the mac_sob-732 *Betaretrovirus* lineage and a high abundance of RT loci from the ERVK-like clade, which are found only in much lower copy numbers in other bats.

Most of the genomes analysed showed an ERV composition that consists mostly of *Gamma-* and *Betaretrovirus* elements, with varying quantities of ERVL-like elements as well as a few elements closely related to betaretroviruses or gammaretroviruses (AB and ERVK-like, and GE clades respectively). An exception was the *M. blainvillei* genome, which was characterised by extremely low numbers of betaretrovirus RT loci

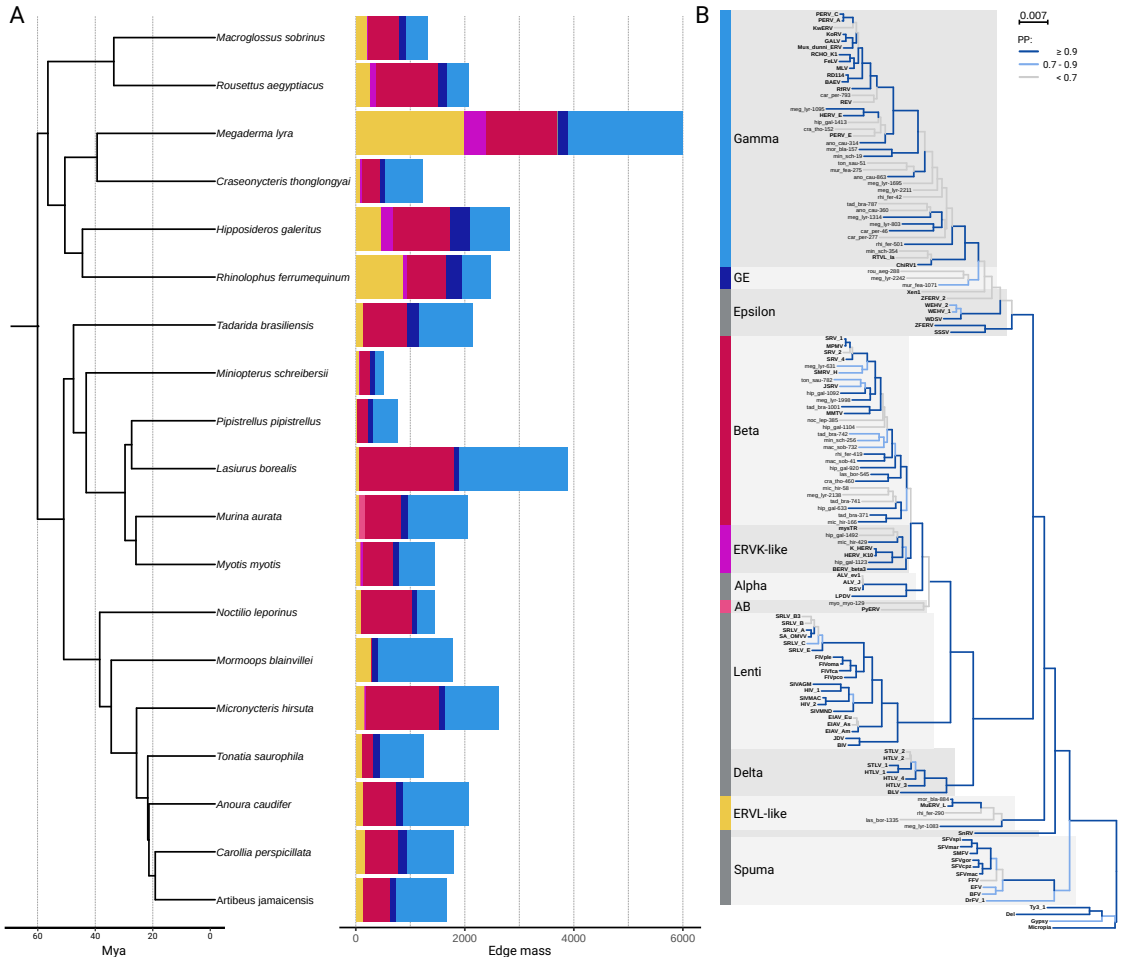


Figure 3.1: Abundance of retroviral clades in the genomes of Chiroptera. **A)** Stacked bar chart with total edge mass for each retroviral clade (based on phylogenetic placement of RT loci on a reference RT tree) for each bat genome. Colours correspond to labelled clades on RT reference tree. **B)** Reference RT tree built using Bayesian inference with branches coloured by posterior probability of associated node.

and other class II retroviruses, with a total edge mass of only 27 within the class II retrovirus group. Additionally, ERVK-like elements appear to be far more abundant within Yinpterochioptera than in Yangochiroptera, and the PyERV-related elements (labelled here as clade AB) appear to be restricted to *M. aurata*, *M. myotis*, and *M. hirsuta*, with the highest abundance in *M. aurata* (106 edge mass).

In order to identify more subtle differences in the ERV composition of each genome, an edge PCA was performed using the phylogenetic placement data. PC1 (69% variance explained) shows that the major differences of ERV composition within the set of genomes is based on the ratio of class I to class II ERVs (Fig. 3.4). Within this variance there is a weak phylogenetic trend, with members of a specific family tending to cluster together. However, at the superfamily level, members of the Noctilionoidea are spread across the

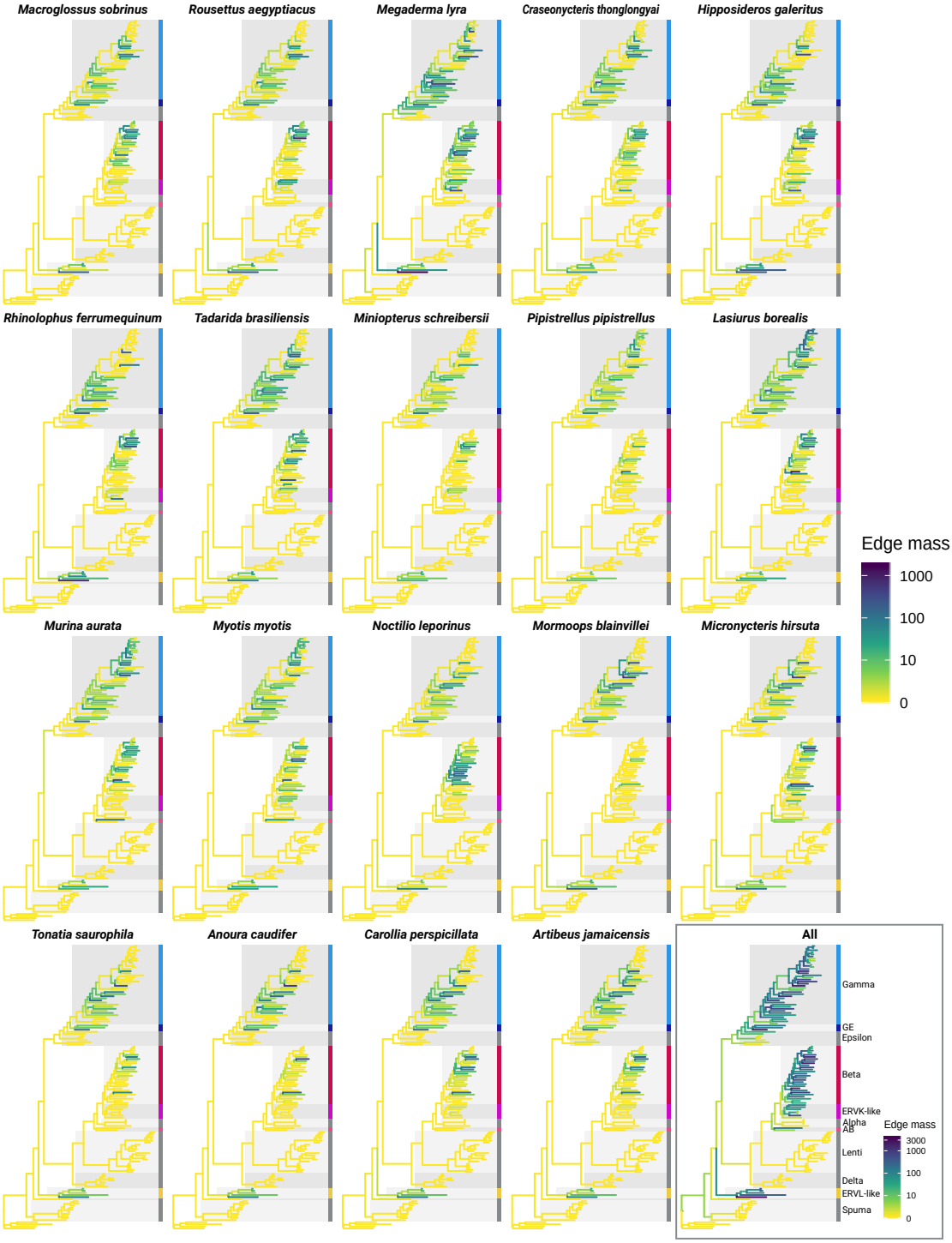


Figure 3.2: Heat map of edge masses across the reference RT tree (based on phylogenetic placement of RT loci) for each bat genome and for all genomes summed.

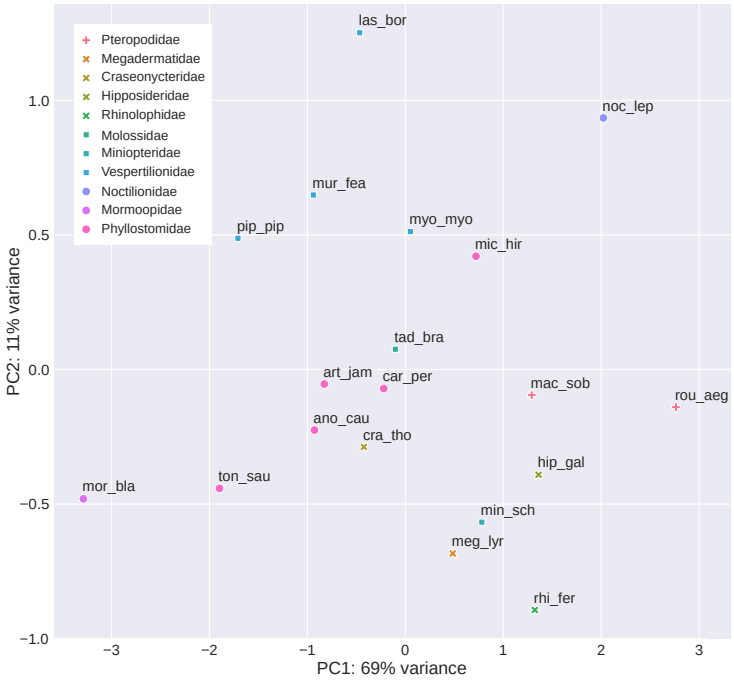


Figure 3.3: Edge PCA plot of bat genomes based on phylogenetic placement of RT loci on the reference RT tree, showing the first two principal components.

PC1 axis (indicating they are diverse in their class I : class II ERV ratio) whereas members of the the Vespertilionoidea are clustered more tightly on this axis (indicating low variability of class I : class II ERV ratio within this clade of bats). PC2 and PC3 explained 11% and 7% variance, respectively, and are both somewhat weighted across edges that define ratios of class III:class I or II ERVs, although each of these PCs clusters the phylogenetic groups of bats in a different way. PC2 appears to mostly highlight the differences in the ratio of class III:class II ERVs, where the grouping of Yinpterochiroptera on the left of the axis shows that relative abundance of class III ERVs is constant within this bat clade. PC3 is mostly weighted on the less basal branches of *Betaretrovirus*, highlighting the expansion of particular betaretroviral lineages in *R. ferrumequinum*, and in *M. lyra*.

3.3.2 Rates of cross-species retrovirus transmission in bats over evolutionary time

To estimate occurrence rates of cross-species transfer of retroviruses amongst bats, the more complete RT sequences identified in the genomes were used to construct trees of RT sequences that were then reconciled with a dated species tree for bats under a maximum parsimony framework (Section 3.2.3 for details). The resulting estimated evolutionary history of the RT sequences in the context of bat divergence provides a predicted sequence of duplications, losses, and horizontal gene transfers (HGTs) that

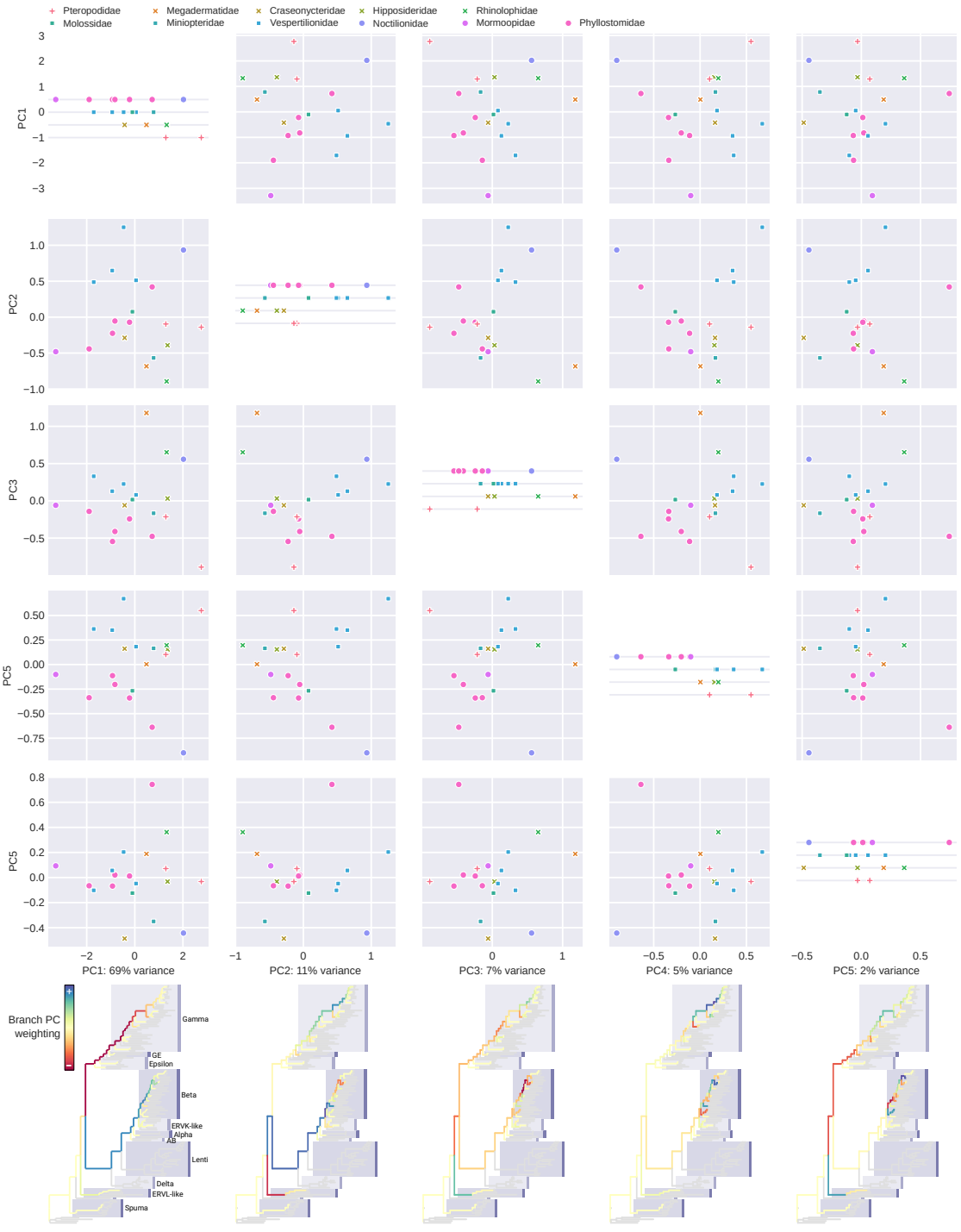


Figure 3.4: Pairwise edge PCA plots of each bat genome based on phylogenetic placement of RT loci on the reference RT tree. Branch weighting for each principal component is visualised on the bottom row.

led to the current distribution of RT loci within the analysed bat species. Fig. 3.5 shows the counts of horizontal ERV transfers, predicted by this reconciliation (with over 90% support), between each pair of branches in the bat phylogeny. This can be reasonably understood as a prediction for the number of retroviruses which were transmitted across a species-pair, and which had descendants that endogenised and became fixed within the species involved. Due to the rarity of such a sequence of events, these counts represent a vast underestimate of the true numbers of retrovirus cross-species transmissions, but do provide insight into their relative rates of occurrence.

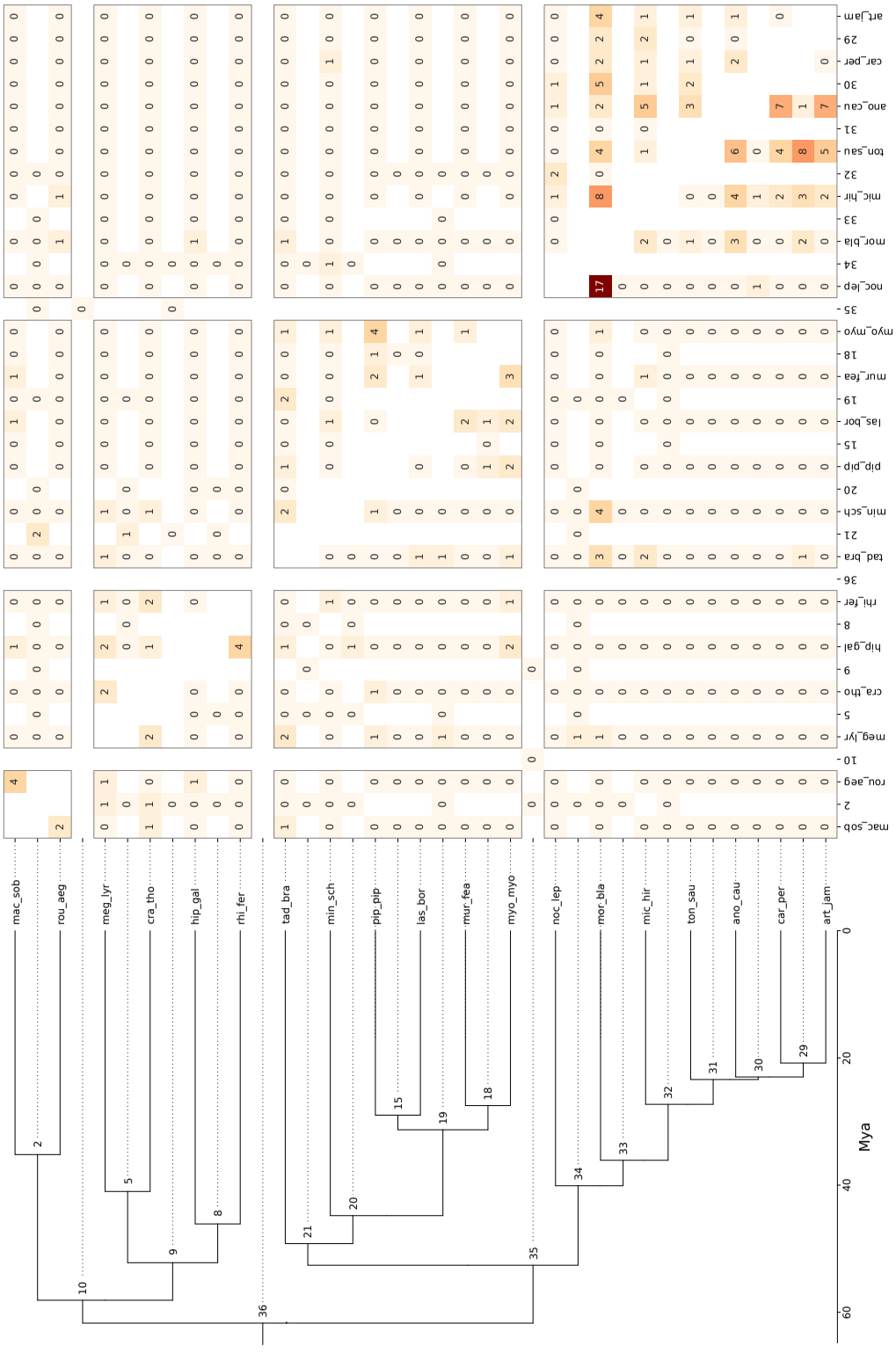


Figure 3.5: Matrix showing counts of inferred horizontal ERV transfers between nodes of the bat phylogeny determined through gene reconciliation analysis (using a 90% support threshold). Blank cells represent branch-pairs which have no chronological overlap.

At the 90% support level, most predicted transmission events (Fig. 3.5) were seen to occur between taxa that were more closely related, or which had overlapping current or historical distributions. Dividing the bat species into the superfamilies (Vespertilionoidea, Noctilionoidea, Rhinolophoidea, and Pteropodoidea - composed of the single family Pteropodidae) shows marked differences in the numbers of transfers taking place between different groups. Specifically, Pteropodidae, as well as Rhinolophoidea and Vespertilionoidea superfamilies, show similar patterns to each other, with low numbers of predicted transfers between the three groups and a relatively higher number of predicted transfers within each group than between. Interestingly, the *T. brasiliensis* branch (within Vespertilionoidea) stands out as having been involved in a total of 21 transmissions across all four of the superfamilies represented in this study.

Noctilionoidea appears to have a substantially higher number of cross-species transfers within the superfamily than the other superfamilies. However, this difference may reflect the over-representation of a single family (Phyllostomidae) within this group, biasing the set of noctilionoid branch pairs towards more closely related pairs than those represented in Rhinolophoidea and Vespertilionoidea. Within Noctilionoidea, the *M. blainvillei* branch has particularly high counts of predicted transfers to phyllostomid branches and a striking 17 transfers to the *N. leporinus* branch. The *M. blainvillei* branch also has an unusually high number of transfers outside its superfamily, having been involved in four transmission events with each of the *T. brasiliensis* and *M. schreibersii* branches.

To further examine the between-superfamily differences in frequency of cross-species ERV transfer, I fitted a Poisson GLM to this data to make quantitative contrasts (Fig. 3.6, Table 3.1, Table 3.2). The model predicts that overall, less phylogenetically related bats have a lower rate of ERV transfer between them – the estimated marginal mean (EMM) coefficient for the trend of *divergence* is -0.0405 (Table 3.1). Additionally, pairwise superfamily comparisons show significant differences in the model estimates for the slope of this trend between Noctilionoidea and each of Vespertilionoidea and Rhinolophoidea (Table 3.2). This suggests that the ERV transfer rate varies more with phylogenetic relatedness of noctilionoid species than with vespertilionoid species or rhinolophoid species and that recently diverged species of noctilionoid bats will have higher ERV transfer rates than recently diverged species of vespertilionoid or rhinolophoid bats. A model constructed with an additional interaction term, *intra vs inter* × *divergence*, was fitted but the extra term did not significantly improve over the first model according to a LRT (*P*-value = 0.9049) or AIC score (ΔAIC = +1.98). This implies that phylogenetic relatedness

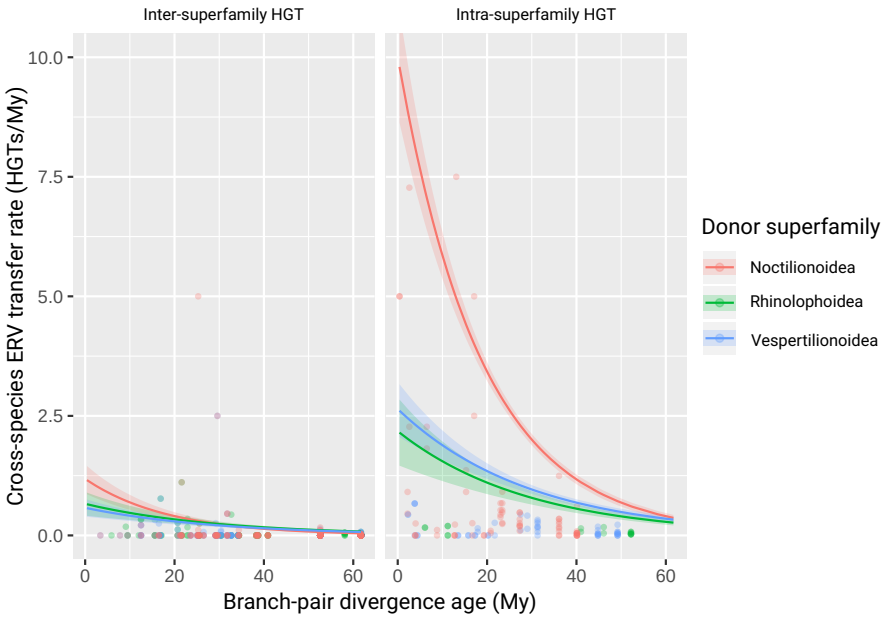


Figure 3.6: Rate of ERV transfers (HGTs) between two branches dependent on divergence age of the branch-pair. Split by the superfamily of the donor branch (colours) and whether the recipient branch is within the same superfamily as the donor branch (intra-superfamily HGT, right panel), or outside the superfamily of the donor branch (inter-superfamily HGT, left panel). Poisson regression model predictions shown as lines (shaded area is standard error). Rates based on inferred HGT counts and branch-pair overlap time shown as points.

Model term (Group)	EMM / Trend	SE
inter vs intra (inter)	-0.954	0.1031
inter vs intra (intra)	0.663	0.0701
superfamily (nocti)	0.0362	0.0968
superfamily (rhino)	-0.2561	0.126
superfamily (vesper)	-0.2175	0.0943
divergence (overall)	-0.0405	0.00334

Table 3.1: Estimated marginal means (EMMs) or trends for each group within each term in the Poisson GLM of HGT_{count} (calculated on a log scale due to the log link function). EMMs represent the estimated marginal average of the response variable for a group within a categorical term in the model, averaged over all other terms. Trends represent the estimated slope of of the response variable trend across a continuous term in the model, averaged over all other terms. Bat superfamily names abbreviated as Noctilionoidea - nocti, Vespertilionoidea - vesper, Rhinolophoidea - rhino.

Group comparison Within group: Contrast groups	Est. difference ratio of HGT_{rate}	Est. difference of log response / trend	Log SE	Z-ratio	*P*-value
Overall difference between inter vs intra:					
overall: inter - intra	0.198	-1.62	0.126	-12.828	**<.0001**
Differences between inter- vs intra-superfamily HGT within each superfamily:					
nocti: inter - intra	0.118	-2.14	0.202	-10.562	**<.0001**
rhino: inter - intra	0.304	-1.19	0.245	-4.865	**<.0001**
vesper: inter - intra	0.219	-1.52	0.205	-7.427	**<.0001**
Differences between each superfamily within inter- or intra-superfamily HGT:					
inter: nocti - rhino	0.835	-0.1798	0.259	-0.694	0.7668
inter: nocti - vesper	0.947	-0.0543	0.247	-0.22	0.9737
inter: rhino - vesper	1.134	0.1256	0.251	0.5	0.8715
intra: nocti - rhino	2.147	0.7643	0.184	4.147	**0.0001**
intra: nocti - vesper	1.753	0.5616	0.130	4.305	**<.0001**
intra: rhino - vesper	0.816	-0.2028	0.193	-1.048	0.5466
Differences of divergence trend between each superfamily:					
nocti - rhino	-	-0.01962	0.00814	-2.409	**0.0423**
nocti - vesper	-	-0.01987	0.0073	-2.724	**0.0177**
rhino - vesper	-	-0.000251	0.009	-0.028	0.9996

Table 3.2: Estimated group differences of estimated marginal means (EMMs) for comparisons within the Poisson GLM of HGT_{count} (calculated on a log scale due to the log link function). Estimated difference ratio is the exponentiated estimated log difference which corresponds to the ratio of the effective mean rates being compared. Bat superfamily names abbreviated as Noctilionoidea - nocti, Vespertilionoidea - vesper, Rhinolophoidea - rhino.

has the same scaling effect on HGT_{rate} regardless of whether the transfer is occurring within a superfamily or between superfamilies.

Comparison of EMMs shows that the rate of intra-superfamily ERV transfers predicted by the model are significantly higher than inter-superfamily ERV transfers across all branches when branch divergence age is accounted for (effective mean difference ratio intra/inter = 5.05, *P*-value < 10^{-4}), however, the size of the effect is dependent on the superfamily of the donor branch (Table 3.2).

Interestingly, the model shows that ERV transfers between branches within the Noctilionoidea occur at a rate ~1.8 times higher than ERV transfers within the Vespertilionoidea, and ~2.1 times higher than within the Rhinolophoidea when branch divergence age is accounted for (*P*-value < 10^{-4} and *P*-value = 10^{-4} respectively). Because of the over-representation of Phyllostomidae and Vespertilionidae families within their respective superfamilies within the dataset, it is likely that a substantial portion of the differences in rates attributed to the these superfamilies reflects the higher number of inter-family branch-pairs within these groups. Although divergence ages (a proxy for phylogenetic relatedness) are accounted for in the model, a fairer comparison is

possible by restricting the comparison to the intra-family ERV transfer rate of Phyllostomidae and Vespertilionidae. This enables a specific comparison of ERV transfer rates between two groups with more comparable phylogenetic diversity. A second Poisson GLM model was fitted to counts of transfers within the he Phyllostomidae and Vespertilionidae families (Section 3.2.4), which predicted ERV transfers between branches within the Phyllostomidae family occur at a rate ~2.0 times higher than ERV transfers within the Vespertilionidae family when averaged over the effect of branch divergence age (P-value = 0.0007).

3.4 Discussion

3.4.1 An unbiased analysis of ERV activity across Chiroptera

In recent years, the number of available bat genome sequences has increased dramatically, in part due to activities of sequencing consortia such as Bat1k and the Vertebrate Genome project (Koepfli et al. 2015; Teeling et al. 2018). Previous studies have comprehensively examined ERV composition within *M. lucifigus*, and *P. vampyrus* (J.A. Hayward et al. 2013; Zhuo et al. 2013; A. Hayward et al. 2015) and others have included ERV analysis of these two genomes as part of broader comparative analyses across vertebrates (A. Hayward et al. 2015). A recent study also carried out a preliminary examination of ERV composition across six bat genomes generated by the Bat1k consortium (Jebb et al. 2020). However, to date, there have been no comparative and comprehensive examinations of ERV composition or ERV activity among the diversity of bat genomes that are now available. This study is thus the first to provide such an analysis by robustly comparing the phylogenetic distribution of retrovirus RT domains present in the genomes of 19 species within 11 families of bats.

As shown in Chapter 2, the amount of detectable ERV sequence within a genome assembly appears to be highly affected by sequencing method and assembly method (also see R.R. Love et al. 2016), which is unsurprising given that unassembled regions are enriched for repetitive elements such as ERVs and other transposable elements. Thus, here I restricted analyses to a set of genomes with comparable sequencing and assembly methods, thereby reducing systematic biases, and allowing fair comparison between genomes. The ERV detection method used in this study relies on the presence of intact or fragmented RT domain sequence in the genome assembly, which allows for fine-grained phylogenetic analysis across the broad diversity of ERVs. As a result, however, the count estimates produced by this method do not represent the total ERV content of

the genomes, since the method does not count ERVs with missing RT domains or ERV-derived solo-LTRs. It is thus important to interpret the counts as a large underestimate of the true ERV content within the genomes studied, and, although the RT proportions are likely representative of total ERV content for younger ERVs, particularly ancient lineages of ERVs may be under-represented. This is underscored by the findings of Zhuo et al. (2013), who showed that within the *M. lucifugus* genome assembly (Myoluc2.0, accession: GCA_000147115.1), class III ERVs constituted 34% of full length ERVs but 64% of the total number of ERV copies when solo-LTRs were counted. This discrepancy is attributed to a particularly ancient lineage of class III ERVs represented mostly by solo-LTRs. In the present study, the RT counts and edge mass distributions across the RT tree therefore describe more recent trends in the activity of different ERV lineages rather than overall ERV content of each genome.

3.4.2 Numerous expansions and extinctions of bat ERV lineages

The total number of ERV-derived RT domains identified in each bat genome ranged from 516 in *M. schreibersii* to 6077 in *M. lyra*, however, most of the genomes analysed had total counts between 1200 and 2200, suggesting that the overall activity of ERVs within most bat genomes is similar, aside from a few outliers. Similarly, the fact that almost all the RT domains identified originate from the same three retrovirus clades (*Gammaretrovirus*, *Betaretrovirus* and ERVL-like retroviruses) with very few exceptions, shows that retroviruses from other genera have not invaded any bat genomes with much success. This result agrees with the previous finding that other groups of retroviruses are also generally seldom found as ERVs in vertebrate species (A. Hayward et al. 2015). The placement of the RT loci detected here that do potentially fall outside these three clades is consistent with previous evidence of single copy endogenous deltaretrovirus ERVs in bats (Hron et al. 2019), as well as the tentative identification of endogenous *Alpharetrovirus*, *Lentivirus* and *Spumaretrovirus* elements in bats (Jebb et al. 2020).

Although gammaretroviral and betaretroviral ERVs dominate in most of the genomes analysed here, the absolute number of ERVL-like RT loci does vary markedly across Chiroptera, and is notably lower in the Vespertilionidae and Miniopteridae families than in the other bat families. This agrees with previous analysis of ERVs in the *M. lucifugus* genome in which only a low number of complete ERVs, but a high number of solo-LTRs, from the ERVL-like clade were identified (Zhuo et al. 2013). This perhaps implies that this lineage of bats has managed to control the replication of ERVL-like ERVs within their genomes. In the Vespertilionidae, the suppression of this group of ERVs has occurred

despite overall ERV activity remaining at similar or increased (in the case of *L. borealis*) levels compared to other bats.

The genomes of *M. blainvillei* and *T. saurophila* also stand out, appearing to have suppressed the expansion of specific ERV clades. Both of these species contain an exceptionally low proportion of RT loci derived from betaretroviruses and other class II retroviruses, indicating that recent activity of class II ERVs is extremely limited. Remarkably, in *M. blainvillei*, the quantity of class II-derived RT loci is so low that a comprehensive extinction of class II ERVs appears to have occurred.

In comparison, a small number of all ERV RT loci was identified in *M. schreibersii* (n = 516), indicating that the activity of all ERV groups is drastically reduced in this species compared to other bats. Interestingly, this coincides with the fact that *Miniopterus* bats have smaller genomes than most other bats (1.9-2.0 Gbp compared to well over 2 Gbp in most other species, Smith et al. 2013; Gregory 2020). It is therefore tempting to speculate that the smaller genome size in this group of bats may be related to this observed lack of ERV activity. However, a previous analysis identified that a total of 3.86% of the closely related *M. natalensis* genome is occupied by ERVs (Platt et al. 2016), which is a similar proportion to bats in both the Mormoopidae and Vespertilionidae families. This suggests that the ERV content of *Miniopterus* species constitutes an expected proportion of their genomes rather than an unusually small proportion. This, in turn, is consistent with the hypothesis that small genome sizes in bats are due to increased rates of DNA loss rather than lower transposable element activity (Kapusta et al. 2017).

In clear contrast, *M. lyra* has a remarkably high quantity of ERVL-like RT loci, possessing an edge mass within the ERVL-like clade over twice as high as any other bat and over 25 times higher than *C. thonglongyai*, its closest relative among the bats analysed in this study. This indicates a that major expansion of ERVL-like ERVs has occurred in the Megadermatidae family since the split from the Craseonycteridae ~41.0 Mya, and, given that observed RT loci are biased towards younger ERV copies, these elements likely continue to be highly active in *M. lyra*. Although less pronounced, the quantity of *Gammaretrovirus* RT loci present in *M. lyra* is also substantially higher than in any other yinpterochiropteran genomes, with the expansions of three particular lineages being responsible (GALV/KORV related, HERV-E/PERV-E related, and meg_lyr-1314 related gammaretroviruses). Furthermore, although the total quantity of class II RT loci identified in *M. lyra* is similar to other yinpterochiropterans, there is a high quantity of loci derived from the ERVK-like clade and a high quantity of loci derived from a specific lineage of betaretroviruses (mac_sob-732), each of which is found only in low

copy numbers in other bats. Taken together, these six unique and ongoing ERV expansions within the *M. lyra* genome across all three classes of retrovirus suggest a broadly unhampered ERV replication environment within this bat species. Similarly, the ERV community within the *L. borealis* genome also appears to have expanded through a number of smaller ERV lineage expansions occurring across the beta- and gammaretroviral clades. Finally, a notable expansion of PyERV-related retroviruses also appears to have occurred in the *M. myotis*, and *M. aurata* lineage.

In summary, a number of ERV lineage expansions and extinctions of varying breadth across different ERV classes have occurred in different bat lineages, however, as with the discovery of the *Helitron* transposable element revolution in Vespertilionidae (Platt et al. 2014; Platt et al. 2016), the specific causes of these widespread changes are currently unclear and will require substantial future work to elucidate.

3.4.3 RT sequence gene reconciliation reconstructs the retrovirus sharing history of Chiroptera

The large body of ERVs within host genomes contains information about past retrovirus infections of those hosts. In this study, I leverage this retrovirus history encoded in bat genomes in order to systematically examine cross-species retrovirus transmission within Chiroptera across evolutionary time. To do this, I used the sequences of the RT loci identified in the previous step to infer the occurrence of horizontal transfers of ERVs between bat species using a gene reconciliation approach. At the 90% confidence level, gene reconciliation (which is based on parsimony) is able to give accurate estimates of the total number of horizontal transfer events (T.-H. Nguyen et al. 2013). At the same time, however, some false positives may arise, potentially due to incorrect tree topologies, a lack of locus detection in a particular genome, or the effects of incomplete lineage sorting (ILS) on the phylogenetic distribution of ERV loci (Stolzer et al. 2012). Despite such false positives, the broader numerical trends should be robust to these scenarios. In the data examined here, each detected ERV transfer corresponds to a retrovirus which was transmitted from one species to another and which had descendants that became endogenised and independently fixed within both the donor and receiver host lineages. Thus, given the rarity of provirus/ERV fixation, the proportion of all historical, cross-species, retrovirus transmissions that can be detected with this method is extremely small. Nonetheless, given the assumption that ERV fixation rates are the same across host species, the relative frequencies of these events provides information about the differences in the true frequency of cross-species retrovirus transmission between different host species.

Scrutinising the estimated counts of ERV transfers in Fig. 3.5 shows outlier branch-pairs that may have inflated estimates. These estimates may be inflated due to ILS, such as the extraordinarily high number of predicted transfers from *M. blainvillei* to *N. leporinus*. In other cases, the cause of an inflated estimate is unclear, but the high number of transfer events for a branch-pair may seem unlikely due to a lack of plausible sympatry between the modern-day or ancestral populations. The main example being the relatively high number of predicted transfers between *M. blainvillei* (Mormoopidae are endemic to South America) and *M. schreibersii* (Miniopteridae have exclusively Old World distributions).

Encouragingly, there is a notable dearth of transfers between the noctilionoid branches (which evolved in the New World) and the other branches analysed in this study, which are mostly of Old World origin, suggesting that inferred transfer events are plausible. Additionally, the branch representing the *T. brasiliensis* lineage is involved in six predicted transfers to the noctilionoid branches as well as a number of predicted transfers to bats with extant and ancestral Old World distributions. Interestingly, while *T. brasiliensis* has a New World distribution, it is likely to have an Old World origin, having split from its African sister taxa ~18Mya (Ammerman et al. 2012). Thus this lineage has plausibly had geographic overlap both with the populations represented in the noctilionoid branches as well as the populations with Old World distributions represented in the branches outside Noctilionoidea. This further illustrates the plausibility of the inferred horizontal ERV transfer events.

3.4.4 Phylogenetic relatedness impacts the rate of retrovirus transmission

By fitting a Poisson-based regression model to the counts of inferred ERV transfer events it was possible to quantitatively assess the hypothesis that more distantly related branches (i.e. branches with a greater time since divergence) would have a lower rate of cross-species retroviral transmission between them due to increased genetic barriers. The model predicted a significant negative correlation between ERV transfer rate and branch divergence. This recapitulates previously observed host-virus evolutionary dynamics in which more closely related species are known to share more viruses between them (Streicker et al. 2010; Olival et al. 2017; Albery et al. 2020).

I also show that after accounting for phylogenetic relatedness of the host species (branch-pair divergence time), my model predicts that the rate of retrovirus transmission between the three superfamilies considered is significantly lower than within

each superfamily (3.2). This is particularly interesting as it suggests that additional factors present within a superfamily determine the rate of retroviral cross-species transmission. Moreover, it could indicate that the major splits at the base of the chiropteran lineage led to particularly pronounced physiological or immunological differences at the superfamily level, which create a barrier for viral transmission between these groups. The relatively low number of inferred transfers between the root branches of each superfamily are perhaps a sign of this, although this could be a result of the bias towards detection of younger ERVs in this analysis, which in turn biases against the detection of ERV transfers between branches closer to the root of the tree.

Previous studies that have examined inter-order viral transmission of extant viruses have predicted that phylogenetic relatedness is only important for determining likelihood of transmission within taxonomic orders, and is not an important factor for likelihood of transmission between taxonomic orders (Albery et al. 2020; Mollentze and Streicker 2020). In this context, the presence of a superfamily effect in Chiroptera suggests that a significant boundary for cross-species retrovirus transmission may occur at the host *superfamily* taxonomic level in addition to, or instead of, at the *order*-level, at least within Chiroptera.

3.4.5 Increased rate of cross-species retrovirus transmission within noctilionoid bats

Finally, the Poisson regression model also predicted differences in the cross-species retrovirus transmission occurring within different superfamilies. The model predicted Noctilionoidea to possess a significantly higher level of cross-species ERV transfers within the superfamily than the Vespertilionoidea or the Rhinolophoidea. A separate Poisson regression model showed that the detected difference was also present at the family level between Phyllostomidae and Vespertilionidae, wherein Phyllostomidae was predicted to have twice the rate of intra-family ERV transfer compared with the Vespertilionidae. Additionally, the first model did not show any significant differences between donor superfamilies for inter-superfamily transfers.

Taken together, this suggests that ERV transfer has generally been occurring at a much higher rate within the Noctilionoidea than within either the Vespertilionoidea or the Rhinolophoidea. However, the propensity for noctilionoid species to transmit viruses across species-barriers does not extend to transmission outside of the superfamily. The most likely explanation for this result is that it simply reflects the geography of the populations examined; notably, most members of the Noctilionoidea are endemic to

the Neotropics with many species in sympatry with one another, whereas the Vespertilionoidea and Rhinolophoidea have much more dispersed ranges found across multiple continents. This geographic effect, and specifically the degree of sympatry, has previously been shown to be one of the most important factors determining cross-species viral sharing (Luis et al. 2015; Olival et al. 2017; Albery et al. 2020; Mollentze and Streicker 2020). Alternatively, such an effect could potentially be explained by the higher proportion of short branches in the noctilionoid phylogeny which could cause inflation of inferred transfer events due to the effects of ILS on the RT reconciliation analysis (Stolzer et al. 2012). A final possibility is that physiological or ecological traits, such as high rates of frugivory within the Noctilionoidea (Rojas et al. 2012), have contributed to an increased rate of cross-species viral transmission within this group, although adequately disentangling such effects from the effects of phylogeny would need much higher sample sizes (also see Luis et al. 2015).

4

Non-retroviral EVEs in bat genomes

Contents

Abstract **91**
4.1 Introduction **91**
4.1.1 Bats harbour a diversity of viruses 91
4.1.2 Non-retrovirus EVEs are rare in vertebrate genomes 92
4.1.3 Bornaviruses and bornavirus-derived elements 92
4.1.4 Filoviruses and filovirus-derived elements 94
4.1.5 Parvoviruses and parvovirus-derived elements 96
4.1.6 Non-retroviral EVEs in bat genomes 97
4.2 Methods **98**
4.2.1 EVE mining 98
4.2.2 Orthology clustering 99
4.2.3 Sequence alignment and phylogenetic tree construction 100
4.2.4 Selection analysis 101
4.2.5 Structural protein modelling 101
4.3 Results **102**
4.3.1 Bornavirus-derived elements in bat genomes 103
4.3.2 chEBL elements with conserved ORFs 107
4.3.3 Filovirus-derived elements in bat genomes 109
4.3.4 Parvovirus-derived elements in bat genomes 114
4.4 Discussion **117**
4.4.1 Ancient cultervirus and carbovirus infections in bats 117
4.4.2 Phylogenetic distribution of bornavirus elements among bats . 118
4.4.3 Phylogenetic distribution of filovirus elements among bats . . 120
4.4.4 Phylogenetic distribution of parvovirus elements among bats . 121
4.4.5 Putatively co-opted EBL and EFL elements in bats 122

Abstract

Endogenous viral elements (EVEs) derived from viruses other than retroviruses are found in low copy numbers within eukaryotic genomes, having resulted from rare genome integration events during viral infections in the ancestors of modern species. In this chapter, I survey the genomes of 36 bat species from 11 bat families for non-retroviral EVEs, to comprehensively characterise the non-retroviral EVE repertoires of bat species from across Chiroptera, to produce the most phylogenetically extensive study of non-retroviral EVEs in bats to date. I identify many novel EVE loci across bats, derived from bornaviruses, filoviruses, and parvoviruses, and find no evidence of non-retroviral EVEs derived from other virus families. Furthermore, I use the identified EVEs to gain insight into the evolutionary history of these three viral families. EVEs derived from the *Bornaviridae* provide new information about the biogeography and host-range of the poorly characterised *Carbovirus* and *Cultervirus* genera, suggesting a long history of association between bats and these viruses, and providing evidence of a potential African origin followed by bat-mediated dispersal of carboviruses throughout southern continents. Additionally, a group of syntenic EVEs derived from *Tetraparvovirus* is found across the Chiroptera, and other members of the Laurasiatheria, implying a minimum age of approximately 76 million years for the *Tetraparvovirus* genus. Finally, I also investigate examples of EVEs with conserved open reading frames (ORFs), and detail a putatively co-opted EVE (named *chEBLM*) present in nine species of pyllostomid and mormoopid bats that is derived from a bornavirus *M* gene.

4.1 Introduction

4.1.1 Bats harbour a diversity of viruses

Bats account for over twenty percent of all mammal species (Burgin et al. 2018), and between them harbour a correspondingly rich diversity of viruses, being host to a particularly high diversity of *Coronaviridae*, *Lyssavirus*, and *Paramyxoviridae* (L. Chen et al. 2014; Beltz 2018). Although the alpha species richness of bats is highest in South America, modern bat species are present across all continents except Antarctica, having spread across the globe and diversified to fill a plethora of ecological niches in an evolutionary radiation that took place during the early Eocene (Teeling et al. 2005; J.J. Shi and Rabosky 2015). The order Chiroptera comprises two suborders: Yinpterochiroptera, found only across the Old World, consisting of the Pteropodidae family and Rhinolophoidea superfamily, and Yangochiroptera, found across both the Old World and New World, consisting of the Rhinolophoidea, Emballonuroidea, Noctilionoidea, and Vespertilionoidea superfamilies (Teeling et al. 2005; Jones and Teeling 2006).

4.1.2 Non-retrovirus EVEs are rare in vertebrate genomes

Endogenous viral elements (EVEs) account for 5-15% of the total DNA content of animal genomes. While the vast majority of EVEs are retrovirus-derived (Smit et al. 2013), low probability events occurring over millions of years have led to the accumulation of small numbers of non-retroviral EVEs in animal genomes. These non-retroviral EVEs rarely exceed twenty or thirty loci within a given genome (Katzourakis and Gifford 2010; Aiewsakun and Katzourakis 2015) and the key difference that leads to this large disparity in frequency is the obligate nature of genomic integration during the retrovirus replication cycle compared with the lack of any such stage in other vertebrate viruses. The integration of non-retrovirus genetic sequence into a host cell genome can occur through two main routes. First, mRNA generated by the virus can be captured by the retrotranscribing machinery of retroelements that are present in the cell, such as LINEs and ERVs. The DNA generated by this reverse transcription may then be concatenated with the fragments of the retroelement, after which, mechanisms of retroelement integration can incorporated this DNA into the host cell genome (Belyi et al. 2010b). Second, in the case of DNA viruses, a second pathway to integration involves direct homologous recombination of the viral genomic DNA with the chromosomal DNA of the host cell (Belyi et al. 2010a). Where such integration events occur in a germline cell, then the integrated sequence will be vertically inherited by offspring and may eventually become fixed in a population. EVEs derived from a diverse range of viral lineages, representing every group of virus in the Baltimore classification system, have been identified across numerous vertebrate genomes (Aiewsakun and Katzourakis 2015). Interestingly, despite over 30 families of viruses having been detected in bats (L. Chen et al. 2014), EVEs from only three families of viruses (excluding retroviruses) have been found in bat genomes to date: *Parvoviridae* (ssDNA), as well as *Filoviridae* and *Bornaviridae* both from the order *Mononegavirales* (-ve RNA) (Liu et al. 2011; Edwards et al. 2018; Horie and Tomonaga 2019; Jebb et al. 2020).

4.1.3 Bornaviruses and bornavirus-derived elements

Bornaviruses (family *Bornaviridae*) are non-segmented, negative-strand RNA viruses, characterised by a viral envelope and a genome of approximately 9 Kbp in length that encodes six genes (see Table 4.1). Uniquely among RNA viruses, Bornaviruses establish a persistent intranuclear infection, which they achieve by attaching to host chromatin (Matsumoto et al. 2012). The family is named after Bornavirus Disease Virus (BDV),

Gene name (Protein)	Function
N (Nucleoprotein)	Nucleoprotein which encapsidates the viral RNA to form the viral nucleocapsid.
L (RNA-directed RNA polymerase L)	RNA-dependent RNA polymerase (RdRp) which is responsible for replication of the RNA genome.
P (Phosphoprotein)	Phosphoprotein which acts as a cofactor to the viral polymerase and also as a phosphorylation decoy involved in pathogenicity.
M (Matrix protein)	Structural protein which lines the inside of virions and attaches the nucleocapsids to the cellular membrane during viral assembly and budding. M proteins are also known to be involved in regulation of RNA replication in other negative-strand RNA viruses.
G (Envelope glycoprotein p57)	Envelope glycoprotein involved in virion attachment to a cell receptor and subsequent fusion of the viral envelope and cell membrane.
X (X protein)	Non-structural protein which is required for viral replication, acting as a regulator of the viral polymerase, and additionally acts as an inhibitor of apoptosis in the central nervous system

Table 4.1: Overview of genes comprising the genomes of the *Bornaviridae*.

which in turn received its name from the town of Borna, Germany where it was originally identified in the nineteenth century as the cause of an epidemic amongst the town's horses (Horie et al. 2013). Viruses within *Bornaviridae* are mostly known to infect mammals and birds, though pythons are reported to be hosts for two species of bornavirus (Hyndman et al. 2018). BDV has also been shown to cause fatal encephalitis in humans (Liesche et al. 2019), and the endemic nature of the disease in livestock in European regions suggests zoonotic transmission from a reservoir species is likely. Recently the family has been expanded due to the discovery of Wuhan sharpbelly bornavirus (WhSBV) in the sharpbelly (a ray-finned fish), thus implying a much broader host range than had been previously sampled (M. Shi et al. 2018). Consequently, the family now comprises three genera: *Orthobornavirus* (within which BDV and most other known bornavirus species are classified); *Carbovirus*; and *Cultervirus* (Amarasinghe et al. 2019). Based on alignments of the L protein, *Cultervirus* appears to be basal to *Carbovirus* and *Orthobornavirus* (Amarasinghe et al. 2019). However, due to the present lack of samples in both *Carbovirus* and *Cultervirus*, little is known about host-range, biogeography, or evolutionary history of these genera.

Previous studies have detected endogenous bornavirus-like elements (EBL elements) in a wide range of vertebrate genomes showing a broad diversity of bornaviruses beyond all sampled exogenous bornaviruses (Katzourakis and Gifford 2010; Horie et al. 2013; Cui and L.-F. Wang 2015; M. Shi et al. 2018). In fact, the discovery of orthologous bornavirus-like elements in the genomes of members of the Afrotheria has provided an estimated minimum age of 65 My for some clades of bornaviruses (Horie et al. 2013). Similarly to other endogenous mononegaviral elements, all EBL insertions detected in genomes to date (to my knowledge) do not contain the entire genome sequence of the bornavirus.

Instead, EBLs all appear to be derived from bornavirus mRNAs, in some cases including even the poly-adenalated tail in the sequence insertion (Katzourakis and Gifford 2010).

Bats specifically appear to have a deep-rooted association with bornaviruses, as highlighted by the number of EBL elements already found within the limited set of bat genomes available at the time of a number of studies (Katzourakis and Gifford 2010; Horie et al. 2013; Cui and L.-F. Wang 2015). Furthermore, by using flanking region alignments to account for orthologous elements, Cui and L.-F. Wang (2015) reconstructed a tree of bornavirus sequences, each of which was putatively derived from a separate exogenous ancient bornavirus. This viral tree matched the host tree topology and divergence time estimates of the bat species they examined suggesting that bat-infecting bornaviruses follow a co-divergence pattern with their hosts. Yet despite the finding that bornaviruses had been integrated into vesper bat genomes relatively recently, as early as ~6 Mya (Cui and L.-F. Wang 2015), a recent study was unable to detect any orthobornavirus infection in a number of European vesper bat species (across seven genera) sampled within Germany (Nobach and Herden 2020).

4.1.4 Filoviruses and filovirus-derived elements

Like the *Bornaviridae,* the *Filoviridae* family of viruses in the order *Mononegavirales* are viruses with linear, non-segmented negative-sense RNA genomes (15-19 Kbp in length) with broadly similar gene order (3′-N-P-M-(G)-L-5′). However, unlike bornaviruses, filoviruses mostly possess filamentous virions (though there is a diversity of shapes), and are much more likely to cause fatal disease in humans. *Filoviridae* comprises five genera, two of which (*Striavirus* and *thamnovirus*) are found only in fish, and three of which (*Cuevavirus, Ebolavirus,* and *Marburgvirus*) are mammal-specific (Kuhn et al. 2019). Notably, ebolaviruses and marburgviruses are highly virulent and are major causes of lethal human disease, with Zaire ebolavirus virus being the cause of over 10,000 deaths during the 2014-16 West African outbreak (World Health Organization 2020). Filoviruses are endemic to Africa, Eastern and South-Eastern Asia, and Eastern and Southern Europe with bats as natural hosts and potential reservoirs of filoviruses of serious public health concerns such as Marburg virus, Ravn virus, and ebola virus (Leroy et al. 2005; Towner et al. 2009). More broadly, bats from Pteropodidae, Rhinolophidae, Megadermatidae, Miniopteridae, and Molossidae families have all been associated with extant filoviruses beyond those that pose a disease threat to humans (Beltz 2018). In addition, filovirus-derived EVEs in bats from the Vespertilionidae family

imply that vesper bats have also historically been host to filoviruses and may continue to be so (D.J. Taylor et al. 2010; Belyi et al. 2010b; Katzourakis and Gifford 2010).

Unlike bornaviruses, filoviruses replicate in the cytoplasm of the cell after being endocytosed into the cell. A key component of the replication cycle is the RNP complex (formed from the association of the nucleoprotein NP, polymerase cofactor VP35, matrix protein VP40, and RdRp L) which is responsible for producing mRNAs to generate viral proteins, as well as replicating the RNA genome. The produced viral proteins, and genome copies are then assembled into new virions in the cytoplasm before budding off at the plasma membrane (Kuhn et al. 2019).

Although filovirus-derived EVEs, or endogenous filovirus-like elements (EFLs), are rarely found in mammal genomes, sequences related to the NP gene have been identified in at least 13 mammalian genera, L in a single genus, and VP35 in seven mammal genera (Belyi et al. 2010b; Katzourakis and Gifford 2010; D.J. Taylor et al. 2010). Importantly, a syntenic filovirus-derived EVE discovered in the mouse and the brown rat are able to provide a dating point for the age of the mammal-specific filovirus genera at a minimum of ~30 My (D.J. Taylor et al. 2010). Also, of particular interest, is an EVE derived from a VP35 gene found in a syntenic position across 16 species of *Myotis* spanning Old World, North American, and South American clades of the genus which contains a conserved ORF of ~280 codons Edwards et al. (2018). Based on the divergence times of the *Myotis* species in which the EVE is found, Edwards et al. (2018) estimated a lower bound for the insertion time of 18 Mya implying long term conservation of this virus-derived ORF, and also showed that purifying selection has acted on this coding sequence, although there is currently no evidence that this locus is transcribed or translated. As well as forming part of the RNP complex responsible for RNA synthesis, Ebola VP35 plays a critical role in immune evasion, and inhibits immune response by blocking the production of type-I interferon. Compared to the Ebola VP35, Edwards et al. (2018) showed that the EVE-encoded VP35 in *Myotis* has lost its ability to bind dsRNA but has retained an inhibitory activity on type I interferon production (albeit more moderate than its viral counterpart). Taken together, this evidence strongly suggests a case of co-option in which the EVE-derived coding region confers an immunity-related selective advantage to *Myotis* bats. In this study I expand the set of known filovirus-derived EVEs in bats and characterise a largely undisrupted filovirus-like nucleoprotein (NP) encoding ORF present in *M. myotis*. Like VP35, NP is part of the RNP complex and has RNA-binding activity and also has a structural function in which it forms helical homo-polymers that encapsulate the virus genome.

4.1.5 Parvoviruses and parvovirus-derived elements

Viruses from the family *Parvoviridae* are small non-enveloped viruses with a linear single stranded DNA genome of only 4-6 Kbp in length. In animal hosts, including humans, they are known to cause diseases that range in severity from sub-clinical to lethal. Due to their limited genome, they are heavily dependent on the host cell machinery within actively dividing cells in order to replicate, and in some cases require co-infection with other viruses (such as adenoviruses) to replicate at all. Historically the family has been divided into two groups: the *Parvovirinae,* a group of vertebrate infecting viruses; and *Densovirinae*, a group of invertebrate infecting viruses. Recently, newly discovered parvovirus sequences obtained from metagenomic studies, has divided the *Parvoviridae* into four monophyletic groups: *Parvovirinae,* two separate groups of *Densovirinae,* and the *Chapparvovirinae* which contain viruses that infect both vertebrate and invertebrate hosts (Pénzes et al. 2019).

Bats may well play a key role in the diversity and transmission of parvoviruses. Within the *Parvovirinae* subfamily, viruses from five of the eight described genera have been detected in bats (*Bocaparvovirus*, *Dependoparvovirus*, *Protoparvovirus*, *Tetraparvovirus*, and *Amdoparvovirus*), indicating that bats host a high diversity of parvoviruses . Furthermore, bocaviruses isolated from bats are as diverse as bocaviruses found in all other animals combined to date, suggesting that bats are the ultimate gene source of most bocaviruses (Lau et al. 2017). Additionally, bat bocaviruses are known to transmit readily between bat species (Lau et al. 2016). Viruses identified by high-throughput sequencing associated with three bat species (*E. helvum, M. davidii, D. rotundus*) were placed within the Chapparvovirus group suggesting bats are also involved in the transmission of this virus group (Pénzes et al. 2019).

The limited genome consists of two major gene cassettes flanked by short (100-500 bp) loosely palindromic regions which form hairpin structures and provide the origins of replications for the virus. The gene cassette closest to the 3'-end codes for the non-structural replication initiator gene (NS, sometimes known as the replicase), which combines sequence-specific DNA-binding and DNA-nicking activity as well as a DNA-helicase domain. The second major gene cassette, closer to the 5'-end of the genome, codes for the capsid protein gene (VP) which forms a capsid with T=1 icosahedral symmetry. Most parvoviruses also code genus-specific ancillary proteins in alternate and overlapping reading frames, which carry out a variety of functions in the viral life cycle. The genome does not include any polymerase genes; instead, parvoviruses rely entirely upon host DNA polymerases (Cotmore et al. 2019).

The reliance on host DNA polymerases for replication forces the majority of the parvovirus replication cycle to occur within the host cell nucleus. The virions containing the intact DNA genome are trafficked into the host cell nucleus where they are uncoated and begin replicating by exploiting host DNA damage response pathways (though the mechanisms of this process have not yet been established). In the case of the more genetically-limited parvoviruses like Adeno-asociated viruses (AAVs), co-infection by other viruses is necessary to evoke a replicative environment within the nucleus. Exploiting damage repair pathways for replication is an imprecise means of copying genetic material and leads to rapid mutation rates (~10^{-2} substitutions per nucleotide per replication), high recombination rates, and therefore rapid generation of strain diversity (Cotmore et al. 2019).

Many parvovirus-derived EVEs, or endogenous parvovirus elements (EPVs), have been discovered in vertebrate genomes, often as incomplete fragments of parvovirus genes but in many cases as nearly complete copies of the entire parvovirus genome. This contrasts with RNA virus-derived EVEs in which the integrated sequences always appear to be derived from viral mRNAs (Belyi et al. 2010a; Katzourakis and Gifford 2010; Liu et al. 2011; Kryukov et al. 2018). Compared with other lineages of DNA viruses, parvovirus-derived EVEs are vastly more frequent in vertebrate genomes and this is likely due to their specific, nuclear based, replication mode, which appears to favour integration, especially in genomic regions prone to double-strand breaks (Liu et al. 2011).

4.1.6 Non-retroviral EVEs in bat genomes

Although a number of non-retroviral EVEs have been identified in bat genomes to date, the diversity of these EVEs represents only three families of viruses (*Parvoviridae*, *Filoviridae*, *Bornaviridae*), compared with the 30 families known to circulate in extant bat species (L. Chen et al. 2014), perhaps underscoring the general rarity of non-retroviral endogenisation events. Interestingly, parvoviruses and bornaviruses, are also the most represented families in of EVEs in other mammalian genomes (Katzourakis and Gifford 2010; Kryukov et al. 2018), and this may be related to the nature of their largely nucleus-located replication cycle (Liu et al. 2011; Matsumoto et al. 2012). However, with >1300 species of extant bats that range across the globe in a remarkable diversity of habitats (Teeling et al. 2018), bats, as a group, provide a high number of opportunities for non-retroviral endogenisation events to occur. Thus, bat genomes could provide a rich source of diverse EVEs, helping to characterise the biogeography and evolutionary history of viral groups.

In this study, I survey 40 genome assemblies from 36 bat species for non-retroviral EVEs, using the DVORFS EVE mining tool I developed (Chapter 2). This represents the most phylogenetically extensive survey of non-retroviral EVEs in bats to date, characterising the EVE repertoire of as yet unexplored bat genomes, and expanding the number of known non-retroviral EVE loci in previously analysed genomes. Despite the broader sampling of bat species, no EVEs derived from viral families other than *Parvoviridae*, *Filoviridae*, *Bornaviridae* are identified by my survey, reinforcing the finding that the entire suite of non-retroviral EVEs in bats is derived only from these three viral families. Within each of these viral families I am able to phylogenetically place the identified bat EVEs within a viral phylogeny in order to gain new insights into the evolutionary history of these viruses. With this, I provide the first evidence of culterviruses and carboviruses infecting bats, and gain insight into the possible biogeographical origins of the *Carbovirus* genus, as well as revealing potential patterns of host-specificity at the genus level of bornaviruses. With the broader phylogenetic sampling of bat species, I am also able to show that filovirus-derived EVEs are present beyond the Myotis genus of bats and can be found across the Vespertilionidae family. Additionally, I identify a syntenic tetroparvovirus derived-EVE present across Chiroptera that is also present within a number of genomes in the Laurasiatheria superorder, which implies a minimum age of ~76 My for the *Tetraparvovirus* genus. Finally, I also explore the potential functional role of the elements with viral protein encoding ORFs that have been evolutionarily conserved for millions of years.

4.2 Methods

4.2.1 EVE mining

To search for non-retroviral EVEs in bat genomes I used the the Disrupted Viral ORF Search (DVORFS) tool that I developed (Chapter 2) with a curated list of PFAM domains (El-Gebali et al. 2019) as a query. The PFAM domains were selected through the following process. A long-list of domains was gathered by a taxonomic query to the PFAM database (version 31.0) that returned all PFAM-A families listed as being present in the genomes of virus species from within any viral genus with known vertebrate hosts. This long-list was then manually reduced by removing all domains that are also found commonly in eukaryotic genes, retroviruses or retroelements, leaving only those specific to non-retroviruses.

The curated PFAM domains were then used to query the bat genomes using DVORFS with HMMER pre-search and a bit-score cutoff of 15.0 and a length cutoff of 20 amino acid matches. Hits from exclusively repetitive regions of the query domains were removed in order to remove spurious and uninformative hits. Manual inspection of the hits revealed that only parvovirus and negative-strand RNA virus domains were consistently present in a number of bat genomes and are thus unlikely to be spurious. Therefore further analysis was conducted only on elements containing domains found in those virus groups. The hits were then merged where necessary into contiguous individual elements (multiple domain hits from the same viral group were combined into a single element if each hit was within 5000nt of the previous hit and the sequence of hits was in the same order as in the corresponding viral genome) and the sequences were extracted including a 1000 bp flanking region on each side.

EVEs which contained domains of any mononegaviral origin were compared with GenBank reference genomes of the order Mononegavirales using the tBLASTx-based ViPTree software (Nishimura et al. 2017). Based on the phylogenetic tree produced by this method, the EVEs were categorised into viral families for downstream analysis. All the elements containing only L gene domains grouped most closely with other elements containing bornavirus-specific domains and bornavirus reference genomes. These elements were therefore categorised as bornavirus-derived elements. All the elements containing filovirus-specific domains were categorised as filovirus-derived elements. All the elements containing parvovirus-specific domains were categorised as parvovirus-derived elements.

4.2.2 Orthology clustering

For each of the three families of non-retroviral EVEs identified, flanking regions were extracted and clustered to provide information about synteny of the elements across the bat genomes. Flanking regions of 2000 bp length, starting at a distance of 500 bp from the start and end of the domain hits, were extracted for each element. The 500 bp gap between the hits and the start of the flanking regions was used in order to avoid including viral sequence (undetected by DVORFS) in the flanking sequence to be analysed; this could otherwise cause elements to be considered syntenic despite only being similar in stretches of viral sequences close to, but outside, the detected domain hits. The 2000 bp flanks of each element were then separately concatenated to form 4000 bp length sequences (some elements were near scaffold edges and therefore did not have full length flanking sequences) which were clustered using the PSI-CD-HIT tool

with the parameters: `-G 1 -c 0.3 -g 0 -aS 0.3 -prog 'blastn'` (W. Li and Godzik 2006). This aligns the sequences using BLASTn and clusters them into groups with a minimum total identity to the longest/first encountered sequence in the group (a greedy algorithm). Here I defined the clustering threshold arbitrarily as minimum 30% identity to the focal sequence across the entire 4000 bp flanking sequence. These groups of putatively syntenic elements ('syngroups') containing more than two elements are labelled by syngroup ID on the tree tips in Fig. 4.2, Fig. 4.5, and Fig. 4.8.

4.2.3 Sequence alignment and phylogenetic tree construction

EVE sequences from the DIGS for EVEs database (Zhu et al. 2018) and genomic sequences of reference exogenous viruses were queried using DVORFS with the curated set of PFAM domains to obtain predicted peptide and nucleotide sequences for these domains. For each family of EVEs (filovirus-, bornavirus- and parvovirus-derived elements), the predicted domain peptide sequences obtained from the analysed bat genomes were aligned with the domain peptide sequences predicted from the DIGS for EVEs database, and exogenous virus sequences, using the E-INS-i alignment method in MAFFT (Katoh and Standley 2013). These peptide alignments were then converted into frameshift-aware codon alignments of the original genomic nucleotide sequence using a custom script. These codon alignments were further refined by manual adjustment using the software AliView (Larsson 2014) and subsequently converted back into peptide alignments.

A maximum-likelihood phylogenetic tree was constructed for each of these peptide alignments using IQ-TREE (version 1.6.12) with default parameters (L.-T. Nguyen et al. 2015). To select the substitution model for each tree, I used the IQ-TREE tool ModelFinder, which compares models based on Bayesian information criterion (BIC) (Kalyaanamoorthy et al. 2017). For trees based on alignments of multiple domains, each domain was separately aligned and substitution models were separately selected for each domain by ModelFinder. A maximum-likelihood tree was then constructed with IQ-TREE using the combined alignment, partitioned by domain and corresponding substitution model. Branch support was assessed for all trees using non-parametric bootstrapping with 1000 replicates.

4.2.4 Selection analysis

To determine whether *chEBLM* is undergoing positive selection I used a maximum-likelihood-based selection test. First, the nine *chEBLM* ORF nucleotide sequences were aligned using MAFFT (default parameters). The alignment was then manually adjusted to ensure codons were correctly aligned and frameshifted regions in *C. perspicillata* and *M. hirsuta* sequences were removed. This codon alignment, along with a species tree of the nine bat species in which *chEBLM* orthologues were detected, were used for the following selection analysis. `codeml` from the PAML software (version 4.9j, Z. Yang 2007) was used to implement a maximum likelihood-based site model test for positive selection. Two models which classify sites in an alignment according to their nonsynonymous/synonymous substitution rate ratio, dN/dS ratio, or ω, were compared: Model M2a, which classes sites as either under purifying selection (where $\omega < 1$), under neutral selection (where $\omega = 1$), or under positive selection (where $\omega > 1$); and model M1a which only classifies sites as under purifying or neutral selection. A likelihood ratio test (LRT) was then conducted to test whether M2a fitted the data significantly better than M1a. This showed a significant chi-square value (*P*-value = 0.0164), suggesting that including the possibility of positively selected sites in the model provided a significantly improved fit. Bayes empirical Bayes (BEB) posterior probabilities of the class of each site in the alignment according to the M2a model were then obtained from `codeml`.

4.2.5 Structural protein modelling

In order to identify potential function of conserved ORFs within *chEFLNp* ands *chEBLM* elements, I used a structural modelling approach where translated ORF sequences were used to produce a structural model using the I-TASSER webserver (J. Yang and Y. Zhang 2015). I-TASSER first identifies appropriate structural templates from the Protein Data Bank (Berman et al. 2003) for the input peptide sequence using a sequence homology search and a panel of protein threading tools to identify structural homology. Fragments of the structural templates are then reassembled and combined with *ab initio* structural modelling before being iteratively refined to give a final structure prediction for the input peptide. The final predicted structures of the putative EVE-derived peptides were then visualised, and compared with the aligned structures of relevant proteins using PyMOL (Schrödinger, LLC 2015).

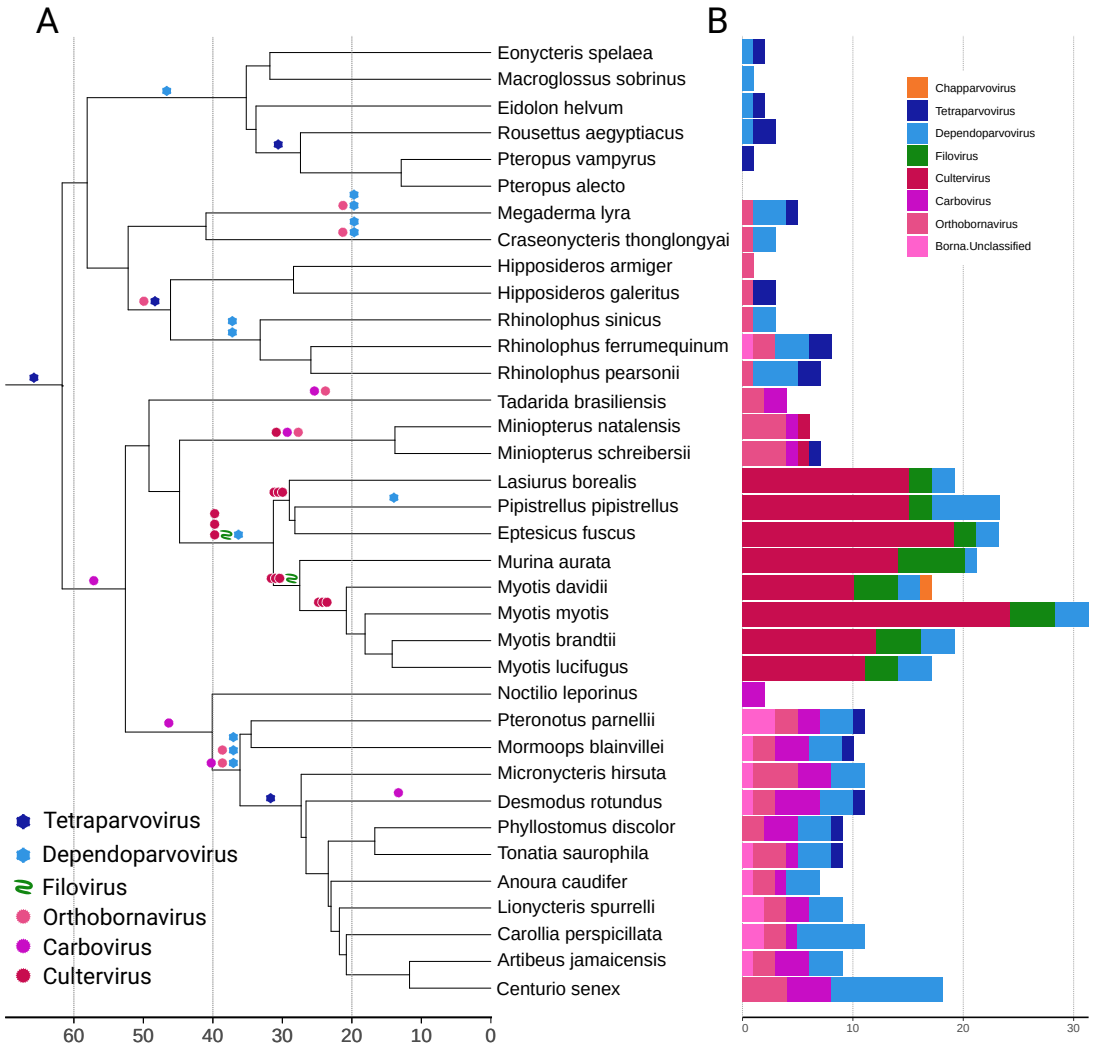


Figure 4.1: Summary of non-retroviral integrations and non-retroviral EVE counts in Chiroptera. A) Dated phylogenetic tree of bat species analysed in this study, divergence dates from TimeTree. Branches labelled with inferred virus integration events, based on phylogenetic analysis, categorised by virus genus. B) Bar plot showing total number of EVEs derived from each clade of virus within each bat genome.

4.3 Results

I conducted a survey of 40 genome assemblies from 36 bat species using the DVORFS tools I developed (Chapter 2) and detected 222 putative bornavirus-derived elements, 27 putative filovirus-derived elements, and 116 putative parvovirus-derived elements (corresponding to 214, 27, and 102 elements, respectively, when duplicates between assemblies of the same species are taken into account). These were further classified into specific genera and into a total of 41 groups of similar elements based on phylogenetic trees built using the EVE sequences (Fig. 4.1).

4.3.1 Bornavirus-derived elements in bat genomes

Across the 40 bat genome assemblies, 222 putative chiropteran EBL elements (chEBL elements) were identified, corresponding to 218 separate EBL elements when identical EVEs from different assemblies of the same species were discounted (Table C.1). Of these, 168 are elements containing bornavirus N domains (EBLN elements), 36 are elements containing bornavirus L domains (EBLL elements), three are elements containing bornavirus P domains (EBLP elements), nine are elements containing M domains (EBLM elements), one is an element containing a bornavirus G domain (EBLG element), and one is a putative element containing a bornavirus X domain (EBLX element). Only five of the identified EBL elements contained more than one detected domain and none contained sequence from more than two genes, suggesting the elements detected are most likely derived from viral mRNAs rather than viral genomic RNA.

Next, I clustered these elements based on the nucleotide similarity of their flanking regions in order to help determine orthologous elements. Based on the clustering thresholds I used (see Section 4.2.2) 171 elements clustered into 40 putatively orthologous groups, 51 elements did not cluster with any others. Finally, I used the predicted amino acid sequences of the identified chEBL elements and all EBL elements from the DIGS for EVES database Zhu et al. (2018) to create separate peptide alignments for each bornavirus PFAM domain (see Section 4.2.3). The alignments were then used to construct maximum-likelihood phylogenetic trees (Fig. 4.2). Based on these phylogenetic trees, the phylogenetic relationships of the bat species, and the predicted orthologous groups, I manually grouped the chEBL elements into 28 groups of related elements (Fig. 4.2 and Fig. 4.3).

Strikingly, the number of detected bornavirus sequence insertions in yinpterochiropterans was extremely low. I was unable to detect any EBLs in the genomes of the Pteropodidae family, and only four bornavirus sequence insertions were detected among the yinpterochiropteran genomes I analysed (Fig. 4.3). However, a tBLASTn search using previously identified chEBLNs (Cui and L.-F. Wang 2015) showed the presence of a single copy of an EBLN element in each of the pteropodid genomes (e-values 2.74e-7 to 7.86e-26) suggesting this element inserted between ~35 Mya and ~58 Mya, a single copy of a related element in four rhinolophoid genomes (*H. armiger, H. galeritus, R. sinicus, R. pearsonii*; e-values range from 3.13e-4 to 3.14e-7) and a single copy of a related element in *M. natalensis* and *M. schreibersii* (e-values 3.91e-6 and 1.22e-6). In comparison to this total of six putative bornavirus insertions and low EBL copy numbers across yinpterochiropteran genomes, yangochiropteran genomes harbour evidence of a multitude

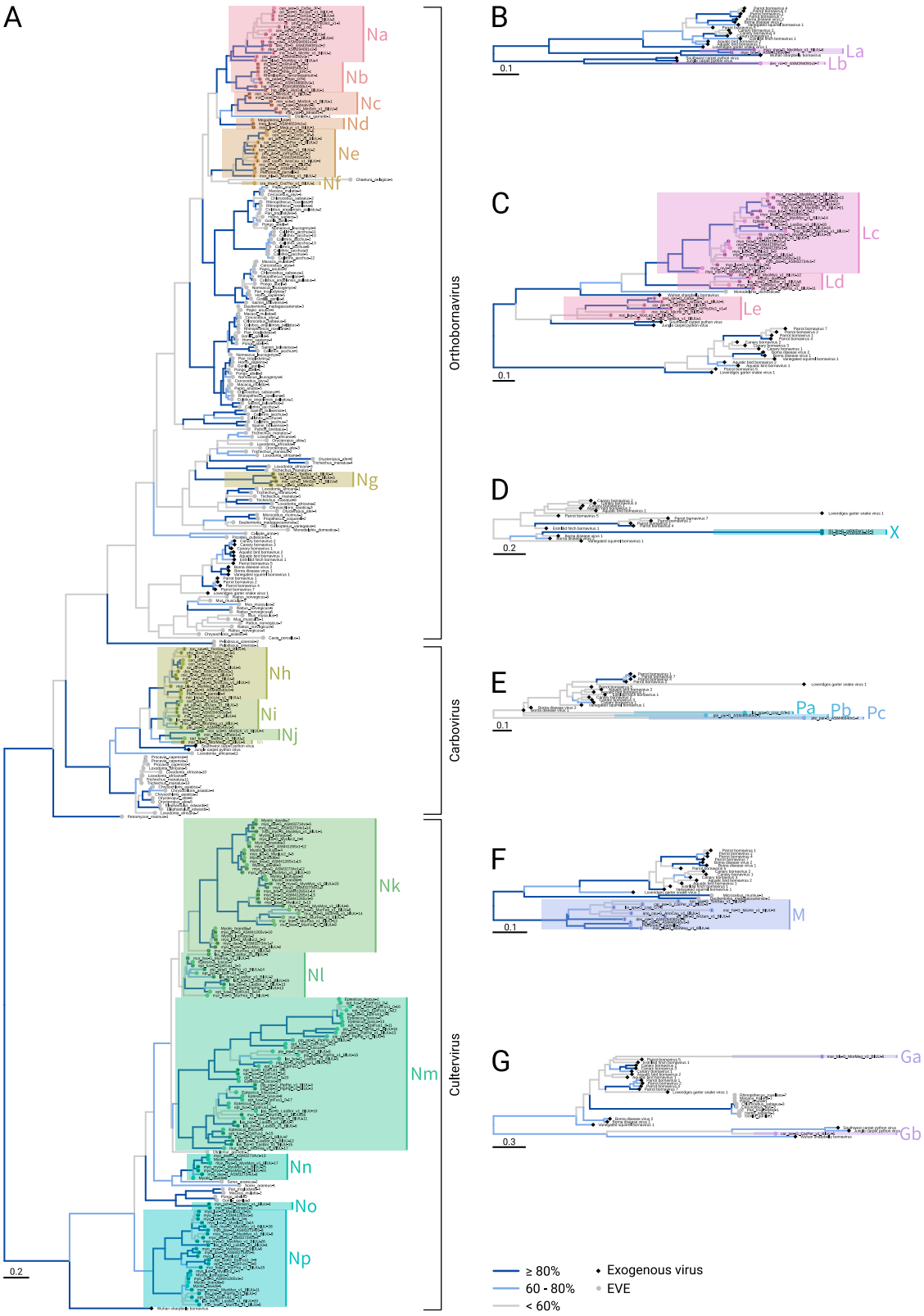


Figure 4.2: Phylogenetic trees (shown midpoint rooted for clarity) of exogenous bornaviruses, EBL elements obtained from the DIGS for EVEs database (Zhu et al. 2018), and chEBL elements with chEBL groups labelled. Trees constructed using alignments of bornavirus PFAM domains: A) BDV_P40 (N); B) Mononeg_RNA_pol (L); C) Mononeg_mRNAcap (L); D) BDV_P10 (X); E) BDV_P24 (P); F) BDV_M (M); BDV_G (G). Branch bootstrap support indicated by branch colour. Tips labelled with syngroup IDs (based on flanking sequence homology).

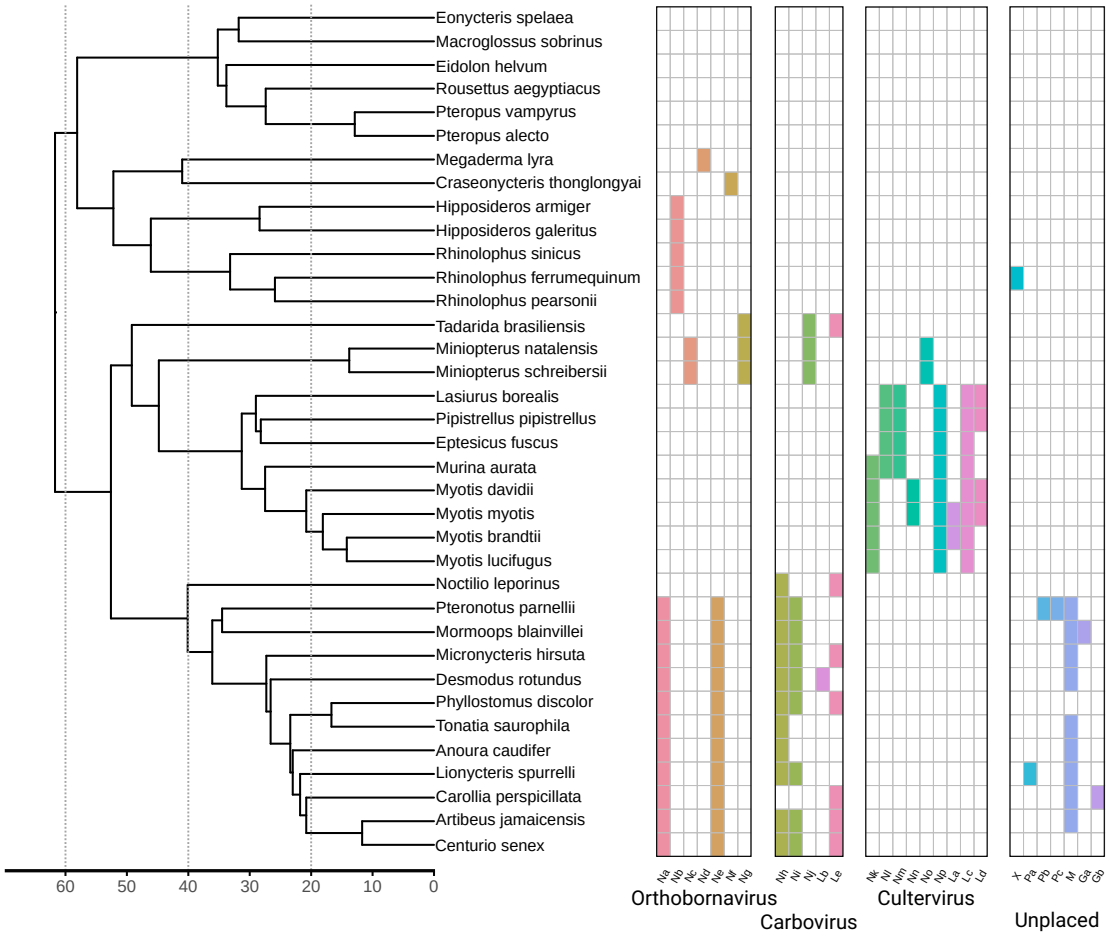


Figure 4.3: Presence of chEBL groups within each bat genome and categorised by bornavirus genus. Groups inferred from virus domain phylogenies in Fig. 4.2. Dated phylogenetic tree of bat species uses divergence dates from TimeTree.

of bornavirus insertions and subsequent duplications. 26 groups of EBL elements were detected across yangochiropteran genomes, of which some groups comprise EBL elements with multiple copies in each genome and in some cases, evidence of numerous separate insertion events from distinct exogenous bornaviruses.

Of the 168 identified EBLN elements, 99 clustered most closely with Wuhan sharpbelly bornavirus (WhSBV) (100% clade bootstrap support), classified as a member of the genus *Cultervirus* (Amarasinghe et al. 2019). Interestingly, these 99 elements (labelled as groups chEBLNk-p) are all found within the bats of the Vespertilionoidea superfamily. This pattern is also repeated in the phylogenetic trees for the Mononeg_RNA_pol and Mononeg_mRNAcap domains, where groups chEBLLa, chEBLLc, and chEBLLd (comprising 28 elements in total) also cluster most closely with WhSBV (99%, 88% and 88% clade bootstrap support respectively) and are found exclusively in Vespertilionoidea bat genomes. In fact, groups chEBLNk, chEBLLa, and chEBLLc all represent the same viral

group, as two sets of orthologous EBLs which contain sequence from both N and L genes (flanking sequence group IDs 65 and 61) are found within these chEBL groups. The placement of groups chEBLNn and chEBLNo relative to other cultervirus-like elements from the DIGS for EVES database suggests potential inter-order viral transmission of ancient culterviruses. Two EBLNs found in the genome of the common shrew (*Sorex araneus,* order Eulipotyphla) form successive outgroups to the chEBLNn group (72% and 75% clade bootstrap support). This implies the viral chain of transmission of an ancient cultervirus included an ancestor of *S. araneus* before infecting and being endogenised in the genome of the ancestor of *M. myotis* and *M. davidii* (between ~20 Mya and ~30 Mya). A group of four EBLN elements found in primate genomes (*Gorilla gorilla*, *Pongo abelii*, *Macaca mulatta*, *Pan troglodytes*) form a closely related sister group to the chEBLNo elements in *Miniopterus,* which suggests an ancient cultervirus lineage with a host range that spans the Boreoeutheria.

Five groups of EBL elements group phylogenetically with the two identified exogenous carboviruses, Southwest carpet python virus and Jungle carpet python virus. chEBLNh, chEBLNi, and chEBLNj are composed of a total of 25 EBLN elements, which potentially form a clade with the exogenous carboviruses (77% clade bootstrap support). Groups chEBLNh and chEBLNi are both found across the majority of the Noctilionoidea, each group comprising likely orthologous (84.55% and 75.85% median flank nucleotide identity respectively) single copy elements (excluding potential duplications of a chEBLNh element in *M. hisrusta* and *C. senex*). Both chEBLNh and chEBLNi groups are each likely derived from a single insertion event and can be dated using estimated bat species divergence times. Basing divergence times on Potter (2019), the estimated insertion dates for *chEBLNh* and *chEBLNi* are between ~40 Mya and ~49 Mya, and between ~36 Mya and ~40 Mya respectively. Despite being relatively phylogenetically close to chEBLNh and chEBLNi groups, elements in the chEBLNj group are found in the phylogenetically-distant Molossidae and Miniopteridae bats. Although the chEBLNj elements in *M. natalensis* and *M. schreibersii* are very likely orthologues (99.6% flank nucleotide identity) derived from a single insertion, it is not possible to determine the orthology of the chEBLNj element in *T. brasiliensis* owing to the short scaffold length on which this element is found in the *T. brasiliensis* genome assembly.

The chEBLGb group consists of a sole element detected in *C. perspicillata* which is placed on the BDV_G phylogeny as the closest outgroup of the exogenous carbovirus clade; however, this is weakly supported (55% clade bootstrap support). chEBLLb is a group that likewise only consists of a single detected element in a phyllostomid bat, but this

element is found in *D. rotundus* and forms a highly supported clade with the exogenous carboviruses (98% clade bootstrap support). Finally, chEBLLe is a group of elements which is placed as the closest outgroup in the 'Mononeg_mRNAcap' domain phylogeny, albeit with very weak bootstrap support (50% clade bootstrap support). This group consists of elements detected in single copies within the phylogenetically disparate genomes of *C. senex, A. jamaicensis, C. perspicillata, P. discolor, M. hirsuta, N. leporinus,* and *T. brasiliensis.* Although all the chEBLLe elements, excepting that present in *T. brasiliensis,* are detected in noctilionoid bats, the grouping is highly paraphyletic. Despite this, the identity of the flanking regions appears high even when including the *T. brasiliensis* element (identity between 65% and 89%), though this may be an artefact of the clustering method used.

A number of EBLN groups (chEBLNa-g) were unable to be reliably placed within the bornavirus phylogeny, though are perhaps related most closely with exogenous orthobornaviruses and a large number of vertebrate EBLNs from the DIGS for EVEs database (55% clade bootstrap support). Of these, chEBLNa, chEBLNb, and chEBLNe are each composed of single copy elements present within monophyletic clades of bat, and with likely orthology within the group and therefore inferable insertion times: chEBLNa and chEBLNe - between ~36 Mya and ~40 Mya; chEBLNb between ~36 Mya and <40 Mya.

4.3.2 chEBL elements with conserved ORFs

Of the numerous EBL elements detected, most have disrupted ORFs when compared with exogenous bornavirus sequences; however, a few of the elements appear to have conserved ORFs for significant lengths of bornavirus protein domains. The long ORFs found in this subset of chEBL elements may be a result of incomplete degradation of the original viral sequence insertion due to the insertion being a recent event, or in some cases, perhaps a result of maintenance of the ORF due to co-option of the viral genes by the host bat.

The most interesting cases of putatively co-opted elements are the nine *chEBLM* elements that I detected across the Phyllostomidae and Mormoopidae families of bats. The elements are all found as single copies in each genome and the flanking regions of these nine elements all cluster together with nucleotide identities of 44-80% indicating syntenic positions in the genomes. Additionally, only two of the genomes in the monophyletic set of eleven Phyllostomidae and Mormoopidae genomes in this study had no detected *chEBLM* element and no *chEBLM* elements were detected in genomes outside

this group. Together, these lines of evidence strongly suggest a single insertion of a bornavirus *M* gene in the ancestor of Phyllostomidae and Mormoopidae as the origin of *chEBLM* elements. Based on the bat clade divergence dates in Potter (2019), this places the age of the insertion between ~36 Mya and ~40 Mya. The phylogeny generated from the BDV_M domain sequence alignment places the group of *chEBLM* elements as an outgroup to both *Orthobornavirus* and *Carbovirus* (100% bootstrap support). In order to determine the phylogenetic position of these elements relative to *Cultervirus* (which has no sequenced *M* gene), I utilised a short, poorly conserved section of the L gene which lies downstream of the M ORF. This region lies outside of the phylogenetically-conserved Mononeg_RNA_pol and Mononeg_mRNAcap domains (thus I did not attempt to align it with the other L-containing EBLs), however BLASTx hits from exogenous orthobornavirus L protein sequences had consistently lower e-values than those from carbovirus or cultervirus L protein sequences to this region. It is therefore reasonably likely that this element originated from a viral lineage related, although distantly, to currently identified exogenous orthobornaviruses.

A peptide alignment of Borna disease virus matrix protein (BDV-M) with the *chEBLM* elements (Fig. C.1), shows that despite the length of time since the insertion event, *P. parnellii* and *M. blainvillei* have both retained an ORF of 85 codons corresponding to positions 57-140 of BDV-M, while *A. caudifer, A. jamaicensis, D. rotundus, L. spurrelli*, and *T. saurophila* have maintained an ORF of 97 codons spanning positions 45-140 in BDV-M. The *chEBLM* elements in *M. hirsuta* and *C. perspicillata* have each acquired frameshift mutations which disrupt the coding sequence in the first third and the final sixth of the coding region, respectively. Although the mutations in these species may have created missense sequence at the beginning and end of the coding region, the frameshifts have not significantly shortened the ORFs as they have not created in-frame early stop codons in the missense sequence. The lack of degradation of this viral ORF in the majority of the Phyllostomidae and Mormoopidae genomes suggests evolutionary conservation of the coding capacity of this virally-derived coding region and, therefore, perhaps co-option of this element at the protein level. To further test the hypothesis that the *chEBLM* locus is co-opted in this lineage of bats, I carried out a maximum likelihood-based selection test to detect positive selection occurring amongst the nine *chEBLM* orthologues (Section 4.2.5). The site model test showed significance for the presence of positively selected sites within *chEBLM* (P-value = 0.0164), and five sites within the coding sequence were estimated with Bayes empirical Bayes (BEB) estimation to have a posterior probability >0.5 of possessing dN/dS > 1 in each case.

BDV-M forms a stable, symmetrical homo-tetramer with two relatively well-characterised molecular functions: membrane binding and RNA binding (Neumann et al. 2009). In order to investigate whether these functions were retained in a putative chEBLM protein, I compared the 3D structure of chEBLM and BDV-M by constructing a homology based model of *A. jamaicensis*-chEBLM structurally aligned to the BDV-M structure model (PDB ID: 3F1J) (4.4) . The structural model of *A. jamaicensis*-chEBLM predicts that alpha-helix α5, and beta-strands β3 and β4 remain intact and in the same position compared with BDV-M, suggesting that the majority of the oligomerisation surface remains intact in chEBLM (4.4C). Interestingly, the five positions in chEBLM that were highlighted as potentially under positive selection by the selection test all occur on, or adjacent to, this oligomerisation surface. Therefore it may be the case that chEBLM has retained the capability to form a homo-oligomer (4.4D), or perhaps is capable of interacting with exogenous bornavirus M proteins such as BDV-M in some fashion (4.4E and F). Conversely, the RNA nucleotide binding pocket formed by PHE-37 and HIS-112 has not been conserved: PHE-37 is in the coding region lost in all nine *chEBLM* elements towards the N-terminus of the original *M* gene, and HIS-112 has been replaced in six of the nine putative chEBLM peptides with non-aromatic residues (4.4C). Additionally, the lost N-terminus region, containing beta-strands β1 and β2, is also responsible for creating much of the highly positively charged surface of the putative membrane binding face of the BDV-M tetramer, thus suggesting this functionality has also been lost in the putative proteins from all of the *chEBLM* elements even if the proteins are able to form tetramers.

4.3.3 Filovirus-derived elements in bat genomes

My survey also yielded 27 putative chiropteran EFL elements (chEFL elements) across these assemblies (see Table C.2). Of these, 22 are elements containing filovirus NP domains (EFLNP elements), and the remaining five elements contain a filovirus VP35 domain (EFL35 elements). Unlike the chEBL and chEPV elements detected in this study, the presence of EFL elements was detected in only one family of bats, the Vespertilionidae.

The identified EFL elements were clustered by flanking region homology into 'syn-groups' and were then aligned with exogenous filoviruses and EFL elements from the DIGS for EVEs database and phylogenetic trees were constructed using the same methods as used for the EBL elements. Based on these analyses I manually grouped

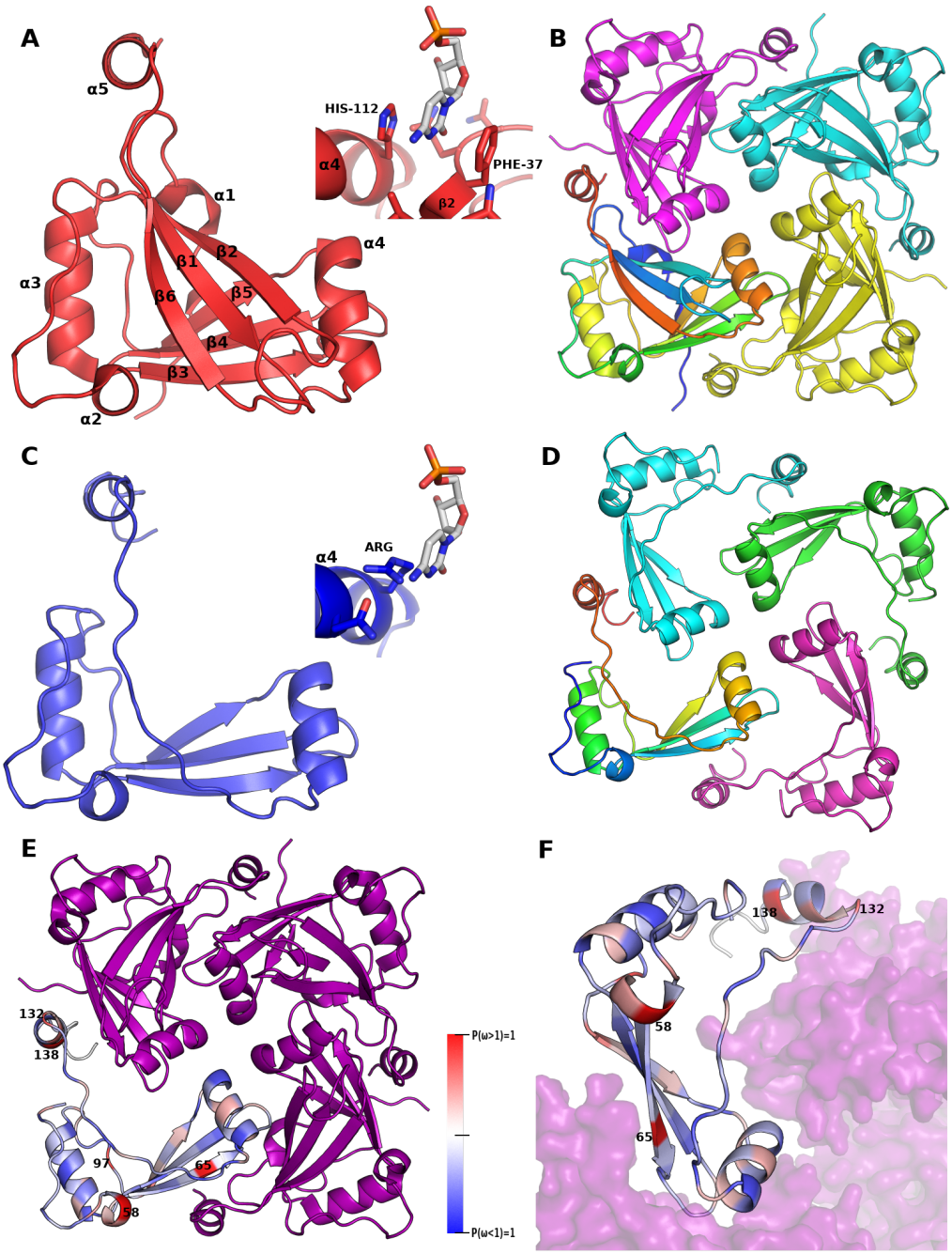


Figure 4.4: Structural differences between BDV-M and *A. jamaicensis* chEBLM. **A)** BDV-M protein structure displayed as ribbon diagram, helices and beta-strands annotated. Nucleotide binding pocket located between α4 and β2 inset with RNA nucleotide ligand positioned between HIS-112 and PHE-37. **B)** Homo-tetramer assembly of BDV-M. **C)** *A. jamaicensis* chEBLM structural homology model displayed as ribbon diagram. Disrupted nucleotide binding pocket inset. **D)** Homo-tetramer assembly of *A. jamaicensis* chEBLM based on positions of BDV-M peptides in BDV-M homo-tetramer. **E & F)** Hetero-tetramer assembly of three BDV-M peptides (purple) and one chEBLM peptide based on positions of BDV-M peptides in BDV-M homo-tetramer. chEBLM peptide coloured by site-wise BEB posterior probabilities of selection across *chEBLM* orthologues – red indicates higher estimated probability of positive selection, blue indicates higher estimated probability of purifying selection. The five sites with BEB > 0.5 for positive selection are labelled by BDV-M position.

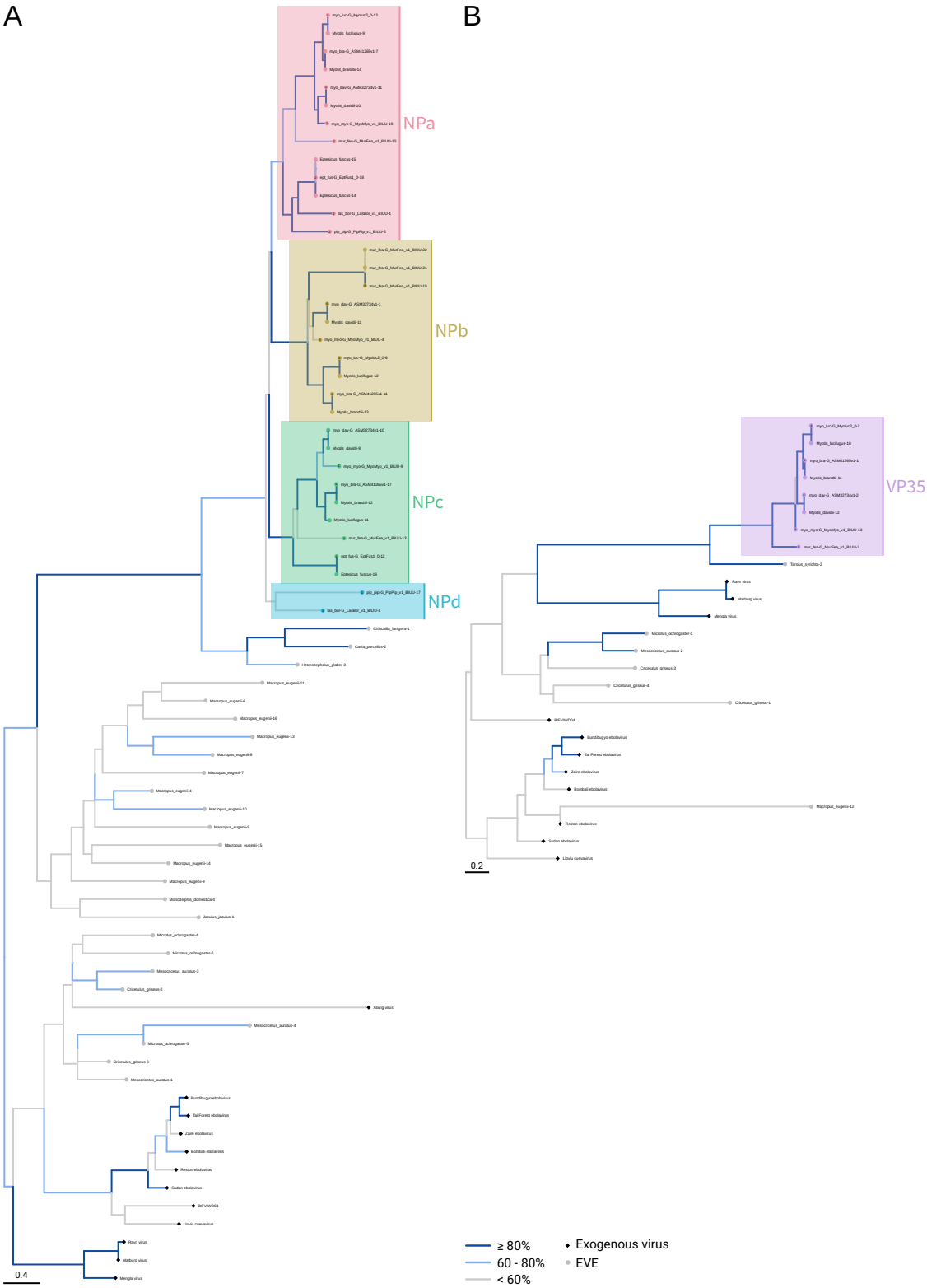


Figure 4.5: Phylogenetic trees (shown midpoint rooted for clarity) of exogenous parvoviruses, EFL elements obtained from the DIGS for EVEs database (Zhu et al. 2018), and chEFL elements identified in this study with chEFL groups labelled. Trees constructed using alignments of parvovirus PFAM domains: A) Ebola_NP (NP); B) Filo_VP35 (VP35). Branch bootstrap support indicated by branch colour. tips labelled with syngroup IDs (based on flanking sequence homology).

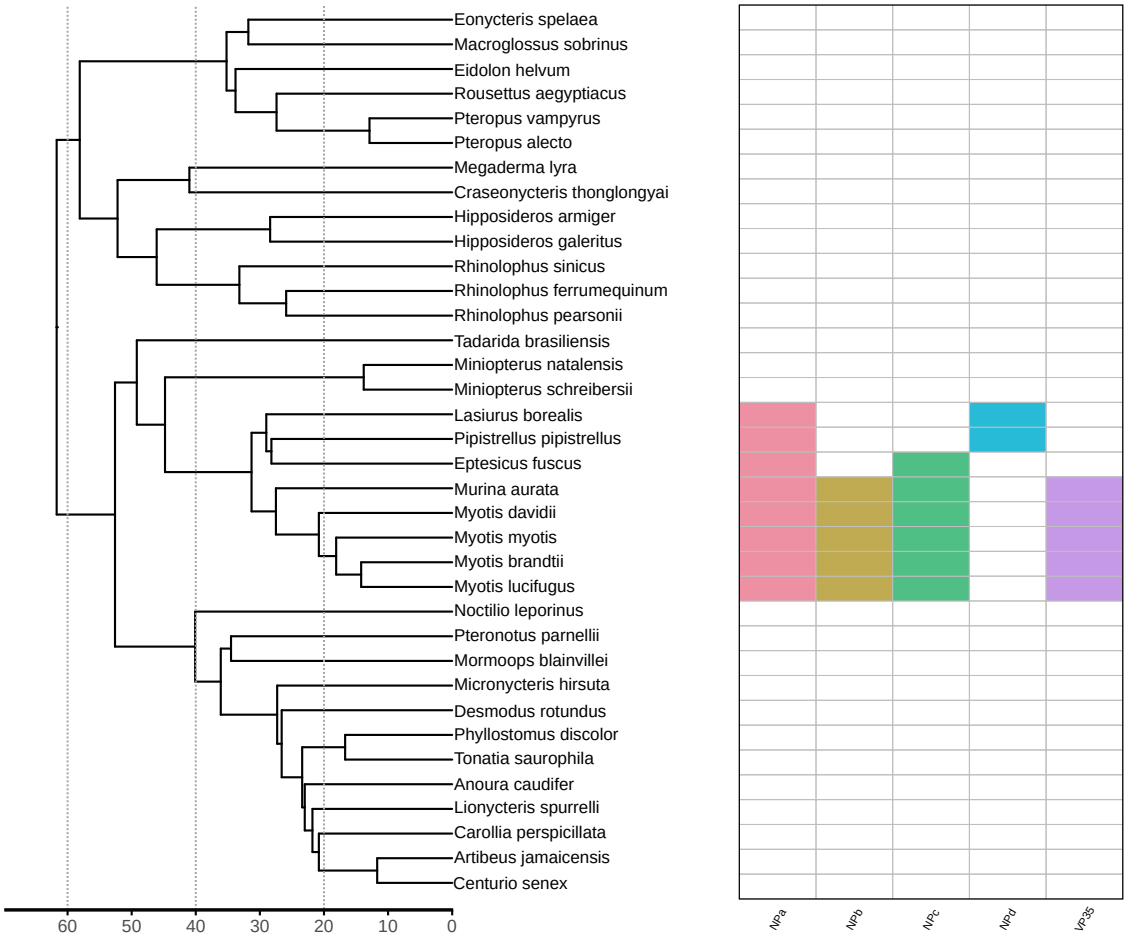


Figure 4.6: Presence of chEFL groups within each bat genome. Groups inferred from virus domain phylogenies in Fig. 4.5. Dated phylogenetic tree of bat species uses divergence dates from TimeTree.

the EFLNP elements into four groups (Fig. 4.5A), and the EFL35 elements into a single group (Fig. 4.5B).

The four chEFLNP groups all cluster together phylogenetically to form a single clade (73% clade bootstrap support) and this clade clusters most closely together with a group of rodent EFLNP elements from the DIGS for EVES database (95% clade bootstrap support). This grouping is a large genetic distance from all other exogenous filoviruses sequences and EFLNP elements in the DIGS for EVES database. Each of the chEFLNPa-c groups are composed of likely syntenic, single copy elements (53.0%, 73.7%, 61.0% median flank nucleotide identity respectively) that each arose from a single insertion. chEFLNPa elements are present in every vespertilionid genome analysed in this study, thus placing the date estimate for the original insertion event between ~31 Mya and ~45 Mya. chEFLNPb elements are present in all the genomes from the *Myotis* genus analysed in this study in addition to the *M. aurata* genome, thus placing the insertion

event that gave rise to this element between ~21 Mya and ~28 Mya. chEFLNPc elements are present in all the genomes from the *Myotis* genus analysed in this study in addition to the *M. aurata* and *E. fuscus* genomes with flanking sequence homology. This is a non-monophyletic distribution among bats and could be explained by either of two scenarios: either the non-detection of a chEFLNPc element in the *L. borealis* and *P. pipistrellus* genomes; or at least two separate losses of the element within the Vespertilionidae (the flanking sequence homology implies horizontal transfer is not a likely explanation). Assuming the former scenario is true, the true phylogenetic distribution of chEFLNPc is the same as that of chEFLNPa and thus has the same insertion date window of between ~31 Mya and ~45 Mya. The chEFLNPd group consists of only two elements, one in *L. borealis* and one in *P. pipistrellus,* which do not phylogenetically cluster with high support to each other (55% clade bootstrap support) but do contain flanking sequence homology.

All the chEFLNP elements detected have disrupted reading frames except for the chEFLNPa element detected in *M. myotis* and *M. davidii* which have undisrupted ORFs 385 codons and 256 codons in length, respectively. In *M. myotis* this undisrupted ORF aligns with the 280 residues closest to the N-terminus of ebolavirus nucleoprotein with 34% sequence identity. However the ORF also includes a region of 99 codons at the C-terminus end of the encoded protein which does not align with ebolavirus nucleoprotein, and moreover does not produce any significant BLASTp hits when searched against the GenBank NR protein database. The I-TASSER webserver was again used to produce a structural homology model of the encoded protein. Based on this model, the aligned region of 280 amino acids proximal to the N-terminus region of the *M. myotis* chEFLNPa protein has a TM-score of 0.966 when compared with the structure of ebolavirus nucleoprotein (PDB ID: 6NUT), indicating near identical predicted structural homology. Analysis using the IUPred2A webserver (Mészáros et al. 2018) shows the 99 residue C-terminus region to have a high probability of being disordered (see Fig. 4.7).

In agreement with previous studies (D.J. Taylor et al. 2011; Kondoh et al. 2017; Edwards et al. 2018), chEFL35 elements with long ORFs were found in all four of the *Myotis* genomes analysed. However, in addition, I detected an undisrupted chEFL35 ORF with flanking sequence homology to other chEFL35 elements in the *M. aurata* genome showing that this element was present in the ancestor of *Myotis* and *Murina* bats. Alignment of the full ORF with *Myotis* chEFL35 ORFs shows the *M. aurata* element to have an 18 codon in-frame insertion, however the insertion does not occur within either the putative homo-oligomerization domain or the IFN inhibitory domain identified by

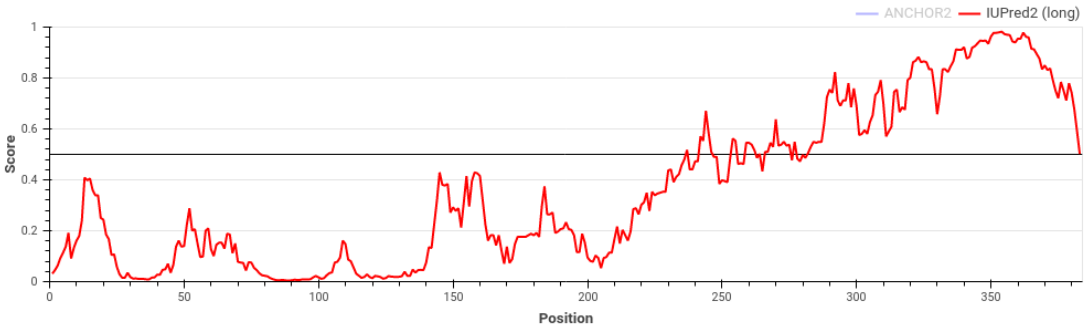


Figure 4.7: Plot of disorder probability (score) by residue position of putative protein encoded by chEFLNPa in *M. myotis*.

Kondoh et al. (2017). Additionally there are non-synonymous differences in 3 sites found to be under purifying selection by Edwards et al. (2018) (TYR-137 to HIS, PRO-182 to HIS, ASP-216 to SER).

4.3.4 Parvovirus-derived elements in bat genomes

The 40 genome assemblies from 36 bat species that I analysed using DVORFS yielded 116 putative chiropteran EPV elements (chEPV elements) across these assemblies, corresponding to 116 separate EPV elements when identical EVEs from different assemblies of the same species are discounted (see Table C.3). The identified EPV elements were clustered by flanking region homology into 'syngroups' and were then aligned with exogenous parvoviruses and EPV elements from the DIGS for EVEs database using the same methods as used for the EBL and EFL elements. As 40% of the elements identified had sequence derived from both NS and VP genes a multi-domain based phylogenetic tree was constructed (see Section 4.2.3 for details). Based on these analyses I manually grouped the EPV elements into 12 groups (Fig. 4.8).

The maximum likelihood tree obtained from the multi-domain alignment of parvovirus domains (Fig. 4.8) has very weak support, with most branches supported by less than 60% of bootstraps. However, based on this topology (which broadly agrees with Liu et al. (2011), Pénzes et al. (2019), and Cotmore et al. (2019)) the parvovirus-derived bat EVEs detected by this study group most closely with exogenous viruses from the genus *Dependoparvovirus*, or from *Tetraparvovirus* (with the exception of a single EVE in *M. davidii* that groups most closely with exogenous *Chapparvovirinae*). Seven of the twelve dependoparvovirus-derived chEPV groups contain genetic sequence from both the VP and NS genes, whereas no trace of the VP gene could be found in any of the elements

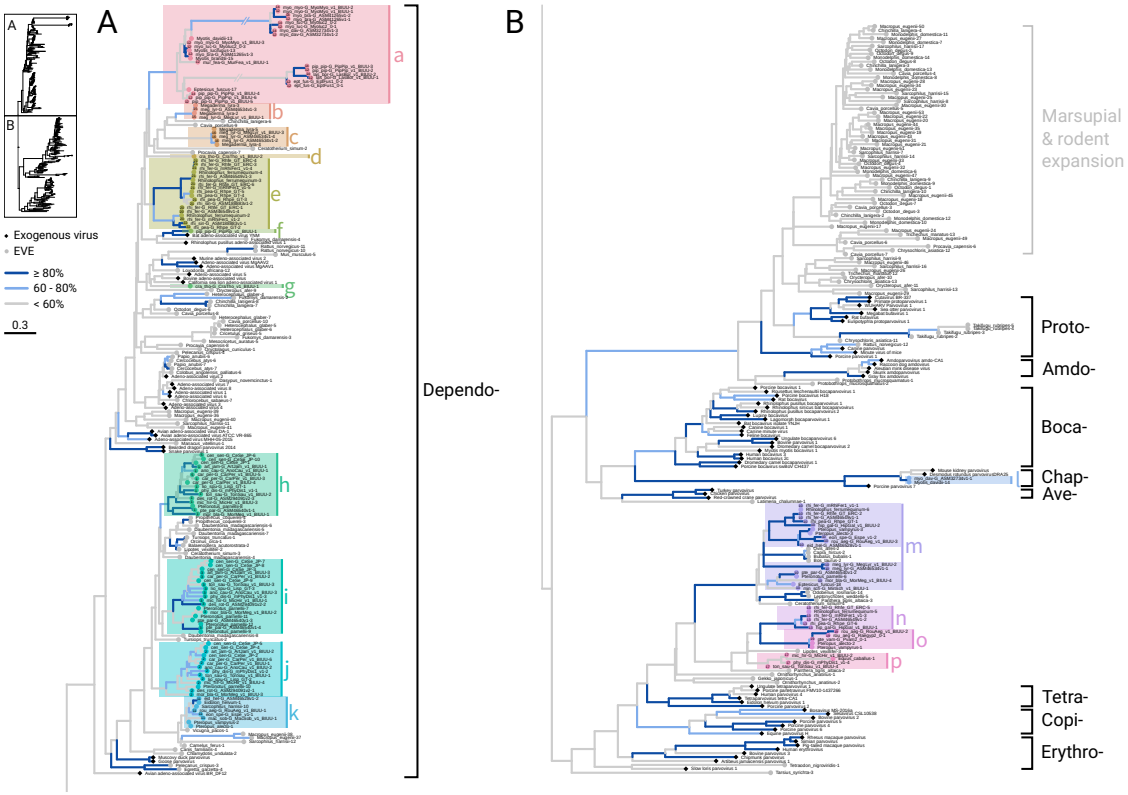


Figure 4.8: Phylogenetic tree (shown midpoint rooted for clarity) of exogenous parvoviruses, EPV elements obtained from the DIGS for EVEs database, and chEPV elements with chEPV groups labelled. Tree constructed using a multi-domain alignment of parvovirus PFAM domains (Rep_N, Parvo_NS1, Parvo_coat_N, Parvo_coat). Branch bootstrap support indicated by branch colour. Tips labelled with syngroup IDs (based on flanking sequence homology).

from the four tetraparvovirus-derived chEPV groups or the lone chapparvovirus-derived element (see Table C.3).

Based on the flanking sequence clustering and topology within each group that resembles the bat phylogeny, each of the chEPV groups excluding chEPVe is composed of elements that are most likely derived from single insertions, which in some cases have undergone duplication within the genome. The comparatively large genetic distance between the two distinct groups of elements within chEPVe and the differences in flanking sequences between these groups suggests two separate insertions of relatively closely related dependoparvoviruses took place within an ancestor of *Rhinolophus*, although an alternative hypothesis is that a single virus insertion was duplicated, after which the sequences diverged rapidly due to mutation. The chEPVa group of elements contains clades with long branch lengths on the phylogenetic tree produced; manual inspection revealed this to be an artefact likely produced by the complex duplicated structure of elements in this group.

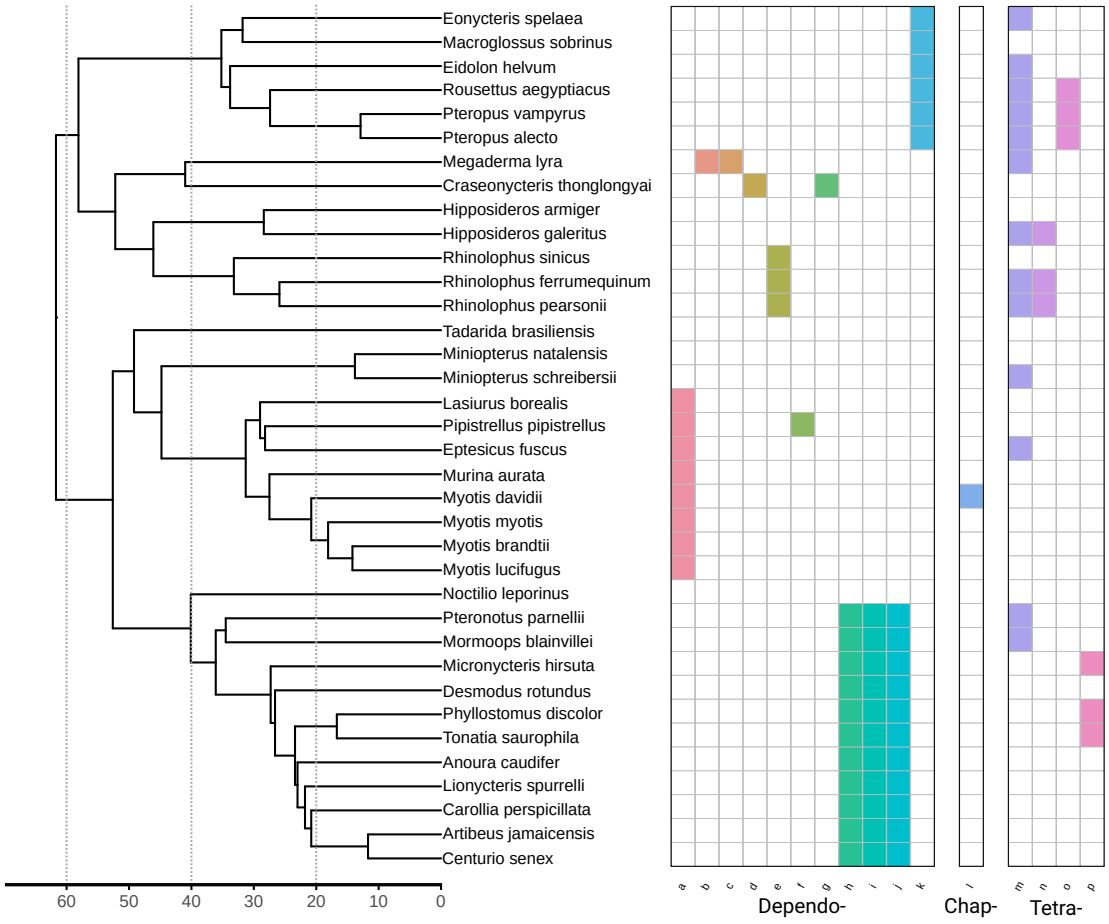


Figure 4.9: Presence of chEPV groups within each bat genome and categorised by parvovirus genus. Groups inferred from multi-domain phylogeny in Fig. 4.8. Dated phylogenetic tree of bat species uses divergence dates from TimeTree.

Groups chEPVm-p all cluster together and form a clade which is a sister group to the genus *Tetraparvovirus*, suggesting these may be ancient relatives of the tetraparvoviruses. The elements in the chEPVm group span the entire bat phylogeny and have high flanking sequence identity (52% median flank identity) confirming that these elements are syntenic and, have therefore, arisen from a single insertion that predates the divergence of Yinpterochiroptera and Yangochiroptera estimated to have occurred ~64 Mya. Interestingly, the chEPVm elements cluster with a number of EPV elements found in the genomes of various members of the Scrotifera clade (80% clade bootstrap support). A nucleotide BLAST search of RefSeq genome assemblies showed the presence of this element, including both right and left flanking sequences (1000 bp in length), in 56 genome assemblies of species within the Scrotifera, as well as a partial sequence (250/480 bp of the element and 727/1000 bp of the left flank) in the star nose mole genome. This suggests the insertion of the element predates the

divergence of the Laurasiatheria, which in turn implies a minimum age of 76.0 Mya for this element (dos Reis et al. 2012).

4.4 Discussion

4.4.1 Ancient cultervirus and carbovirus infections in bats

Due to the recent identification, classification, and sequence availability of viruses in the two newly classified bornavirus genera (*Carbovirus* and *Cultervirus*), I was able to classify 200 of the 214 chEBL elements I detected with DVORFS into each of the three genera (56 of these were previously described in Katzourakis and Gifford (2010), Horie et al. (2013), and Cui and L.-F. Wang (2015)) . I thus show that the historical relationship of bats and bornaviruses spans all three genera of *Bornaviridae*.

The phylogenetic trees constructed from EBLN and EBLL elements (Fig. 4.2) reveal a large group of elements that is exclusive to members of the Vespertilionoidea superfamily (Fig. 4.3), and which clusters most closely with the newly discovered Wuhan sharpbelly bornavirus (WhSBV). Although to date the only known host of this viral genus is a ray-finned fish, these elements provide evidence that bats and other mammals have been infected by ancient relatives of this viral clade from as early as 45 Mya and have continued to be infected as recently as 15 Mya or later. It is therefore highly likely that culterviruses continue to infect mammals despite the fact that extant mammal-infecting species as yet remain undetected. The phylogenetic placement of these chEBLN and chEBLL elements aligns with the two previously identified bat EBL elements that were clustered with WhSBV in ICTV TaxoProp 2018.016M.A.v1.Cultervirus (Amarasinghe et al. 2019), and expands the known presence of cultervirus-derived EBL elements to a broad group of vespertilionid bats as well as to the common shrew and a small number of primate species.

Interestingly, the only bats with detected cultervirus-derived EBL elements are from the Miniopteridae and Vespertilionidae families within the Vespertilionoidea superfamily. Therefore, based on the existence of cultervirus-derived EBL elements in other orders of mammal together with evidence of a putative cross-species transmission event from shrews to vesper bats, and the presence of an extant cultervirus in ray-finned fish, culterviruses clearly have a broad host range and are likely to be more than capable of host-switching across large host-phylogenetic distances.

The phylogenetic trees constructed from EBL elements in this study also show an expanded host range of carboviruses. All EBL elements analysed in this study that cluster with the two known exogenous carboviruses (detected in Australian pythons) are found in bats and other mammals of African origin as well as in South American bats from the Noctilionoidea superfamily. This suggests carboviruses may be endemic to Africa and the southern continents.

4.4.2 Phylogenetic distribution of bornavirus elements among bats

The groups of EBL elements detected in this study fall roughly into two distinct categories, ancient integrations that occurred 40 Mya or earlier in an ancestral bat species, and those composed of numerous recent integrations in closely related groups of extant bat species. Notably, orthobornavirus- and carbovirus-derived elements exclusively comprise the older integrations found in the Rhinolophoidea and Noctilionoidea superfamilies, whereas cultervirus-derived elements are found only as numerous newer integrations within the Vespertilionoidea. Furthermore, each separate genus of the *Bornaviridae* appears to have a unique relationship with each superfamily of bats.

The groups of bats identified here as possessing carbovirus-derived elements in their genomes are predominately members of the Noctillionoidea superfamily (endemic to South America) as well as *T. brasiliensis* (also endemic to South America), two species of *Miniopterus* (with an ancestral range that may include Africa), and a number of African mammals. This phylogenetic distribution of carbovirus-derived elements could perhaps be best explained by biogeography. There is some palaeontological and biogeographic evidence for dispersal events of bats and other fauna occurring during the Eocene, in which species originating on the African continent were able to reach Australia and Antarctica through land bridges and eventually reach South America through warm, ice-free corridors across Antarctica (Gunnell et al. 2014; Rojas et al. 2016). This is hypothesised to be the origin of the noctilionoid radiation in South America that ultimately led to a large diversity of Neotropical bat species. The date of the earliest carbovirus-derived elements characterised in this study (40-49 Mya) and the geographical range implied by the phylogenetic distribution of carbovirus elements amongst bats and other mammals, could be explained by an African origin of carboviruses followed by dispersal by bats and other vertebrates across the southern continents during the Eocene. Furthermore, the only exogenous carboviruses characterised to date are found in Australian pythons, perhaps lending more weight to this hypothesis. An alternative hypothesis is that carboviruses exhibit

specificity for infecting yinpterochiropterans, but are either unable to infect bats in the Vespertilionidae family or that these bats are able to prevent or selectively remove carbovirus integrations.

In contrast to the relatively broad phylogenetic distribution of orthobornavirus- and carbovirus-derived EBL elements, cultervirus elements have a very narrow phylogenetic distribution among bats: they are found only in bats of the Vespertilionidae and Miniopteridae families. This is surprising because cultervirus-derived EBL elements are present in other mammalian orders, and also because vesper bats have a cosmopolitan distribution and thus have ranges that overlap with those of many other groups of bats. This implies a high occurrence of opportunities for vesper bat-infecting culterviruses to jump to other species of bat, yet no EVE-based record of such jumps exists. The restriction to vesper bats also correlates with a recent expansion of helitron transposable elements within the vesper bats as well as a high count of recently integrated SINEs in vesper bat genomes (Platt et al. 2016; Jebb et al. 2020). It is perhaps possible, therefore, that a mechanism related to the activity of helitron or SINE transposons is responsible for increasing the rate of integration of viral genetic material during infections.

The high total number of EBL elements and the co-presence of EFL elements in the vesper bats suggests a generally increased rate of viral integration in this bat family. However, the lack of orthobornavirus- and carbovirus-derived EBL elements suggests a corresponding lack of orthobornavirus and carbovirus infections. This is supported by the findings of Nobach and Herden (2020), which showed a lack of detectable orthobornavirus infections in a number of species of vespertilionid bats sampled in Germany.

Of the species analysed in this study, *M. natalensis* and *M. schreibersii* stand out as having the broadest diversity of EBL elements, with elements derived from each of the three bornavirus genera. Beyond the broad geographic range of this family of bats, it is unclear what might cause them to be uniquely diverse in their EBL repertoire.

Taken together, the EBL elements mined from bat genomes in this study show a strong overall phylogenetic pattern in the relationship of different bat clades with different bornavirus clades. Each bat superfamily or family has a distinct pattern of EBL elements despite high levels of sympatry between species from different bat families. This suggests that bornavirus diversity is primarily driven through co-speciation dynamics, and that host-switching events over large phylogenetic distances are likely rare amongst bat-infecting bornaviruses. It is however worth noting that all the phylogenetic, and especially the apparent biogeographic, patterns discussed

above are relatively speculative. In this study I rely exclusively on a single, conservative database of vertebrate EVEs and sequences mined from the genomes of a single order of mammals. With broader searches for endogenous carbovirus elements and exogenous carboviruses, the true biogeographic pattern may be revealed to be entirely different. This is further compounded by the difficulty of accurately determining ancestral ranges and historical dispersion events of bats given their poor fossil record and their current cosmopolitan ranges.

4.4.3 Phylogenetic distribution of filovirus elements among bats

All the EFL elements detected in this study were found within the genomes of bats exclusively from the Vespertilionidae family. The previously well characterised VP35 element found in a large number of *Myotis* bats (Kondoh et al. 2017; Edwards et al. 2018) was also found in this study to be present in *M. aurata* in a syntenic position on the genome and also found to contain a conserved ORF. This pushes back the estimate of the earliest possible date at which the element was integrated in an ancestral genome from ~21 Mya to ~26 Mya (Ruedi et al. 2013).

The other EFL elements identified by this study are found in a wider range of bat species, albeit still within the Vespertilionidae. Each of the four chEFLNP groups (chEFLNPa-d) are composed of syntenic elements. Given the low branch support of the chEFLNPd group elements, and the non-monophyletic distribution yet potentially syntenic positions of the chEFLNPc and chEFLNPd, it seems reasonable to speculate that chEFLNPc and chEFLNPd elements are orthologous and originate from a single insertion. Based on the high nucleotide identity and the fact that the chEFLNP groups are each most closely related to the other chEFLNP groups than any other EVEs or exogenous viruses in the reference datasets, it may be the case that these groups actually represent paralogous sequences all derived from a single viral sequence insertion in a vespertilionid ancestor, that has undergone two sets of duplications (one duplication in the vespertilionid ancestor and a further duplication in the ancestor of *Myotis* and *Murina*). Should this be the case, it suggests a parallel situation to the cultervirus elements discussed in Section 4.4.2, wherein the cultervirus-derived elements are found duplicated many times within vesper bat genomes. This in turn implies that there is a link between duplication of these mononegaviral elements and the highly-active helitron elements and SINE expansion in this bat family.

Finally, because the chEFLNPa group of elements is likely derived from a single insertion that occurred between ~31 Mya and ~45 Mya, it provides a new dating point for

the existence of mammal-infecting filoviruses. This is at least 10 My older than the date estimated from the syntenic filovirus-derived EVEs in the mouse and brown rat genomes (D.J. Taylor et al. 2010). This age estimate highlights the long-term association filoviruses have had with bats and, perhaps specifically, the vesper bats.

4.4.4 Phylogenetic distribution of parvovirus elements among bats

Although the phylogenetic tree generated from the EPV alignments (Fig. 4.8) has weak support for most branches, the topology broadly agrees with Liu et al. (2011) and Pénzes et al. (2019) and exogenous viruses are grouped into same genera as Cotmore et al. (2019), furthermore, chEPV groupings are in almost every case concordant with flanking sequence similarity and bat phylogeny. It is thus likely a fair assumption that even in the absence of strong clade bootstrap support, chEPV elements are correctly grouped with the genus of parvovirus from which they are derived.

Interestingly, only EVEs derived from dependoparvoviruses and tetraparvoviruses were detected in bat genomes in this study and all EPVs present in the DIGS for EVEs database are also found within these viral lineages. This is surprising because bats have been found to host a broad diversity of parvoviruses and have a deep association with bocaparvoviruses, suggesting that the specific replication strategy of these two virus genera may be more likely to lead to genome integration. Here, the lack of characterisation of the biology of tetraparvoviruses limits hypotheses about similarities between tetraparvoviruses and dependo viruses that could explain this pattern. The distribution of tetraparvovirus- and dependoparvovirus-derived elements across the Chiroptera (Fig. 4.9) shows that both of these viral genera have been infecting and have continued to infect a broad diversity of bat species, suggesting a lack of host specificity. This echoes the high rate of cross-species bocaparvovirus transmission detected amongst bats (Lau et al. 2016).

The chEPVm group of elements were found in syntenic positions in both Yinpterochiroptera and Yangochiroptera in my survey. Furthermore, a BLAST search revealed the presence of this element across multiple orders in the Scrotifera and in a syntenic position in the star nose mole genome, suggesting that the insertion of this element predates the divergence of the Laurasiatheria. This indicates a minimum age of ~76 My (dos Reis et al. 2012) or ~95 My (X. Zhou et al. 2012) for this element. This in turn implies tetraparvoviruses (themselves deeply embedded within the *Parvovirinae* phylogeny) are at least this old, and therefore the *Parvovirinae* may be significantly older than the previous estimate of ~98 My (Liu et al. 2011).

4.4.5 Putatively co-opted EBL and EFL elements in bats

Mormoopid and phyllostomid chEBLM features a long conserved ORF which may act as a viral restriction factor

Many EBL elements which appear to have a functional role in the host have been previously reported, for example, the *itEBLN* gene in ground squirrels appears to function as a viral restriction factor (Fujino et al. 2014; Kim et al. 2020). These are most commonly derived from the N gene, and in at least one case, from the L gene (Horie et al. 2016; Kobayashi et al. 2016; Horie 2017). In this study I characterised a novel EBL element (named *chEBLM*) derived from a bornavirus *M* gene and present in at least nine bat species (within the Mormoopidae and Phyllostomidae families). Remarkably, *chEBLM* retains a highly conserved ORF, and appears to be evolving under positive selection, suggesting co-option and protein coding functionality. Although bornavirus M protein appears to posses functions related to RNA binding, the putative chEBLM protein is truncated in comparison and is missing the RNA binding pocket likely to be crucial for this function (Neumann et al. 2009). Interestingly, the predicted structure of the putative chEBLM protein suggest it has retained the oligomerisation surface present in bornavirus M, and furthermore, a number of residues positioned directly on, or adjacent to, this surface are found to be under positive selection. It thus appears plausible that the function of chEBLM involves binding to itself or to other bornavirus M proteins. This suggests it could function as a restriction factor by binding with the matrix proteins produced by infecting bornaviruses, and therefore inhibiting their capsid assembly or RNA replication functions. Potential mechanisms for such inhibition could be: 1) chEBLM proteins binding to bornavirus matrix proteins could disrupt the tetrameric symmetry, thereby preventing full assembly of the tetramer and thus any functions that relies a tetrameric assembly; or 2) a fully or partially assembled hetero-tetramer has severely reduced RNA binding or membrane binding functions as a result of the chEBLM protein's missing structural elements. Although sequence homology is low, BDV-M has significant structural similarity to the ebola matrix protein, VP40, (Neumann et al. 2009) which invites speculation that the chEBLM protein may possess restriction activity against a wider diversity of negative-strand RNA viruses. This would be a scenario similar to *Fv1* genes found in rodents, which are co-opted EVEs derived from murine lukemia virus (MuLV), but which appear to act as viral restriction factors against a broader range of retroviruses, perhaps by exploiting the conserved structures of its viral targets (Yap et al. 2014).

Despite the strong evidence of protein coding functionality, I was unable to detect any RNA transcription of this sequence in the RNA-seq datasets analysed in Chapter 5. The homology-based modelling which suggests that the oligomerisation surface remains structurally intact is also subject to uncertainties relating to this prediction. Further work demonstrating the transcription of the element or functional assays to ascertain the role of this putative protein are therefore necessary to confirm its functionality and status as a co-opted EVE.

A long conserved ORF in chEFLNPa in *M. myotis* and *M. davidii*

The group of chEFLNPa elements found across the Vespertilionidae appear to be homologous and syntenic, having likely been derived from a single insertion event in the ancestor of these bats between ~31 Mya and ~45 Mya. The *chEFLNPa* elements in *L. borealis, E. fuscus,* and *P. pipistrellus* are highly disrupted containing a number of frameshifts, stop codons, and a ~1 Kbp insertion. *M. aurata, M. davidii, M. brandtii,* and *M. lucifugus* (in the sister clade to *L. borealis, E. fuscus,* and *P. pipistrellus*) each contain *chEFLNPa* elements with far fewer disruptions, and in the case of *M. myotis* and *M. davidii*, long undisrupted ORFs 385 codons and 256 codons in length respectively. Remarkably, the structural similarity of the first 286 residues of the *M. myotis* EVE-encoded protein is extremely high when compared with exogenous ebola virus NP. The remaining 99 residues likely form a disordered region which is unlikely to possess any function. Based on the state of degradation of this EVE in the other bats which possess it, it is possible that this was once co-opted and possessed a functional role in one of the ancestors of *Myotis* but has since lost its function and degraded in all but a few of the *Myotis* species. However, due to the high structural similarity of this putative protein with the ebola NP protein, it is difficult to speculate which of the original virus-evolved properties would be exploited by a bat cell expressing this gene.

5

Patterns of EVE expression and co-option across the Chiroptera

Contents

Abstract **125**
5.1 Introduction **125**
5.1.1 Bats are a speciose clade with unique physiological traits . . . 125
5.1.2 ERVs are important sources of genetic novelty 126
5.1.3 ERV loci play an important role in antiviral immunity 129
5.2 Methods **130**
5.2.1 Spleen transcriptome sequencing 130
5.2.2 EVE annotation 131
5.2.3 EVE orthology analysis 131
5.2.4 Transcript prediction and abundance estimation 132
5.2.5 Differential expression analysis 133
5.2.6 Statistical comparison of expressed and unexpressed ERVs . . 133
5.3 Results **133**
5.3.1 Transcribed non-retroviral EVEs 134
5.3.2 Differences in degradation between transcribed and untranscribed ERV loci 134
5.3.3 Immune response RNA-seq datasets show significant differential expression of ERV loci 137
5.3.4 Searching for co-opted EVEs using evolutionary conservation . 141
5.4 Discussion **145**
5.4.1 Transcribed ERV loci are often less degraded than untranscribed ERV loci 145
5.4.2 Transcriptomic immune responses in bats include the up- and downregulation of ERV-derived transcripts 147
5.4.3 Candidate EVE loci for further investigation 148

Abstract

Endogenous viral elements (EVEs) are the genetic remnants of past viral infections found integrated in host genomes. Intriguingly, EVEs (particularly those derived from retroviruses) have been identified as important sources of genetic novelty, and are regularly co-opted to provide beneficial functions to the host species. Here, I scan the genomes of 35 bat species from 11 bat families in search of retroviral and non-retroviral EVEs and combine this effort with a large aggregation of transcriptomic data to examine their expression patterns. I show that the majority of expressed retroviral EVEs lack coding potential, but are more conserved on average than their non-expressed counterparts, and that elements containing conserved open reading frames (ORFs) appear to be enriched within the set of retroviral EVEs that are transcribed. Using this data I also identify putatively co-opted protein coding EVEs based on ORF conservation, estimated integration dates, and expression activity. Bats are important reservoirs for a number of zoonotic viruses of public health concern, making knowledge of bat–virus interactions, including an understanding of bat immune systems, an important goal for global public health research. Notably, bats have previously been shown to possess atypical antiviral response pathways, though these remain poorly characterised. Here, I identify EVE loci that may play a role in these pathways by utilising publicly available experimental data from three species of bat (*Eidolon helvum*, *Pteropus alecto*, *Artibeus jamaicensis*). As a whole, this study represents the first systematic and comprehensive survey of bat EVEs to integrate transcriptomic data with accurate genomic EVE annotations, identifying several intriguing putatively co-opted EVE loci.

5.1 Introduction

5.1.1 Bats are a speciose clade with unique physiological traits

Over 1300 species of bats have been recorded across more than 200 genera, making the Chiroptera the second-most speciose of the mammalian orders (Burgin et al. 2018). Unique amongst mammals in their capability of powered flight, bats have undergone an evolutionary radiation since their early divergence at approximately 65 Mya from the other orders of the Scrotifera clade within the superorder Laurasiatheria (X. Zhou et al. 2012; Jebb et al. 2020). This 'flight-powered' radiation has led to a global distribution in which bats occupy a broad diversity of ecological niches, having evolved into generalists as well as specialists, with diets that span insects, small vertebrates, blood, nectar, fruit, and pollen (Teeling et al. 2005; Jones and Teeling 2006; J.J. Shi and Rabosky 2015; Teeling et al. 2018). The order Chiroptera consists of 21 families placed into two suborders: Yinpterochiroptera, consisting of the Pteropodidae family and Rhinolophoidea superfamily; and Yangochiroptera consisting of the Emballonuroidea, Noctilionoidea, and Vespertilionoidea superfamilies.

In addition to a range of derived specialisms that have evolved within individual bat lineages during their radiation, a number of bat-specific physiological traits are found across the order. These traits – which include extreme longevity for their body size and metabolic activity (Healy et al. 2014), a low rate of cancer (Olds et al. 2015), a small genome size (Kapusta et al. 2017), and a unique immune system that appears to be adapted to dampen self-mediated immune activation and inflammation (Ahn et al. 2016; Xie et al. 2018; Ahn et al. 2019) – are all likely to have evolved due to the high metabolic demands of powered flight. Additionally, due to their high species richness and ecological diversity, bats are host to an extremely broad diversity of viruses (Mollentze and Streicker 2020) and are implicated as reservoirs for numerous zoonotic viruses of major public health concern such as Nipah and Hendra viruses, Ebola virus, Sosuga virus, and the global outbreak of SARS-CoV-2 (Halpin et al. 2011; Olival and Hayman 2014; Amman et al. 2015; Latinne et al. 2020). A better understanding of bat immune systems and bat–virus interactions is therefore of particular interest, not only due to the unique nature of these interactions, but also to aid more informed management of zoonotic public health risks from this important group of mammals.

5.1.2 ERVs are important sources of genetic novelty

Endogenous retroviruses (ERVs) are the result of successful insertions of retrovirus genomes into the genome of an infected host which are then transmitted in the host germ line (Stocking and Kozak 2008). Retroviruses possess single stranded positive-sense RNA genomes of around 10Kbp in length that encode three main polyproteins essential to the retrovirus life cycle (Gag, Pol, and Env, see Table 5.1 for detail). These genes are flanked by two long terminal repeats (LTRs) that work as regulatory regions for transcription of the provirus once integration has occurred (Wicker et al. 2007; Feschotte and Gilbert 2012). ERVs are ubiquitous in vertebrate genomes; however, although some ERV copies within a genome are the direct result of an infection by an exogenous retrovirus, many more copies are the products of retrotransposition of ERV loci within the cell, meaning that ERVs also act as transposable elements (TEs) (Magiorkinis et al. 2012). After insertion of an ERV in a germline cell (as a result of direct retroviral infection or ERV retrotransposition), the ERV locus can become fixed in the host population. Over time this causes a significant number of ERVs to be accumulated within a genome over the host's evolutionary history, to the extent that ERVs typically comprise 5% to 15% of the total genomic DNA in most mammals (Smit et al. 2013). Interestingly, despite the metabolic constraints of flight limiting the

Gene	Mature protein	Function
gag	Matrix (MA)	The retroviral matrix protein targets the gag polyprotein to the cell plasma membrane during capsid assembly and budding.
gag	Capsid (CA)	Many capsid protein subunits comprise the retroviral capsid, which forms the hydrophobic core of the virion and encapsulates the viral genomic RNA.
gag	Nucleocapsid (NC)	Retroviral nucleocapsid proteins package the viral genomic RNA through nucleic acid binding activity (provided by zinc-finger structures) and aggregation activity which results in condensation of the viral nucleocapsid in the mature virion.
gag/pro/pol	dUTPase (DUT)	Retroviral dUTPase is involved in viral replication in non-dividing cells, and likely prevents G–A transitions in the viral genome. This protein is only present in some genera of retrovirus, and its role is not well elucidated. The position of the coding region in the viral genome varies by genus.
gag/pro/pol	Aspartic Protease (AP or PR)	Retroviral protease cleaves synthesised retroviral polyproteins into their mature functional units.
pol	Reverse Transcriptase (RT)	Retroviral reverse transcriptase contains an RT domain which acts as a polymerase, capable of copying both RNA and DNA. IT also contains a ribonucleases H (RNase H) domain which hydrolyses the RNA template in RNA–DNA complex formed during reverse transcription.
pol	Integrase (IN or INT)	Viral integrase integrates synthesized viral DNA into the cell genomic DNA.
env	Env glycoprotein (Env)	The Env glycoprotein is composed of the transmembrane (TM) and surface (SU) subunits. TM anchors the glycoprotein to the virion envelope while SU is exposed and mediates host cell entry.

Table 5.1: Overview of genomic structure and function of retroviruses, genes and protein coding sequences listed in order of genomic position in 5'–3' direction. Genes in the retroviral genome encode polyproteins which are cleaved into mature proteins by host cell or viral proteases. Although the order of coding sequences is mostly the same across *Retroviridae*, genera differ in which domains are contained within each polyprotein gene.

genome size of bats, the activity of TE families, including ERVs, does not appear to be diminished in bat genomes (Kapusta et al. 2017).

For the host organism, unchecked ERV activity (as with other classes of TE) can cause the majorly deleterious and often oncogenic effects of gene disruption, aberrant expression of neighbouring genes, and genomic instability (Feschotte and Gilbert 2012). As a result, host organisms have evolved mechanisms to counter ERV activity through transcriptional silencing mediated by heterochromatin structure, which is in turn established by histone modification and DNA methylation machinery (Groh and Schotta 2017; Hurst and Magiorkinis 2017). This strict silencing means that most ERV loci are typically of negligible evolutionary impact, with their sequences mutating under a neutral selection regime. This neutrality leads to rapid pseudogenisation, such that the open reading frames (ORFs) found in the coding regions of most ERV loci are highly disrupted.

Importantly, however, a number of examples have been found that show that ERV loci can sometimes provide the basis for beneficial adaptations in the host organism, in a number of different ways. Firstly, ERVs are rich in *cis*-regulatory sequences (particularly within the LTR regions), which the provirus uses to recruit cellular transcriptional

machinery for its own replication, but which have the potential to be co-opted for transcriptional control of adjacent genes (Chuong et al. 2017). A notable result of this is that promoter/enhancer activity of lineage-specific ERVs has independently contributed to the transcriptional networks underlying interferon (IFN) response across diverse mammalian genomes (Chuong et al. 2016).

Secondly, ERVs (and other families of TEs) regularly give rise to RNA transcripts through the process of exonisation, whereby the insertion of a transposable element causes the generation of exonic sequence where previously none existed. Insertion in the intronic regions of existing genes can lead to new splice patterns caused by the splice site-like structures in the TE, and result in a novel exon being incorporated into the mature transcript (Sela et al. 2010; Schmitz and Brosius 2011). Due to their potential to disrupt reading frames, however, selection heavily biases the distribution of these insertions to occurrence in non-coding transcripts or, seldom, in the UTR regions of protein coding transcripts (Kapusta et al. 2013). Additionally, TEs – including ERVs – contain their own transcription start sites and so are also frequent generators of *de novo* long non-coding RNAs (lncRNAs). In some cases, the lncRNA consists mostly of the TE sequence and in others cases the TE causes adjacent intergenic or intronic DNA to be transcribed (Kapusta et al. 2013).

When taken together, TEs likely contribute to the majority of all lncRNAs across vertebrate genomes, with TE sequence fragments present in ~75%, ~68%, and ~66% of lncRNAs in humans, mouse, and zebrafish genomes, respectively (Kapusta et al. 2013). Interestingly, ERVs are an overrepresented family of TEs within human lncRNAs (Kelley and Rinn 2012; Kapusta et al. 2013). Although the functions of lncRNAs are much less broadly characterised than those of protein coding genes (Kapusta and Feschotte 2014; Necsulea et al. 2014), examples of functionally characterised ERV-derived lncRNAs include: lncRNA_ES3, generated by a transcription start site and splice structure of a HERV locus, found to be a *trans*-acting regulator of pluripotency in human embryonic stem cells (Ng et al. 2012); and lnc-EPAV a *trans*-acting positive regulator of virus induced cytokine response in human cells (B. Zhou et al. 2019).

Finally, the *env*, *pol*, and *gag* genes present within ERVs provide protein coding sequences that have the potential to be co-opted to produce functional proteins beneficial to the host organism. Most documented examples of co-opted protein-coding ERV genes are examples of *env* co-option. For example, the well studied syncytins are a group of *env*-derived glycoproteins which carry out cell–cell fusion functions during the placental development of mammalian and viviparous lizard

species (Mi et al. 2000; Cornelis et al. 2017). Remarkably, this group of proteins is the result of 17 independent co-option events throughout the evolution of vertebrates, involving different ERVs in each case (Broecker and Moelling 2019). Other *env*-derived genes have been documented across a broad range of vertebrates and are often responsible for conferring viral resistance (Katzourakis and Aswad 2016; Frank and Feschotte 2017; Broecker and Moelling 2019), for example, a full length Env protein encoded by a HERV has been observed to interfere with human immunodeficiency virus (HIV) production (Terry et al. 2017). Fewer examples exist of well described *gag*- and *pol*-derived genes, however, these include: *CGIN1*, which contains retroviral RNaseH and INT domains from a retroviral *pol* gene and may help confer antiviral resistance through regulating the ubiquitination of viral proteins (Marco and Marín 2009); and the *gag*-derived *Fv1* gene in mice, which acts as a restriction factor for murine lukemia virus (MuLV), perhaps through its interaction with the MuLV capsid protein (Best et al. 1996). Recently, a systematic phylogenomic study identified 93 and 84 putatively co-opted *gag*- and *env*-derived genes, respectively, from wide sampling of vertebrate genomes, suggesting that ERV *gag*, as well as ERV *env*, genes are regularly co-opted (J. Wang and Han 2020).

5.1.3 ERV loci play an important role in antiviral immunity

Notably, many of the known examples of ERV co-option – whether they be regulatory sequences, lncRNAs, or functional proteins – are related to immune function, pointing to a predisposition for ERV loci to become co-opted for immune functions in a phenomenon termed EVE-derived immunity (EDI) (Aswad and Katzourakis 2012; Broecker and Moelling 2019). This is likely due to the fact that viruses are already adapted to interfacing with and exploiting host immune responses; thus, viral genomes are loaded with features that can readily 'plug in' to host immune pathways. This is particularly well demonstrated by the finding that ERVs have deeply shaped the interferon response transcriptional network in humans through their promoter/enhancer sequences (Chuong et al. 2016; M. Wang et al. 2020). Furthermore, broad differential expression of ERV loci has been observed in response to infections by numerous virus species (Contreras-Galindo et al. 2007; Hsiao et al. 2009; Toufaily et al. 2011; M. Wang et al. 2020), and this also tallies with findings that lncRNAs (a large proportion of which are ERV-derived) are involved with human interferon signalling pathways (Peng et al. 2010). In addition, because the main genomic sequences of endogenised viruses are similar to their exogenous counterparts, ERVs are an ideal source of sequence for production of RNA

molecules that are complementary to viral RNA. These RNA molecules can have immune roles either through sensing viral RNA or by directly interfering with it (Aswad and Katzourakis 2012; Roberts et al. 2014).

At the protein coding level, production of counterfeit viral proteins encoded by ERVs and other endogenous viral elements (EVEs) could contribute to innate immunity of the host through two further potential mechanisms. Firstly, the ERV-derived proteins could function to block entry of exogenous viruses into cells, thus providing an innate immune mechanism against a viral infection. The *Fv4* gene found in mice is an *env*-derived gene which helps to provide immunity to MuLV by coding for an Env protein which, when expressed, competes with the exogenous virus in binding to the cell surface receptor used for viral entry (G.M. Taylor et al. 2001). Secondly, a protein coding EDI gene can also confer immunity to a particular virus through interference with viral capsid assembly or other intracellular stages of viral replication. A well-studied example is that of the *Fv1* gene in mice, which is a *gag*-derived gene that produces a protein that binds to, and blocks, retroviral capsid cores shortly after cell entry (Best et al. 1996; Boso et al. 2018).

Previous studies have identified a number of putatively co-opted protein coding ERV-derived genes across chiropteran genomes (Skirmuntt and Katzourakis 2019; J. Wang and Han 2020). Here I perform the most comprehensive study to date of EVEs and their expression patterns in the Chiroptera, and, in turn, determine possible functionality for putatively co-opted instances of both protein coding and non-coding EVE loci. I first extend previous screens for EVEs using the newly-developed DVORFS tool described in Chapter 2 to accurately identify ERV domains in 35 bat genomes – which is a more sensitive method for EVE mining compared with the previous BLAST-based approaches. I then combine these annotations with those generated for non-retrovirus EVEs in Chapter 3. Finally, I combine these EVE annotations with a set of 189 RNA-seq libraries from 21 species of bats.

5.2 Methods

5.2.1 Spleen transcriptome sequencing

Eight RNA-seq read libraries were generated for this study in order to fill in phylogenetic gaps in available public RNA-seq data. Spleen tissue was sampled due to its major role in

immune function of vertebrates. Male adult individuals from four species (*Artibeus jamaicensis, Carollia perspicillata, Desmodus rotundus, Phyllostomus discolor*) were collected under permit in Costa Rica as part of a previous study (Sadier et al. 2018). Dissection was performed immediately after euthansia by isoflurane overdose, and whole spleens were placed in RNAlater and incubated at 4°C overnight and then frozen. Four *Rousettus aegyptiacus* individuals (two adult females and two adult males) were also collected from Copenhagen Zoo by members of the RossiterLab research group (S.J.R., J.P., R.D.) and euthanised by overdose of ether. Dissection was performed immediately afterwards and spleens were prepared and stored in the same way as the other four samples. Dissected spleen tissue was shipped in RNAlater on dry ice to Novogene, Hong Kong where RNA isolation, cDNA library building, and sequencing of the samples was performed. cDNA libraries with an insert size of 250-300 bp were built after isolated samples were enriched for mRNA. cDNA libraries were then sequenced with Illumina HiSeqX to give 150 bp paired-end (PE) reads. Raw reads were then adapter trimmed and quality filtered to produce clean read libraries (each 6.3-8.8 Gbp in size).

5.2.2 EVE annotation

EVEs were annotated in 35 bat genomes (Appendix D.3.1) using a pipeline based on the DVORFS tool (details in Chapter 2). ERV regions were first identified using Repeatmasker loaded with the custom repeat library previously described (Chapter 2). ERV domains were then annotated within these regions using DVORFS with protein HMM profiles from GyDB (Llorens et al. 2011) as queries (GyDB domains: Acc, AP, DUT, ENV, GAG, INT, RNaseH, RT) with bit-cutoff of 20.0 and length-cutoff of 30 amino acids. For non-retroviral EVE domains, the annotations generated in Chapter 3 were used. Briefly, these were generated using DVORFS with HMMER pre-search and using a manually curated set of protein HMM profiles from the PFAM database (El-Gebali et al. 2019) as queries (details in Chapter 3). Hits from exclusively repetitive regions of the query domains were removed in order to remove spurious and uninformative hits.

5.2.3 EVE orthology analysis

Annotated EVEs across the 35 genomes were grouped into orthologous groups determined by pairwise sequence similarity between elements. First, annotated EVE domains were merged into combined elements if they were located within 2000bp of each other. Next, the sequence of each element, including 2000bp of upstream and

downstream flanking sequence, was extracted to form an EVE library for each genome. `pblat` (M. Wang and Kong 2019) was then used to align each EVE library against all other genomes assemblies within the bat superfamily from which the library was generated. A pair of EVEs was considered orthologous if each member was the top-hit for the other's genomic position in a reciprocal BLAST, and also the pairwise alignment had >80% coverage of the EVE region and >80% coverage of either the upstream or downstream flanking region (at >95% identity). In order to connect orthologous pairs of EVEs into orthologous groups, a graph was constructed with each vertex representing an EVE, and edges between vertices representing pairwise orthology between the connected EVEs. Each connected subgraph was then considered to be an orthologous group of EVEs.

5.2.4 Transcript prediction and abundance estimation

A set of 17 public RNA-seq datasets (comprising a total of 189 read libraries) in addition to the eight read libraries generated from spleen samples in this study, was used to identify expressed EVEs in the 35 bat genomes (Table D.3). To accurately quantify locus-specific expression levels for EVE-derived transcripts, an approach was used based on the method implemented in SQuIRE (W.R. Yang et al. 2019). First, each RNA-seq library was aligned to the appropriate species' genome using STAR in two-pass mode (Dobin et al. 2013). Specifically, all RNA-seq libraries from the same species were used together in the first pass to generate splice junction annotations (STAR version 2.7.0f, default parameters), these splice junction annotations were then used in the second pass mapping for each individual library, using lenient multimapping parameters (STAR version 2.7.0f, parameters: `--winAnchorMultimapNmax 100 --outFilterMultimapNmax 100 --alignEndsType EndToEnd --alignEndsProtrude 100 DiscordantPair --outFilterScoreMinOverLread 0.4 --outFilterMatchNminOverLread 0.4`). Next, StringTie (Pertea et al. 2015) was used to assemble the mapped reads for each library into transcripts and count the number of mapped reads for each transcript, again using lenient multimapping parameters (StringTie version v1.3.6, parameters: `-M 1.0`). Finally, the read counts for EVE-containing transcripts (those transcripts overlapping annotated EVE domains) were adjusted using the expectation maximisation algorithm implemented in the 'Count' module from SQuIRE which redistributes multimapped reads amongst the transcripts to which they are mapped (a modified version of the module was used in order to fix existing bugs and allow input of custom TE annotations).

5.2.5 Differential expression analysis

Four of the published RNA-seq datasets used were from studies that characterised transcriptional changes in individual animals or in bat cell cultures in response to immune challenge or immune stimulation (Table D.3). These datasets were used for differential expression analysis. For each of these four datasets, read counts determined by StringTie (for non-EVE-derived transcripts) and SQuIRE Count (for EVE-derived transcripts) in the previous step were used as input for the DESeq2 R package (M.I. Love et al. 2014). DESeq2 was then used to plot principal component analysis (PCA) plots of EVE-derived transcript expression for each dataset to allow for qualitative assessment of expression differences. For each dataset, DESeq2 was then used to compare expression of all transcripts between relevant sample groups, EVE-derived transcripts that were differentially expressed (with adjusted *P*-value <0.05) were output. Read counts for these transcripts were normalised using pseudo-log normalisation and plotted as heatmaps.

5.2.6 Statistical comparison of expressed and unexpressed ERVs

In order to quantitatively compare the distributions of maximum ORF lengths between expressed and unexpressed ERV domain loci, a statistical hypothesis test was carried out. The transcript abundance data were split up by dataset and, within this, species. For each of these datasets, ERV domain loci were considered expressed if three or more reads were mapped to the locus from any single RNA-seq library within the dataset. Because of the multimodal nature of the maximum ORF length distributions, non-parametric tests were used for the following statistical comparisons. Specifically, a Mann-Whitney U test – as implemented by the `wilcox.test()` function in the R standard library (R Core Team 2019) – was carried out to compare the distributions of maximum ORF lengths between expressed and unexpressed ERV domain loci for each ERV domain type within each dataset (total 192 comparisons). To adjust for the family-wise error rate (FWER), a Bonferroni correction was applied to adjust the resulting *P*-values.

5.3 Results

A pipeline, based on the highly sensitive DVORFS tool (see chapter 2), was used to annotate ERV domains in 35 genome assemblies corresponding to 35 bat species from across Chiroptera. This resulted in a dataset containing 296,289 ERV domain loci across

these 35 genomes that could be further analysed (Table D.2). Additionally, I assembled a set of 189 RNA-seq libraries from 21 species of bat, comprising: 17 public RNA-seq datasets, two unpublished RNA-seq datasets (S. J. Rossiter, unpublished data), and a further eight RNA-seq libraries from spleens which I generated for this study (Table D.3). Each RNA-seq library was mapped to a corresponding genome assembly in order to generate transcript predictions and abundance estimates for each transcript. This large set of RNA samples enabled me to characterise a broad set of ERV-derived transcripts across 20 species of bats containing a total of 10,181 ERV domains transcribed in different tissues and conditions. Notably, the RNA-seq libraries from the two datasets which were mapped to the *R. ferrumequinum* genome, contained no reads which mapped to any ERVs or non-retroviral EVEs.

5.3.1 Transcribed non-retroviral EVEs

The annotations of non-retroviral bat EVEs generated in Chapter 3 were used in combination with the transcriptomic data assembled in this study to explore whether any of the previously identified non-retroviral EVEs showed expression. In total, 11 chEBLN elements were found to be expressed (three or more reads in any RNA-seq library mapping to the locus) from six orthologous element groups (Table D.1) across seven species, and one chEPV element was found to be expressed in multiple *R. aegyptiacus* tissues. Of the expressed chEBLN elements, two were orthologues from the chEBLNc group with long undisrupted ORFs (386 codons in length) and were found to be expressed in *M. natalensis* and *M. schreibersii* embryo limbs. Of note were two elements from within the cultervirus-derived chEBLNm group that were found to be expressed in *E. fuscus* cells. *E. fuscus* possesses seven highly disrupted copies of *chEBLNm* and six relatively intact copies that contain ORFs over 200 codons in length. Of these, the copy with the longest intact ORF (325 codons in length) and another relatively intact element were found to be expressed.

5.3.2 Differences in degradation between transcribed and untranscribed ERV loci

To characterise the patterns of ERV ORF degradation across the bat genomes in this study, the maximum length ORF that overlapped with each ERV locus was determined and used as a metric for degradation. Although the annotation pipeline uses DVORFS, which is able to predict the location of frameshifts and premature stop codon mutations

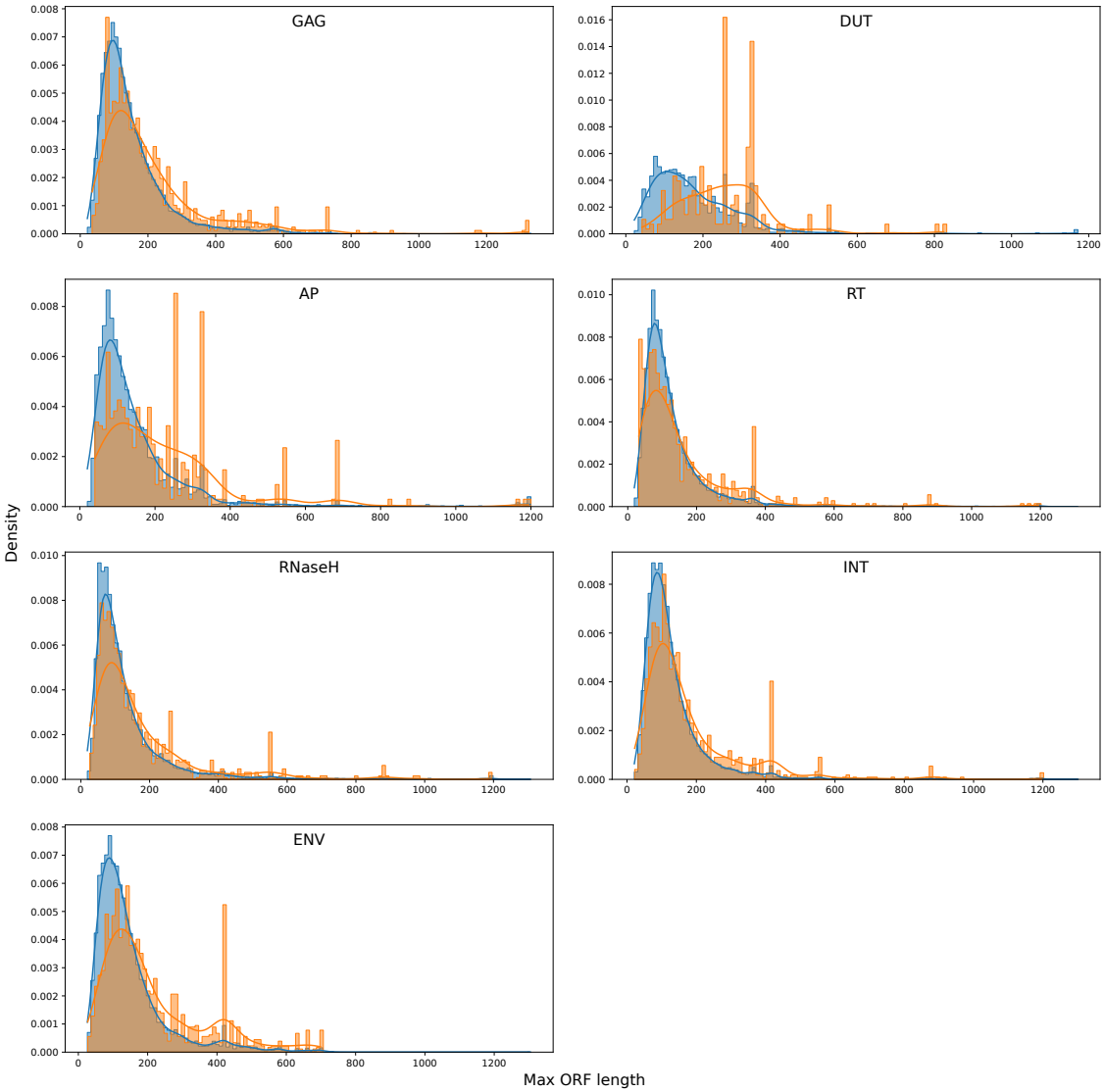


Figure 5.1: Histograms of maximum ORF length at each ERV domain locus by domain type. For each domain type, relative frequencies of maximum ORF lengths of expressed ERV loci (orange) are compared with the background distribution of maximum ORF lengths of all ERV loci (blue).

in EVEs, the high degree of genetic divergence and the degraded nature of ERVs means such alignments may be prone to inaccuracy when determining the precise numbers of ORF disrupting mutations. Instead, maximum ORF length at an identified ERV locus provides a measure of ORF degradation that is independent of the completeness and accuracy of the alignment used to identify the locus, where shorter maximum ORF length implies more degradation and longer maximum ORF length implies less degradation.

The distributions of maximum ORF lengths at ERV loci are plotted for each type of ERV domain in Fig. 5.1. For every type of ERV domain, the majority of loci contain only short ORFs as a result of numerous frame disruptions caused by frameshifting

ERV Domain	No. of datasets (comparisons)	H_0 rejected	H_0 rejected (Bonferroni adj.)
GAG	28	20	14
DUT	24	12	2
AP	28	23	12
RT	28	18	11
RNaseH	28	14	5
INT	28	24	15
ENV	28	21	16

Table 5.2: Counts of significant Mann-Whitney U test comparisons of the distribution of maximum ORF lengths for expressed vs unexpressed ERV domain loci.

indel mutations and by substitutions resulting in premature stop codons. This is visible as the dominant, left-most peak in the ORF length distributions. However, these ORF length distributions are multimodal, with small, right-hand peaks that correspond to the expected ORF length for each domain as found in exogenous beta- and gammaretroviruses. For example, the Pol polyprotein ORF contains: RT, RNaseH, and INT domains in betaretroviruses, measuring approximately 850 codons in length; and AP, RT, RNaseH, and INT domains in gammaretroviridae, measuring approximately 1200 codons in length. Small peaks in the distribution of maximum ORF lengths for these domains can be seen at these lengths. Intriguingly, other peaks at shorter lengths (aside from the main left-most peak) are also present.

To examine whether ERV domain loci that are transcribed in bats are more likely to contain conserved ORFs, each ERV domain locus was categorised as an expressed locus (if a predicted transcript or transcript exon containing the locus had an estimated read count of three or more), or otherwise was classed as an 'unexpressed' locus. The distribution of maximum ORF lengths for expressed ERV domain loci compared with all ERV domain loci for each type of ERV domain was compared (Fig. 5.1). Small peaks corresponding to the intact ORF length of exogenous retroviruses for each type of locus were seen to increase in relative size compared with the background distribution, suggesting that conserved loci with intact ORFs are enriched amongst the set of ERV domain loci that undergo transcription (Fig. 5.1).

Furthermore, the overall ORF length distribution of transcribed ERV loci appears to be shifted to the right compared with the background distribution (to greater and lesser extents) for each domain type. This suggests that ERV domains which were detected as expressed are more likely to have less degraded ORFs than the background set of ERV domain loci. To test this statistically, I carried out Mann-Whitney U tests for each ERV domain type to test whether the distribution of maximum ORF lengths for expressed

ERV domain loci was significantly shifted compared with those in which no expression was detected across the datasets.

When loci from all genomes (for which expression datasets were included) were pooled into one distribution (per domain type), a significant difference was detected for each domain type (P-value < 10^{-29} for each of the seven domain types). Given the reasonable assumption that evolutionary processes will be acting upon these loci in broadly the same way from genome to genome within a single mammalian order, this statistical result suggests the observed pattern may be generally true across all ERV domain loci in bats. However, in order to ensure that these results were not being driven solely by a few datasets with particularly skewed results, I carried out separate Mann-Whitney U tests for each separate dataset and species. Of the 192 comparisons made, 132 were considered significant with an unadjusted P-value threshold of 0.05, of which 75 were considered significant after adjustment for family-wise error rate (FWER) using Bonferroni correction (Table 5.2). For Gag, INT, and Env domains, expression data supported rejecting the null hypothesis in at least half of the experiments, when taken individually after FWER correction.

5.3.3 Immune response RNA-seq datasets show significant differential expression of ERV loci

To identify putative bat EDI genes, I used four published RNA-seq datasets from studies that characterised the transcriptional changes of bat individuals or bat cell cultures in response to immune challenge or immune stimulation. These datasets include: a comparison of transcriptomes from *Artibeus jamaicensis* liver, spleen, and kidney tissue between a control group of individuals (n=2) and a group of individuals of the same species that was experimentally infected with Tacaribe virus (TCRV) (n=2) (Gerrard et al. 2017); a comparison of transcriptomes across three timepoints of *Eidolon helvum* cell cultures (n=3 for each timepoint) inoculated with Zaire ebolavirus (EBOV) (Wynne et al. 2017); a comparison of transcriptomes across five timepoints of *Pteropus alecto* cell cultures (n=3 for each timepoint) after stimulation with IFNα (Cruz-Rivera et al. 2018); and a comparison of transcriptomes across three timepoints of *Rousettus aegyptiacus* cell cultures inoculated with Sendai Cantell virus (SeV) and a control group (n=3 for each timepoint for each treatment type) (Pavlovich et al. 2018). Although two additional published datasets with similar experimental designs were available (Wynne et al. 2014; Hölzer et al. 2016), these did not have more than one replicate for each treatment, precluding robust statistical analysis of differentially expressed transcripts.

For this analysis, transcripts derived from both ERVs and non-retroviral EVEs were examined using the ERV and non-retroviral EVE annotations generated earlier. First, each RNA-seq read library was mapped to the appropriate genome, then used to predict transcripts and determine read counts for each transcript. Next, EVE-derived transcripts were identified using the previously generated EVE annotations. Due to the large proportion of multimapping reads in these transcripts (caused by the high copy numbers of many ERVs), read counts were adjusted for EVE-derived transcripts in order to improve their accuracy. This was carried out using SQuIRE count which redistributes reads using an expectation maximisation algorithm (W.R. Yang et al. 2019). The unadjusted read counts for non-EVE-derived transcripts and these adjusted counts for EVE-derived transcripts were then used for downstream analysis.

To qualitatively assess whether there were overall changes in expression of EVE-derived transcripts and, therefore, whether there were any signs of an EDI response to immune stimulation, PCA plots of log normalised read counts for all EVE-derived transcripts were generated for each of the four comparative transcriptomic datasets (Fig. 5.2). The PCA plots for the *E. helvum*–EBOV experiment experiment showed strong clustering of samples by timepoint after infection, suggesting that the overall transcriptional profile of EVE loci changes over the course of this viral infection. Although fewer replicates were carried out in the *A. jamaicensis*–TCRV experiment, clustering by tissue seems evident as well as a notable difference, along the PC1 axis between spleen samples from infected individuals and spleen samples from control individuals. This suggests overall EVE locus transcription is different across different tissues in *A. jamaicensis*, and that the overall transcriptional profile of EVE loci in the spleen changes in response to immune challenge. No clear clustering of samples based on condition are apparent for the samples in the *P. alecto*–IFNα experiment or the *R. aegyptiacus*–SeV experiment.

Next, to identify specific EVE loci that showed significant differential expression in response to immune stimulus, I carried out differential expression analysis for each of these four datasets.

For the *A. jamaicensis*–TCRV dataset, expression was compared between the infected and control group for each tissue type individually, resulting in a total of 38 EVE-derived transcripts differentially expressed within at least one tissue type (19 in at least two tissue types, eight in all three tissue types) of these, 20 had increased expression in the infected group and 18 had decreased expression in the infected group (Fig. D.1).

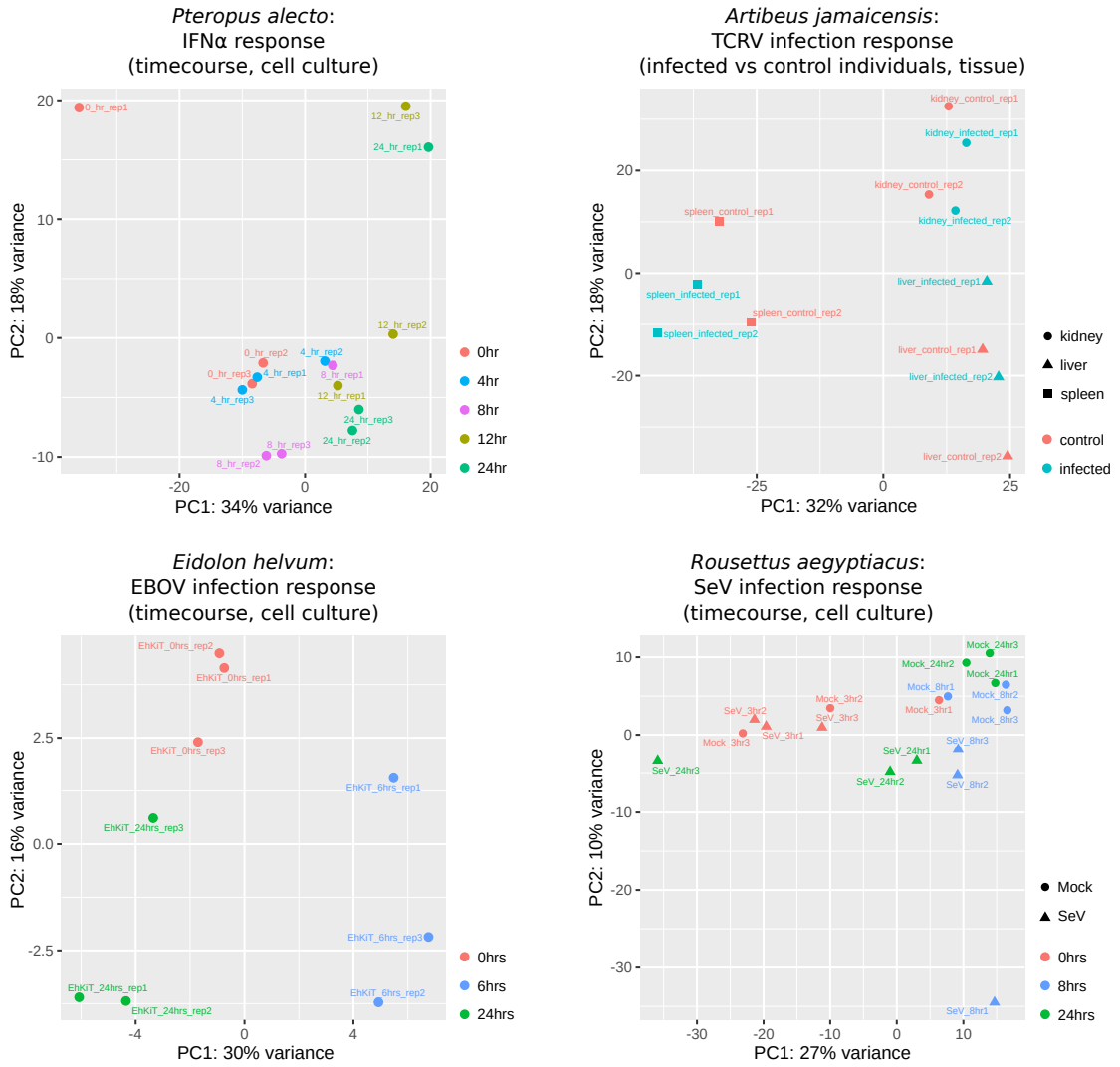


Figure 5.2: Principal component analysis plots for each of the four comparative immune response datasets.

For the *P. alecto*–IFNα dataset, expression was compared between the 0hr timepoint and each subsequent timepoint individually, resulting in a total of 48 EVE-derived transcripts differentially expressed during at least one post inoculation timepoint with respect to the 0hr timepoint (Fig. D.2). Of these, 44 had increased expression and four had decreased expression in the later timepoints compared with the 0hr timepoint.

For the *E. helvum*–EBOV dataset, expression was compared between the 0hr timepoint and the two subsequent timepoints individually, resulting in a total of 30 EVE-derived transcripts differentially expressed during at least one post inoculation timepoint with respect to the 0hr timepoint. Of these, 17 had increased expression and 13 had decreased expression in the later timepoints compared with the 0hr timepoint (Fig. D.3). Of these 30 differentially expressed transcripts, none appear to be orthologues to the 48

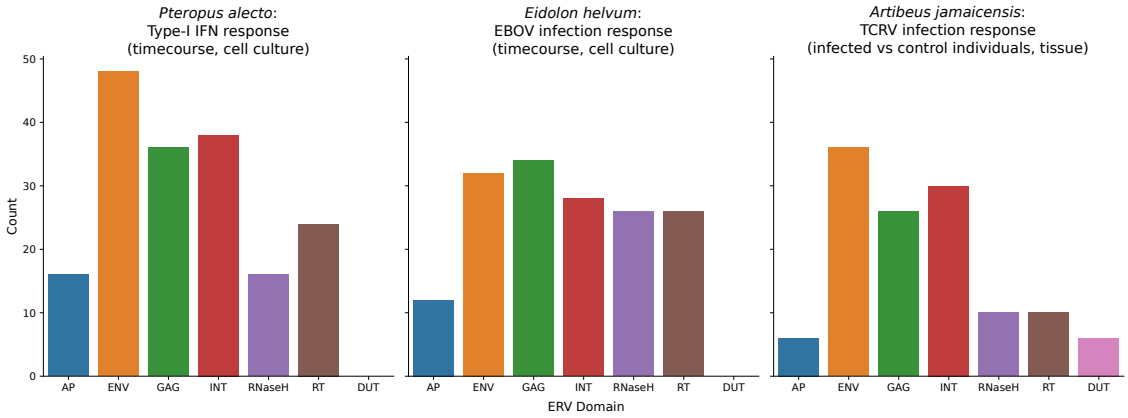


Figure 5.3: Counts of ERV domains, by domain type, present in differentially expressed transcripts from each of the three comparative immune response datasets to show differentially expressed EVE-derived transcripts.

differentially expressed EVE-derived transcripts in *P. alecto*, despite both *P. alecto* and *E. helvum* being members of the Pteropodidae family.

Finally, for the *R. aegyptiacus*–SeV dataset, expression was compared individually for each timepoint between the control group and the infected group. However no EVE-derived transcripts were individually detected to be differentially expressed at the 0.05 significance level for any of these comparisons in this dataset despite the apparent pattern of overall EVE-derived transcript expression visible in the PCA plot (Fig. 5.2).

The types of domains contained within the all three sets of differentially expressed transcripts were determined based on the previously generated EVE annotations; interestingly, no non-retroviral EVE domains were found within these transcripts. For each experiment, the number of each type of ERV domain present in the differentially expressed ERV-derived transcripts was counted. This showed an elevated number of Env, Gag, and IN domains to be present compared to other ERV domains in all three sets of differentially expressed ERV-derived transcripts (Fig. 5.3).

Of the total set of 116 EVE-derived differentially expressed transcripts, 57 corresponded to loci that were located at least 10,000bp away from the edge of a genomic scaffold, this enabled useful manual inspection of their genomic neighbourhoods. The manual inspection of each genomic neighbourhood determined that: three transcripts were not EVE-derived, but spurious hits to cellular genes; 14 were cellular gene loci that were potentially disrupted by the insertion of an ERV; and 33 were transcripts derived solely from ERV loci. Furthermore, most of these non-spurious transcripts appeared to contain only disrupted ERV ORFs with only five transcripts containing a somewhat conserved ERV ORF.

One of these differentially expressed ERV loci, found in *P. alecto*, is shown in Fig. 5.5A, where expression of this locus was found to be significantly upregulated 12hrs and 24hrs after treatment with IFNα (*P. alecto*–IFNα experiment), suggesting this locus is linked to interferon signalling pathways. Interestingly, the expressed transcripts at this locus span across two adjacent ERVs and all domains present within the transcripts appear to have disrupted ORFs suggesting a lack of coding potential. Based on the EVE orthology analysis (Section 5.2.3), no ERVs syntenic to this locus were discovered in other bat genomes.

5.3.4 Searching for co-opted EVEs using evolutionary conservation

Aside from identifying expression patterns that indicate functionality, the co-option of an EVE can also be inferred based on whether it displays high levels of evolutionary conservation. To identify putative co-opted EVE loci, I identified orthologous groups of EVE loci across genomes within each of the four superfamilies of Chiroptera. This was carried out using an approach based on nucleotide similarity of the loci and their flanking regions (Section 5.2.3). This resulted in 5,305 orthologous groups of EVEs. Next, I used divergence dates obtained from TimeTree (Kumar et al. 2017) and Potter (2019) for each of the bat species sampled in this study to provide a minimum age for each locus based on the species in which it was present. Using the combined criteria of long ORF length and ancient locus age allows identification of EVE ORFs which have been maintained over significant periods of evolutionary time and are thus good candidates as functional, co-opted, EVE-derived genes (5.4).

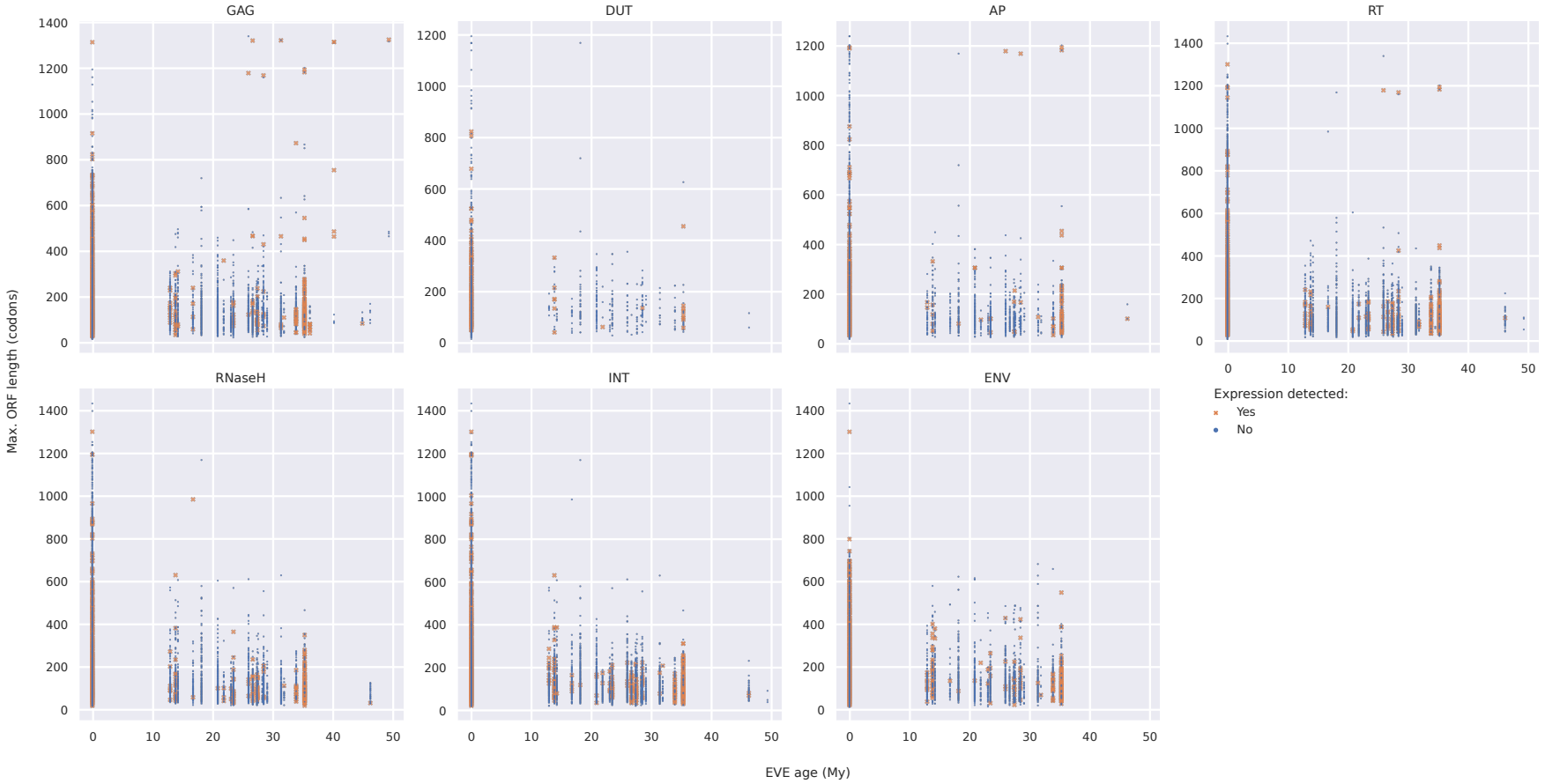


Figure 5.4: Scatter plot of maximum ORF length against estimated minimum locus age for each ERV domain locus. Minimum locus age estimated using divergence dates of species in which each locus is present (a minimum locus age of 0 implies no orthologues for a particular locus were discovered).

Based on these criteria, a number of EVEs were examined manually to determine particularly interesting candidates of co-option using the DVORFS annotations and StringTie transcript annotations generated previously, as well as BLAST searches against the GenBank nr database (Clark et al. 2016). An example of such an EVE is shown in Fig. 5.5B which is found in the *P. alecto* genome. This gammaretroviral ERV insertion is also found in a syntenic location in *E. helvum* and *Macroglossus sobrinus* genomes implying insertion before the divergence of the Pteropodidae family 35.2 Mya. In *P. alecto,* the RNaseH domain of this ERV sits within an intact ORF (351 codons in length) and is present within some isoforms of the transcript derived from this locus, whereas all other domains in the ERV appear to be disrupted. Interestingly, this ORF is not conserved in *E. helvum* and *M. sobrinus,* and, although the locus is found to be transcribed in *E. helvum* kidney cell culture, only transcripts that did not contain the RNaseH domain were detected.

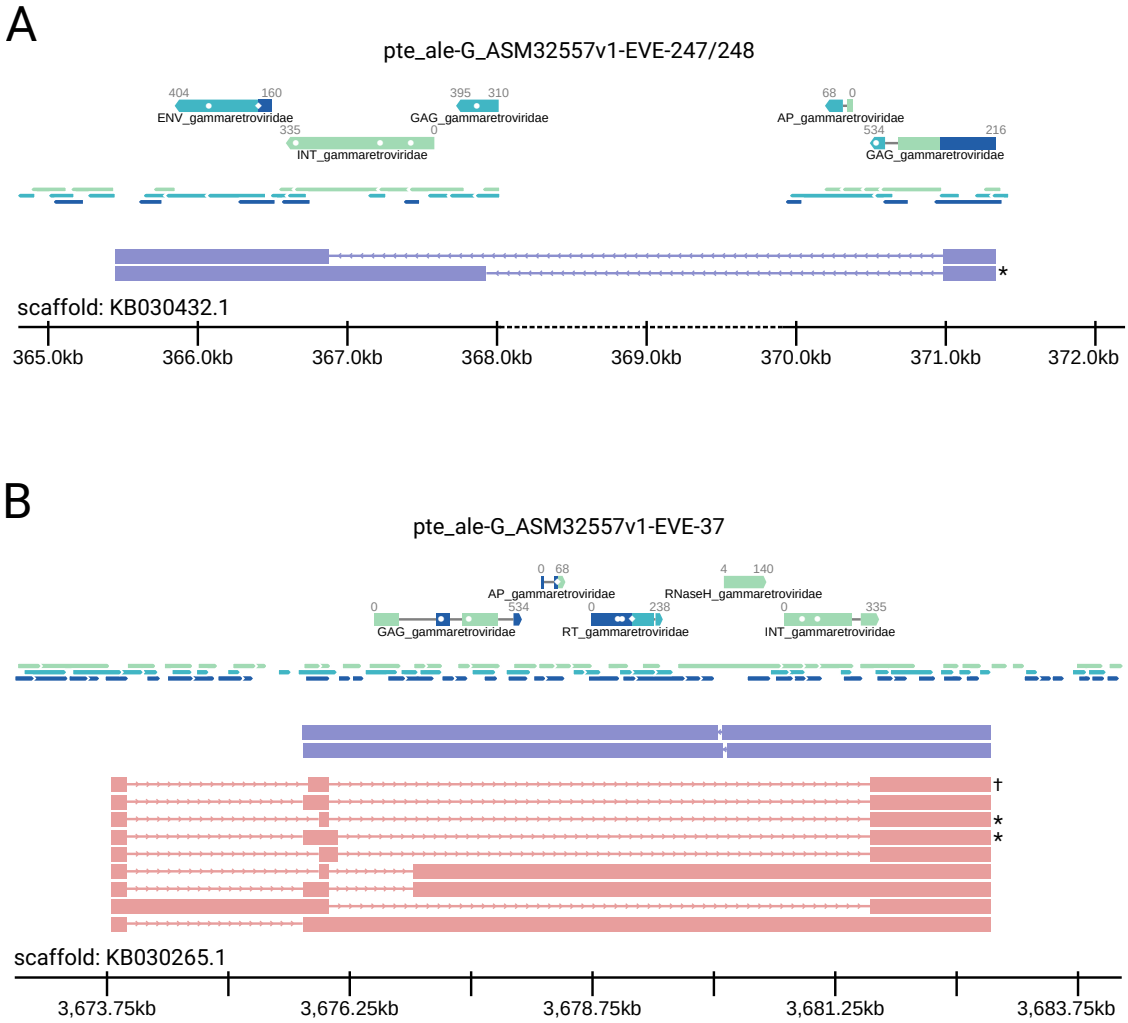


Figure 5.5: Visualisation of two ERV loci of interest showing DVORFS annotations, ORF locations, StringTie predicted transcripts, and genomic co-ordinates. DVORFS hits matching ERV domains are coloured by reading frame, frameshifts within domain coding sequence are marked with diamonds, and premature stop codons within domain coding sequence are marked with circles. Locations of ORFs >33 codons in length are coloured by reading frame, and predicted transcript intron-exon structure are coloured by predicted strand orientation. **A)** visualisation of an ERV locus of interest in *P. alecto*, differentially expressed in *P. alecto*–IFNα experiment with transcript annotations generated using RNA-seq data from Cruz-Rivera et al. (2018). Dashed region on the co-ordinate line indicates ~2kb region of missing sequence in the genomic scaffold. **B)** visualisation of an ERV locus of interest in *P. alecto*, with intact RNaseH domain, and orthologues in *E. helvum* and *M. sobrinus*. Transcript predictions from Cruz-Rivera et al. (2018). * Transcript isoform detected in both Cruz-Rivera et al. (2018) and Wynne et al. (2014) datasets. † Similar transcript found in *E. helvum* in Wynne et al. (2017) dataset.

5.4 Discussion

In this study I sought to compile the largest survey to date of bat ERVs and expressed bat ERVs using 35 genome assemblies, and 189 RNA-seq read libraries. To achieve this, a pipeline was created in order to carry out annotation of ERV domains in each genome using the frameshift-aware alignment approach implemented in DVORFS (see Chapter 2 for benchmarking and details of the approach employed for the creation of this tool), as well as mapping of RNA-seq reads and estimation of ERV-derived transcript abundance. The 296,289 ERV domain annotations generated and the 10,181 expressed ERV domains detected by this pipeline could then be used for downstream analyses of ERV expression patterns and identification of putatively co-opted ERVs and non-retroviral EVEs.

5.4.1 Transcribed ERV loci are often less degraded than untranscribed ERV loci

One signature of a co-opted, protein coding, EVE-derived gene is a conserved ORF and active transcription of the locus. Therefore, by identifying ERV domains with long conserved ORFs which are also included in expressed transcripts, potentially co-opted ERV domains can be identified. The presence of such conserved ORFs is confirmed by the small peaks present in the distribution of maximum ORF lengths for ERV domain loci which correspond to the original ORF length found in exogenous retroviral genomes (Fig. 5.1). Notably, when the distribution of maximum ORF length for only expressed ERV loci is plotted, the peaks at these original ORF lengths increase, implying that loci with original length ORFs are enriched in the set of expressed ERV loci. This is consistent with the presence of intact ERV genes that are transcriptionally active and produce protein-coding transcripts. These genes could either be part of currently active ERVs in the process of replication (through retrotransposition or virion construction), or they could be part of ERVs which are no longer able to replicate but which have nonetheless retained an intact ORF. In the latter case, if the ERV is not young, it is highly likely that the ORF encodes a functional protein of benefit to the host (i.e. it has been co-opted) and has thus been conserved through purifying selection.

To narrow the search to only loci which are more ancient, and thus in which long ORF lengths are more indicative of co-option, minimum age of insertion for each locus was determined based on presence of orthologues within other species (Fig. 5.4). Although this was a particularly conservative means of age determination, it does favour the discovery of co-opted, protein coding EVE loci. This is because an EVE locus providing

a beneficial function to the host species would be expected to have a higher level of conservation among the descendent species of that ancestral host.

The maximum ORF length is also a useful metric for the level of degradation of an ERV domain locus; after inactivation of an ERV, the locus will degrade over time as it accrues mutations, of which some will be frameshifting indels and others will be substitutions leading to premature stop codons. This in turn leads to shorter and shorter sections of unbroken reading frames, and thus, when compared with the original ORF length, lower maximum ORF lengths correspond to more degraded loci. The large left-hand peaks that dominate the distributions of ORF lengths in Fig. 5.1 reflect the fact that an overwhelming proportion of ERVs are pseudogenised and highly degraded; this peak, caused by random mutations and a lack of selection on these loci, has a shape that is likely determined by the ages of the loci present and the neutral mutation rate of the host (and could potentially be modelled as a sum of Poisson-like distributions).

When the distribution of only expressed ERV domain loci is plotted, this large leftmost peak is still present, suggesting that the majority of ERV-derived transcripts are non-coding, which in turn implies that bat ERVs play a similarly important role in lncRNA repertoires as they do in other vertebrates (Kapusta et al. 2013; Craig et al. 2018). Interestingly, the overall distribution of ORF lengths for expressed ERV domain loci is significantly shifted to the right compared with ERV domain loci for which no expression was detected, for each type of ERV domain (as determined by the Mann-Whitney U test). Some part of this rightward shift is due to the increased relative proportion of intact ORFs amongst the expressed ERV loci. However for some types of domain, the large leftmost peak is itself slightly shifted rightwards. This suggests that ERV-derived transcripts without coding potential are derived from less degraded ERVs than the background set of all ERV loci present within bat genomes. This observation could be explained by two possible scenarios. First, younger (i.e. less degraded) ERVs are more likely to give rise to transcripts, despite showing pseudogenisation due to the retention of intact promoter/enhancer regions. Or, alternatively, a substantial number of ERV-derived transcripts are functional lncRNAs and hence experience a slower mutation rate, leading to sequences appearing less degraded when measured with the maximum ORF length metric. Although it is now well-demonstrated that ERVs and other TEs are the progenitors of the majority of lncRNAs found in vertebrates, the functionality (or perhaps non-functionality) of lncRNAs are much less well characterised in any organism due to the volatility of their evolution (Ponjavic et al. 2007; Kapusta et al. 2013; Johnsson et al. 2014; Necsulea et al. 2014; Craig et al. 2018). This makes it difficult to speculate which of these two hypotheses is more likely.

5.4.2 Transcriptomic immune responses in bats include the up- and down-regulation of ERV-derived transcripts

Another hallmark of a co-opted EVE-derived gene – whether protein-coding or not – is the existence of a facultative expression pattern that suggests inclusion of the EVE-derived gene in the host's regulatory networks. EDI genes are a particularly common example of this, where EVE loci become co-opted for immune functionality in the host and thus exhibit an immune response-specific expression pattern. To this end, in order to identify putative bat EDI genes, I sought to take advantage of four datasets from previous transcriptomic studies which examined the transcriptional changes in bats in response to immune challenges or immune stimulation (Gerrard et al. 2017; Wynne et al. 2017; Cruz-Rivera et al. 2018; Pavlovich et al. 2018). For the *E. helvum*–EBOV experiment which used bat cell cultures inoculated with an RNA virus, the qualitative assessment of overall EVE expression patterns using PCA plots showed strong clustering of samples by timepoint post inoculation (Fig. 5.2). Additionally, the *A. jamaicensis*–TCRV experiment which compared different tissues from individuals infected with an RNA virus also showed a notable difference between overall EVE locus expression in the spleen (an organ heavily involved in immune response) of healthy vs infected bats. Both of these observations therefore suggest that immune response to RNA viruses involves broad changes to the expression of EVE loci in bats from across Chiroptera although the *R. aegyptiacus*–SeV does not show a similar pattern. Neither is a clear pattern evident in the PCA plot of EVE expression for the *P. alecto*–IFNα experiment.

To identify specific EVE loci which may play a functional role in these immune responses, I carried out differential expression analysis for each of the four datasets. Across the four datasets, I identified 116 EVE-derived differentially expressed transcripts, none of which contained viral domains originating from outside the *Retroviridae*. Notably, none of the immune challenges in these experiments were carried out using viruses that are related to the EVEs present within the genomes of the four bat species examined (*Parvoviridae, Bornaviridae,* and *Filoviridae*). This perhaps fits with the pattern of known examples of co-opted non-retroviral EVE loci only conferring specific immunity against viruses related to the EVE itself (Fujino et al. 2014; Kim et al. 2020), whereas co-opted ERVs seem to be broadly linked to more general innate immunity pathways (Chuong et al. 2016).

Of the 54 manually examined differentially expressed ERV loci, only five contained ORFs that were relatively undisrupted, implying that if these genes are functional, most possess non-coding rather than coding functionality. This is expected, given that ERVs are

seldom found in protein coding transcripts (Kapusta et al. 2013) and examples of ERV-derived functional lncRNAs have been previously observed (Durruthy-Durruthy et al. 2016; Frank and Feschotte 2017). Additionally, in agreement with the high proportion of lncRNAs in this set of loci, none of the loci appear to be orthologous between the three species, which is also in line with findings that lncRNAs undergo a high rate of evolutionary turnover and are most commonly lineage specific (Johnsson et al. 2014; Kapusta and Feschotte 2014).

In both *E. helvum*–EBOV and *A. jamaicensis*–TCRV experiments, the number of upregulated differentially expressed ERV-derived transcripts was roughly equal to the number of downregulated differentially expressed ERV-derived transcripts, suggesting that bat EDIs could be equally likely to act as immune inhibitors as immune enhancers. Interestingly, in the *P. alecto*–IFNα experiment, no obvious clustering in the PCA plot was observed, but 48 differentially expressed ERV-derived transcripts were observed. This implies that the majority of expressed ERV loci in *P. alecto* cells do not respond to IFNα, but that a small number have evolved to do so. This, together with the fact that 44 out of 48 differentially expressed ERV-derived transcripts were upreguated rather than downregulated in response to IFNα, points to a similar scenario to that observed in human cells, in which ERV co-option is frequently related to the interferon-inducible enhancer activity of the ERV sequence (Chuong et al. 2016).

Although determining the functionality of these putative EDI genes requires much further work, it is tempting to speculate that these co-opted genes could form part of the enhanced interferon response found in bats (J.A. Hayward et al. 2018). A further possibility is that these loci may be involved in activation of genes that are found to be generally activated through interferon signalling in other mammals, but which are hypothesised to be activated through alternative routes in bats (Glennon et al. 2015). A third possibility is that a number of these loci may simply neighbour immune-related cellular genes and are co-expressed alongside them as has been observed in human cells responding to Dengue virus infection (M. Wang et al. 2020).

5.4.3 Candidate EVE loci for further investigation

The various analyses carried out in this study provide the most thorough survey to date for identifying co-opted EVEs in bats. The long-list of EVE loci with signatures of functionality and co-option across Chiroptera which was generated by this study are therefore of particular interest for further investigation. However due to the scope of this

study I have picked three example loci which are illustrative in addition to possessing particularly strong signatures of co-option making them especially worthy of attention.

The first of these exemplary EVEs is one of the three paralogous copies of *chEBLNc* found in each of *M. natalensis* and *M. schreibersii* genomes. This EVE is a strong candidate as a co-opted gene due to the conservation of its expression in the embryos of both of these *Miniopterus* species. In addition, this EVE contains a conserved coding region with an undisrupted ORF that spans 386 codons in both species, whereas the two other paralogous EVEs present in each of the two species both contain disrupted ORFs with no detected expression in either of the two RNA-seq datasets. This suggests that this insertion is not too young to have been pseudogenised under relaxed selection, given that the paralogous copies have done so. Instead, it supports the hypothesis that purifying selection has acted to maintain the ORF at this locus in both species, as well as maintain regulatory sequences that enable transcription of the locus. Together, these pieces of evidence point to this copy of *chEBLNc* functioning as a beneficial protein-coding gene in *Miniopterus,* having been derived from an orthobornavirus nucleocapsid gene (Chapter 3 for details). A similar example of a co-opted protein-coding bornavirus nucleocapsid gene (*itEBLN*) has been characterised in the ground squirrel, where it was shown to confer innate immunity against bornavirus infection through interfering with the replication of the exogenous virus (Fujino et al. 2014; Kim et al. 2020). It may therefore be the case that, like *itEBLN*, *chEBLNc* is an EDI gene and similarly provides some form of innate immunity in *Miniopterus*. However, functional assays would be required to determine if this hypothesis is correct.

The second notable putative co-opted EVE is also of interest due to its protein-coding potential. The ERV shown in Fig. 5.5B (labelled as 'pte_ale-G_ASM32557v1-EVE-37'), is present in the *P. alecto* , *E. helvum* and *M. sobrinus* genomes at syntenic locations, suggesting integration occurred in the ancestor of the Pteropodidae over 35.2Mya. Strikingly, the RNaseH coding region of this ERV is conserved and transcribed in *P. alecto*, although is not conserved in *E. helvum* and *M. sobrinus*. Furthermore, although a transcript derived from this ERV locus was detected in *E. helvum*, none of the isoforms contained the RNaseH sequence. According to the same logic as in the case of *chEBLNc*, the degradation of related coding regions (both the other ERV genes surrounding RNaseH, and the copies present in *E. helvum* and *M. sobrinus*) in contrast to the conserved RNaseH ORF suggest the action of purifying selection at this locus and, therefore, co-option. Although further work is necessary to determine whether a functional protein product is generated, it may be a rare example of a co-opted retroviral RNaseH domain. Two

other RNaseH co-option event are noted in the literature: an intact gypsy-type retrotransposon (with an intact RNaseH coding region) was discovered in human, sheep, and mouse genomes (Lynch and Tristem 2003); and the *CGIN1* gene of retroviral origin, which contains an intact RNaseH domain in addition to an IN and other domains, and appears to play a role in anitviral immunity through its involvement in the ubiquitination of viral proteins (Marco and Marín 2009). In both examples, the function of the RNaseH domain is unclear and therefore the function of this putatively co-opted RNaseH gene is hard to speculate upon. It may the case that it plays a role in antiviral immunity by leveraging its original function of RNA hydrolysis or RNA binding activity, or perhaps it hampers retrovirus replication through interference in capsid assembly in the vein of other co-opted retroviral genes such as *Fv1* (Yap et al. 2014).

The final notable example is also found in the *P. alecto* genome and is shown in Fig. 5.5A (labelled as 'pte_ale-G_ASM32557v1-EVE-247/248'). In comparison to the previous two examples, this locus does not possess undisrupted ORFs. In fact, all the coding regions of the pair of tandem ERVs that comprise this locus, are disrupted and syntenic loci were not found in the other bat genomes in this study (although this may be in part due to the ~2kb region of missing sequence between the two ERVs). Interestingly, however, transcription of this locus was found to be significantly upregulated in response to IFNα, which suggests a functional role for the lncRNA derived from this locus as part of the interferon regulatory network. This is particularly intriguing given the expanded interferon-stimulated gene (ISG) repertoire found in *P. alecto* (Cruz-Rivera et al. 2018), and the fact that ERVs have been found to easily 'hook in' to interferon signalling networks in other species of mammal (Chuong et al. 2016).

6

General Discussion

Endogenous viral elements (EVEs) are curious genomic features embedded within eukaryotic genomes. They represent the 'fossilised' genetic remnants of viral infections that occurred in the ancestor of the host, often millions of years in the past. EVEs therefore provide a window – albeit a somewhat murky one – into host–virus interactions occurring deep in the past. Because bats occupy such a broad range of ecological niches, exhibit unique physiological traits, and harbour a high diversity of viruses, their interactions with viruses are of particular interest in two main areas. Firstly, bats are of particular importance for human and veterinary health due to their implication as viral reservoirs for a number of zoonotic viruses of major concern, including the globally significant SARS-CoV-2 pandemic (Latinne et al. 2020). Secondly, the evolutionary radiation that generated a wide diversity of ecological traits across bat species provides a useful model for examining the relationship between ecological host traits and virus ecology as well as co-speciation dynamics (Luis et al. 2015; Mélade et al. 2016). In this project, I have leveraged the information encoded within bat EVEs to glean new insights into general patterns of bat–virus interactions as well as specific bat–virus associations.

In order to accurately identify EVEs across many gigabases of genomic sequence, I developed a new tool, named 'Disrupted Viral ORF Search' (DVORFS), which employs the frameshift-aware, pHMM-based, alignment algorithm of the GeneWise software (Birney et al. 2004) to accurately recover EVE sequences despite the frequent presence of frameshifting mutations within them. This novel EVE-mining approach appears to improve EVE detection over previous EVE-mining methods, which are not as sensitive

to EVEs that are distantly related to sequences in reference databases, and also often require intensive manual curation to extract more than just small fragments of a given EVE sequence (Zhu et al. 2018; Skirmuntt and Katzourakis 2019). Interestingly, the benchmarking of the DVORFS tool on a panel of genome assemblies uncovered the substantial differences between genome assembly methods in the quantity of endogenous retroviruses (ERVs) present within an assembly. This result likely extends to other families of transposable elements, and has important implications for studies which compare transposable element content across genomes, highlighting the need to account for the assembly method when making meaningful comparisons.

By applying DVORFS to identify ERV reverse transcriptase (RT) domains in 19 bat genome assemblies from across the Chiroptera - all of which were generated using consistent methods - allowed a detailed and unbiased comparison of the ERV composition between the various bat lineages represented. This revealed that beta- and gammaretroviruses dominate the ERV landscape in most bat species as is typical in vertebrates (A. Hayward et al. 2015), and the total abundance of ERVs is mostly consistent across Chiroptera. Intriguingly, specific bat species appeared to be major outliers in terms of ERV abundance and phylogenetic composition. A particularly extreme example is the *Megaderma lyra* genome, in which lineages from all three classes of ERVs have undergone major expansions leading to threefold the total number of ERVs compared with the mean across the 19 genomes. In contrast, other species such as *Pipistrellus pipistrellus* and *Miniopterus schreibersii*, have a much smaller complement of ERVs compared with the majority of the species examined. In general, a minority of bats appear to contain ERV lineages that have managed to escape the regulatory mechanisms of the host and undergone significant expansion, whereas other select species have managed to dramatically reduce overall ERV activity or pushed specific ERV lineages to extinction. The occurrences of these ERV lineage expansions and extinctions do not follow any obvious pattern, however, future genomic analyses may be able to uncover the specific mechanisms responsible in each case. Just as with the increasingly apparent species-specific nature of adaptations within bat immune systems (Banerjee, Baker, et al. 2020), specific unique adaptations to the internal threat of transposable elements may be similarly present across the Chiroptera.

Because ERV lineages within a genome are the result of retrovirus infections in the host, the phylogenetic relationships between the ERVs identified across the genomes also contains information about the long term history of cross-species retrovirus transmission amongst bats. Using phylogenetic reconciliation it was therefore possible to

infer specific retroviral transmission events between bat lineages across the evolutionary history of Chiroptera in a systematic way. Furthermore, this enabled a comparison of the rates of cross-species retrovirus transmissions in different clades of bat, showing that cross-species transmission has been occurring at an increased rate within the Noctilionoidea superfamily. The Noctilionoidea are predominantly endemic to the Neotropics and thus many of the lineages within the clade have evolved in sympatry with one another whereas the Vespertilionoidea and Rhinolophoidea clades are much more geographically spread out, having dispersed across numerous continents throughout their history (Teeling et al. 2005). This is consistent with previous findings that sympatry (as might be expected) is one of the most important factors determining cross-species viral transmission (Luis et al. 2015; Olival et al. 2017; Albery et al. 2020; Mollentze and Streicker 2020).

Although my finding that level of range overlap most likely explains differences in cross-species transmission rates is not new, taken together with smaller patterns within the inferred data that are consistent with known historical biogeography, it suggests that the systematic method used here is robust and scalable across broad host phylogenies. This is despite the somewhat tricky nature of ERV identification and retrovirus phylogeny building that underlies the inferred transmission events. Future work applying this method with greater samples of bat genomes may reveal more detail about the historic patterns of retrovirus transmission amongst bats. Additionally, this method could be applied to other host clades or a broader phylogenetic scope in order to inform our understanding of cross-species viral transmission more generally. For example, an analysis of this sort could also help to reveal specific unexpected viral transmission routes or perhaps aid in inferring historical biogeography of host species. However, caveats to extrapolating trends from retroviruses to other viral clades do apply, given that there can be major differences in viral transmission mechanisms between even closely related viral clades (Latinne et al. 2020; Mollentze and Streicker 2020).

DVORFS was also employed to more generally search for retroviral and non-retroviral domains within the genomes of a total of 35 bat species. This analysis revealed a number of EVEs derived from the *Bornaviridae*, *Filoviridae*, and *Parvoviridae* scattered throughout the genomes, in addition to the abundant ERVs. Interestingly, a tetraparvovirus-derived EVE was identified that was syntenic between bat species across both the Yinpterochiroptera and Yangochiroptera, and furthermore this EVE was found to be present in an additional 56 mammalian genomes across the Scrotifera clade as well as partially present in the genome of the star nosed mole. These observations imply an integration event that predates the divergence of the Laurasiatheria and provides a minimum age

of ~76 My for the *Tetraparvovirus* genus. Other elements revealed by this search were related to the recently discovered *Carbovirus* and *Cultervirus* genera of bornaviruses and substantially expand our knowledge of these so far poorly characterised groups of viruses. The presence of cultervirus-derived elements in vespertilionid bats expands the known host range of the *Cultervirus* genus and provides a minimum age of ~31 My for this genus. The geographic distribution of bat and other mammalian species in which carbovirus-derived EVEs were detected, along with both the inferred minimum age of the carbovirus EVEs in bats (~40 My), and the range of the python species in which extant carboviruses have been identified, points to an intriguing biogeographic history of these viruses. Together, the observations point to a potential African origin of carboviruses followed by dispersal mediated by bats and other vertebrates across the southern continents during the Eocene, leading to an inferred modern day endemism across Africa and the southern continents. These observations therefore lend further weight to the hypothesis that African species were able to reach Australia and Antarctica through land bridges and eventually reach South America through warm corridors across Antarctica during the Eocene (Gunnell et al. 2014; Rojas et al. 2016).

Beyond the historical information that can be gleaned from viral elements, another aspect of EVEs that is particularly intriguing is their occasional co-option. In such situations, it is fascinating to observe the effects of selection on genes that have spent the first part of their life evolving within a virus before integrating into a cellular organism and abruptly beginning to evolve under entirely different selection pressures. The results of the EVE search allowed me to locate particularly interesting instances of EVEs with ancient integration dates that have maintained long ORFs, suggesting that co-option of these elements has occurred. Perhaps the most interesting single example identified was that of *chEBLM*, found across the Phyllostomidae and the Mormoopidae family. This EVE, derived from a gene encoding bornavirus Matrix protein, contains a conserved ORF in eight of the nine genomes in which it is present, suggesting selection has acted to maintain the coding potential of this element. Intriguingly, homology modelling and selection analysis points to functionality that may involve a chEBLM protein binding to the Matrix proteins of infecting bornaviruses, perhaps acting as a restriction factor by hampering capsid assembly or some other part of the viral replication cycle. By additionally combining the results of the EVE search with a sizeable aggregation of RNA-seq data, it was also possible to include information about expression activity of the identified EVEs. Using these data, the list of putatively co-opted EVEs could be refined further to those with conserved ORFs as well as detected expression – another hallmark of functionality and therefore co-option.

Specific RNA-seq datasets also enabled the identification, via differential expression analysis, of a set of ERVs that may be co-opted for immune function within certain bat species. Interestingly, this set of ERVs appears to exclusively produce transcripts without coding potential, illustrating the important connection between the major role of ERVs in generating lncRNA repertoires and their propensity to 'hook in' to immune signalling networks (Kapusta et al. 2013; Chuong et al. 2016). This is of particular interest in relation to evidence that bats may possess expanded sets of interferon-stimulated genes and potential alternative activation pathways (Cruz-Rivera et al. 2018; Glennon et al. 2015), which leads to tempting speculation that bat ERVs may have been crucial to the development of these derived traits in specific species of bat.

In summary, this thesis presents the development of a novel EVE-mining approach, which enables the first large-scale, comprehensive, and comparative, survey of bat EVEs. This survey provides new evidence relating to the host associations and historical biogeography of specific viral groups infecting bats, and also describes broader patterns of cross-species transmission within bats through evolutionary time. Finally, my survey uncovers putative new cases of EVE co-option within bats, and highlights the potential for ERVs to have contributed to the derived immune responses of certain bat species.

References

Ahmed, M. and Liang, P. (2012). Transposable Elements Are a Significant Contributor to Tandem Repeats in the Human Genome. Comparative and Functional Genomics *2012*. Y. Aizawa, ed., 947089. DOI: 10.1155/2012/947089.

Ahn, M., Cui, J., Irving, A.T., and Wang, L.-F. (2016). Unique Loss of the PYHIN Gene Family in Bats Amongst Mammals: Implications for Inflammasome Sensing. Scientific Reports *6*. DOI: 10.1038/srep21722.

Ahn, M. et al. (2019). Dampened NLRP3-Mediated Inflammation in Bats and Implications for a Special Viral Reservoir Host. Nature Microbiology *4*.5 (5), 789–799. DOI: 10.1038/s41564-019-0371-3.

Aiewsakun, P. and Katzourakis, A. (2015). Endogenous Viruses: Connecting Recent and Ancient Viral Evolution. Virology. 60th Anniversary Issue *479–480*, 26–37. DOI: 10.1016/j.virol.2015.02.011.

Aiewsakun, P. and Katzourakis, A. (2016). Time-Dependent Rate Phenomenon in Viruses. Journal of Virology *90*.16, 7184–7195. DOI: 10.1128/JVI.00593-16.

Aiewsakun, P. and Katzourakis, A. (2017). Marine Origin of Retroviruses in the Early Palaeozoic Era. Nature Communications *8*.1 (1), 13954. DOI: 10.1038/ncomms13954.

Aiewsakun, P. and Simmonds, P. (2018). The Genomic Underpinnings of Eukaryotic Virus Taxonomy: Creating a Sequence-Based Framework for Family-Level Virus Classification. Microbiome *6*, 38. DOI: 10.1186/s40168-018-0422-7.

Albery, G.F., Eskew, E.A., Ross, N., and Olival, K.J. (2020). Predicting the Global Mammalian Viral Sharing Network Using Phylogeography. Nature Communications *11*. DOI: 10.1038/s41467-020-16153-4.

Altschul, S.F., Madden, T.L., Schäffer, A.A., Zhang, J., Zhang, Z., Miller, W., and Lipman, D.J. (1997). Gapped BLAST and PSI-BLAST: A New Generation of Protein Database Search Programs. Nucleic Acids Research *25*.17, 3389–3402. DOI: 10.1093/nar/25.17.3389.

Amarasinghe, G.K. et al. (2019). Taxonomy of the Order Mononegavirales: Update 2019. Archives of Virology *164*.7, 1967–1980. DOI: 10.1007/s00705-019-04247-4.

Ammerman, L.K., Lee, D.N., and Tipps, T.M. (2012). First Molecular Phylogenetic Insights into the Evolution of Free-Tailed Bats in the Subfamily Molossinae (Molossidae, Chiroptera). Journal of Mammalogy *93*.1, 12–28. DOI: 10.1644/11-MAMM-A-103.1.

Amman, B.R. et al. (2015). A Recently Discovered Pathogenic Paramyxovirus, Sosuga Virus, Is Present in Rousettus Aegyptiacus Fruit Bats at Multiple Locations in Uganda. Journal of wildlife diseases *51*.3, 774–779. DOI: 10.7589/2015-02-044.

Aswad, A. and Katzourakis, A. (2012). Paleovirology and Virally Derived Immunity. Trends in Ecology & Evolution *27*.11, 627–636. DOI: 10.1016/j.tree.2012.07.007.

Banerjee, A., Rapin, N., Bollinger, T., and Misra, V. (2017). Lack of Inflammatory Gene Expression in Bats: A Unique Role for a Transcription Repressor. Scientific Reports *7*.1 (1), 2232. DOI: 10.1038/s41598-017-01513-w.

Banerjee, A., Baker, M.L., Kulcsar, K., Misra, V., Plowright, R., and Mossman, K. (2020). Novel Insights Into Immune Systems of Bats. Frontiers in Immunology *11*. DOI: 10.3389/fimmu.2020.00026.

Banerjee, A., Zhang, X., Yip, A., Schulz, K.S., Irving, A.T., Bowdish, D., Golding, B., Wang, L.-F., and Mossman, K. (2020). Positive Selection of a Serine Residue in Bat IRF3 Confers Enhanced Antiviral Protection. iScience *23*.3. DOI: 10.1016/j.isci.2020.100958.

Bao, W., Kojima, K.K., and Kohany, O. (2015). Repbase Update, a Database of Repetitive Elements in Eukaryotic Genomes. Mobile DNA *6*.1, 11. DOI: 10.1186/s13100-015-0041-9.

Barzilai, A., Rotman, G., and Shiloh, Y. (2002). ATM Deficiency and Oxidative Stress: A New Dimension of Defective Response to DNA Damage. DNA Repair *1*.1, 3–25. DOI: 10.1016/S1568-7864(01)00007-6.

Barbera, P., Kozlov, A.M., Czech, L., Morel, B., Darriba, D., Flouri, T., and Stamatakis, A. (2019). EPA-Ng: Massively Parallel Evolutionary Placement of Genetic Sequences. Systematic Biology *68*.2, 365–369. DOI: 10.1093/sysbio/syy054.

Belyi, V.A., Levine, A.J., and Skalka, A.M. (2010a). 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. DOI: 10.1128/JVI.01789-10.

Belyi, V.A., Levine, A.J., and Skalka, A.M. (2010b). Unexpected Inheritance: Multiple Integrations of Ancient Bornavirus and Ebolavirus/Marburgvirus Sequences in Vertebrate Genomes. PLOS Pathogens *6*.7, e1001030. DOI: 10.1371/journal.ppat.1001030.

Beltz, L.A. (2018). Bats and Human Health Ebola, SARS, Rabies and Beyond (Hoboken: New Jersey : Wiley Blackwell).

Berman, H., Henrick, K., and Nakamura, H. (2003). Announcing the Worldwide Protein Data Bank. Nature Structural & Molecular Biology *10*.12 (12), 980–980. DOI: 10.1038/nsb1203-980.

Berger, S.A., Krompass, D., and Stamatakis, A. (2011). Performance, Accuracy, and Web Server for Evolutionary Placement of Short Sequence Reads under Maximum Likelihood. Systematic Biology *60*.3, 291–302. DOI: 10.1093/sysbio/syr010.

Best, S., Tissier, P.L., Towers, G., and Stoye, J.P. (1996). Positional Cloning of the Mouse Retrovirus Restriction Gene Fvl. Nature *382*.6594 (6594), 826–829. DOI: 10.1038/382826a0.

Birney, E., Clamp, M., and Durbin, R. (2004). GeneWise and Genomewise. Genome Research *14*.5, 988–995. DOI: 10.1101/gr.1865504.

Boso, G., Buckler-White, A., and Kozak, C.A. (2018). Ancient Evolutionary Origin and Positive Selection of the Retroviral Restriction Factor Fv1 in Muroid Rodents. Journal of Virology *92*.18. DOI: 10.1128/JVI.00850-18.

Bratsch, S., Wertz, N., Chaloner, K., Kunz, T.H., and Butler, J.E. (2011). The Little Brown Bat, M. Lucifugus, Displays a Highly Diverse VH, DH and JH Repertoire but Little Evidence of Somatic Hypermutation. Developmental & Comparative Immunology *35*.4, 421–430. DOI: 10.1016/j.dci.2010.06.004.

Broecker, F. and Moelling, K. (2019). Evolution of Immune Systems From Viruses and Transposable Elements. Frontiers in Microbiology *10*. DOI: 10.3389/fmicb.2019.00051.

Brook, C.E. et al. (2020). Accelerated Viral Dynamics in Bat Cell Lines, with Implications for Zoonotic Emergence. eLife *9*. D. Haydon, ed., e48401. DOI: 10.7554/eLife.48401.

BROAD Institute (2015). *DISCOVAR: Assemble Genomes, Find Variants.* URL: https://software.broadinstitute.org/software/discovar/blog/ (visited on [visited on 10/08/2018]).

Burgin, C.J., Colella, J.P., Kahn, P.L., and Upham, N.S. (2018). How Many Species of Mammals Are There? Journal of Mammalogy *99*.1, 1–14. DOI: 10.1093/jmammal/gyx147.

Camacho, C., Coulouris, G., Avagyan, V., Ma, N., Papadopoulos, J., Bealer, K., and Madden, T.L. (2009). BLAST+: Architecture and Applications. BMC Bioinformatics *10*, 421. DOI: 10.1186/1471-2105-10-421.

Carlson, C.J., Zipfel, C.M., Garnier, R., and Bansal, S. (2019). Global Estimates of Mammalian Viral Diversity Accounting for Host Sharing. Nature Ecology & Evolution *3*.7 (7), 1070–1075. DOI: 10.1038/s41559-019-0910-6.

Chen, L., Liu, B., Yang, J., and Jin, Q. (2014). DBatVir: The Database of Bat-Associated Viruses. Database: The Journal of Biological Databases and Curation *2014*. DOI: 10.1093/database/bau021.

Chen, M. and Cui, J. (2019). Discovery of Endogenous Retroviruses with Mammalian Envelopes in Avian Genomes Uncovers Long-Term Bird-Mammal Interaction. Virology *530*, 27–31. DOI: 10.1016/j.virol.2019.02.005.

Chionh, Y.T., Cui, J., Koh, J., Mendenhall, I.H., Ng, J.H.J., Low, D., Itahana, K., Irving, A.T., and Wang, L.-F. (2019). High Basal Heat-Shock Protein Expression in Bats Confers Resistance to Cellular Heat/Oxidative Stress. Cell Stress and Chaperones *24*.4, 835–849. DOI: 10.1007/s12192-019-01013-y.

Chuong, E.B., Elde, N.C., and Feschotte, C. (2016). Regulatory Evolution of Innate Immunity through Co-Option of Endogenous Retroviruses. Science *351*.6277, 1083–1087. DOI: 10.1126/science.aad5497.

Chuong, E.B., Elde, N.C., and Feschotte, C. (2017). Regulatory Activities of Transposable Elements: From Conflicts to Benefits. Nature Reviews Genetics *18*.2, 71–86. DOI: 10.1038/nrg.2016.139.

Clark, K., Karsch-Mizrachi, I., Lipman, D.J., Ostell, J., and Sayers, E.W. (2016). GenBank. Nucleic Acids Research *44*.D1, D67–D72. DOI: 10.1093/nar/gkv1276.

Cleaveland, S., Laurenson, M.K., and Taylor, L.H. (2001). Diseases of Humans and Their Domestic Mammals: Pathogen Characteristics, Host Range and the Risk of Emergence. Philosophical Transactions of the Royal Society B: Biological Sciences *356*.1411, 991–999. DOI: 10.1098/rstb.2001.0889.

Contreras-Galindo, R., López, P., Vélez, R., and Yamamura, Y. (2007). HIV-1 Infection Increases the Expression of Human Endogenous Retroviruses Type K (HERV-K) in Vitro. AIDS Research and Human Retroviruses *23*.1, 116–122. DOI: 10.1089/aid.2006.0117.

Cornelis, G. et al. (2017). An Endogenous Retroviral Envelope Syncytin and Its Cognate Receptor Identified in the Viviparous Placental Mabuya Lizard. Proceedings of the National Academy of Sciences *114*.51, E10991–E11000. DOI: 10.1073/pnas.1714590114.

Cotmore, S.F. et al. (2019). ICTV Virus Taxonomy Profile: Parvoviridae. Journal of General Virology, *100*.3, 367–368. DOI: 10.1099/jgv.0.001212.

Craig, R.J., Suh, A., Wang, M., and Ellegren, H. (2018). Natural Selection beyond Genes: Identification and Analyses of Evolutionarily Conserved Elements in the Genome of the Collared Flycatcher (Ficedula Albicollis). Molecular Ecology, n/a–n/a. DOI: 10.1111/mec.14462.

Cruz-Rivera, P.C.D.L., Kanchwala, M., Liang, H., Kumar, A., Wang, L.-F., Xing, C., and Schoggins, J.W. (2018). The IFN Response in Bats Displays Distinctive IFN-Stimulated Gene Expression Kinetics with Atypical RNASEL Induction. The Journal of Immunology *200*.1, 209–217. DOI: 10.4049/jimmunol.1701214.

Cui, J., Tachedjian, M., Wang, L., Tachedjian, G., Wang, L.-F., and Zhang, S. (2012). Discovery of Retroviral Homologs in Bats: Implications for the Origin of Mammalian Gammaretroviruses. Journal of Virology *86*.8, 4288–4293. DOI: 10.1128/JVI.06624-11.

Cui, J., Tachedjian, G., and Wang, L.-F. (2015). Bats and Rodents Shape Mammalian Retroviral Phylogeny. Scientific Reports *5*, 16561. DOI: 10.1038/srep16561.

Cui, J. and Wang, L.-F. (2015). Genomic Mining Reveals Deep Evolutionary Relationships between Bornaviruses and Bats. Viruses *7*.11, 5792–5800. DOI: 10.3390/v7112906.

Czech, L. and Stamatakis, A. (2019). Scalable Methods for Analyzing and Visualizing Phylogenetic Placement of Metagenomic Samples. PLOS ONE *14*.5, e0217050. DOI: 10.1371/journal.pone.0217050.

Czech, L., Barbera, P., and Stamatakis, A. (2020). Genesis and Gappa: Processing, Analyzing and Visualizing Phylogenetic (Placement) Data. Bioinformatics *36*.10, 3263–3265. DOI: 10.1093/bioinformatics/btaa070.

Dang, C.C., Lefort, V., Le, V.S., Le, Q.S., and Gascuel, O. (2011). ReplacementMatrix: A Web Server for Maximum-Likelihood Estimation of Amino Acid Replacement Rate Matrices. Bioinformatics *27*.19, 2758–2760. DOI: 10.1093/bioinformatics/btr435.

Diehl, W.E., Patel, N., Halm, K., and Johnson, W.E. (2016). Tracking Interspecies Transmission and Long-Term Evolution of an Ancient Retrovirus Using the Genomes of Modern Mammals. eLife *5*, e12704. DOI: 10.7554/eLife.12704.

Dobin, A., Davis, C.A., Schlesinger, F., Drenkow, J., Zaleski, C., Jha, S., Batut, P., Chaisson, M., and Gingeras, T.R. (2013). STAR: Ultrafast Universal RNA-Seq Aligner. Bioinformatics *29*.1, 15–21. DOI: 10.1093/bioinformatics/bts635.

Dos Reis, M., Inoue, J., Hasegawa, M., Asher, R.J., Donoghue, P.C.J., and Yang, Z. (2012). Phylogenomic Datasets Provide Both Precision and Accuracy in Estimating the Timescale of Placental

Mammal Phylogeny. Proceedings. Biological Sciences *279*.1742, 3491–3500. DOI: 10.1098/rspb.2012.0683.

Durruthy-Durruthy, J. et al. (2016). The Primate-Specific Noncoding RNA HPAT5 Regulates Pluripotency during Human Preimplantation Development and Nuclear Reprogramming. Nature Genetics *48*.1 (1), 44–52. DOI: 10.1038/ng.3449.

Eckalbar, W.L. et al. (2016). Transcriptomic and Epigenomic Characterization of the Developing Bat Wing. Nature Genetics *48*.5 (5), 528–536. DOI: 10.1038/ng.3537.

Eddy, S.R. (2009). A New Generation of Homology Search Tools Based on Probabilistic Inference. Genome Informatics. International Conference on Genome Informatics *23*.1, 205–211.

Eddy, S.R. (2011). Accelerated Profile HMM Searches. PLOS Computational Biology *7*.10, e1002195. DOI: 10.1371/journal.pcbi.1002195.

Edwards, M.R. et al. (2018). Conservation of Structure and Immune Antagonist Functions of Filoviral VP35 Homologs Present in Microbat Genomes. Cell Reports *24*.4, 861–872.e6. DOI: 10.1016/j.celrep.2018.06.045.

El-Gebali, S. et al. (2019). The Pfam Protein Families Database in 2019. Nucleic Acids Research *47*.D1, D427–D432. DOI: 10.1093/nar/gky995.

Escalera-Zamudio, M., Zepeda-Mendoza, M.L., Loza-Rubio, E., Rojas-Anaya, E., Méndez-Ojeda, M.L., Arias, C.F., and Greenwood, A.D. (2015). The Evolution of Bat Nucleic Acid-Sensing Toll-like Receptors. Molecular Ecology *24*.23, 5899–5909. DOI: 10.1111/mec.13431.

Feschotte, C. and Gilbert, C. (2012). Endogenous Viruses: Insights into Viral Evolution and Impact on Host Biology. Nature Reviews Genetics *13*.4, 283–296. DOI: 10.1038/nrg3199.

Frank, J.A. and Feschotte, C. (2017). Co-Option of Endogenous Viral Sequences for Host Cell Function. Current Opinion in Virology. Animal Models for Viral Diseases • Paleovirology *25*, 81–89. DOI: 10.1016/j.coviro.2017.07.021.

Fujino, K., Horie, M., Honda, T., Merriman, D.K., and 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. DOI: 10.1073/pnas.1407046111.

Fushan, A.A. et al. (2015). Gene Expression Defines Natural Changes in Mammalian Lifespan. Aging Cell *14*.3, 352–365. DOI: 10.1111/acel.12283.

Geib, T., Sauder, C., Venturelli, S., Hässler, C., Staeheli, P., and Schwemmle, M. (2003). Selective Virus Resistance Conferred by Expression of Borna Disease Virus Nucleocapsid Components. Journal of Virology *77*.7, 4283–4290. DOI: 10.1128/JVI.77.7.4283-4290.2003.

Gerrard, D.L., Hawkinson, A., Sherman, T., Modahl, C.M., Hume, G., Campbell, C.L., Schountz, T., and Frietze, S. (2017). Transcriptomic Signatures of Tacaribe Virus-Infected Jamaican Fruit Bats. mSphere *2*.5, e00245–17. DOI: 10.1128/mSphere.00245-17.

Gifford, R.J., Blomberg, J., Coffin, J.M., Fan, H., Heidmann, T., Mayer, J., Stoye, J., Tristem, M., and Johnson, W.E. (2018). Nomenclature for Endogenous Retrovirus (ERV) Loci. Retrovirology *15*.1, 59. DOI: 10.1186/s12977-018-0442-1.

Giles, J.R., Eby, P., Parry, H., Peel, A.J., Plowright, R.K., Westcott, D.A., and McCallum, H. (2018). Environmental Drivers of Spatiotemporal Foraging Intensity in Fruit Bats and Implications for Hendra Virus Ecology. Scientific Reports *8*.1 (1), 9555. DOI: 10.1038/s41598-018-27859-3.

Glennon, N.B., Jabado, O., Lo, M.K., and Shaw, M.L. (2015). Transcriptome Profiling of the Virus-Induced Innate Immune Response in Pteropus Vampyrus and Its Attenuation by Nipah Virus Interferon Antagonist Functions. Journal of Virology *89*.15, 7550–7566. DOI: 10.1128/JVI.00302-15.

Gough, J., Karplus, K., Hughey, R., and Chothia, C. (2001). Assignment of Homology to Genome Sequences Using a Library of Hidden Markov Models That Represent All Proteins of Known structure11Edited by G. Von Heijne. Journal of Molecular Biology *313*.4, 903–919. DOI: 10.1006/jmbi.2001.5080.

Gregory, T.R. (2020). *Animal Genome Size Database*.
URL: http://www.genomesize.com/ (visited on [visited on 09/16/2020]).

Groh, S. and Schotta, G. (2017). Silencing of Endogenous Retroviruses by Heterochromatin. Cellular and Molecular Life Sciences *74*.11, 2055–2065. DOI: 10.1007/s00018-017-2454-8.

Gunnell, G.F., Simmons, N.B., and Seiffert, E.R. (2014). New Myzopodidae (Chiroptera) from the Late Paleogene of Egypt: Emended Family Diagnosis and Biogeographic Origins of Noctilionoidea. PLOS ONE *9*.2, e86712. DOI: 10.1371/journal.pone.0086712.

Halpin, K. et al. (2011). Pteropid Bats Are Confirmed as the Reservoir Hosts of Henipaviruses: A Comprehensive Experimental Study of Virus Transmission. The American Journal of Tropical Medicine and Hygiene *85*.5, 946–951. DOI: 10.4269/ajtmh.2011.10-0567.

Harrison, P.M., Hegyi, H., Balasubramanian, S., Luscombe, N.M., Bertone, P., Echols, N., Johnson, T., and Gerstein, M. (2002). Molecular Fossils in the Human Genome: Identification and Analysis of the Pseudogenes in Chromosomes 21 and 22. Genome Research *12*.2, 272–280. DOI: 10.1101/gr.207102.

Hayward, A., Grabherr, M., and Jern, P. (2013). Broad-Scale Phylogenomics Provides Insights into Retrovirus–Host Evolution. Proceedings of the National Academy of Sciences *110*.50, 20146–20151. DOI: 10.1073/pnas.1315419110.

Hayward, J.A., Tachedjian, M., Cui, J., Field, H., Holmes, E.C., Wang, L.-F., and Tachedjian, G. (2013). Identification of Diverse Full-Length Endogenous Betaretroviruses in Megabats and Microbats. Retrovirology *10*.1, 35. DOI: 10.1186/1742-4690-10-35.

Hayward, A., Cornwallis, C.K., and Jern, P. (2015). Pan-Vertebrate Comparative Genomics Unmasks Retrovirus Macroevolution. Proceedings of the National Academy of Sciences *112*.2, 464–469. DOI: 10.1073/pnas.1414980112.

Hayward, J.A., Tachedjian, M., Cui, J., Cheng, A.Z., Johnson, A., Baker, M.L., Harris, R.S., Wang, L.-F., Tachedjian, G., and Teeling, E. (2018). Differential Evolution of Antiretroviral Restriction Factors in Pteropid Bats as Revealed by APOBEC3 Gene Complexity. Molecular Biology and Evolution. DOI: 10.1093/molbev/msy048.

Hayman, D.T.S. (2016). Bats as Viral Reservoirs. Annual Review of Virology *3*.1, 77–99. DOI: 10.1146/annurev-virology-110615-042203.

He, X., Korytář, T., Schatz, J., Freuling, C.M., Müller, T., and Köllner, B. (2014). Anti-Lyssaviral Activity of Interferons κ and ω from the Serotine Bat, Eptesicus Serotinus. Journal of Virology *88*.10, 5444–5454. DOI: 10.1128/JVI.03403-13.

Healy, K., Guillerme, T., Finlay, S., Kane, A., Kelly, S.B.A., McClean, D., Kelly, D.J., Donohue, I., Jackson, A.L., and Cooper, N. (2014). Ecology and Mode-of-Life Explain Lifespan Variation in Birds and Mammals. Proceedings of the Royal Society of London B: Biological Sciences *281*.1784, 20140298. DOI: 10.1098/rspb.2014.0298.

Hoang, D.T., Chernomor, O., von Haeseler, A., Minh, B.Q., and Vinh, L.S. (2018). UFBoot2: Improving the Ultrafast Bootstrap Approximation. Molecular Biology and Evolution *35*.2, 518–522. DOI: 10.1093/molbev/msx281.

Hock, R.J. (1951). The Metabolic Rates and Body Temperatures of Bats. The Biological Bulletin *101*.3, 289–299. DOI: 10.2307/1538547.

Hölzer, M. et al. (2016). Differential Transcriptional Responses to Ebola and Marburg Virus Infection in Bat and Human Cells. Scientific Reports *6*, 34589. DOI: 10.1038/srep34589.

Holmes, E.C. and Duchêne, S. (2019). Evolutionary Stasis of Viruses? Nature Reviews Microbiology *17*.5 (5), 329–329. DOI: 10.1038/s41579-019-0168-7.

Horie, M., Kobayashi, Y., Suzuki, Y., and Tomonaga, K. (2013). Comprehensive Analysis of Endogenous Bornavirus-like Elements in Eukaryote Genomes. Philosophical Transactions of the Royal Society B: Biological Sciences *368*.1626, 20120499. DOI: 10.1098/rstb.2012.0499.

Horie, M. et al. (2016). An RNA-Dependent RNA Polymerase Gene in Bat Genomes Derived from an Ancient Negative-Strand RNA Virus. Scientific Reports *6*.1, 1–9. DOI: 10.1038/srep25873.

Horie, M. and Tomonaga, K. (2019). Paleovirology of Bornaviruses: What Can Be Learned from Molecular Fossils of Bornaviruses. Virus Research. Integration of Viral Sequences into Eukaryotic Host Genomes: Legacy of Ancient Infections *262*, 2–9. DOI: 10.1016/j.virusres.2018.04.006.

Horie, M. (2017). The Biological Significance of Bornavirus-Derived Genes in Mammals. Current Opinion in Virology. Paleovirology *25*, 1–6. DOI: 10.1016/j.coviro.2017.06.004.

Hron, T., Elleder, D., and Gifford, R.J. (2019). Deltaretroviruses Have Circulated since at Least the Paleogene and Infected a Broad Range of Mammalian Species. Retrovirology *16*.1, 33. DOI: 10.1186/s12977-019-0495-9.

Hsiao, F.C., Tai, A.K., Deglon, A., Sutkowski, N., Longnecker, R., and Huber, B.T. (2009). EBV LMP-2A Employs a Novel Mechanism to Transactivate the HERV-K18 Superantigen through Its ITAM. Virology *385*.1, 261–266. DOI: 10.1016/j.virol.2008.11.025.

Hurst, T.P. and Magiorkinis, G. (2017). Epigenetic Control of Human Endogenous Retrovirus Expression: Focus on Regulation of Long-Terminal Repeats (LTRs). Viruses *9*.6, 130. DOI: 10.3390/v9060130.

Hyndman, T.H., Shilton, C.M., Stenglein, M.D., and Jr, J.F.X.W. (2018). Divergent Bornaviruses from Australian Carpet Pythons with Neurological Disease Date the Origin of Extant Bornaviridae Prior to the End-Cretaceous Extinction. PLOS Pathogens *14*.2, e1006881. DOI: 10.1371/journal.ppat.1006881.

Iv, F.A.M. and Evans, S.N. (2013). Edge Principal Components and Squash Clustering: Using the Special Structure of Phylogenetic Placement Data for Sample Comparison. PLOS ONE *8*.3, e56859. DOI: 10.1371/journal.pone.0056859.

Jacox, E., Chauve, C., Szöllősi, G.J., Ponty, Y., and Scornavacca, C. (2016). ecceTERA: Comprehensive Gene Tree-Species Tree Reconciliation Using Parsimony. Bioinformatics *32*.13, 2056–2058. DOI: 10.1093/bioinformatics/btw105.

Jebb, D. et al. (2020). Six Reference-Quality Genomes Reveal Evolution of Bat Adaptations. Nature *583*.7817 (7817), 578–584. DOI: 10.1038/s41586-020-2486-3.

Johnsson, P., Lipovich, L., Grandér, D., and Morris, K.V. (2014). Evolutionary Conservation of Long Non-Coding RNAs; Sequence, Structure, Function. Biochimica et Biophysica Acta (BBA) - General Subjects *1840*.3, 1063–1071. DOI: 10.1016/j.bbagen.2013.10.035.

Jones, G. and Teeling, E.C. (2006). The Evolution of Echolocation in Bats. Trends in Ecology & Evolution *21*.3, 149–156. DOI: 10.1016/j.tree.2006.01.001.

Kacprzyk, J., Hughes, G.M., Palsson-McDermott, E.M., Quinn, S.R., Puechmaille, S.J., O'neill, L.A.J., and Teeling, E.C. (2017). A Potent Anti-Inflammatory Response in Bat Macrophages May Be Linked to Extended Longevity and Viral Tolerance. Acta Chiropterologica *19*.2, 219–228. DOI: 10.3161/15081109ACC2017.19.2.001.

Kalyaanamoorthy, S., Minh, B.Q., Wong, T.K.F., von Haeseler, A., and Jermiin, L.S. (2017). ModelFinder: Fast Model Selection for Accurate Phylogenetic Estimates. Nature Methods *14*.6 (6), 587–589. DOI: 10.1038/nmeth.4285.

Kapusta, A., Kronenberg, Z., Lynch, V.J., Zhuo, X., Ramsay, L., Bourque, G., Yandell, M., and Feschotte, C. (2013). Transposable Elements Are Major Contributors to the Origin, Diversification, and Regulation of Vertebrate Long Noncoding RNAs. PLOS Genetics *9*.4, e1003470. DOI: 10.1371/journal.pgen.1003470.

Kapusta, A. and Feschotte, C. (2014). Volatile Evolution of Long Noncoding RNA Repertoires: Mechanisms and Biological Implications. Trends in Genetics *30*.10, 439–452. DOI: 10.1016/j.tig.2014.08.004.

Kapusta, A., Suh, A., and Feschotte, C. (2017). Dynamics of Genome Size Evolution in Birds and Mammals. Proceedings of the National Academy of Sciences *114*.8, E1460–E1469. DOI: 10.1073/pnas.1616702114.

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

Katoh, K. and Standley, D.M. (2013). MAFFT Multiple Sequence Alignment Software Version 7: Improvements in Performance and Usability. Molecular Biology and Evolution *30*.4, 772–780. DOI: 10.1093/molbev/mst010.

Katzourakis, A., Aiewsakun, P., Jia, H., Wolfe, N.D., LeBreton, M., Yoder, A.D., and Switzer, W.M. (2014). Discovery of Prosimian and Afrotherian Foamy Viruses and Potential Cross Species Transmissions amidst Stable and Ancient Mammalian Co-Evolution. Retrovirology *11*.1, 61. DOI: 10.1186/1742-4690-11-61.

Katzourakis, A. and Aswad, A. (2016). Evolution: Endogenous Viruses Provide Shortcuts in Antiviral Immunity. Current Biology *26*.10, R427–R429. DOI: 10.1016/j.cub.2016.03.072.

Kelley, D. and Rinn, J. (2012). Transposable Elements Reveal a Stem Cell-Specific Class of Long Noncoding RNAs. Genome Biology *13*.11, R107. DOI: 10.1186/gb-2012-13-11-r107.

Kim, K.S., Yamamoto, Y., Nakaoka, S., Tomonaga, K., Iwami, S., and Honda, T. (2020). Modeling Borna Disease Virus In Vitro Spread Reveals the Mode of Antiviral Effect Conferred by an Endogenous Bornavirus-Like Element. Journal of Virology *94*.21. DOI: 10.1128/JVI.01204-20.

Kirsip, H. and Abroi, A. (2019). Protein Structure-Guided Hidden Markov Models (HMMs) as A Powerful Method in the Detection of Ancestral Endogenous Viral Elements. Viruses *11*.4, 320. DOI: 10.3390/v11040320.

Kobayashi, Y., Horie, M., Nakano, A., Murata, K., Itou, T., and Suzuki, Y. (2016). Exaptation of Bornavirus-Like Nucleoprotein Elements in Afrotherians. PLOS Pathogens *12*.8, e1005785. DOI: 10.1371/journal.ppat.1005785.

Koepfli, K.-P., Paten, B., and O'Brien, S.J. (2015). The Genome 10K Project: A Way Forward. Annual Review of Animal Biosciences *3*.1, 57–111. DOI: 10.1146/annurev-animal-090414-014900.

Konkel, M.K. and Batzer, M.A. (2010). A Mobile Threat to Genome Stability: The Impact of Non-LTR Retrotransposons upon the Human Genome. Seminars in cancer biology *20*.4, 211–221. DOI: 10.1016/j.semcancer.2010.03.001.

Kondo, H., Chiba, S., and Suzuki, N. (2015). "Detection and Analysis of Non-Retroviral RNA Virus-Like Elements in Plant, Fungal, and Insect Genomes". Plant Virology Protocols: New Approaches to Detect Viruses and Host Responses, I. Uyeda and C. Masuta, ed. Methods in Molecular Biology. (New York, NY: Springer), pp. 73–88. DOI: 10.1007/978-1-4939-1743-3_7.

Kondoh, T. et al. (2017). Putative Endogenous Filovirus VP35-like Protein Potentially Functions as an IFN Antagonist but Not a Polymerase Cofactor. PLOS ONE *12*.10, e0186450. DOI: 10.1371/journal.pone.0186450.

Krauel, J.J. and McCracken, G.F. (2013). "Recent Advances in Bat Migration Research". Bat Evolution, Ecology, and Conservation, R.A. Adams and S.C. Pedersen, ed. (New York, NY: Springer), pp. 293–313. DOI: 10.1007/978-1-4614-7397-8_15.

Kryukov, K., Ueda, M.T., Imanishi, T., and Nakagawa, S. (2018). Systematic Survey of Non-Retroviral Virus-like Elements in Eukaryotic Genomes. Virus Research. DOI: 10.1016/j.virusres.2018.02.002.

Kuhn, J.H. et al. (2019). ICTV Virus Taxonomy Profile: Filoviridae. Journal of General Virology, *100*.6, 911–912. DOI: 10.1099/jgv.0.001252.

Kumar, S. and Subramanian, S. (2002). Mutation Rates in Mammalian Genomes. Proceedings of the National Academy of Sciences *99*.2, 803–808. DOI: 10.1073/pnas.022629899.

Kumar, S., Stecher, G., Suleski, M., and Hedges, S.B. (2017). TimeTree: A Resource for Timelines, Timetrees, and Divergence Times. Molecular Biology and Evolution *34*.7, 1812–1819. DOI: 10.1093/molbev/msx116.

Kunz, T.H. (1982). "Roosting Ecology of Bats". Ecology of Bats, T.H. Kunz, ed. (Boston, MA: Springer US), pp. 1–55. DOI: 10.1007/978-1-4613-3421-7_1.

Lack, J.B. and Van Den Bussche, R.A. (2010). Identifying the Confounding Factors in Resolving Phylogenetic Relationships in Vespertilionidae. Journal of Mammalogy *91*.6, 1435–1448. DOI: 10.1644/09-MAMM-A-354.1.

Laing, E.D., Sterling, S.L., Weir, D.L., Beauregard, C.R., Smith, I.L., Larsen, S.E., Wang, L.-F., Snow, A.L., Schaefer, B.C., and Broder, C.C. (2019). Enhanced Autophagy Contributes to Reduced Viral Infection in Black Flying Fox Cells. Viruses *11*.3 (3), 260. DOI: 10.3390/v11030260.

Larsson, A. (2014). AliView: A Fast and Lightweight Alignment Viewer and Editor for Large Datasets. Bioinformatics *30*.22, 3276–3278. DOI: 10.1093/bioinformatics/btu531.

Latinne, A. et al. (2020). Origin and Cross-Species Transmission of Bat Coronaviruses in China. Nature Communications *11*.1 (1), 4235. DOI: 10.1038/s41467-020-17687-3.

Lau, S.K.P. et al. (2016). Identification and Interspecies Transmission of a Novel Bocaparvovirus among Different Bat Species in China. Journal of General Virology, *97*.12, 3345–3358. DOI: 10.1099/jgv.0.000645.

Lau, S.K.P. et al. (2017). Bats Host Diverse Parvoviruses as Possible Origin of Mammalian Dependoparvoviruses and Source for Bat–Swine Interspecies Transmission. Journal of General Virology *98*.12, 3046–3059. DOI: 10.1099/jgv.0.000969.

Lee, A.K. et al. (2015). De Novo Transcriptome Reconstruction and Annotation of the Egyptian Rousette Bat. BMC Genomics *16*.1, 1033. DOI: 10.1186/s12864-015-2124-x.

Lei, M. and Dong, D. (2016). Phylogenomic Analyses of Bat Subordinal Relationships Based on Transcriptome Data. Scientific Reports *6*.1 (1), 27726. DOI: 10.1038/srep27726.

Lenth, R. (2020). *Emmeans: Estimated Marginal Means, Aka Least-Squares Means*. manual. ().

Lerat, E. and Ochman, H. (2004). Ψ-Φ: Exploring the Outer Limits of Bacterial Pseudogenes. Genome Research *14*.11, 2273–2278. DOI: 10.1101/gr.2925604.

Leroy, E.M., Kumulungui, B., Pourrut, X., Rouquet, P., Hassanin, A., Yaba, P., Délicat, A., Paweska, J.T., Gonzalez, J.-P., and Swanepoel, R. (2005). Fruit Bats as Reservoirs of Ebola Virus. Nature *438*.7068 (7068), 575–576. DOI: 10.1038/438575a.

Letko, M., Seifert, S.N., Olival, K.J., Plowright, R.K., and Munster, V.J. (2020). Bat-Borne Virus Diversity, Spillover and Emergence. Nature Reviews Microbiology, 1–11. DOI: 10.1038/s41579-020-0394-z.

Lewin, H.A. et al. (2018). Earth BioGenome Project: Sequencing Life for the Future of Life. Proceedings of the National Academy of Sciences *115*.17, 4325–4333. DOI: 10.1073/pnas.1720115115.

Li, W. and Godzik, A. (2006). Cd-Hit: A Fast Program for Clustering and Comparing Large Sets of Protein or Nucleotide Sequences. Bioinformatics *22*.13, 1658–1659. DOI: 10.1093/bioinformatics/btl158.

Li, R. et al. (2010). De Novo Assembly of Human Genomes with Massively Parallel Short Read Sequencing. Genome Research *20*.2, 265–272. DOI: 10.1101/gr.097261.109.

Liesche, F. et al. (2019). The Neuropathology of Fatal Encephalomyelitis in Human Borna Virus Infection. Acta Neuropathologica *138*.4, 653–665. DOI: 10.1007/s00401-019-02047-3.

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

Llorens, C., Fares, M.A., and Moya, A. (2008). Relationships of Gag-Pol Diversity between Ty3/Gypsy and Retroviridae LTR Retroelements and the Three Kings Hypothesis. BMC Evolutionary Biology *8*, 276. DOI: 10.1186/1471-2148-8-276.

Llorens, C. et al. (2011). The Gypsy Database (GyDB) of Mobile Genetic Elements: Release 2.0. Nucleic Acids Research *39* (suppl_1), D70–D74. DOI: 10.1093/nar/gkq1061.

Love, M.I., Huber, W., and Anders, S. (2014). Moderated Estimation of Fold Change and Dispersion for RNA-Seq Data with DESeq2. Genome Biology *15*, 550. DOI: 10.1186/s13059-014-0550-8.

Love, R.R., Weisenfeld, N.I., Jaffe, D.B., Besansky, N.J., and Neafsey, D.E. (2016). Evaluation of DISCOVAR de Novo Using a Mosquito Sample for Cost-Effective Short-Read Genome Assembly. BMC Genomics *17*.1, 187. DOI: 10.1186/s12864-016-2531-7.

Luis, A.D. et al. (2013). A Comparison of Bats and Rodents as Reservoirs of Zoonotic Viruses: Are Bats Special? Proc. R. Soc. B *280*.1756, 20122753. DOI: 10.1098/rspb.2012.2753.

Luis, A.D., O'Shea, T.J., Hayman, D.T.S., Wood, J.L.N., Cunningham, A.A., Gilbert, A.T., Mills, J.N., and Webb, C.T. (2015). Network Analysis of Host–Virus Communities in Bats and Rodents Reveals Determinants of Cross-Species Transmission. Ecology Letters *18*.11, 1153–1162. DOI: 10.1111/ele.12491.

Lynch, C. and Tristem, M. (2003). A Co-Opted Gypsy-Type LTR-Retrotransposon Is Conserved in the Genomes of Humans, Sheep, Mice, and Rats. Current Biology *13*.17, 1518–1523. DOI: 10.1016/S0960-9822(03)00618-3.

Ma, S. et al. (2016). Cell Culture-Based Profiling across Mammals Reveals DNA Repair and Metabolism as Determinants of Species Longevity. eLife *5*. A.J. Wagers, ed., e19130. DOI: 10.7554/eLife.19130.

Magiorkinis, G., Gifford, R.J., Katzourakis, A., Ranter, J.D., and Belshaw, R. (2012). Env-Less Endogenous Retroviruses Are Genomic Superspreaders. Proceedings of the National Academy of Sciences *109*.19, 7385–7390. DOI: 10.1073/pnas.1200913109.

Mao, X., Tsagkogeorga, G., Bailey, S.E., and Rossiter, S.J. (2017). Genomics of Introgression in the Chinese Horseshoe Bat (Rhinolophus Sinicus) Revealed by Transcriptome Sequencing. Biological Journal of the Linnean Society *121*.3, 698–710. DOI: 10.1093/biolinnean/blx017.

Marco, A. and Marín, I. (2009). CGIN1: A Retroviral Contribution to Mammalian Genomes. Molecular Biology and Evolution *26*.10, 2167–2170. DOI: 10.1093/molbev/msp127.

Matsumoto, Y. et al. (2012). Bornavirus Closely Associates and Segregates with Host Chromosomes to Ensure Persistent Intranuclear Infection. Cell Host & Microbe *11*.5, 492–503. DOI: 10.1016/j.chom.2012.04.009.

McCracken, G.F. (2003). "Estimates of Population Sizes in Summer Colonies of Brazilian Free-tailed Bats (Tadarida Brasiliensis)". Monitoring Trends in Bat Populations in the United States and Territories: Problems and Prospects, T.J. O'Shea and M.A. Bogan, ed. (US Geolological Survey, Biological Resources Discipline, Information and Technology report), pp. 21–30.

Mélade, J., Wieseke, N., Ramasindrazana, B., Flores, O., Lagadec, E., Gomard, Y., Goodman, S.M., Dellagi, K., and Pascalis, H. (2016). An Eco-Epidemiological Study of Morbilli-Related Paramyxovirus Infection in Madagascar Bats Reveals Host-Switching as the Dominant Macro-Evolutionary Mechanism. Scientific Reports *6*.1 (1), 23752. DOI: 10.1038/srep23752.

Memish, Z.A. et al. (2013). Middle East Respiratory Syndrome Coronavirus in Bats, Saudi Arabia. Emerging Infectious Diseases *19*.11, 1819–1823. DOI: 10.3201/eid1911.131172.

Mészáros, B., Erdős, G., and Dosztányi, Z. (2018). IUPred2A: Context-Dependent Prediction of Protein Disorder as a Function of Redox State and Protein Binding. Nucleic Acids Research *46*.W1, W329–W337. DOI: 10.1093/nar/gky384.

Metegnier, G., Becking, T., Chebbi, M.A., Giraud, I., Moumen, B., Schaack, S., Cordaux, R., and Gilbert, C. (2015). Comparative Paleovirological Analysis of Crustaceans Identifies Multiple Widespread Viral Groups. Mobile DNA *6*.1, 16. DOI: 10.1186/s13100-015-0047-3.

Mi, S. et al. (2000). Syncytin Is a Captive Retroviral Envelope Protein Involved in Human Placental Morphogenesis. Nature *403*.6771, 785–789. DOI: 10.1038/35001608.

Mollentze, N. and Streicker, D.G. (2020). Viral Zoonotic Risk Is Homogenous among Taxonomic Orders of Mammalian and Avian Reservoir Hosts. Proceedings of the National Academy of Sciences *117*.17, 9423–9430. DOI: 10.1073/pnas.1919176117.

Mühldorfer, K., Speck, S., Kurth, A., Lesnik, R., Freuling, C., Müller, T., Kramer-Schadt, S., and Wibbelt, G. (2011). Diseases and Causes of Death in European Bats: Dynamics in Disease Susceptibility and Infection Rates. PLOS ONE *6*.12, e29773. DOI: 10.1371/journal.pone.0029773.

Nakagawa, S. and Takahashi, M.U. (2016). gEVE: A Genome-Based Endogenous Viral Element Database Provides Comprehensive Viral Protein-Coding Sequences in Mammalian Genomes. Database *2016*. DOI: 10.1093/database/baw087.

Nascimento, F.F., Reis, M. dos, and Yang, Z. (2017). A Biologist's Guide to Bayesian Phylogenetic Analysis. Nature Ecology & Evolution *1*.10 (10), 1446–1454. DOI: 10.1038/s41559-017-0280-x.

Necsulea, A., Soumillon, M., Warnefors, M., Liechti, A., Daish, T., Zeller, U., Baker, J.C., Grützner, F., and Kaessmann, H. (2014). The Evolution of lncRNA Repertoires and Expression Patterns in Tetrapods. Nature *505*.7485, 635. DOI: 10.1038/nature12943.

Neumann, P., Lieber, D., Meyer, S., Dautel, P., Kerth, A., Kraus, I., Garten, W., and Stubbs, M.T. (2009). Crystal Structure of the Borna Disease Virus Matrix Protein (BDV-M) Reveals ssRNA Binding Properties. Proceedings of the National Academy of Sciences of the United States of America *106*.10, 3710–3715. DOI: 10.1073/pnas.0808101106.

Ng, S.-Y., Johnson, R., and Stanton, L.W. (2012). Human Long Non-Coding RNAs Promote Pluripotency and Neuronal Differentiation by Association with Chromatin Modifiers and Transcription Factors. The EMBO journal *31*.3, 522–533. DOI: 10.1038/emboj.2011.459.

Nguyen, T.-H., Ranwez, V., Berry, V., and Scornavacca, C. (2013). Support Measures to Estimate the Reliability of Evolutionary Events Predicted by Reconciliation Methods. PLOS ONE *8*.10, e73667. DOI: 10.1371/journal.pone.0073667.

Nguyen, L.-T., Schmidt, H.A., von Haeseler, A., and 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. DOI: 10.1093/molbev/msu300.

Nishimura, Y., Yoshida, T., Kuronishi, M., Uehara, H., Ogata, H., and Goto, S. (2017). ViPTree: The Viral Proteomic Tree Server. Bioinformatics *33*.15, 2379–2380. DOI: 10 . 1093 / bioinformatics/btx157.

Nobach, D. and Herden, C. (2020). No Evidence for European Bats Serving as Reservoir for Borna Disease Virus 1 or Other Known Mammalian Orthobornaviruses. Virology Journal *17*.1, 11. DOI: 10.1186/s12985-020-1289-3.

Novák, P., Neumann, P., and Macas, J. (2010). Graph-Based Clustering and Characterization of Repetitive Sequences in next-Generation Sequencing Data. BMC Bioinformatics *11*.1, 378. DOI: 10.1186/1471-2105-11-378.

Olds, J.E., Burrough, E.R., Fales-Williams, A.J., Lehmkuhl, A., Madson, D., Patterson, A.J., and Yaeger, M.J. (2015). Retrospective Evaluation of Cases of Neoplasia in a Captive Population of Egyptian Fruit Bats (Rousettus Aegyptiacus). Journal of Zoo and Wildlife Medicine *46*.2, 325–332. DOI: 10.1638/2014-0069R2.1.

Olival, K.J. and Hayman, D.T.S. (2014). Filoviruses in Bats: Current Knowledge and Future Directions. Viruses *6*.4, 1759–1788. DOI: 10.3390/v6041759.

Olival, K.J., Hosseini, P.R., Zambrana-Torrelio, C., Ross, N., Bogich, T.L., and Daszak, P. (2017). Host and Viral Traits Predict Zoonotic Spillover from Mammals. Nature *546*.7660, 646–650. DOI: 10.1038/nature22975.

Ophinni, Y., Palatini, U., Hayashi, Y., and Parrish, N.F. (2019). piRNA-Guided CRISPR-like Immunity in Eukaryotes. Trends in Immunology *40*.11, 998–1010. DOI: 10.1016/j.it.2019.09.003.

Parrish, N.F. et al. (2015). piRNAs Derived from Ancient Viral Processed Pseudogenes as Transgenerational Sequence-Specific Immune Memory in Mammals. RNA *21*.10, 1691–1703. DOI: 10.1261/rna.052092.115.

Park, J., Karplus, K., Barrett, C., Hughey, R., Haussler, D., Hubbard, T., and Chothia, C. (1998). Sequence Comparisons Using Multiple Sequences Detect Three Times as Many Remote Homologues as Pairwise methods11Edited by J. Thornton. Journal of Molecular Biology *284*.4, 1201–1210. DOI: 10.1006/jmbi.1998.2221.

Pasquesi, G.I.M. et al. (2018). Squamate Reptiles Challenge Paradigms of Genomic Repeat Element Evolution Set by Birds and Mammals. Nature Communications *9*.1 (1), 2774. DOI: 10.1038/s41467-018-05279-1.

Pavlovich, S.S. et al. (2018). The Egyptian Rousette Genome Reveals Unexpected Features of Bat Antiviral Immunity. Cell *173*.5, 1098–1110.e18. DOI: 10.1016/j.cell.2018.03.070.

Peng, X. et al. (2010). Unique Signatures of Long Noncoding RNA Expression in Response to Virus Infection and Altered Innate Immune Signaling. mBio *1*.5, e00206–10. DOI: 10.1128/mBio.00206-10.

Pénzes, J.J., de Souza, W.M., Agbandje-McKenna, M., and Gifford, R.J. (2019). An Ancient Lineage of Highly Divergent Parvoviruses Infects Both Vertebrate and Invertebrate Hosts. Viruses *11*.6, 525. DOI: 10.3390/v11060525.

Pertea, M., Pertea, G.M., Antonescu, C.M., Chang, T.-C., Mendell, J.T., and Salzberg, S.L. (2015). StringTie Enables Improved Reconstruction of a Transcriptome from RNA-Seq Reads. Nature Biotechnology *33*.3 (3), 290–295. DOI: 10.1038/nbt.3122.

Platt, R.N., Vandewege, M.W., Kern, C., Schmidt, C.J., Hoffmann, F.G., and Ray, D.A. (2014). Large Numbers of Novel miRNAs Originate from DNA Transposons and Are Coincident with a Large Species Radiation in Bats. Molecular Biology and Evolution *31*.6, 1536–1545. DOI: 10.1093/molbev/msu112.

Platt, R.N., Mangum, S.F., and Ray, D.A. (2016). Pinpointing the Vesper Bat Transposon Revolution Using the Miniopterus Natalensis Genome. Mobile DNA *7*, 12. DOI: 10.1186/s13100-016-0071-y.

Plowright, R.K., Foley, P., Field, H.E., Dobson, A.P., Foley, J.E., Eby, P., and Daszak, P. (2011). Urban Habituation, Ecological Connectivity and Epidemic Dampening: The Emergence of

Hendra Virus from Flying Foxes (Pteropus Spp.) Proceedings of the Royal Society B: Biological Sciences *278*.1725, 3703–3712. DOI: 10.1098/rspb.2011.0522.

Plowright, R.K., Peel, A.J., Streicker, D.G., Gilbert, A.T., McCallum, H., Wood, J., Baker, M.L., and Restif, O. (2016). Transmission or Within-Host Dynamics Driving Pulses of Zoonotic Viruses in Reservoir–Host Populations. PLOS Neglected Tropical Diseases *10*.8, e0004796. DOI: 10.1371/journal.pntd.0004796.

Ponjavic, J., Ponting, C.P., and Lunter, G. (2007). Functionality or Transcriptional Noise? Evidence for Selection within Long Noncoding RNAs. Genome Research *17*.5, 556–565. DOI: 10.1101/gr.6036807.

Potter, J. (2019). "A Genome-Wide Study of the Molecular Adaptations Underpinning Dietary Specialisations in New World Leaf-Nosed Bats (Phyllostomidae, Chiroptera)". Doctor of Philosophy. Queen Mary University of London. 258 pp.

Quinlan, A.R. and Hall, I.M. (2010). BEDTools: A Flexible Suite of Utilities for Comparing Genomic Features. Bioinformatics *26*.6, 841–842. DOI: 10.1093/bioinformatics/btq033.

R Core Team (2019). *R: A Language and Environment for Statistical Computing*. manual. R Foundation for Statistical Computing. (Vienna, Austria: R Foundation for Statistical Computing).

Racey, P.A. (2015). "The Uniqueness of Bats". Bats and Viruses: A New Frontier of Emerging Infectious Diseases, (John Wiley & Sons), pp. 1–22.

Reitan, T. and Nielsen, A. (2016). Do Not Divide Count Data with Count Data; A Story from Pollination Ecology with Implications Beyond. PLoS ONE *11*.2. DOI: 10.1371/journal.pone.0149129.

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

Roberts, J.T., Cardin, S.E., and Borchert, G.M. (2014). Burgeoning Evidence Indicates That microRNAs Were Initially Formed from Transposable Element Sequences. Mobile Genetic Elements *4*. DOI: 10.4161/mge.29255.

Rojas, D., Vale, Á., Ferrero, V., and Navarro, L. (2012). The Role of Frugivory in the Diversification of Bats in the Neotropics. Journal of Biogeography *39*.11, 1948–1960. DOI: 10.1111/j.1365-2699.2012.02709.x.

Rojas, D., Warsi, O.M., and Dávalos, L.M. (2016). Bats (Chiroptera: Noctilionoidea) Challenge a Recent Origin of Extant Neotropical Diversity. Systematic Biology *65*.3, 432–448. DOI: 10.1093/sysbio/syw011.

Ruedi, M., Stadelmann, B., Gager, Y., Douzery, E.J.P., Francis, C.M., Lin, L.-K., Guillén-Servent, A., and Cibois, A. (2013). Molecular Phylogenetic Reconstructions Identify East Asia as the Cradle for the Evolution of the Cosmopolitan Genus Myotis (Mammalia, Chiroptera). Molecular Phylogenetics and Evolution *69*.3, 437–449. DOI: 10.1016/j.ympev.2013.08.011.

Sadier, A., Davies, K.T., Yohe, L.R., Yun, K., Donat, P., Hedrick, B.P., Dumont, E.R., Dávalos, L.M., Rossiter, S.J., and Sears, K.E. (2018). Multifactorial Processes Underlie Parallel Opsin Loss in Neotropical Bats. eLife *7*. P.J. Wittkopp, ed., e37412. DOI: 10.7554/eLife.37412.

Sanjuán, R. (2012). From Molecular Genetics to Phylodynamics: Evolutionary Relevance of Mutation Rates Across Viruses. PLOS Pathogens *8*.5, e1002685. DOI: 10.1371/journal.ppat.1002685.

Schmitz, J. and Brosius, J. (2011). Exonization of Transposed Elements: A Challenge and Opportunity for Evolution. Biochimie. "Coding or Non-Coding: Need They Be Exclusive?" *93*.11, 1928–1934. DOI: 10.1016/j.biochi.2011.07.014.

Schoggins, J.W., Wilson, S.J., Panis, M., Murphy, M.Y., Jones, C.T., Bieniasz, P., and Rice, C.M. (2011). A Diverse Range of Gene Products Are Effectors of the Type I Interferon Antiviral Response. Nature *472*.7344, 481–485. DOI: 10.1038/nature09907.

Schad, J. and Voigt, C.C. (2016). Adaptive Evolution of Virus-Sensing Toll-like Receptor 8 in Bats. Immunogenetics *68*.10, 783–795. DOI: 10.1007/s00251-016-0940-z.

Schuh, A.J., Amman, B.R., Jones, M.E.B., Sealy, T.K., Uebelhoer, L.S., Spengler, J.R., Martin, B.E., Coleman-McCray, J.A.D., Nichol, S.T., and Towner, J.S. (2017). Modelling Filovirus Maintenance in Nature by Experimental Transmission of Marburg Virus between Egyptian Rousette Bats. Nature Communications *8*.1 (1), 14446. DOI: 10.1038/ncomms14446.

Scheben, A., Ramos, O.M., Kramer, M., Goodwin, S., Oppenheim, S., Becker, D.J., Schatz, M.C., Simmons, N.B., Siepel, A., and McCombie, W.R. (2020). Unraveling Molecular Mechanisms of Immunity and Cancer-Resistance Using the Genomes of the Neotropical Bats Artibeus Jamaicensis and Pteronotus Mesoamericanus. bioRxiv, 2020.09.09.290502. DOI: 10.1101/2020.09.09.290502.

Schrödinger, LLC (2015). "The PyMOL Molecular Graphics System, Version 1.8".

Sela, N., Mersch, B., Hotz-Wagenblatt, A., and Ast, G. (2010). Characteristics of Transposable Element Exonization within Human and Mouse. PLOS ONE *5*.6, e10907. DOI: 10.1371/journal.pone.0010907.

Shaw, A.E. et al. (2017). Fundamental Properties of the Mammalian Innate Immune System Revealed by Multispecies Comparison of Type I Interferon Responses. PLOS Biology *15*.12, e2004086. DOI: 10.1371/journal.pbio.2004086.

Shi, J.J. and Rabosky, D.L. (2015). Speciation Dynamics during the Global Radiation of Extant Bats. Evolution *69*.6, 1528–1545. DOI: 10.1111/evo.12681.

Shi, M. et al. (2018). The Evolutionary History of Vertebrate RNA Viruses. Nature *556*.7700, 197–202. DOI: 10.1038/s41586-018-0012-7.

Silvestre-Ryan, J., Wang, Y., Sharma, M., Lin, S., Shen, Y., Dider, S., and Holmes, I. (2020). Machine Boss: Rapid Prototyping of Bioinformatic Automata. bioRxiv, 2020.02.13.945071. DOI: 10.1101/2020.02.13.945071.

Simmons, N.B., Seymour, K.L., Habersetzer, J., and Gunnell, G.F. (2008). Primitive Early Eocene Bat from Wyoming and the Evolution of Flight and Echolocation. Nature *451*.7180 (7180), 818–821. DOI: 10.1038/nature06549.

Simmonds, P., Aiewsakun, P., and Katzourakis, A. (2019a). Prisoners of War — Host Adaptation and Its Constraints on Virus Evolution. Nature Reviews Microbiology *17*.5 (5), 321–328. DOI: 10.1038/s41579-018-0120-2.

Simmonds, P., Aiewsakun, P., and Katzourakis, A. (2019b). Reply to 'Evolutionary Stasis of Viruses?' Nature Reviews Microbiology *17*.5 (5), 329–330. DOI: 10.1038/s41579-019-0169-6.

Skewes-Cox, P., Sharpton, T.J., Pollard, K.S., and DeRisi, J.L. (2014). Profile Hidden Markov Models for the Detection of Viruses within Metagenomic Sequence Data. PLOS ONE *9*.8, e105067. DOI: 10.1371/journal.pone.0105067.

Skirmuntt, E.C. and Katzourakis, A. (2019). The Evolution of Endogenous Retroviral Envelope Genes in Bats and Their Potential Contribution to Host Biology. Virus Research *270*, 197645. DOI: 10.1016/j.virusres.2019.197645.

Smit, A., Hubley, R., and Green, P. (2013). *RepeatMasker Open-4.0*. URL: http://www.repeatmasker.org/ (visited on [visited on 10/08/2018]).

Smith, J.D., Bickham, J.W., and Gregory, T.R. (2013). Patterns of Genome Size Diversity in Bats (Order Chiroptera). Genome *56*.8, 457–472. DOI: 10.1139/gen-2013-0046.

Sperber, G.O., Airola, T., Jern, P., and Blomberg, J. (2007). Automated Recognition of Retroviral Sequences in Genomic Data—RetroTector©. Nucleic Acids Research *35*.15, 4964–4976. DOI: 10.1093/nar/gkm515.

Stocking, C. and Kozak, C.A. (2008). Murine Endogenous Retroviruses. Cellular and molecular life sciences : CMLS *65*.21, 3383–3398. DOI: 10.1007/s00018-008-8497-0.

Stolzer, M., Lai, H., Xu, M., Sathaye, D., Vernot, B., and Durand, D. (2012). Inferring Duplications, Losses, Transfers and Incomplete Lineage Sorting with Nonbinary Species Trees. Bioinformatics *28*.18, i409–i415. DOI: 10.1093/bioinformatics/bts386.

Streicker, D.G., Turmelle, A.S., Vonhof, M.J., Kuzmin, I.V., McCracken, G.F., and Rupprecht, C.E. (2010). Host Phylogeny Constrains Cross-Species Emergence and Establishment of Rabies Virus in Bats. Science *329*.5992, 676–679. DOI: 10.1126/science.1188836.

Suchard, M.A. and Redelings, B.D. (2006). BAli-Phy: Simultaneous Bayesian Inference of Alignment and Phylogeny. Bioinformatics *22*.16, 2047–2048. DOI: 10.1093/bioinformatics/btl175.

Suh, A., Weber, C.C., Kehlmaier, C., Braun, E.L., Green, R.E., Fritz, U., Ray, D.A., and Ellegren, H. (2014). Early Mesozoic Coexistence of Amniotes and Hepadnaviridae. PLOS Genetics *10*.12, e1004559. DOI: 10.1371/journal.pgen.1004559.

Switzer, W.M. et al. (2005). Ancient Co-Speciation of Simian Foamy Viruses and Primates. Nature *434*.7031 (7031), 376–380. DOI: 10.1038/nature03341.

Taylor, G.M., Gao, Y., and Sanders, D.A. (2001). Fv-4: Identification of the Defect in Env and the Mechanism of Resistance to Ecotropic Murine Leukemia Virus. Journal of Virology *75*.22, 11244–11248. DOI: 10.1128/JVI.75.22.11244-11248.2001.

Taylor, D.J., Leach, R.W., and Bruenn, J. (2010). Filoviruses Are Ancient and Integrated into Mammalian Genomes. BMC Evolutionary Biology *10*.1, 193. DOI: 10.1186/1471-2148-10-193.

Taylor, D.J., Dittmar, K., Ballinger, M.J., and Bruenn, J.A. (2011). Evolutionary Maintenance of Filovirus-like Genes in Bat Genomes. BMC Evolutionary Biology *11*, 336. DOI: 10.1186/1471-2148-11-336.

Teeling, E.C., Springer, M.S., Madsen, O., Bates, P., O'Brien, S.J., and Murphy, W.J. (2005). A Molecular Phylogeny for Bats Illuminates Biogeography and the Fossil Record. Science *307*.5709, 580–584. DOI: 10.1126/science.1105113.

Teeling, E.C., Vernes, S.C., Dávalos, L.M., Ray, D.A., Gilbert, M.T.P., Myers, E., and Consortium, B. (2018). Bat Biology, Genomes, and the Bat1K Project: To Generate Chromosome-Level Genomes for All Living Bat Species. Annual Review of Animal Biosciences *6*.1, 23–46. DOI: 10.1146/annurev-animal-022516-022811.

Terry, S.N., Manganaro, L., Cuesta-Dominguez, A., Brinzevich, D., Simon, V., and Mulder, L.C.F. (2017). Expression of HERV-K108 Envelope Interferes with HIV-1 Production. Virology *509*, 52–59. DOI: 10.1016/j.virol.2017.06.004.

Ti, S., A, S., Wc, C., L, L., A, H., Tc, G., R, A., and T, S. (2012). Transcriptome Sequencing and Annotation for the Jamaican Fruit Bat (Artibeus Jamaicensis). Plos one *7*.11, e48472–e48472. DOI: 10.1371/journal.pone.0048472.

To, T.-H., Jacox, E., Ranwez, V., and Scornavacca, C. (2015). A Fast Method for Calculating Reliable Event Supports in Tree Reconciliations via Pareto Optimality. BMC Bioinformatics *16*.1, 384. DOI: 10.1186/s12859-015-0803-x.

Toufaily, C., Landry, S., Leib-Mosch, C., Rassart, E., and Barbeau, B. (2011). Activation of LTRs from Different Human Endogenous Retrovirus (HERV) Families by the HTLV-1 Tax Protein and T-Cell Activators. Viruses *3*.11 (11), 2146–2159. DOI: 10.3390/v3112146.

Towner, J.S. et al. (2009). Isolation of Genetically Diverse Marburg Viruses from Egyptian Fruit Bats. PLOS Pathogens *5*.7, e1000536. DOI: 10.1371/journal.ppat.1000536.

Turmelle, A.S. and Olival, K.J. (2010). Correlates of Viral Richness in Bats (Order Chiroptera). EcoHealth *6*.4, 522–539. DOI: 10.1007/s10393-009-0263-8.

Virtue, E.R., Marsh, G.A., Baker, M.L., and Wang, L.-F. (2011). Interferon Production and Signaling Pathways Are Antagonized during Henipavirus Infection of Fruit Bat Cell Lines. PLOS ONE *6*.7, e22488. DOI: 10.1371/journal.pone.0022488.

Wang, T., Zeng, J., Lowe, C.B., Sellers, R.G., Salama, S.R., Yang, M., Burgess, S.M., Brachmann, R.K., and Haussler, D. (2007). Species-Specific Endogenous Retroviruses Shape the Transcriptional Network of the Human Tumor Suppressor Protein P53. Proceedings of the National Academy of Sciences *104*.47, 18613–18618. DOI: 10.1073/pnas.0703637104.

Wang, J. et al. (2014). Primate-Specific Endogenous Retrovirus-Driven Transcription Defines Naive-like Stem Cells. Nature *516*.7531 (7531), 405–409. DOI: 10.1038/nature13804.

Wang, Z., Dai, M., Wang, Y., Cooper, K.L., Zhu, T., Dong, D., Zhang, J., and Zhang, S. (2014). Unique Expression Patterns of Multiple Key Genes Associated with the Evolution of Mammalian Flight. Proceedings of the Royal Society B: Biological Sciences *281*.1783, 20133133. DOI: 10.1098/rspb.2013.3133.

Wang, L.-F. and Anderson, D.E. (2019). Viruses in Bats and Potential Spillover to Animals and Humans. Current Opinion in Virology. Emerging Viruses: Intraspecies Transmission • Viral Immunology *34*, 79–89. DOI: 10.1016/j.coviro.2018.12.007.

Wang, M. and Kong, L. (2019). Pblat: A Multithread Blat Algorithm Speeding up Aligning Sequences to Genomes. BMC Bioinformatics *20*.1, 28. DOI: 10.1186/s12859-019-2597-8.

Wang, J. and Han, G.-Z. (2020). Frequent Retroviral Gene Co-Option during the Evolution of Vertebrates. Molecular Biology and Evolution. DOI: 10.1093/molbev/msaa180.

Wang, K., Tian, S., Galindo-González, J., Dávalos, L.M., Zhang, Y., and Zhao, H. (2020). Molecular Adaptation and Convergent Evolution of Frugivory in Old World and Neotropical Fruit Bats. Molecular Ecology *29*.22, 4366–4381. DOI: 10.1111/mec.15542.

Wang, M., Qiu, Y., Liu, H., Liang, B., Fan, B., Zhou, X., and Liu, D. (2020). Transcription Profile of Human Endogenous Retroviruses in Response to Dengue Virus Serotype 2 Infection. Virology *544*, 21–30. DOI: 10.1016/j.virol.2020.01.014.

Wicker, T. et al. (2007). A Unified Classification System for Eukaryotic Transposable Elements. Nature Reviews Genetics *8*.12, 973. DOI: 10.1038/nrg2165.

Willoughby, A.R., Phelps, K.L., Consortium, P., and Olival, K.J. (2017). A Comparative Analysis of Viral Richness and Viral Sharing in Cave-Roosting Bats. Diversity *9*.3 (3), 35. DOI: 10.3390/d9030035.

Wong, A.C.P., Li, X., Lau, S.K.P., and Woo, P.C.Y. (2019). Global Epidemiology of Bat Coronaviruses. Viruses *11*.2 (2), 174. DOI: 10.3390/v11020174.

World Health Organization (2020). *Ebola Virus Disease [Fact Sheet]*.

Wynne, J.W. et al. (2014). Proteomics Informed by Transcriptomics Reveals Hendra Virus Sensitizes Bat Cells to TRAIL-Mediated Apoptosis. Genome Biology *15*.11, 532. DOI: 10.1186/s13059-014-0532-x.

Wynne, J.W. et al. (2017). Comparative Transcriptomics Highlights the Role of the Activator Protein 1 Transcription Factor in the Host Response to Ebolavirus. Journal of Virology *91*.23. DOI: 10.1128/JVI.01174-17.

Xie, J., Li, Y., Shen, X., Goh, G., Zhu, Y., Cui, J., Wang, L.-F., Shi, Z.-L., and Zhou, P. (2018). Dampened STING-Dependent Interferon Activation in Bats. Cell Host & Microbe *0*.0. DOI: 10.1016/j.chom.2018.01.006.

Xu, X., Zhao, H., Gong, Z., and Han, G.-Z. (2018). Endogenous Retroviruses of Non-Avian/Mammalian Vertebrates Illuminate Diversity and Deep History of Retroviruses. PLOS Pathogens *14*.6, e1007072. DOI: 10.1371/journal.ppat.1007072.

Yang, J. and Zhang, Y. (2015). I-TASSER Server: New Development for Protein Structure and Function Predictions. Nucleic Acids Research *43*.W1, W174–W181. DOI: 10.1093/nar/gkv342.

Yang, W.R., Ardeljan, D., Pacyna, C.N., Payer, L.M., and Burns, K.H. (2019). SQuIRE Reveals Locus-Specific Regulation of Interspersed Repeat Expression. Nucleic Acids Research *47*.5, e27–e27. DOI: 10.1093/nar/gky1301.

Yang, Z. (2007). PAML 4: Phylogenetic Analysis by Maximum Likelihood. Molecular Biology and Evolution *24*.8, 1586–1591. DOI: 10.1093/molbev/msm088.

Yap, M.W., Colbeck, E., Ellis, S.A., and Stoye, J.P. (2014). Evolution of the Retroviral Restriction Gene Fv1: Inhibition of Non-MLV Retroviruses. PLOS Pathogens *10*.3, e1003968. DOI: 10.1371/journal.ppat.1003968.

Young, C.C.W. and Olival, K.J. (2016). Optimizing Viral Discovery in Bats. PLOS ONE *11*.2, e0149237. DOI: 10.1371/journal.pone.0149237.

Zhang, G. et al. (2013). Comparative Analysis of Bat Genomes Provides Insight into the Evolution of Flight and Immunity. Science, 1230835. DOI: 10.1126/science.1230835.

Zhou, P., Cowled, C., Todd, S., Crameri, G., Virtue, E.R., Marsh, G.A., Klein, R., Shi, Z., Wang, L.-F., and Baker, M.L. (2011). Type III IFNs in Pteropid Bats: Differential Expression Patterns Provide Evidence for Distinct Roles in Antiviral Immunity. The Journal of Immunology *186*.5, 3138–3147. DOI: 10.4049/jimmunol.1003115.

Zhou, X., Xu, S., Xu, J., Chen, B., Zhou, K., and Yang, G. (2012). Phylogenomic Analysis Resolves the Interordinal Relationships and Rapid Diversification of the Laurasiatherian Mammals. Systematic Biology *61*.1, 150–150. DOI: 10.1093/sysbio/syr089.

Zhou, P., Cowled, C., Mansell, A., Monaghan, P., Green, D., Wu, L., Shi, Z., Wang, L.-F., and Baker, M.L. (2014). IRF7 in the Australian Black Flying Fox, Pteropus Alecto: Evidence for a Unique Expression Pattern and Functional Conservation. PLOS ONE *9*.8, e103875. DOI: 10.1371/journal.pone.0103875.

Zhou, P. et al. (2016). Contraction of the Type I IFN Locus and Unusual Constitutive Expression of IFN-α in Bats. Proceedings of the National Academy of Sciences *113*.10, 2696–2701. DOI: 10.1073/pnas.1518240113.

Zhou, B. et al. (2019). Endogenous Retrovirus-Derived Long Noncoding RNA Enhances Innate Immune Responses via Derepressing RELA Expression. mBio *10*.4, e00937–19. DOI: 10.1128/mBio.00937-19.

Zhuo, X., Rho, M., and Feschotte, C. (2013). Genome-Wide Characterization of Endogenous Retroviruses in the Bat Myotis Lucifugus Reveals Recent and Diverse Infections. Journal of Virology *87*.15, 8493–8501. DOI: 10.1128/JVI.00892-13.

Zhuo, X. and Feschotte, C. (2015). Cross-Species Transmission and Differential Fate of an Endogenous Retrovirus in Three Mammal Lineages. PLOS Pathogens *11*.11, e1005279. DOI: 10.1371/journal.ppat.1005279.

Zhu, H., Dennis, T., Hughes, J., and Gifford, R.J. (2018). Database-Integrated Genome Screening (DIGS): Exploring Genomes Heuristically Using Sequence Similarity Search Tools and a Relational Database. bioRxiv, 246835. DOI: 10.1101/246835.

Appendices

A

Chapter 2 supplementary material

A.1 Supplementary tables

Table A.1: Number of putative ERV RT loci detected by each search method

Assembly	GeneWise		HMMER		RetroTector		tBLASTn		combined
rhi_fer ASM46549v1	757	67.8%	709	63.5%	242	21.7%	1048	93.8%	1117
eid_hel ASM46528v1	741	68.0%	688	63.1%	317	29.1%	1006	92.3%	1090
pte_ale ASM32557v1	842	69.4%	808	66.6%	485	40.0%	1062	87.5%	1214
des_rot ASM294091v2	695	54.5%	657	51.5%	647	50.7%	961	75.3%	1276
rou_aeg RouAeg_v1_BIUU	2084	84.0%	2042	82.3%	556	22.4%	1647	66.4%	2482
rhi_fer RhiFer_v1_BIUU	2496	89.9%	2528	91.0%	722	26.0%	1873	67.4%	2777
ton_sau TonSau_v1_BIUU	1256	73.9%	1238	72.9%	606	35.7%	1308	77.0%	1699
eon_spe Espe.v1	1364	78.9%	1257	72.7%	869	50.3%	1209	69.9%	1729
phy_dis mPhyDis1_v1.p	1030	69.3%	963	64.8%	827	55.7%	1109	74.6%	1486
rhi_fer mRhiFer1_v1.p	1356	86.7%	1308	83.6%	1030	65.9%	1225	78.3%	1564

A.2 Supplementary methods

A.2.1 Genome assemblies used in custom ERV library construction

Accessions for each of the 15 genome assemblies used during the construction of a custom bat ERV library to be used with RepeatMasker: *Hipposideros armiger* (GCA_001890085.1), *Megaderma lyra* (GCA_000465345.1), *Pteronotus parnellii* (GCA_000465405.1), *Desmodus rotundus* (GCA_002940915.2), *Pteropus vampyrus* (GCA_000151845.2), *Eidolon helvum* (GCA_000465285.1), *Pteropus alecto*

(GCF_000325575.1), *Rousettus aegyptiacus* (GCA_001466805.2), *Rhinolophus ferrumequinum* (GCA_000465495.1), *Rhinolophus sinicus* (GCA_001888835.1), *Eptesicus fuscus* (GCA_000308155.1), *Myotis lucifugus* (GCA_000147115.1), *Myotis brandtii* (GCA_000412655.1), *Myotis davidii* (GCF_000327345.1), *Miniopterus natalensis* (GCA_001595765.1).

B

Chapter 3 supplementary material

B.1 Supplementary methods

B.1.1 Genome assemblies used in Chapter 3

Accessions of the 19 genome assemblies analysed in Chapter 3: *Anoura caudifer* (GCA_004027475.1), *Artibeus jamaicensis* (GCA_004027435.1), *Carollia perspicillata* (GCA_004027735.1), *Craseonycteris thonglongyai* (GCA_004027555.1), *Hipposideros galeritus* (GCA_004027415.1), *Lasiurus borealis* (GCA_004026805.1), *Macroglossus sobrinus* (GCA_004027375.1), *Megaderma lyra* (GCA_004026885.1), *Micronycteris hirsuta* (GCA_004026765.1), *Miniopterus schreibersii* (GCA_004026525.1), *Mormoops blainvillei* (GCA_004026545.1), *Murina aurata* (GCA_004026665.1), *Myotis myotis* (GCA_004026985.1), *Noctilio leporinus* (GCA_004026585.1), *Pipistrellus pipistrellus* (GCA_004026625.1), *Rhinolophus ferrumequinum* (GCA_007922735.1), *Rousettus aegyptiacus* (GCA_004024865.1), *Tadarida brasiliensis* (GCA_004025005.1), *Tonatia saurophila* (GCA_004024845.1).

C

Chapter 4 supplementary material

C.1 Supplementary tables

Table C.1: Putative EBL elements detected in this study. Stops and frame-shifts counted only within aligned hit region. *Total codon length of ORF with greatest overlap with hit.

EBL group Bat Species	Max. ORF*	Stops	F.shifts	EBL ID	Hit ID
chEBLNa (BDV_P40)					
ano_cau	132	3	4	ano_cau-G_AnoCau_v1_BIUU-2	G_AnoCau_v1_BIUU-BDV_P40-6
art_jam	104	1	3	art_jam-G_ArtJam_v1_BIUU-1	G_ArtJam_v1_BIUU-BDV_P40-1
car_per	125	2	2	car_per-G_CarPer_v1_BIUU-4	G_CarPer_v1_BIUU-BDV_P40-25
cen_sen	152	0	1	cen_sen-G_CeSe_JP-3	G_CeSe_JP-BDV_P40-21
cen_sen	83	0	0	cen_sen-G_CeSe_JP-1	G_CeSe_JP-BDV_P40-7
des_rot	123	2	5	des_rot-G_ASM294091v2-5	G_ASM294091v2-BDV_P40-19
des_rot	65	0	0	des_rot-G_ASM294091v2-2	G_ASM294091v2-BDV_P40-2
des_rot	51	0	0	des_rot-G_ASM294091v2-8	G_ASM294091v2-BDV_P40-34
lio_spu	116	3	5	lio_spu-G_Lisp_GT-1	G_Lisp_GT-BDV_P40-13
mic_hir	130	1	6	mic_hir-G_MicHir_v1_BIUU-1	G_MicHir_v1_BIUU-BDV_P40-6
mor_bla	128	1	5	mor_bla-G_MorMeg_v1_BIUU-4	G_MorMeg_v1_BIUU-BDV_P40-16
phy_dis	125	2	2	phy_dis-G_mPhyDis1_v1-3	G_mPhyDis1_v1-BDV_P40-20
pte_par	54	3	1	pte_par-G_ASM46540v1-7	G_ASM46540v1-BDV_P40-44
ton_sau	164	0	1	ton_sau-G_TonSau_v1_BIUU-4	G_TonSau_v1_BIUU-BDV_P40-13
ton_sau	104	2	3	ton_sau-G_TonSau_v1_BIUU-3	G_TonSau_v1_BIUU-BDV_P40-12
chEBLNb (BDV_P40)					
hip_arm	127	3	3	hip_arm-G_ASM189008v1-1	G_ASM189008v1-BDV_P40-1
hip_gal	134	3	4	hip_gal-G_HipGal_v1_BIUU-1	G_HipGal_v1_BIUU-BDV_P40-30
rhi_fer	73	6	5	rhi_fer-G_ASM46549v1-1	G_ASM46549v1-BDV_P40-7
rhi_fer	65	6	5	rhi_fer-G_Rhfe_GT_ERC-1	G_Rhfe_GT_ERC-BDV_P40-11
rhi_fer	65	6	5	rhi_fer-G_mRhiFer1_v1-2	G_mRhiFer1_v1-BDV_P40-10
rhi_pea	97	2	6	rhi_pea-G_Rhpe_GT-1	G_Rhpe_GT-BDV_P40-35
rhi_sin	90	4	5	rhi_sin-G_ASM188883v1-1	G_ASM188883v1-BDV_P40-1
chEBLNc (BDV_P40)					
min_nat	386	0	0	min_nat-G_Mnatv1-2	G_Mnatv1-BDV_P40-2
min_nat	126	1	2	min_nat-G_Mnatv1-6	G_Mnatv1-BDV_P40-6
min_nat	83	1	3	min_nat-G_Mnatv1-4	G_Mnatv1-BDV_P40-4
min_sch	386	0	0	min_sch-G_MinSch_v1_BIUU-2	G_MinSch_v1_BIUU-BDV_P40-2
min_sch	126	1	2	min_sch-G_MinSch_v1_BIUU-1	G_MinSch_v1_BIUU-BDV_P40-1
min_sch	73	1	8	min_sch-G_MinSch_v1_BIUU-6	G_MinSch_v1_BIUU-BDV_P40-10

chEBLNd (BDV_P40)					
meg_lyr	151	0	4	meg_lyr-G_MegLyr_v1_BIUU-1	G_MegLyr_v1_BIUU-BDV_P40-4
meg_lyr	146	0	4	meg_lyr-G_ASM46534v1-1	G_ASM46534v1-BDV_P40-30
chEBLNe (BDV_P40)					
ano_cau	106	0	5	ano_cau-G_AnoCau_v1_BIUU-4	G_AnoCau_v1_BIUU-BDV_P40-20
art_jam	100	1	3	art_jam-G_ArtJam_v1_BIUU-2	G_ArtJam_v1_BIUU-BDV_P40-2
car_per	152	0	3	car_per-G_CarPer_v1_BIUU-2	G_CarPer_v1_BIUU-BDV_P40-9
cen_sen	115	1	3	cen_sen-G_CeSe_JP-5	G_CeSe_JP-BDV_P40-33
cen_sen	99	0	2	cen_sen-G_CeSe_JP-6	G_CeSe_JP-BDV_P40-32
des_rot	167	2	1	des_rot-G_ASM294091v2-3	G_ASM294091v2-BDV_P40-15
lio_spu	121	3	2	lio_spu-G_Lisp_GT-2	G_Lisp_GT-BDV_P40-19
mic_hir	100	1	4	mic_hir-G_MicHir_v1_BIUU-2	G_MicHir_v1_BIUU-BDV_P40-10
mor_bla	171	2	2	mor_bla-G_MorMeg_v1_BIUU-2	G_MorMeg_v1_BIUU-BDV_P40-13
phy_dis	127	0	5	phy_dis-G_mPhyDis1_v1-2	G_mPhyDis1_v1-BDV_P40-18
pte_par	107	2	6	pte_par-G_ASM46540v1-2	G_ASM46540v1-BDV_P40-27
ton_sau	121	3	4	ton_sau-G_TonSau_v1_BIUU-2	G_TonSau_v1_BIUU-BDV_P40-7
chEBLNf (BDV_P40)					
cra_tho	81	0	4	cra_tho-G_CraTho_v1_BIUU-1	G_CraTho_v1_BIUU-BDV_P40-5
chEBLNg (BDV_P40)					
min_nat	67	0	1	min_nat-G_Mnatv1-5	G_Mnatv1-BDV_P40-5
min_sch	66	0	1	min_sch-G_MinSch_v1_BIUU-5	G_MinSch_v1_BIUU-BDV_P40-11
tad_bra	84	0	1	tad_bra-G_TadBra_v1_BIUU-3	G_TadBra_v1_BIUU-BDV_P40-7
tad_bra	84	0	1	tad_bra-G_TadBra_v1_BIUU-4	G_TadBra_v1_BIUU-BDV_P40-8
chEBLNh (BDV_P40)					
ano_cau	82	0	0	ano_cau-G_AnoCau_v1_BIUU-3	G_AnoCau_v1_BIUU-BDV_P40-17
art_jam	141	0	0	art_jam-G_ArtJam_v1_BIUU-6	G_ArtJam_v1_BIUU-BDV_P40-15
cen_sen	169	0	0	cen_sen-G_CeSe_JP-8	G_CeSe_JP-BDV_P40-35
cen_sen	109	0	0	cen_sen-G_CeSe_JP-7	G_CeSe_JP-BDV_P40-34
des_rot	269	0	0	des_rot-G_ASM294091v2-1	G_ASM294091v2-BDV_P40-1
lio_spu	210	0	0	lio_spu-G_Lisp_GT-5	G_Lisp_GT-BDV_P40-29
mic_hir	172	0	0	mic_hir-G_MicHir_v1_BIUU-6	G_MicHir_v1_BIUU-BDV_P40-18
mic_hir	172	0	0	mic_hir-G_MicHir_v1_BIUU-7	G_MicHir_v1_BIUU-BDV_P40-22
mor_bla	126	0	0	mor_bla-G_MorMeg_v1_BIUU-5	G_MorMeg_v1_BIUU-BDV_P40-18
mor_bla	107	0	0	mor_bla-G_MorMeg_v1_BIUU-3	G_MorMeg_v1_BIUU-BDV_P40-14
noc_lep	78	0	0	noc_lep-G_NocLep_v1_BIUU-1	G_NocLep_v1_BIUU-BDV_P40-1
phy_dis	308	0	0	phy_dis-G_mPhyDis1_v1-1	G_mPhyDis1_v1-BDV_P40-1
pte_par	139	0	0	pte_par-G_ASM46540v1-6	G_ASM46540v1-BDV_P40-42
ton_sau	120	0	0	ton_sau-G_TonSau_v1_BIUU-1	G_TonSau_v1_BIUU-BDV_P40-1
chEBLNi (BDV_P40)					
art_jam	140	0	0	art_jam-G_ArtJam_v1_BIUU-3	G_ArtJam_v1_BIUU-BDV_P40-7
cen_sen	159	0	0	cen_sen-G_CeSe_JP-4	G_CeSe_JP-BDV_P40-23
des_rot	132	0	0	des_rot-G_ASM294091v2-4	G_ASM294091v2-BDV_P40-17
lio_spu	111	0	0	lio_spu-G_Lisp_GT-3	G_Lisp_GT-BDV_P40-27
mic_hir	94	0	0	mic_hir-G_MicHir_v1_BIUU-4	G_MicHir_v1_BIUU-BDV_P40-15
mor_bla	75	0	0	mor_bla-G_MorMeg_v1_BIUU-1	G_MorMeg_v1_BIUU-BDV_P40-2
phy_dis	140	0	0	phy_dis-G_mPhyDis1_v1-5	G_mPhyDis1_v1-BDV_P40-39
pte_par	66	1	1	pte_par-G_ASM46540v1-5	G_ASM46540v1-BDV_P40-39
chEBLNj (BDV_P40)					
min_nat	209	0	0	min_nat-G_Mnatv1-1	G_Mnatv1-BDV_P40-1
min_sch	174	0	0	min_sch-G_MinSch_v1_BIUU-4	G_MinSch_v1_BIUU-BDV_P40-9
tad_bra	157	0	0	tad_bra-G_TadBra_v1_BIUU-2	G_TadBra_v1_BIUU-BDV_P40-5
chEBLNk (BDV_P40)					
mur_fea	240	0	1	mur_fea-G_MurFea_v1_BIUU-12	G_MurFea_v1_BIUU-BDV_P40-21
mur_fea	230	1	2	mur_fea-G_MurFea_v1_BIUU-9	G_MurFea_v1_BIUU-BDV_P40-17
mur_fea	125	1	3	mur_fea-G_MurFea_v1_BIUU-1	G_MurFea_v1_BIUU-BDV_P40-3
mur_fea	117	0	4	mur_fea-G_MurFea_v1_BIUU-14	G_MurFea_v1_BIUU-BDV_P40-26
mur_fea	112	2	1	mur_fea-G_MurFea_v1_BIUU-6	G_MurFea_v1_BIUU-BDV_P40-9
mur_fea	98	0	2	mur_fea-G_MurFea_v1_BIUU-3	G_MurFea_v1_BIUU-BDV_P40-6
myo_bra	371	0	0	myo_bra-G_ASM41265v1-12	G_ASM41265v1-BDV_P40-13
myo_bra	266	1	0	myo_bra-G_ASM41265v1-15	G_ASM41265v1-BDV_P40-21
myo_bra	205	1	5	myo_bra-G_ASM41265v1-13	G_ASM41265v1-BDV_P40-19
myo_bra	187	0	1	myo_bra-G_ASM41265v1-10	G_ASM41265v1-BDV_P40-10
myo_bra	112	1	4	myo_bra-G_ASM41265v1-9	G_ASM41265v1-BDV_P40-9
myo_bra	98	0	2	myo_bra-G_ASM41265v1-14	G_ASM41265v1-BDV_P40-20
myo_dav	238	1	0	myo_dav-G_ASM32734v1-12	G_ASM32734v1-BDV_P40-15
myo_dav	220	0	1	myo_dav-G_ASM32734v1-7	G_ASM32734v1-BDV_P40-7
myo_dav	204	0	3	myo_dav-G_ASM32734v1-6	G_ASM32734v1-BDV_P40-6
myo_dav	178	0	2	myo_dav-G_ASM32734v1-3	G_ASM32734v1-BDV_P40-4
myo_dav	175	0	1	myo_dav-G_ASM32734v1-14	G_ASM32734v1-BDV_P40-22

myo_luc	373	0	0	myo_luc-G_Myoluc2_0-5	G_Myoluc2_0-BDV_P40-8
myo_luc	311	0	1	myo_luc-G_Myoluc2_0-3	G_Myoluc2_0-BDV_P40-5
myo_luc	222	0	4	myo_luc-G_Myoluc2_0-13	G_Myoluc2_0-BDV_P40-23
myo_luc	219	0	1	myo_luc-G_Myoluc2_0-8	G_Myoluc2_0-BDV_P40-11
myo_myo	369	0	0	myo_myo-G_MyoMyo_v1_BIUU-10	G_MyoMyo_v1_BIUU-BDV_P40-18
myo_myo	231	0	2	myo_myo-G_MyoMyo_v1_BIUU-1	G_MyoMyo_v1_BIUU-BDV_P40-1
myo_myo	194	0	1	myo_myo-G_MyoMyo_v1_BIUU-2	G_MyoMyo_v1_BIUU-BDV_P40-7
myo_myo	127	0	6	myo_myo-G_MyoMyo_v1_BIUU-20	G_MyoMyo_v1_BIUU-BDV_P40-40
myo_myo	84	2	1	myo_myo-G_MyoMyo_v1_BIUU-6	G_MyoMyo_v1_BIUU-BDV_P40-13
chEBLNl (BDV_P40)					
ept_fus	200	0	1	ept_fus-G_EptFus1_0-3	G_EptFus1_0-BDV_P40-7
ept_fus	152	1	3	ept_fus-G_EptFus1_0-23	G_EptFus1_0-BDV_P40-33
ept_fus	101	1	1	ept_fus-G_EptFus1_0-14	G_EptFus1_0-BDV_P40-19
las_bor	262	0	2	las_bor-G_LasBor_v1_BIUU-9	G_LasBor_v1_BIUU-BDV_P40-17
las_bor	157	0	4	las_bor-G_LasBor_v1_BIUU-2	G_LasBor_v1_BIUU-BDV_P40-3
las_bor	129	0	3	las_bor-G_LasBor_v1_BIUU-13	G_LasBor_v1_BIUU-BDV_P40-21
las_bor	93	1	3	las_bor-G_LasBor_v1_BIUU-19	G_LasBor_v1_BIUU-BDV_P40-33
mur_fea	124	1	0	mur_fea-G_MurFea_v1_BIUU-5	G_MurFea_v1_BIUU-BDV_P40-8
mur_fea	88	1	4	mur_fea-G_MurFea_v1_BIUU-4	G_MurFea_v1_BIUU-BDV_P40-7
pip_pip	165	0	2	pip_pip-G_PipPip_v1_BIUU-12	G_PipPip_v1_BIUU-BDV_P40-21
pip_pip	160	0	7	pip_pip-G_PipPip_v1_BIUU-14	G_PipPip_v1_BIUU-BDV_P40-23
chEBLNm (BDV_P40)					
ept_fus	325	0	1	ept_fus-G_EptFus1_0-2	G_EptFus1_0-BDV_P40-8
ept_fus	324	0	1	ept_fus-G_EptFus1_0-11	G_EptFus1_0-BDV_P40-16
ept_fus	306	0	2	ept_fus-G_EptFus1_0-1	G_EptFus1_0-BDV_P40-6
ept_fus	288	0	0	ept_fus-G_EptFus1_0-15	G_EptFus1_0-BDV_P40-20
ept_fus	217	0	2	ept_fus-G_EptFus1_0-22	G_EptFus1_0-BDV_P40-32
ept_fus	209	1	2	ept_fus-G_EptFus1_0-17	G_EptFus1_0-BDV_P40-27
ept_fus	139	1	3	ept_fus-G_EptFus1_0-5	G_EptFus1_0-BDV_P40-10
ept_fus	132	2	0	ept_fus-G_EptFus1_0-4	G_EptFus1_0-BDV_P40-9
ept_fus	129	1	1	ept_fus-G_EptFus1_0-20	G_EptFus1_0-BDV_P40-30
ept_fus	127	1	1	ept_fus-G_EptFus1_0-16	G_EptFus1_0-BDV_P40-26
ept_fus	105	0	3	ept_fus-G_EptFus1_0-7	G_EptFus1_0-BDV_P40-11
ept_fus	105	1	0	ept_fus-G_EptFus1_0-6	G_EptFus1_0-BDV_P40-12
ept_fus	96	1	1	ept_fus-G_EptFus1_0-19	G_EptFus1_0-BDV_P40-29
las_bor	197	0	4	las_bor-G_LasBor_v1_BIUU-6	G_LasBor_v1_BIUU-BDV_P40-13
las_bor	188	1	1	las_bor-G_LasBor_v1_BIUU-15	G_LasBor_v1_BIUU-BDV_P40-22
las_bor	173	0	2	las_bor-G_LasBor_v1_BIUU-7	G_LasBor_v1_BIUU-BDV_P40-14
las_bor	112	1	2	las_bor-G_LasBor_v1_BIUU-12	G_LasBor_v1_BIUU-BDV_P40-19
las_bor	109	1	0	las_bor-G_LasBor_v1_BIUU-8	G_LasBor_v1_BIUU-BDV_P40-15
las_bor	94	0	4	las_bor-G_LasBor_v1_BIUU-10	G_LasBor_v1_BIUU-BDV_P40-18
mur_fea	285	0	1	mur_fea-G_MurFea_v1_BIUU-17	G_MurFea_v1_BIUU-BDV_P40-31
mur_fea	213	2	1	mur_fea-G_MurFea_v1_BIUU-11	G_MurFea_v1_BIUU-BDV_P40-20
mur_fea	89	2	4	mur_fea-G_MurFea_v1_BIUU-18	G_MurFea_v1_BIUU-BDV_P40-32
pip_pip	252	0	0	pip_pip-G_PipPip_v1_BIUU-16	G_PipPip_v1_BIUU-BDV_P40-25
pip_pip	242	0	1	pip_pip-G_PipPip_v1_BIUU-8	G_PipPip_v1_BIUU-BDV_P40-14
pip_pip	174	0	1	pip_pip-G_PipPip_v1_BIUU-6	G_PipPip_v1_BIUU-BDV_P40-12
pip_pip	150	0	1	pip_pip-G_PipPip_v1_BIUU-3	G_PipPip_v1_BIUU-BDV_P40-7
pip_pip	150	0	1	pip_pip-G_PipPip_v1_BIUU-4	G_PipPip_v1_BIUU-BDV_P40-8
pip_pip	137	0	2	pip_pip-G_PipPip_v1_BIUU-9	G_PipPip_v1_BIUU-BDV_P40-15
pip_pip	127	0	3	pip_pip-G_PipPip_v1_BIUU-13	G_PipPip_v1_BIUU-BDV_P40-22
pip_pip	117	0	3	pip_pip-G_PipPip_v1_BIUU-15	G_PipPip_v1_BIUU-BDV_P40-24
pip_pip	103	1	3	pip_pip-G_PipPip_v1_BIUU-7	G_PipPip_v1_BIUU-BDV_P40-13
pip_pip	88	0	1	pip_pip-G_PipPip_v1_BIUU-18	G_PipPip_v1_BIUU-BDV_P40-40
chEBLNn (BDV_P40)					
myo_dav	275	0	0	myo_dav-G_ASM32734v1-13	G_ASM32734v1-BDV_P40-21
myo_dav	66	0	3	myo_dav-G_ASM32734v1-9	G_ASM32734v1-BDV_P40-11
myo_myo	306	1	0	myo_myo-G_MyoMyo_v1_BIUU-17	G_MyoMyo_v1_BIUU-BDV_P40-33
myo_myo	283	0	1	myo_myo-G_MyoMyo_v1_BIUU-24	G_MyoMyo_v1_BIUU-BDV_P40-46
myo_myo	283	0	1	myo_myo-G_MyoMyo_v1_BIUU-25	G_MyoMyo_v1_BIUU-BDV_P40-47
chEBLNo (BDV_P40)					
min_nat	145	0	2	min_nat-G_Mnatv1-3	G_Mnatv1-BDV_P40-3
min_sch	145	0	2	min_sch-G_MinSch_v1_BIUU-3	G_MinSch_v1_BIUU-BDV_P40-4
chEBLNp (BDV_P40)					
ept_fus	112	0	2	ept_fus-G_EptFus1_0-9	G_EptFus1_0-BDV_P40-13
ept_fus	101	0	0	ept_fus-G_EptFus1_0-8	G_EptFus1_0-BDV_P40-14
ept_fus	56	5	3	ept_fus-G_EptFus1_0-13	G_EptFus1_0-BDV_P40-18
las_bor	92	1	2	las_bor-G_LasBor_v1_BIUU-20	G_LasBor_v1_BIUU-BDV_P40-34
las_bor	71	0	3	las_bor-G_LasBor_v1_BIUU-3	G_LasBor_v1_BIUU-BDV_P40-5

mur_fea	212	0	2	mur_fea-G_MurFea_v1_BIUU-15	G_MurFea_v1_BIUU-BDV_P40-27
mur_fea	120	1	0	mur_fea-G_MurFea_v1_BIUU-8	G_MurFea_v1_BIUU-BDV_P40-15
myo_bra	397	0	2	myo_bra-G_ASM41265v1-8	G_ASM41265v1-BDV_P40-8
myo_bra	112	0	1	myo_bra-G_ASM41265v1-2	G_ASM41265v1-BDV_P40-2
myo_dav	141	1	1	myo_dav-G_ASM32734v1-5	G_ASM32734v1-BDV_P40-5
myo_dav	106	0	8	myo_dav-G_ASM32734v1-4	G_ASM32734v1-BDV_P40-3
myo_dav	102	1	4	myo_dav-G_ASM32734v1-8	G_ASM32734v1-BDV_P40-9
myo_luc	350	0	3	myo_luc-G_Myoluc2_0-14	G_Myoluc2_0-BDV_P40-22
myo_luc	158	0	4	myo_luc-G_Myoluc2_0-15	G_Myoluc2_0-BDV_P40-24
myo_luc	141	0	2	myo_luc-G_Myoluc2_0-7	G_Myoluc2_0-BDV_P40-10
myo_luc	136	0	3	myo_luc-G_Myoluc2_0-1	G_Myoluc2_0-BDV_P40-3
myo_luc	112	1	2	myo_luc-G_Myoluc2_0-4	G_Myoluc2_0-BDV_P40-7
myo_myo	158	0	4	myo_myo-G_MyoMyo_v1_BIUU-26	G_MyoMyo_v1_BIUU-BDV_P40-48
myo_myo	139	0	2	myo_myo-G_MyoMyo_v1_BIUU-5	G_MyoMyo_v1_BIUU-BDV_P40-12
myo_myo	102	0	4	myo_myo-G_MyoMyo_v1_BIUU-28	G_MyoMyo_v1_BIUU-BDV_P40-50
myo_myo	81	0	2	myo_myo-G_MyoMyo_v1_BIUU-11	G_MyoMyo_v1_BIUU-BDV_P40-20
myo_myo	74	1	3	myo_myo-G_MyoMyo_v1_BIUU-8	G_MyoMyo_v1_BIUU-BDV_P40-16
pip_pip	108	1	3	pip_pip-G_PipPip_v1_BIUU-2	G_PipPip_v1_BIUU-BDV_P40-6
chEBLX (BDV_P10)					
rhi_fer	52	0	0	rhi_fer-G_ASM46549v1-2	G_ASM46549v1-BDV_P10-8
rhi_fer	52	0	0	rhi_fer-G_mRhiFer1_v1-1	G_mRhiFer1_v1-BDV_P10-5
chEBLPa (BDV_P24)					
lio_spu	76	0	0	lio_spu-G_Lisp_GT-4	G_Lisp_GT-BDV_P24-28
chEBLPb (BDV_P24)					
pte_par	120	0	0	pte_par-G_ASM46540v1-1	G_ASM46540v1-BDV_P24-20
chEBLPc (BDV_P24)					
pte_par	70	0	0	pte_par-G_ASM46540v1-4	G_ASM46540v1-BDV_P24-29
chEBLM (BDV_M)					
ano_cau	135	0	0	ano_cau-G_AnoCau_v1_BIUU-1	G_AnoCau_v1_BIUU-BDV_M-1
art_jam	102	0	0	art_jam-G_ArtJam_v1_BIUU-4	G_ArtJam_v1_BIUU-BDV_M-8
car_per	93	0	1	car_per-G_CarPer_v1_BIUU-3	G_CarPer_v1_BIUU-BDV_M-20
des_rot	97	0	0	des_rot-G_ASM294091v2-6	G_ASM294091v2-BDV_M-31
lio_spu	94	0	0	lio_spu-G_Lisp_GT-6	G_Lisp_GT-BDV_M-37
mic_hir	138	0	0	mic_hir-G_MicHir_v1_BIUU-3	G_MicHir_v1_BIUU-BDV_M-11
mor_bla	85	0	0	mor_bla-G_MorMeg_v1_BIUU-6	G_MorMeg_v1_BIUU-BDV_M-20
pte_par	85	0	0	pte_par-G_ASM46540v1-3	G_ASM46540v1-BDV_M-28
ton_sau	91	0	0	ton_sau-G_TonSau_v1_BIUU-5	G_TonSau_v1_BIUU-BDV_M-21
chEBLGb (BDV_G)					
car_per	27	0	1	car_per-G_CarPer_v1_BIUU-5	G_CarPer_v1_BIUU-BDV_G-35
chEBLLa (Mononeg_RNA_pol)					
myo_bra	84	1	0	myo_bra-G_ASM41265v1-4	G_ASM41265v1-Mononeg_RNA_pol-6
myo_myo	163	0	0	myo_myo-G_MyoMyo_v1_BIUU-6	G_MyoMyo_v1_BIUU-Mononeg_RNA_pol-15
chEBLLb (Mononeg_RNA_pol)					
des_rot	144	0	1	des_rot-G_ASM294091v2-7	G_ASM294091v2-Mononeg_RNA_pol-32
chEBLLc (Mononeg_mRNAcap)					
las_bor	141	0	0	las_bor-G_LasBor_v1_BIUU-16	G_LasBor_v1_BIUU-Mononeg_mRNAcap-30
las_bor	127	2	0	las_bor-G_LasBor_v1_BIUU-18	G_LasBor_v1_BIUU-Mononeg_mRNAcap-32
mur_fea	101	1	3	mur_fea-G_MurFea_v1_BIUU-7	G_MurFea_v1_BIUU-Mononeg_mRNAcap-10
myo_bra	361	0	0	myo_bra-G_ASM41265v1-3	G_ASM41265v1-Mononeg_mRNAcap-4
myo_bra	212	0	1	myo_bra-G_ASM41265v1-6	G_ASM41265v1-Mononeg_mRNAcap-3
myo_bra	134	0	1	myo_bra-G_ASM41265v1-5	G_ASM41265v1-Mononeg_mRNAcap-5
myo_bra	103	0	1	myo_bra-G_ASM41265v1-10	G_ASM41265v1-Mononeg_mRNAcap-11
myo_dav	128	0	3	myo_dav-G_ASM32734v1-7	G_ASM32734v1-Mononeg_mRNAcap-8
myo_luc	167	0	2	myo_luc-G_Myoluc2_0-3	G_Myoluc2_0-Mononeg_mRNAcap-6
myo_luc	140	0	0	myo_luc-G_Myoluc2_0-9	G_Myoluc2_0-Mononeg_mRNAcap-19

myo_luc	123	0	1	myo_luc-G_Myoluc2_0-11	G_Myoluc2_0-Mononeg_mRNAcap-17
myo_myo	511	0	0	myo_myo-G_MyoMyo_v1_BIUU-7	G_MyoMyo_v1_BIUU-Mononeg_mRNAcap-14
myo_myo	234	0	1	myo_myo-G_MyoMyo_v1_BIUU-15	G_MyoMyo_v1_BIUU-Mononeg_mRNAcap-29
myo_myo	157	0	2	myo_myo-G_MyoMyo_v1_BIUU-3	G_MyoMyo_v1_BIUU-Mononeg_mRNAcap-9
myo_myo	104	0	3	myo_myo-G_MyoMyo_v1_BIUU-2	G_MyoMyo_v1_BIUU-Mononeg_mRNAcap-8
myo_myo	99	0	0	myo_myo-G_MyoMyo_v1_BIUU-21	G_MyoMyo_v1_BIUU-Mononeg_mRNAcap-41
myo_myo	99	1	0	myo_myo-G_MyoMyo_v1_BIUU-22	G_MyoMyo_v1_BIUU-Mononeg_mRNAcap-43
myo_myo	99	0	0	myo_myo-G_MyoMyo_v1_BIUU-27	G_MyoMyo_v1_BIUU-Mononeg_mRNAcap-49
myo_myo	70	0	1	myo_myo-G_MyoMyo_v1_BIUU-23	G_MyoMyo_v1_BIUU-Mononeg_mRNAcap-44
myo_myo	66	0	0	myo_myo-G_MyoMyo_v1_BIUU-31	G_MyoMyo_v1_BIUU-Mononeg_mRNAcap-59
myo_myo	64	0	0	myo_myo-G_MyoMyo_v1_BIUU-14	G_MyoMyo_v1_BIUU-Mononeg_mRNAcap-30
pip_pip	103	0	2	pip_pip-G_PipPip_v1_BIUU-1	G_PipPip_v1_BIUU-Mononeg_mRNAcap-5
chEBLLd (Mononeg_mRNAcap)					
las_bor	57	0	1	las_bor-G_LasBor_v1_BIUU-5	G_LasBor_v1_BIUU-Mononeg_mRNAcap-6
myo_myo	62	1	0	myo_myo-G_MyoMyo_v1_BIUU-12	G_MyoMyo_v1_BIUU-Mononeg_mRNAcap-21
myo_myo	37	0	0	myo_myo-G_MyoMyo_v1_BIUU-18	G_MyoMyo_v1_BIUU-Mononeg_mRNAcap-34
pip_pip	123	1	2	pip_pip-G_PipPip_v1_BIUU-11	G_PipPip_v1_BIUU-Mononeg_mRNAcap-19
chEBLLe (Mononeg_mRNAcap)					
art_jam	70	0	0	art_jam-G_ArtJam_v1_BIUU-5	G_ArtJam_v1_BIUU-Mononeg_mRNAcap-13
car_per	100	0	1	car_per-G_CarPer_v1_BIUU-1	G_CarPer_v1_BIUU-Mononeg_mRNAcap-1
cen_sen	70	0	0	cen_sen-G_CeSe_JP-2	G_CeSe_JP-Mononeg_mRNAcap-19
mic_hir	67	0	1	mic_hir-G_MicHir_v1_BIUU-5	G_MicHir_v1_BIUU-Mononeg_mRNAcap-16
noc_lep	84	0	1	noc_lep-G_NocLep_v1_BIUU-2	G_NocLep_v1_BIUU-Mononeg_mRNAcap-3
phy_dis	88	0	1	phy_dis-G_mPhyDis1_v1-4	G_mPhyDis1_v1-Mononeg_mRNAcap-29
tad_bra	134	1	1	tad_bra-G_TadBra_v1_BIUU-1	G_TadBra_v1_BIUU-Mononeg_mRNAcap-2

Table C.2: Putative EFL elements detected in this study. Stops and frame-shifts counted only within aligned hit region. *Total codon length of ORF with greatest overlap with hit.

EFL group Bat Species	Max. ORF*	Stops	F.shifts	EFL ID	Hit ID
chEFLNPa (Ebola_NP)					
ept_fus	218	0	2	ept_fus-G_EptFus1_0-18	G_EptFus1_0-Ebola_NP-28
las_bor	115	0	0	las_bor-G_LasBor_v1_BIUU-1	G_LasBor_v1_BIUU-Ebola_NP-2
mur_fea	130	1	2	mur_fea-G_MurFea_v1_BIUU-10	G_MurFea_v1_BIUU-Ebola_NP-19
myo_bra	224	0	1	myo_bra-G_ASM41265v1-7	G_ASM41265v1-Ebola_NP-7
myo_dav	256	0	0	myo_dav-G_ASM32734v1-11	G_ASM32734v1-Ebola_NP-14
myo_luc	183	0	1	myo_luc-G_Myoluc2_0-12	G_Myoluc2_0-Ebola_NP-20
myo_myo	385	0	0	myo_myo-G_MyoMyo_v1_BIUU-19	G_MyoMyo_v1_BIUU-Ebola_NP-35
pip_pip	111	1	1	pip_pip-G_PipPip_v1_BIUU-5	G_PipPip_v1_BIUU-Ebola_NP-10
chEFLNPb (Ebola_NP)					
mur_fea	116	0	2	mur_fea-G_MurFea_v1_BIUU-19	G_MurFea_v1_BIUU-Ebola_NP-33
mur_fea	76	0	1	mur_fea-G_MurFea_v1_BIUU-21	G_MurFea_v1_BIUU-Ebola_NP-46
mur_fea	66	1	1	mur_fea-G_MurFea_v1_BIUU-22	G_MurFea_v1_BIUU-Ebola_NP-47
myo_bra	139	0	1	myo_bra-G_ASM41265v1-11	G_ASM41265v1-Ebola_NP-12
myo_dav	127	0	2	myo_dav-G_ASM32734v1-1	G_ASM32734v1-Ebola_NP-1

myo_luc	138	0	1	myo_luc-G_Myoluc2_0-6	G_Myoluc2_0-Ebola_NP-9
myo_myo	127	0	2	myo_myo-G_MyoMyo_v1_BIUU-4	G_MyoMyo_v1_BIUU-Ebola_NP-10
chEFLNPc (Ebola_NP)					
ept_fus	147	1	2	ept_fus-G_EptFus1_0-12	G_EptFus1_0-Ebola_NP-17
mur_fea	140	0	1	mur_fea-G_MurFea_v1_BIUU-13	G_MurFea_v1_BIUU-Ebola_NP-25
myo_bra	148	2	2	myo_bra-G_ASM41265v1-17	G_ASM41265v1-Ebola_NP-23
myo_dav	147	1	2	myo_dav-G_ASM32734v1-10	G_ASM32734v1-Ebola_NP-13
myo_myo	222	1	1	myo_myo-G_MyoMyo_v1_BIUU-9	G_MyoMyo_v1_BIUU-Ebola_NP-17
chEFLNPd (Ebola_NP)					
las_bor	89	0	1	las_bor-G_LasBor_v1_BIUU-4	G_LasBor_v1_BIUU-Ebola_NP-4
pip_pip	81	0	2	pip_pip-G_PipPip_v1_BIUU-17	G_PipPip_v1_BIUU-Ebola_NP-32
chEFL35 (Filo_VP35)					
mur_fea	308	0	0	mur_fea-G_MurFea_v1_BIUU-2	G_MurFea_v1_BIUU-Filo_VP35-5
myo_bra	287	0	0	myo_bra-G_ASM41265v1-1	G_ASM41265v1-Filo_VP35-1
myo_dav	277	0	0	myo_dav-G_ASM32734v1-2	G_ASM32734v1-Filo_VP35-2
myo_luc	287	0	0	myo_luc-G_Myoluc2_0-2	G_Myoluc2_0-Filo_VP35-4
myo_myo	290	0	0	myo_myo-G_MyoMyo_v1_BIUU-13	G_MyoMyo_v1_BIUU-Filo_VP35-23

Table C.3: Putative EPV elements detected in this study. Stops and frame-shifts counted only within aligned hit region. *Total codon length of ORF with greatest overlap with hit.

EPV group Bat Species	EPV ID PFAM domain	Max. ORF*	Stops	F.shifts	Hit ID
chEPVa					
ept_fus	ept_fus-G_EptFus1_0-1				
	Rep_N	130	0	4	G_EptFus1_0-Rep_N-23
	Parvo_NS1	130	1	3	G_EptFus1_0-Parvo_NS1-21
	Parvo_coat_N	122	0	0	G_EptFus1_0-Parvo_coat_N-24
	Parvo_coat	166	1	8	G_EptFus1_0-Parvo_coat-22
ept_fus	ept_fus-G_EptFus1_0-2				
	Parvo_coat	102	0	1	G_EptFus1_0-Parvo_coat-25
las_bor	las_bor-G_LasBor_v1_BIUU-1				
	Rep_N	107	1	3	G_LasBor_v1_BIUU-Rep_N-10
	Parvo_NS1	151	3	0	G_LasBor_v1_BIUU-Parvo_NS1-8
	Parvo_coat_N	177	0	1	G_LasBor_v1_BIUU-Parvo_coat_N-11
	Parvo_coat	91	2	4	G_LasBor_v1_BIUU-Parvo_coat-9
las_bor	las_bor-G_LasBor_v1_BIUU-2				
	Parvo_coat	67	0	1	G_LasBor_v1_BIUU-Parvo_coat-12
mur_fea	mur_fea-G_MurFea_v1_BIUU-1				
	Rep_N	87	1	2	G_MurFea_v1_BIUU-Rep_N-13
	Parvo_NS1	258	1	1	G_MurFea_v1_BIUU-Parvo_NS1-11
	Parvo_coat_N	127	0	0	G_MurFea_v1_BIUU-Parvo_coat_N-14
	Parvo_coat	170	0	5	G_MurFea_v1_BIUU-Parvo_coat-12
myo_bra	myo_bra-G_ASM41265v1-1				
	Parvo_coat	202	0	3	G_ASM41265v1-Parvo_coat-15
myo_bra	myo_bra-G_ASM41265v1-2				
	Parvo_coat	60	0	1	G_ASM41265v1-Parvo_coat-18
myo_bra	myo_bra-G_ASM41265v1-3				
	Rep_N	145	0	2	G_ASM41265v1-Rep_N-16
	Parvo_NS1	311	1	3	G_ASM41265v1-Parvo_NS1-14
	Parvo_coat_N	97	0	0	G_ASM41265v1-Parvo_coat_N-17
myo_dav	myo_dav-G_ASM32734v1-2				
	Rep_N	68	1	3	G_ASM32734v1-Rep_N-18
	Parvo_NS1	264	0	3	G_ASM32734v1-Parvo_NS1-16
	Parvo_coat_N	96	0	0	G_ASM32734v1-Parvo_coat_N-19
	Parvo_coat	192	0	2	G_ASM32734v1-Parvo_coat-17
myo_dav	myo_dav-G_ASM32734v1-3				
	Parvo_coat	60	0	1	G_ASM32734v1-Parvo_coat-20
myo_luc	myo_luc-G_Myoluc2_0-1				
	Parvo_coat	112	0	5	G_Myoluc2_0-Parvo_coat-13
myo_luc	myo_luc-G_Myoluc2_0-2				
	Parvo_coat	60	0	1	G_Myoluc2_0-Parvo_coat-16
myo_luc	myo_luc-G_Myoluc2_0-3				
	Rep_N	104	1	2	G_Myoluc2_0-Rep_N-14
	Parvo_NS1	136	0	2	G_Myoluc2_0-Parvo_NS1-12
	Parvo_coat_N	115	0	0	G_Myoluc2_0-Parvo_coat_N-15
myo_myo	myo_myo-G_MyoMyo_v1_BIUU-1				
	Parvo_coat	192	0	5	G_MyoMyo_v1_BIUU-Parvo_coat-3

myo_myo	myo_myo-G_MyoMyo_v1_BIUU-2				
	Parvo_coat	60	0	1	G_MyoMyo_v1_BIUU-Parvo_coat-6
myo_myo	myo_myo-G_MyoMyo_v1_BIUU-3				
	Rep_N	145	0	2	G_MyoMyo_v1_BIUU-Rep_N-4
	Parvo_NS1	136	0	3	G_MyoMyo_v1_BIUU-Parvo_NS1-2
	Parvo_coat_N	79	0	0	G_MyoMyo_v1_BIUU-Parvo_coat_N-5
pip_pip	pip_pip-G_PipPip_v1_BIUU-2				
	Parvo_coat	142	0	3	G_PipPip_v1_BIUU-Parvo_coat-28
pip_pip	pip_pip-G_PipPip_v1_BIUU-3				
	Parvo_coat	67	0	1	G_PipPip_v1_BIUU-Parvo_coat-31
pip_pip	pip_pip-G_PipPip_v1_BIUU-4				
	Rep_N	62	0	3	G_PipPip_v1_BIUU-Rep_N-30
	Parvo_NS1	94	3	5	G_PipPip_v1_BIUU-Parvo_NS1-27
	Parvo_coat_N	173	0	1	G_PipPip_v1_BIUU-Parvo_coat_N-29
pip_pip	pip_pip-G_PipPip_v1_BIUU-5				
	Rep_N	71	0	2	G_PipPip_v1_BIUU-Rep_N-36
pip_pip	pip_pip-G_PipPip_v1_BIUU-6				
	Rep_N	54	0	3	G_PipPip_v1_BIUU-Rep_N-39
chEPVb					
meg_lyr	meg_lyr-G_ASM46534v1-3				
	Parvo_NS1	80	1	1	G_ASM46534v1-Parvo_NS1-27
	Parvo_coat_N	182	0	0	G_ASM46534v1-Parvo_coat_N-26
	Parvo_coat	308	0	0	G_ASM46534v1-Parvo_coat-25
meg_lyr	meg_lyr-G_MegLyr_v1_BIUU-1				
	Parvo_NS1	77	1	2	G_MegLyr_v1_BIUU-Parvo_NS1-3
	Parvo_coat_N	200	0	0	G_MegLyr_v1_BIUU-Parvo_coat_N-2
	Parvo_coat	321	0	0	G_MegLyr_v1_BIUU-Parvo_coat-1
chEPVc					
meg_lyr	meg_lyr-G_ASM46534v1-2				
	Parvo_coat	85	0	0	G_ASM46534v1-Parvo_coat-23
meg_lyr	meg_lyr-G_ASM46534v1-4				
	Parvo_coat	72	1	0	G_ASM46534v1-Parvo_coat-32
meg_lyr	meg_lyr-G_MegLyr_v1_BIUU-3				
	Parvo_coat	330	1	4	G_MegLyr_v1_BIUU-Parvo_coat-10
chEPVd					
cra_tho	cra_tho-G_CraTho_v1_BIUU-2				
	Rep_N	138	1	0	G_CraTho_v1_BIUU-Rep_N-4
chEPVe					
rhi_fer	rhi_fer-G_ASM46549v1-3				
	Rep_N	107	1	0	G_ASM46549v1-Rep_N-16
	Parvo_NS1	197	1	5	G_ASM46549v1-Parvo_NS1-15
	Parvo_coat_N	114	0	0	G_ASM46549v1-Parvo_coat_N-17
	Parvo_coat	285	0	2	G_ASM46549v1-Parvo_coat-14
rhi_fer	rhi_fer-G_ASM46549v1-4				
	Parvo_coat_N	97	0	0	G_ASM46549v1-Parvo_coat_N-19
	Parvo_coat	279	4	3	G_ASM46549v1-Parvo_coat-18
rhi_fer	rhi_fer-G_mRhiFer1_v1-2				
	Parvo_coat_N	145	0	0	G_mRhiFer1_v1-Parvo_coat_N-4
	Parvo_coat	279	4	3	G_mRhiFer1_v1-Parvo_coat-3
rhi_fer	rhi_fer-G_mRhiFer1_v1-4				
	Rep_N	111	1	1	G_mRhiFer1_v1-Rep_N-20
	Parvo_NS1	197	1	5	G_mRhiFer1_v1-Parvo_NS1-19
	Parvo_coat_N	165	0	0	G_mRhiFer1_v1-Parvo_coat_N-22
	Parvo_coat	325	0	4	G_mRhiFer1_v1-Parvo_coat-17
rhi_fer	rhi_fer-G_mRhiFer1_v1-5				
	Rep_N	107	1	0	G_mRhiFer1_v1-Rep_N-21
	Parvo_NS1	197	1	5	G_mRhiFer1_v1-Parvo_NS1-18
rhi_fer	rhi_fer-G_Rhfe_GT_ERC-1				
	Parvo_coat_N	139	0	0	G_Rhfe_GT_ERC-Parvo_coat_N-3
	Parvo_coat	279	4	3	G_Rhfe_GT_ERC-Parvo_coat-2
rhi_fer	rhi_fer-G_Rhfe_GT_ERC-3				
	Rep_N	101	0	1	G_Rhfe_GT_ERC-Rep_N-10
rhi_fer	rhi_fer-G_Rhfe_GT_ERC-4				
	Rep_N	111	1	1	G_Rhfe_GT_ERC-Rep_N-8
	Parvo_NS1	197	1	5	G_Rhfe_GT_ERC-Parvo_NS1-7
	Parvo_coat_N	165	0	0	G_Rhfe_GT_ERC-Parvo_coat_N-9
	Parvo_coat	325	0	4	G_Rhfe_GT_ERC-Parvo_coat-6
rhi_fer	rhi_fer-G_Rhfe_GT_ERC-6				
	Rep_N	107	1	0	G_Rhfe_GT_ERC-Rep_N-15

	Parvo_NS1	200	1	5	G_Rhfe_GT_ERC-Parvo_NS1-14
rhi_pea	rhi_pea-G_Rhpe_GT-2				
	Rep_N	162	0	0	G_Rhpe_GT-Rep_N-13
	Parvo_NS1	122	2	4	G_Rhpe_GT-Parvo_NS1-11
	Parvo_coat_N	244	0	0	G_Rhpe_GT-Parvo_coat_N-12
	Parvo_coat	134	3	3	G_Rhpe_GT-Parvo_coat-10
rhi_pea	rhi_pea-G_Rhpe_GT-3				
	Parvo_coat	217	0	4	G_Rhpe_GT-Parvo_coat-19
rhi_pea	rhi_pea-G_Rhpe_GT-4				
	Rep_N	108	1	0	G_Rhpe_GT-Rep_N-24
	Parvo_NS1	262	0	6	G_Rhpe_GT-Parvo_NS1-22
rhi_pea	rhi_pea-G_Rhpe_GT-5				
	Rep_N	64	0	0	G_Rhpe_GT-Rep_N-25
	Parvo_NS1	122	0	3	G_Rhpe_GT-Parvo_NS1-23
rhi_sin	rhi_sin-G_ASM188883v1-1				
	Rep_N	76	1	0	G_ASM188883v1-Rep_N-5
	Parvo_NS1	101	1	4	G_ASM188883v1-Parvo_NS1-3
	Parvo_coat_N	100	0	1	G_ASM188883v1-Parvo_coat_N-4
	Parvo_coat	136	3	3	G_ASM188883v1-Parvo_coat-2
rhi_sin	rhi_sin-G_ASM188883v1-2				
	Rep_N	74	1	1	G_ASM188883v1-Rep_N-8
	Parvo_NS1	180	0	5	G_ASM188883v1-Parvo_NS1-7
	Parvo_coat_N	168	0	0	G_ASM188883v1-Parvo_coat_N-9
	Parvo_coat	382	0	4	G_ASM188883v1-Parvo_coat-6
chEPVf					
pip_pip	pip_pip-G_PipPip_v1_BIUU-1				
	Rep_N	103	1	0	G_PipPip_v1_BIUU-Rep_N-4
	Parvo_NS1	449	0	0	G_PipPip_v1_BIUU-Parvo_NS1-2
	Parvo_coat_N	113	1	0	G_PipPip_v1_BIUU-Parvo_coat_N-3
	Parvo_coat	638	0	1	G_PipPip_v1_BIUU-Parvo_coat-1
chEPVg					
cra_tho	cra_tho-G_CraTho_v1_BIUU-1				
	Parvo_coat	66	0	1	G_CraTho_v1_BIUU-Parvo_coat-2
chEPVh					
ano_cau	ano_cau-G_AnoCau_v1_BIUU-1				
	Rep_N	178	0	2	G_AnoCau_v1_BIUU-Rep_N-5
	Parvo_NS1	159	0	4	G_AnoCau_v1_BIUU-Parvo_NS1-2
	Parvo_coat_N	77	0	1	G_AnoCau_v1_BIUU-Parvo_coat_N-4
	Parvo_coat	73	7	0	G_AnoCau_v1_BIUU-Parvo_coat-3
art_jam	art_jam-G_ArtJam_v1_BIUU-1				
	Rep_N	133	0	2	G_ArtJam_v1_BIUU-Rep_N-5
	Parvo_NS1	68	3	5	G_ArtJam_v1_BIUU-Parvo_NS1-3
	Parvo_coat_N	133	0	1	G_ArtJam_v1_BIUU-Parvo_coat_N-6
	Parvo_coat	91	7	2	G_ArtJam_v1_BIUU-Parvo_coat-4
car_per	car_per-G_CarPer_v1_BIUU-3				
	Rep_N	85	1	1	G_CarPer_v1_BIUU-Rep_N-15
	Parvo_NS1	110	3	4	G_CarPer_v1_BIUU-Parvo_NS1-12
	Parvo_coat_N	53	0	2	G_CarPer_v1_BIUU-Parvo_coat_N-14
	Parvo_coat	94	6	2	G_CarPer_v1_BIUU-Parvo_coat-13
car_per	car_per-G_CarPer_v1_BIUU-4				
	Rep_N	109	0	2	G_CarPer_v1_BIUU-Rep_N-17
car_per	car_per-G_CarPer_v1_BIUU-5				
	Rep_N	109	0	1	G_CarPer_v1_BIUU-Rep_N-30
cen_sen	cen_sen-G_CeSe_JP-1				
	Parvo_coat_N	77	0	0	G_CeSe_JP-Parvo_coat_N-5
cen_sen	cen_sen-G_CeSe_JP-10				
	Parvo_coat	149	2	2	G_CeSe_JP-Parvo_coat-31
cen_sen	cen_sen-G_CeSe_JP-6				
	Parvo_coat	84	3	1	G_CeSe_JP-Parvo_coat-24
des_rot	des_rot-G_ASM294091v2-3				
	Parvo_coat_N	84	0	0	G_ASM294091v2-Parvo_coat_N-21
	Parvo_coat	140	6	0	G_ASM294091v2-Parvo_coat-20
lio_spu	lio_spu-G_Lisp_GT-1				
	Rep_N	136	1	3	G_Lisp_GT-Rep_N-23
	Parvo_NS1	110	1	1	G_Lisp_GT-Parvo_NS1-20
	Parvo_coat_N	76	0	1	G_Lisp_GT-Parvo_coat_N-22
	Parvo_coat	94	6	2	G_Lisp_GT-Parvo_coat-21
mic_hir	mic_hir-G_MicHir_v1_BIUU-3				
	Rep_N	82	0	0	G_MicHir_v1_BIUU-Rep_N-9

	Parvo_NS1	242	3	2	G_MicHir_v1_BIUU-Parvo_NS1-7
	Parvo_coat	191	2	0	G_MicHir_v1_BIUU-Parvo_coat-8
mor_bla	mor_bla-G_MorMeg_v1_BIUU-1				
	Rep_N	82	0	4	G_MorMeg_v1_BIUU-Rep_N-5
	Parvo_NS1	213	2	3	G_MorMeg_v1_BIUU-Parvo_NS1-3
	Parvo_coat_N	123	0	0	G_MorMeg_v1_BIUU-Parvo_coat_N-6
	Parvo_coat	259	4	6	G_MorMeg_v1_BIUU-Parvo_coat-4
phy_dis	phy_dis-G_mPhyDis1_v1-1				
	Rep_N	134	0	3	G_mPhyDis1_v1-Rep_N-7
	Parvo_NS1	77	4	1	G_mPhyDis1_v1-Parvo_NS1-5
	Parvo_coat_N	60	0	1	G_mPhyDis1_v1-Parvo_coat_N-8
	Parvo_coat	71	6	0	G_mPhyDis1_v1-Parvo_coat-6
pte_par	pte_par-G_ASM46540v1-1				
	Rep_N	64	2	1	G_ASM46540v1-Rep_N-6
	Parvo_NS1	281	2	2	G_ASM46540v1-Parvo_NS1-3
	Parvo_coat_N	141	0	0	G_ASM46540v1-Parvo_coat_N-5
	Parvo_coat	126	5	3	G_ASM46540v1-Parvo_coat-4
ton_sau	ton_sau-G_TonSau_v1_BIUU-2				
	Rep_N	76	1	2	G_TonSau_v1_BIUU-Rep_N-11
	Parvo_NS1	164	2	2	G_TonSau_v1_BIUU-Parvo_NS1-8
	Parvo_coat_N	77	0	0	G_TonSau_v1_BIUU-Parvo_coat_N-10
	Parvo_coat	131	2	1	G_TonSau_v1_BIUU-Parvo_coat-9
chEPVi					
ano_cau	ano_cau-G_AnoCau_v1_BIUU-3				
	Rep_N	136	1	3	G_AnoCau_v1_BIUU-Rep_N-16
	Parvo_NS1	73	1	1	G_AnoCau_v1_BIUU-Parvo_NS1-14
	Parvo_coat_N	280	0	0	G_AnoCau_v1_BIUU-Parvo_coat_N-15
	Parvo_coat	136	0	7	G_AnoCau_v1_BIUU-Parvo_coat-13
art_jam	art_jam-G_ArtJam_v1_BIUU-3				
	Rep_N	140	0	0	G_ArtJam_v1_BIUU-Rep_N-20
	Parvo_NS1	154	1	1	G_ArtJam_v1_BIUU-Parvo_NS1-18
	Parvo_coat_N	156	1	0	G_ArtJam_v1_BIUU-Parvo_coat_N-19
car_per	car_per-G_CarPer_v1_BIUU-2				
	Parvo_NS1	161	0	1	G_CarPer_v1_BIUU-Parvo_NS1-7
	Parvo_coat_N	76	1	0	G_CarPer_v1_BIUU-Parvo_coat_N-8
	Parvo_coat	199	0	6	G_CarPer_v1_BIUU-Parvo_coat-6
cen_sen	cen_sen-G_CeSe_JP-3				
	Parvo_coat_N	44	1	1	G_CeSe_JP-Parvo_coat_N-13
cen_sen	cen_sen-G_CeSe_JP-7				
	Parvo_coat	135	1	5	G_CeSe_JP-Parvo_coat-26
cen_sen	cen_sen-G_CeSe_JP-8				
	Parvo_NS1	154	1	1	G_CeSe_JP-Parvo_NS1-28
	Parvo_coat	99	1	3	G_CeSe_JP-Parvo_coat-27
cen_sen	cen_sen-G_CeSe_JP-9				
	Rep_N	121	0	0	G_CeSe_JP-Rep_N-29
des_rot	des_rot-G_ASM294091v2-2				
	Rep_N	100	1	1	G_ASM294091v2-Rep_N-8
	Parvo_NS1	131	0	2	G_ASM294091v2-Parvo_NS1-5
	Parvo_coat_N	208	0	0	G_ASM294091v2-Parvo_coat_N-6
	Parvo_coat	136	0	4	G_ASM294091v2-Parvo_coat-4
lio_spu	lio_spu-G_Lisp_GT-3				
	Parvo_NS1	278	0	0	G_Lisp_GT-Parvo_NS1-35
	Parvo_coat_N	83	1	0	G_Lisp_GT-Parvo_coat_N-36
	Parvo_coat	201	1	4	G_Lisp_GT-Parvo_coat-34
mic_hir	mic_hir-G_MicHir_v1_BIUU-1				
	Rep_N	99	2	1	G_MicHir_v1_BIUU-Rep_N-4
	Parvo_NS1	66	1	3	G_MicHir_v1_BIUU-Parvo_NS1-2
	Parvo_coat_N	169	0	0	G_MicHir_v1_BIUU-Parvo_coat_N-3
	Parvo_coat	185	1	4	G_MicHir_v1_BIUU-Parvo_coat-1
mor_bla	mor_bla-G_MorMeg_v1_BIUU-2				
	Rep_N	188	0	0	G_MorMeg_v1_BIUU-Rep_N-8
	Parvo_NS1	109	4	1	G_MorMeg_v1_BIUU-Parvo_NS1-7
phy_dis	phy_dis-G_mPhyDis1_v1-3				
	Rep_N	112	0	1	G_mPhyDis1_v1-Rep_N-16
	Parvo_NS1	194	0	0	G_mPhyDis1_v1-Parvo_NS1-12
	Parvo_coat_N	76	0	2	G_mPhyDis1_v1-Parvo_coat_N-15
	Parvo_coat	208	1	3	G_mPhyDis1_v1-Parvo_coat-14
pte_par	pte_par-G_ASM46540v1-3				
	Parvo_coat_N	193	0	1	G_ASM46540v1-Parvo_coat_N-33

	Parvo_coat	212	2	4	G_ASM46540v1-Parvo_coat-32
pte_par	pte_par-G_ASM46540v1-4				
	Rep_N	146	1	0	G_ASM46540v1-Rep_N-38
	Parvo_NS1	85	3	3	G_ASM46540v1-Parvo_NS1-36
	Parvo_coat_N	248	0	0	G_ASM46540v1-Parvo_coat_N-37
	Parvo_coat	164	2	1	G_ASM46540v1-Parvo_coat-35
ton_sau	ton_sau-G_TonSau_v1_BIUU-3				
	Rep_N	68	0	0	G_TonSau_v1_BIUU-Rep_N-17
	Parvo_NS1	193	0	0	G_TonSau_v1_BIUU-Parvo_NS1-15
	Parvo_coat_N	120	0	1	G_TonSau_v1_BIUU-Parvo_coat_N-16
	Parvo_coat	128	0	6	G_TonSau_v1_BIUU-Parvo_coat-14
chEPVj					
ano_cau	ano_cau-G_AnoCau_v1_BIUU-2				
	Rep_N	166	0	1	G_AnoCau_v1_BIUU-Rep_N-10
	Parvo_NS1	127	2	5	G_AnoCau_v1_BIUU-Parvo_NS1-8
	Parvo_coat_N	52	0	1	G_AnoCau_v1_BIUU-Parvo_coat_N-11
	Parvo_coat	138	3	5	G_AnoCau_v1_BIUU-Parvo_coat-9
art_jam	art_jam-G_ArtJam_v1_BIUU-2				
	Rep_N	168	0	1	G_ArtJam_v1_BIUU-Rep_N-11
	Parvo_NS1	122	1	3	G_ArtJam_v1_BIUU-Parvo_NS1-10
	Parvo_coat_N	53	0	2	G_ArtJam_v1_BIUU-Parvo_coat_N-12
	Parvo_coat	169	3	4	G_ArtJam_v1_BIUU-Parvo_coat-9
car_per	car_per-G_CarPer_v1_BIUU-1				
	Rep_N	80	1	2	G_CarPer_v1_BIUU-Rep_N-4
	Parvo_NS1	95	2	4	G_CarPer_v1_BIUU-Parvo_NS1-2
	Parvo_coat_N	39	1	1	G_CarPer_v1_BIUU-Parvo_coat_N-5
	Parvo_coat	130	2	5	G_CarPer_v1_BIUU-Parvo_coat-3
car_per	car_per-G_CarPer_v1_BIUU-6				
	Parvo_coat	101	0	0	G_CarPer_v1_BIUU-Parvo_coat-31
cen_sen	cen_sen-G_CeSe_JP-2				
	Rep_N	126	0	0	G_CeSe_JP-Rep_N-8
cen_sen	cen_sen-G_CeSe_JP-4				
	Parvo_NS1	140	1	4	G_CeSe_JP-Parvo_NS1-17
	Parvo_coat_N	36	0	1	G_CeSe_JP-Parvo_coat_N-18
cen_sen	cen_sen-G_CeSe_JP-5				
	Parvo_coat	101	2	1	G_CeSe_JP-Parvo_coat-20
des_rot	des_rot-G_ASM294091v2-1				
	Rep_N	171	0	0	G_ASM294091v2-Rep_N-9
	Parvo_NS1	181	2	4	G_ASM294091v2-Parvo_NS1-3
	Parvo_coat_N	70	0	0	G_ASM294091v2-Parvo_coat_N-7
lio_spu	lio_spu-G_Lisp_GT-2				
	Rep_N	117	1	1	G_Lisp_GT-Rep_N-32
	Parvo_NS1	144	2	4	G_Lisp_GT-Parvo_NS1-30
	Parvo_coat_N	155	1	1	G_Lisp_GT-Parvo_coat_N-33
	Parvo_coat	159	1	6	G_Lisp_GT-Parvo_coat-31
mic_hir	mic_hir-G_MicHir_v1_BIUU-4				
	Rep_N	91	2	1	G_MicHir_v1_BIUU-Rep_N-14
	Parvo_NS1	206	2	2	G_MicHir_v1_BIUU-Parvo_NS1-12
	Parvo_coat	97	4	3	G_MicHir_v1_BIUU-Parvo_coat-13
mor_bla	mor_bla-G_MorMeg_v1_BIUU-3				
	Rep_N	122	0	3	G_MorMeg_v1_BIUU-Rep_N-11
	Parvo_NS1	126	1	1	G_MorMeg_v1_BIUU-Parvo_NS1-10
	Parvo_coat_N	80	1	1	G_MorMeg_v1_BIUU-Parvo_coat_N-12
phy_dis	phy_dis-G_mPhyDis1_v1-2				
	Parvo_NS1	87	5	2	G_mPhyDis1_v1-Parvo_NS1-11
	Parvo_coat	148	1	2	G_mPhyDis1_v1-Parvo_coat-13
ton_sau	ton_sau-G_TonSau_v1_BIUU-1				
	Rep_N	138	2	1	G_TonSau_v1_BIUU-Rep_N-4
	Parvo_NS1	125	1	3	G_TonSau_v1_BIUU-Parvo_NS1-2
	Parvo_coat_N	118	0	1	G_TonSau_v1_BIUU-Parvo_coat_N-5
	Parvo_coat	210	1	4	G_TonSau_v1_BIUU-Parvo_coat-3
chEPVk					
eid_hel	eid_hel-G_ASM46528v1-2				
	Parvo_NS1	109	1	1	G_ASM46528v1-Parvo_NS1-6
eon_spe	eon_spe-G_Espe_v1-1				
	Parvo_NS1	85	0	1	G_Espe_v1-Parvo_NS1-1
mac_sob	mac_sob-G_MacSob_v1_BIUU-1				
	Parvo_NS1	127	0	1	G_MacSob_v1_BIUU-Parvo_NS1-4
rou_aeg	rou_aeg-G_RouAeg_v1_BIUU-1				

	Parvo_NS1	97	0	1	G_RouAeg_v1_BIUU-Parvo_NS1-1
chEPVl					
myo_dav	myo_dav-G_ASM32734v1-1				
	Parvo_NS1	498	0	0	G_ASM32734v1-Parvo_NS1-12
chEPVm					
eid_hel	eid_hel-G_ASM46528v1-1				
	Parvo_NS1	83	1	3	G_ASM46528v1-Parvo_NS1-5
eon_spe	eon_spe-G_Espe_v1-2				
	Parvo_NS1	83	0	0	G_Espe_v1-Parvo_NS1-12
hip_gal	hip_gal-G_HipGal_v1_BIUU-2				
	Parvo_NS1	111	0	2	G_HipGal_v1_BIUU-Parvo_NS1-18
meg_lyr	meg_lyr-G_ASM46534v1-1				
	Parvo_NS1	92	0	1	G_ASM46534v1-Parvo_NS1-8
meg_lyr	meg_lyr-G_MegLyr_v1_BIUU-2				
	Parvo_NS1	92	0	0	G_MegLyr_v1_BIUU-Parvo_NS1-9
min_sch	min_sch-G_MinSch_v1_BIUU-1				
	Parvo_NS1	116	1	3	G_MinSch_v1_BIUU-Parvo_NS1-8
mor_bla	mor_bla-G_MorMeg_v1_BIUU-4				
	Parvo_NS1	79	1	2	G_MorMeg_v1_BIUU-Parvo_NS1-19
pte_par	pte_par-G_ASM46540v1-2				
	Parvo_NS1	30	2	0	G_ASM46540v1-Parvo_NS1-25
rhi_fer	rhi_fer-G_ASM46549v1-1				
	Parvo_NS1	63	2	2	G_ASM46549v1-Parvo_NS1-9
rhi_fer	rhi_fer-G_mRhiFer1_v1-1				
	Parvo_NS1	63	2	2	G_mRhiFer1_v1-Parvo_NS1-1
rhi_fer	rhi_fer-G_Rhfe_GT_ERC-2				
	Parvo_NS1	63	2	2	G_Rhfe_GT_ERC-Parvo_NS1-4
rhi_pea	rhi_pea-G_Rhpe_GT-1				
	Parvo_NS1	86	0	1	G_Rhpe_GT-Parvo_NS1-9
rou_aeg	rou_aeg-G_RouAeg_v1_BIUU-3				
	Parvo_NS1	83	0	1	G_RouAeg_v1_BIUU-Parvo_NS1-8
chEPVn					
hip_gal	hip_gal-G_HipGal_v1_BIUU-1				
	Parvo_NS1	77	0	0	G_HipGal_v1_BIUU-Parvo_NS1-14
rhi_fer	rhi_fer-G_ASM46549v1-2				
	Parvo_NS1	59	0	1	G_ASM46549v1-Parvo_NS1-13
rhi_fer	rhi_fer-G_mRhiFer1_v1-3				
	Parvo_NS1	59	0	1	G_mRhiFer1_v1-Parvo_NS1-16
rhi_fer	rhi_fer-G_Rhfe_GT_ERC-5				
	Parvo_NS1	59	0	1	G_Rhfe_GT_ERC-Parvo_NS1-13
rhi_pea	rhi_pea-G_Rhpe_GT-6				
	Parvo_NS1	51	0	1	G_Rhpe_GT-Parvo_NS1-36
chEPVo					
pte_vam	pte_vam-G_Pvam2_0-1				
	Parvo_NS1	95	0	1	G_Pvam2_0-Parvo_NS1-2
rou_aeg	rou_aeg-G_Raegyp2_0-1				
	Parvo_NS1	64	0	1	G_Raegyp2_0-Parvo_NS1-1
rou_aeg	rou_aeg-G_RouAeg_v1_BIUU-2				
	Parvo_NS1	64	0	0	G_RouAeg_v1_BIUU-Parvo_NS1-3
chEPVp					
mic_hir	mic_hir-G_MicHir_v1_BIUU-2				
	Parvo_NS1	40	0	0	G_MicHir_v1_BIUU-Parvo_NS1-5
phy_dis	phy_dis-G_mPhyDis1_v1-4				
	Parvo_NS1	95	0	0	G_mPhyDis1_v1-Parvo_NS1-33
ton_sau	ton_sau-G_TonSau_v1_BIUU-4				
	Parvo_NS1	56	0	0	G_TonSau_v1_BIUU-Parvo_NS1-18

C.2 Supplementary Figures

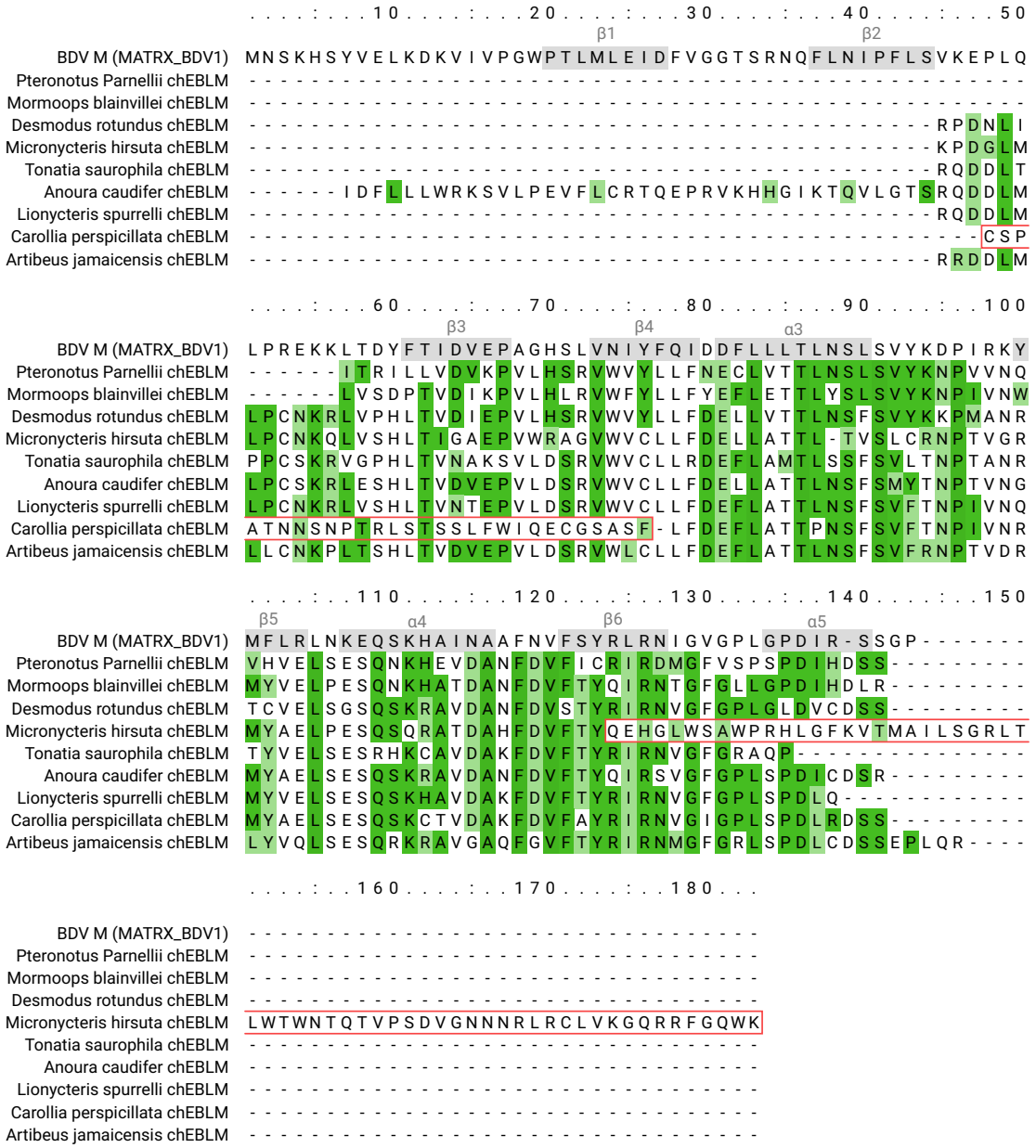


Figure C.1: Multiple sequence alignment of BDV M protein and chEBLM ORF-encoded peptide sequences. Residues identical to BDV M: dark green; residues similar to BDV M: light green. Frameshifted regions indicated by red boxes. Secondary structure of BDV-M labelled in grey.

C.3 Supplementary methods

C.3.1 Genome assemblies used in Chapter 5

Accessions of the 40 genome assemblies analysed in Chapter 4: *Eonycteris spelaea* (GCA_003508835.1), *Macroglossus sobrinus* (GCA_004027375.1), *Eidolon helvum* (GCA_000465285.1), *Rousettus aegyptiacus* (GCA_004024865.1, GCA_001466805.2), *Pteropus vampyrus* (GCA_000151845.2), *Pteropus alecto* (GCA_000325575.1), *Megaderma lyra* (GCA_004026885.1, GCA_000465345.1), *Craseonycteris thonglongyai* (GCA_004027555.1), *Hipposideros armiger* (GCA_001890085.1), *Hipposideros galeritus* (GCA_004027415.1), *Rhinolophus sinicus* (GCA_001888835.1), *Rhinolophus ferrumequinum* (S.J. Rossiter unpublished data, GCA_004115265.1, GCA_000465495.1), *Rhinolophus pearsonii* (S.J. Rossiter unpublished data), *Tadarida brasiliensis* (GCA_004025005.1), *Miniopterus natalensis* (GCA_001595765.1), *Miniopterus schreibersii* (GCA_004026525.1), *Lasiurus borealis* (GCA_004026805.1), *Pipistrellus pipistrellus* (GCA_004026625.1), *Eptesicus fuscus* (GCA_000308155.1), *Murina aurata* (GCA_004026665.1), *Myotis lucifugus* (GCA_000147115.1), *Myotis myotis* (GCA_004026985.1), *Myotis brandtii* (GCA_000412655.1), *Myotis davidii* (GCA_000327345.1), *Noctilio leporinus* (GCA_004026585.1), *Mormoops blainvillei* (GCA_004026545.1), *Pteronotus parnellii* (GCA_000465405.1), *Desmodus rotundus* (GCA_002940915.2), *Micronycteris hirsuta* (GCA_004026765.1), *Phyllostomus discolor* (GCA_004126475.1), *Tonatia saurophila* (GCA_004024845.1), *Anoura caudifer* (GCA_004027475.1), *Lionycteris spurrelli* (S.J. Rossiter unpublished data), *Carollia perspicillata* (GCA_004027735.1), *Artibeus jamaicensis* (GCA_004027435.1), *Centurio senex* (S.J. Rossiter unpublished data).

D

Chapter 5 supplementary material

D.1 Supplementary tables

Table D.1: Non-retroviral EVE loci found to be expressed amongst 189 RNA-seq libraries.
* Orthogroups correspond to orthologous EVE groups assigned in chapter 4.

Species	Orthogroup*	Viral Domain	Max ORF length	Expression detected in	Hit ID
car_per	chEBLNa	bornavirus N	125	multiple tissues	G_CarPer_v1_BIUU-BDV_P40-25
min_nat	chEBLNc	bornavirus N	386	embryo limbs	G_Mnatv1-BDV_P40-2
min_sch	chEBLNc	bornavirus N	386	embryo limbs	G_MinSch_v1_BIUU-BDV_P40-2
des_rot	chEBLNe	bornavirus N	167	multiple tissues	G_ASM294091v2-BDV_P40-15
art_jam	chEBLNe	bornavirus N	100	kidney, liver, spleen	G_ArtJam_v1_BIUU-BDV_P40-2
ton_sau	chEBLNe	bornavirus N	121	multiple tissues	G_TonSau_v1_BIUU-BDV_P40-7
min_sch	chEBLNj	bornavirus N	174	embryo limbs	G_MinSch_v1_BIUU-BDV_P40-9
ept_fus	chEBLNm	bornavirus N	217	cell culture	G_EptFus1_0-BDV_P40-32
ept_fus	chEBLNm	bornavirus N	325	cell culture	G_EptFus1_0-BDV_P40-8
min_sch	chEBLNo	bornavirus N	145	embryo limbs	G_MinSch_v1_BIUU-BDV_P40-4
min_nat	chEBLNo	bornavirus N	145	embryo limbs	G_Mnatv1-BDV_P40-3
rou_aeg	chEPVo	parvovirus NS	64	multiple tissues	G_RouAeg_v1_BIUU-Parvo_NS1-3

Table D.2: Total counts of ERV domain loci annotated, and number detected as expressed amongst 189 RNA-seq libraries, for each bat species in this study. An ERV domain locus was classified as expressed if any individual RNA-seq library contained three or more reads mapping to that locus.

Species	GAG		DUT		AP		RT		RNaseH		INT		ENV	
	Expr.	Total	Expr.	Total	Expr.	Total	Expr.	Total	Expr.	Total	Expr.	Total	Expr.	Total
ano_cau	-	2082	-	353	-	634	-	2078	-	1249	-	2678	-	1171
art_jam	279	1879	23	235	85	538	164	1679	153	1197	245	2146	439	1982
car_per	268	2780	9	569	53	813	102	1809	99	1262	343	3442	294	2621
cen_sen	-	2009	-	337	-	0	-	1916	-	1190	-	2466	-	1094
cra_tho	-	4138	-	209	-	614	-	1248	-	1340	-	3639	-	1678
des_rot	63	549	7	92	18	256	32	701	55	620	76	907	52	297
eid_hel	45	551	2	26	20	185	41	748	40	546	62	766	44	436
eon_spe	252	1663	84	220	103	467	235	1374	168	1333	316	1865	345	1634
ept_fus	9	736	0	95	6	271	4	669	9	585	24	1013	7	510
hip_arm	103	695	16	178	32	381	124	1108	80	821	168	1093	141	639
hip_gal	-	2608	-	932	-	1001	-	2840	-	1853	-	2954	-	2510
las_bor	-	4305	-	1156	-	1441	-	3936	-	3363	-	6355	-	2527
lio_spu	-	704	-	88	-	347	-	879	-	740	-	1165	-	389
mac_sob	-	1351	-	219	-	383	-	1326	-	1035	-	1569	-	1758
meg_lyr	17	5678	1	822	11	1100	13	6204	24	2973	53	7414	117	7450
mic_hir	-	2982	-	931	-	1175	-	2645	-	1559	-	2933	-	1991
min_nat	55	265	21	57	40	172	107	424	67	341	97	444	62	214
min_sch	16	291	6	68	12	182	27	520	21	467	47	702	24	294
mor_bla	5	1663	0	24	6	685	3	1786	3	1271	19	2396	14	470
mur_fea	-	3211	-	359	-	1020	-	2080	-	1687	-	3541	-	1383
myo_bra	-	864	-	114	-	314	-	782	-	583	-	1141	-	661
myo_dav	-	756	-	106	-	295	-	647	-	538	-	1052	-	561
myo_luc	18	1140	1	228	32	590	4	888	4	732	41	1622	43	782
myo_myo	-	2829	-	631	-	1383	-	1469	-	1295	-	3312	-	1326
noc_lep	12	535	0	64	18	251	3	1572	6	627	12	1540	5	243
phy_dis	39	789	5	116	22	342	27	1046	26	835	44	1173	29	399
pip_pip	-	994	-	95	-	319	-	773	-	1291	-	1616	-	712
pte_ale	141	778	5	29	44	231	96	857	106	631	164	927	127	609
pte_par	11	457	0	46	6	175	6	696	7	540	16	762	3	189
pte_vam	-	923	-	46	-	338	-	915	-	841	-	1126	-	865
rhi_fer	0	660	0	213	0	423	0	1363	0	779	0	1268	0	629
rhi_pea	29	666	7	192	14	376	26	1670	21	810	53	1372	73	647
rou_aeg	976	1956	137	337	213	553	572	2089	632	1329	950	1844	719	1386
tad_bra	43	2938	4	636	27	1331	54	2163	51	2261	84	3937	71	1882
ton_sau	102	1032	6	176	47	378	82	1260	57	787	133	1441	131	838

Table D.3: Overview of RNA-seq datasets used this study.

Species	Description	Sample type	Comparison	Project accession	Associated article	Total no. libs.	Sequencing run accessions
H. armiger	Transcriptome sequencing for gene prediction	brain, heart, kidney, liver, lung, muscle, spleen	-	PRJNA267626	-	7	SRR1657901, SRR1657902, SRR1657903, SRR1657904, SRR1657905, SRR1657906, SRR1657907
M. lyra	Brain transcriptome sequencing investigating bat phylogeny	brain	-	PRJNA290538	Lei and Dong 2016	1	SRR2153218
M. natalensis	Differential expression in M. natalensis embryo limbs	embryo fore-limb, embryo hind-limb	-	PRJNA270639	Eckalbar et al. 2016	18	SRR1719013, SRR1719014, SRR1719015, SRR1719016, SRR1719017, SRR1719018, SRR1719204, SRR1719206, SRR1719207, SRR1719208, SRR1719209, SRR1719211, SRR1719212, SRR1719213, SRR1719214, SRR1719241, SRR1719242, SRR1719266
M. schreibersii	Transcriptome sequencing of embryonic autopods of M. schreibersii	embryo limbs and digits	-	PRJNA218524	Z. Wang et al. 2014	14	SRR974728, SRR974729, SRR974730, SRR974731, SRR974732, SRR974733, SRR974734, SRR974735, SRR974736, SRR974737, SRR974738, SRR974739, SRR974740, SRR974741
T. brasiliensis	Transcriptome sequencing of tissues investigating DNA repair mechanisms across mammal species	liver	-	PRJNA184055	Fushan et al. 2015	3	SRR636883, SRR636884, SRR636885
M. blainvillei	Tissue sequencing: heart	heart	-	unpublished	provided by KD	4	DR091_MOBL_H_L001, DR091_MOBL_H_L002, DR091_MOBL_H_L003, DR091_MOBL_H_L004
P. parnellii	Tissue sequencing: heart	heart	-	unpublished	provided by KD	16	DR038_PTPA_H_L001, DR038_PTPA_H_L002, DR038_PTPA_H_L003, DR038_PTPA_H_L004, DR073_PTPA_H_L001, DR073_PTPA_H_L002, DR073_PTPA_H_L003, DR073_PTPA_H_L004, DR089_PTPA_H_L001, DR089_PTPA_H_L002, DR089_PTPA_H_L003, DR089_PTPA_H_L004, DR090_PTPA_H_L001, DR090_PTPA_H_L002, DR090_PTPA_H_L003, DR090_PTPA_H_L004
N. leporinus	Tissue sequencing: heart	heart	-	unpublished	provided by KD	4	DR101_NOLE_H_L001, DR101_NOLE_H_L002, DR101_NOLE_H_L003, DR101_NOLE_H_L004
A. jamaicensis	Transcriptome profiling of a pathogen response in Artibeus jamaicensis experimentally infected with tacaribe virus	spleen, kidney, liver	infected vs uninfected individuals	PRJNA305413	Gerrard et al. 2017	12	SRR2982097, SRR2982098, SRR2982099, SRR2982100, SRR2982101, SRR2982102, SRR2982103, SRR2982104, SRR2982105, SRR2982106, SRR2982107, SRR2982108
A. jamaicensis	Transcriptome sequencing for gene annotation of Artibeus jamaicensis genome	spleen sample	-	PRJNA61227	Ti et al. 2012	1	SRR539297
A. jamaicensis	Tissue sequencing: spleen	spleen	-	unpublished	generated for this study	1	LS060_sp
C. perspicillata	Tissue sequencing: spleen	spleen	-	unpublished	generated for this study	1	LS070_sp
C. perspicillata	Tissue sequencing: gallbladder, intestine, kidney, liver, pancreas, spleen, stomach, tongue	multiple tissues	-	unpublished	provided by JP	8	Cape_ga1, Cape_in1, Cape_ki1, Cape_li1, Cape_pa1, Cape_sp1, Cape_st1, Cape_to1

D. rotundus	Tissue sequencing: gallbladder, intestine, kidney, liver, pancreas, spleen, stomach, tongue	multiple tissues	-	unpublished	provided by JP	8	Dero_ga1, Dero_in1, Dero_ki1, Dero_li1, Dero_pa1, Dero_sp1, Dero_st1, Dero_to1
D. rotundus	Tissue sequencing: spleen	spleen	-	unpublished	generated for this study	1	LS101_sp
P. discolor	Tissue sequencing: spleen	spleen	-	unpublished	generated for this study	1	LS110_sp
T. saurophila	Tissue sequencing: intestine, kidney, liver, pancreas, spleen, stomach, tongue	multiple tissues	-	unpublished	provided by JP	7	Tosa_in1, Tosa_ki1, Tosa_li1, Tosa_pa1, Tosa_sp1, Tosa_st1, Tosa_to1
E. helvum	Transcriptome analysis of human, pig and bat cell lines infected with Zaire ebolavirus	cell culture	time course after inoculation	PRJNA305831	Wynne et al. 2017	9	SRR6134360, SRR6134358, SRR6134357, SRR6134354, SRR6134353, SRR6134352, SRR6134351, SRR6134348, SRR6134344
E. spelaea	Heat-shock protein expression in bats	liver, lung, spleen	-	PRJNA530519	Chionh et al. 2019	6	SRR8836180, SRR8836181, SRR8836182, SRR8836183, SRR8836184, SRR8836185
P. alecto	Transcriptional responses to infection by HeV of human and bat cells	cell culture	-	PRJNA255833	Wynne et al. 2014	3	SRR1524840, SRR1531161, SRR1531544
P. alecto	Transcriptional response to type I IFN in cells from a broad range of mammalian species	cell culture	time course after treatment	PRJNA397372	Cruz-Rivera et al. 2018	15	SRR5904911, SRR5904912, SRR5904913, SRR5904914, SRR5904915, SRR5904916, SRR5904917, SRR5904918, SRR5904919, SRR5904920, SRR5904921, SRR5904922, SRR5904923, SRR5904924, SRR5904925
R. aegyptiacus	De-novo transcriptome characterisation	mutiple tissues	-	PRJNA300284	Lee et al. 2015	12	SRR2913598, SRR2914059, SRR2914282, SRR2914295, SRR2914359, SRR2914360, SRR2914366, SRR2914368, SRR2914369, SRR2914370, SRR2914371, SRR2914372
R. aegyptiacus	Transcriptional responses to Ebola and Marburg virus infection in bat and human cells	cell culture	-	PRJNA429171	Hölzer et al. 2016	9	SRR6453216, SRR6453215, SRR6453214, SRR6453213, SRR6453212, SRR6453211, SRR6453210, SRR6453209, SRR6453208
R. aegyptiacus	Transcriptional response of R. aegyptiacus cells during infection of Sendai Cantell virus	cell culture	time course after inoculation	PRJNA429211	Pavlovich et al. 2018	17	SRR7609218, SRR7609219, SRR7609220, SRR7609221, SRR7609222, SRR7609223, SRR7609224, SRR7609225, SRR7609226, SRR7609227, SRR7609228, SRR7609229, SRR7609230, SRR7609231, SRR7609232, SRR7609233, SRR7609234
R. aegyptiacus	Tissue sequencing: spleen	spleen	-	unpublished	generated for this study	4	CPN003_sp, CPN005_sp, CPN008_sp, CPN009_sp
R. ferrumequinum	Differential expression in R. ferrumequinum liver during torpid vs active state	liver	-	PRJNA298663	-	2	SRR2754983, SRR2757329
R. ferrumequinum	Transcriptome sequencing investigating Horseshoe bat introgression	pooled muscle, heart, liver	-	PRJNA369304	Mao et al. 2017	1	SRR5219076
R. pearsonii	Transcriptome sequencing investigating Horseshoe bat introgression	pooled muscle, heart, liver	-	PRJNA369304	Mao et al. 2017	1	SRR5219064

E. fuscus	Transcriptome sequencing of cell cultures investigating DNA repair mechanisms across mammal species	cell culture	-	PRJNA343174	Ma et al. 2016	1	SRR4249968
M. lucifugus	Transcriptome sequencing of cell cultures investigating DNA repair mechanisms across mammal species	cell culture	-	PRJNA343174	Ma et al. 2016	2	SRR4249979, SRR4249988

D.2 Supplementary figures

Figure D.1: Heatmap showing log normalised read counts for the 38 EVE-derived transcripts found to be differentially expressed within at least one tissue type between control and infected groups in the *A. jamaicensis*–TCRV experiment (Gerrard et al. 2017).

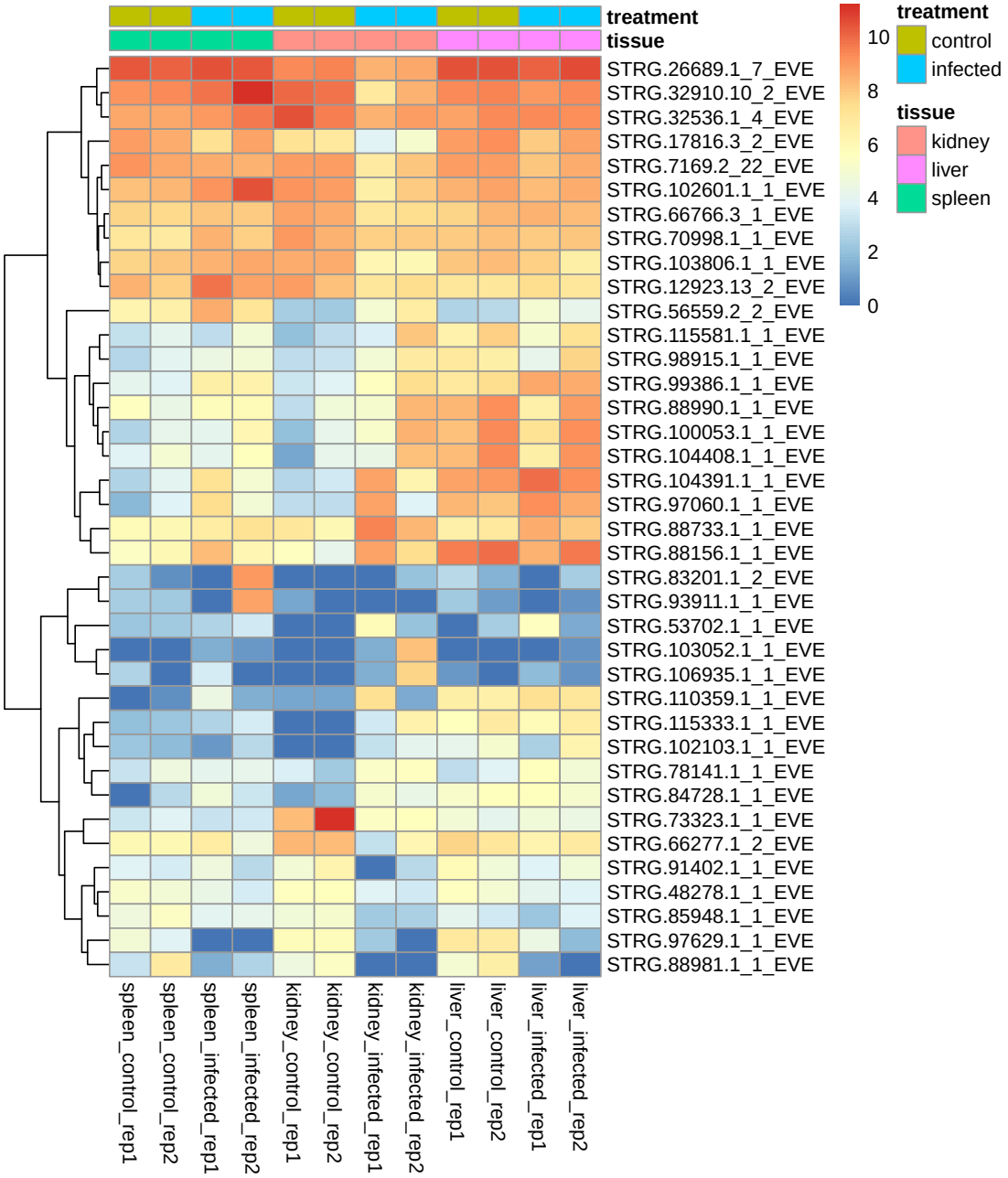


Figure D.2: Heatmap showing log normalised read counts for the 48 EVE-derived transcripts found to be differentially expressed between the 0hr timepoint and any following timepoint in the *P. alecto*–IFNα experiment(Cruz-Rivera et al. 2018).

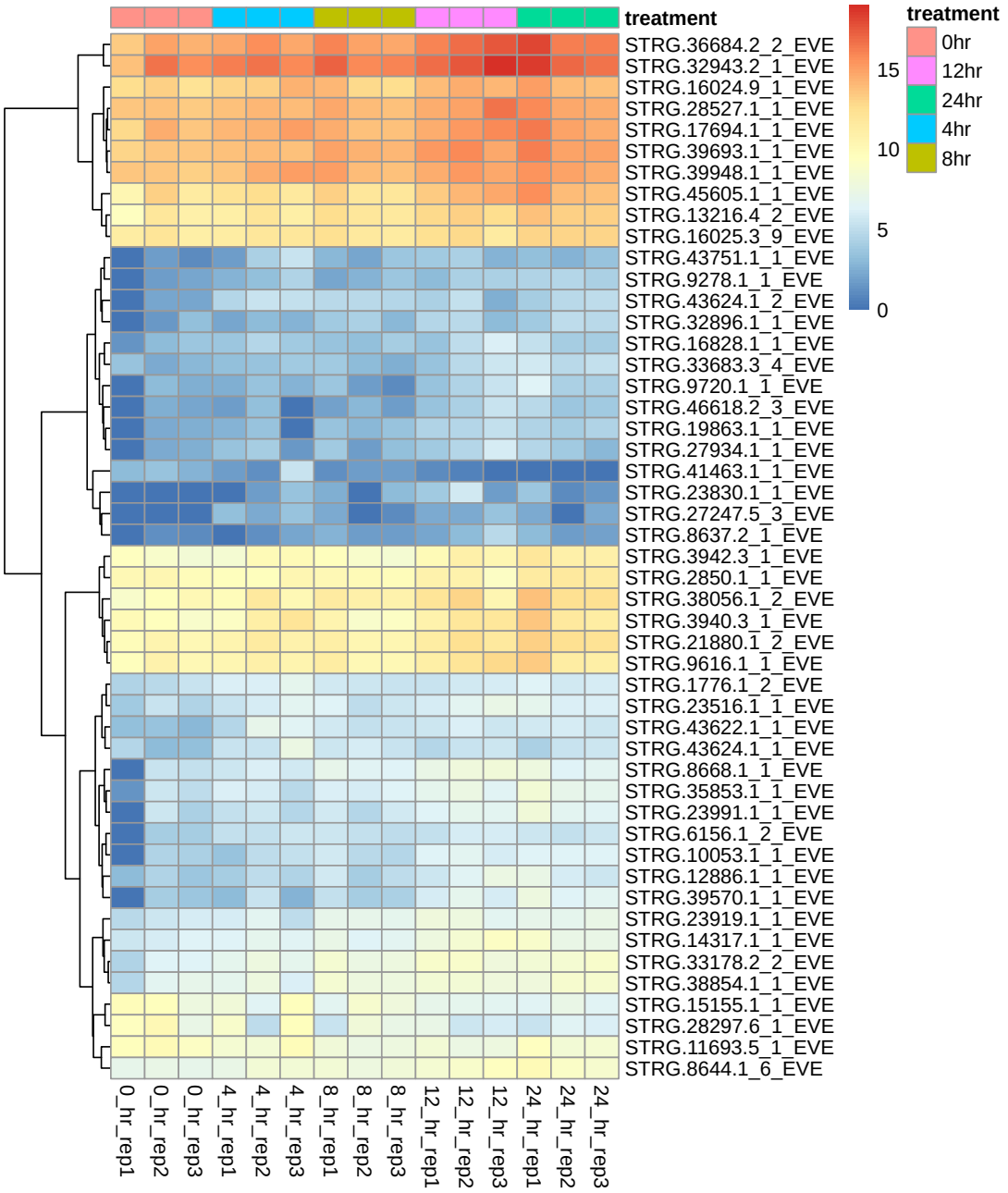
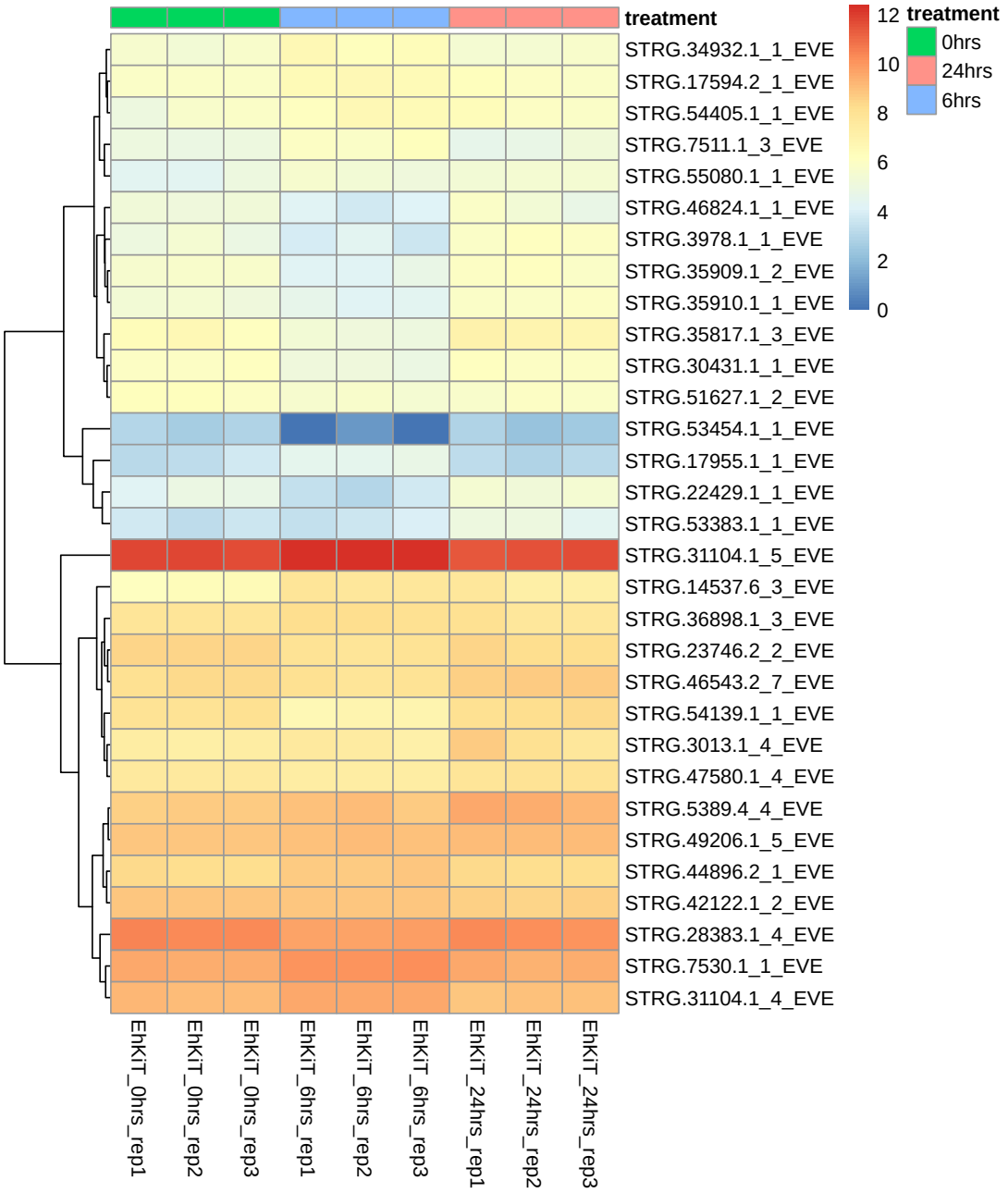


Figure D.3: Heatmap showing log normalised read counts for the 30 EVE-derived transcripts found to be differentially expressed between the 0hr timepoint and either of the following timepoints in the *E. helvum*–EBOV experiment (Wynne et al. 2017).



D.3 Supplementary methods

D.3.1 Genome assemblies used in Chapter 5

Accessions of the 35 genome assemblies analysed in Chapter 5: *Anoura caudifer* (GCA004027475.1), *Artibeus jamaicensis* (GCA004027435.1), *Carollia perspicillata* (GCA004027735.1), *Centurio senex* (S.J. Rossiter unpublished data), *Craseonycteris thonglongyai* (GCA004027555.1), *Desmodus rotundus* (GCA002940915.2), *Eidolon helvum* (GCA000465285.1), *Eonycteris spelaea* (GCA003508835.1), *Eptesicus fuscus* (GCA000308155.1), *Hipposideros armiger* (GCA001890085.1), *Hipposideros galeritus* (GCA004027415.1), *Lasiurus borealis* (GCA004026805.1), *Lionycteris spurrelli* (S.J. Rossiter unpublished data), *Macroglossus sobrinus* (GCA004027375.1), *Megaderma lyra* (GCA004026885.1), *Micronycteris hirsuta* (GCA004026765.1), *Miniopterus natalensis* (GCA001595765.1), *Miniopterus schreibersii* (GCA004026525.1), *Mormoops blainvillei* (GCA004026545.1), *Murina aurata* (GCA004026665.1), *Myotis lucifugus* (GCA000147115.1), *Myotis myotis* (GCA004026985.1), *Myotis brandtii* (GCA000412655.1), *Myotis davidii* (GCA000327345.1), *Noctilio leporinus* (GCA004026585.1), *Phyllostomus discolor* (GCA004126475.1), *Pipistrellus pipistrellus* (GCA004026625.1), *Pteronotus parnellii* (GCA000465405.1), *Pteropus vampyrus* (GCA000151845.2), *Pteropus alecto* (GCA000325575.1), *Rhinolophus ferrumequinum* (GCA004115265.1), *Rhinolophus pearsonii* (S.J. Rossiter unpublished data), *Rousettus aegyptiacus* (GCA004024865.1), *Tadarida brasiliensis* (GCA004025005.1), *Tonatia saurophila* (GCA004024845.1).