Paleovirology:

Connecting recent and ancient viral evolution



Pakorn Aiewsakun

Balliol College

University of Oxford

A thesis submitted for the degree of

Doctor of Philosophy

2016

Pakorn Aiewsakun
Balliol College

Doctor of Philosophy
Trinity Term, 2016

Abstract

Paleovirology: Connecting recent and ancient viral evolution

Endogenous viral elements, or *viral genomic fossils*, have proven extremely valuable in the study of the macroevolution of viruses, providing important, and otherwise unobtainable, insights into the ancient origin of viruses, and how their ancestors might have co-evolved with their hosts in the distant past. This type of investigation falls within the realm of paleovirology—the study of ancient viruses. Investigations of extant viruses and paleovirological analyses, however, often give conflicting results, especially those concerning viral evolutionary rates and timescales. Reconciling these two types of analyses is a necessary step towards a better understanding of the overall long-term evolutionary dynamics of viruses. The main study system of this thesis is foamy viruses (FVs). FVs are characterised by their stable co-speciation history with their hosts, allowing their evolutionary dynamics to be modelled and investigated over various timescales. This unique evolutionary feature makes FVs one of the best subjects for connecting recent and ancient viral evolution. The work here reports the discovery of several endogenous mammalian FVs, and examines how mammalian FVs co-evolve with their hosts. Analyses reveal a co-diversifying history of the two that could be dated back to the basal radiation of eutherians more than 100 million years ago. However, a small number of ancient FV cross-species transmissions could still be found, mostly involving New World monkey FVs. Based on this extended FV-mammal co-speciation pattern, this thesis investigates the long-term evolutionary rate dynamics of FVs, and shows that the rate estimates of FV evolution appear to decrease continuously as the rate measurement timescale increases, following a power-law decay function. The work presented here also shows that this so-called '*time-dependent rate phenomenon*' is in fact a pervasive evolutionary feature of all viruses, and surprisingly, the rate estimates of evolution of all viruses seem to decay at the same speed, decreasing by approximately half for every 3-fold increase in the measurement timescale. Based on this power-law rate-decay pattern, we could infer evolutionary timescales of modern-day lentiviruses that are consistent with paleovirological analyses for the first time. Finally, this thesis reports the discovery of basal FV-like endogenous retroviruses (FLERVs) in amphibian and fish genomes. Phylogenetic analyses reveal that the progenitors of ray-finned fish FLERVs co-diversify broadly with their fish hosts, but also suggest that there might have been several ancient viral cross-class transmissions, involving lobe-finned fish, shark, and frog FLERVs. Again, by using the power-law rate-decay model, analyses in this thesis suggest that this major retroviral clade has an ancient Ordovician marine origin, originating together with their jawed vertebrate hosts more than 450 million years ago. This finding implies that the origin of retroviruses as a whole must be in the early Paleozoic Era, if not earlier. The results presented here bridge ancient and recent viral evolution.

Acknowledgements

This thesis could not possibly exist without the support of Dr. Aris Katzourakis. I am really grateful for his insightful, patient, and encouraging supervision of my work over the past four years. It is my pleasure to work with him, and his training makes me more professional and independent. I also would like to thank William M Switzer, who provided some of the crucial data for my work. The financial support from the zoology department, University of Oxford, and the DPST (development and promotion of the gifted in science and technology) project, the Royal Thai government, has put me in this privileged position, and for this, I am very thankful.

The ideas presented here were greatly enriched by discussions and interactions with many people, especially those in Paleovirology group, University of Oxford, past and present, including Amr Aswad, Patrick Gemmell, Nico Kist, Thanos Mourikis, Fabricia Nascimento, Cindy Santander, Audrey Lin, Samuel Campbell, Emilia Skirmuntt, and Charlie Ince. My thesis also greatly benefited from the comments of various anonymous reviewers.

Finally, I would like to thank my friends and family. Eli, Keshav, Leonie, Hannah, and Luke have made living in Oxford a very joyful experience. My family members and friends in Thailand have also been very supportive and encouraging during my stay overseas. Thank you guys!

Table of Contents

ABSTRACT ..... **I**
ACKNOWLEDGEMENTS ..... **II**
TABLE OF CONTENTS ..... **III**

CHAPTER ZERO: INTRODUCTION
0.1 ABSTRACT **2**
0.2 VIRAL CLASSIFICATION **3**
0.3 RETROVIRUS **4**
0.4 PALEOVIROLOGY **8**
0.5 EVE DISCOVERY **11**
0.6 EVE DISCOVERY HAS VASTLY EXPANDED VIRAL HOST RANGE **15**
0.7 DATING EVES AND GENERATING ROBUST VIRAL EVOLUTIONARY TIMESCALES **18**
0.8 EVES REVEAL ANCIENT VIRAL-HOST INTERACTION AND TRANSMISSION ROUTES **22**
0.9 DISCREPANCY BETWEEN VIRAL EVOLUTIONARY TIMESCALES AND RATES INFERRED FROM EVES AND MODERN-DAY VIRAL SEQUENCES **25**
0.10 THESIS OVERVIEW **32**
0.11 REFERENCES **35**

CHAPTER ONE: ANCIENT CO-EVOLUTION OF FOAMY VIRUSES AND THEIR MAMMALIAN HOSTS
1.1 ABSTRACT **58**
1.2 INTRODUCTION **59**
1.3 RESULTS **64**
1.3.1 CHARACTERISATION OF THE COMPLETE GENOME OF THE EXTANT PSFVGAL FROM A GALAGO 64
1.3.2 DISCOVERY AND CHARACTERISATION OF PSFVAYE AND CHREFV 67
1.3.3 PHYLOGENETIC RECONSTRUCTION AND RECONCILIATION ANALYSIS 72
1.3.4 THE EFFECTS OF NEUTRAL GENETIC CHANGES ON PHYLOGENETIC INFERENCE 77
1.3.5 FV-HOST DIVERGENCE CORRELATION ANALYSIS 77
1.4 DISCUSSION **79**
1.5 MATERIALS AND METHODS **87**
1.5.1 DISCOVERY AND CHARACTERISATION OF CHREFV AND PSFVAYE 87

1.5.2 ESTIMATING THE INTEGRATION DATES OF PSFVAYE AND CHREFV WITH MONTE CARLO SIMULATION 87
1.5.3 PHYLOGENETIC RECONSTRUCTION 89
1.5.4 PHYLOGENETIC RECONCILIATION ANALYSIS 90
1.5.5 APPROXIMATE UNBIASED (AU) TEST 90
1.5.6 EXAMINATION OF THE EFFECTS OF HOST NEUTRAL GENETIC CHANGES ON PHYLOGENETIC INFERENCE 91
1.5.7 FV-HOST DIVERGENCE CORRELATION ANALYSIS 91
1.6 REFERENCES 93

CHAPTER TWO: CODIVERGENCE AND CROSS-SPECIES TRANSMISSIONS OF NEW WORLD MONKEY FOAMY VIRUSES
2.1 ABSTRACT 103
2.2 INTRODUCTION 104
2.3 RESULTS 105
2.3.1 PHYLOGENETIC RECONCILIATION ANALYSIS 105
2.3.2 EVALUATING AND REFINING THE CO-SPECIATION MODEL 111
2.3.3 NWM-FV DIVERGENCE CORRELATION ANALYSIS 115
2.4 DISCUSSION 116
2.5 CONCLUSIONS 117
2.6 MATERIALS AND METHODS 117
2.6.1 PHYLOGENETIC RECONSTRUCTION 117
2.6.2 PHYLOGENETIC RECONCILIATION ANALYSIS 118
2.6.3 ESTIMATING THE TIMESCALES OF NWM FVS 119
2.6.4 NWM-FV DIVERGENCE CORRELATION ANALYSIS 120
2.7 REFERENCES 121

CHAPTER THREE: TIME DEPENDENCY OF FOAMY VIRUS EVOLUTIONARY RATE ESTIMATES
3.1 ABSTRACT 127
3.2 INTRODUCTION 128
3.3 RESULTS 130
3.3.1 FV NUCLEOTIDE SUBSTITUTION RATE ESTIMATE DECREASES WITH MEASUREMENT TIMESCALE 130
3.3.2 EMPIRICAL DESCRIPTION OF FV NUCLEOTIDE SUBSTITUTION RATE DECAY 133
3.3.3 LEAVE-ONE-OUT CROSS VALIDATION ANALYSES 138

3.3.4 Examining the performance of currently available molecular clocks in TDRP correction 141
3.3.4.1 Aggregated-node calibration scheme 143
3.3.4.2 Dispersed-node calibration scheme 144
3.3.5 Inferring missing FV divergence dates by using the PL model 145
3.4 DISCUSSION 146
3.4.1 What are the causes of the TDRP? 146
3.4.2 The TDRP as a possible explanation of the short-term and long-term rate discrepancy 147
3.4.3 Implications of the TDRP 148
3.4.4 Possible evolutionary rate dynamics heterogeneity among viral lineages 150
3.5 CONCLUSIONS 151
3.6 MATERIALS AND METHODS 151
3.6.1 Phylogenetic reconstruction 151
3.6.2 Inferring FV node-to-tip total per-lineage substitutions (s estimates), evolutionary timescales (t estimates), and node-to-tip average evolutionary rates (r estimates) over various time frames, and assessing the correlation between r and t estimates 153
3.6.3 Describing the temporal dynamics of r 155
3.6.4 Validating the performance of the models under the context of t inference by using the leave-one-out cross validation (LOOCV) technique 155
3.6.5 Examining the effect of the TDRP on t inference 156
3.6.6 Inferring missing FV divergence dates by using the PL model 157
3.7 REFERENCES 158

CHAPTER FOUR: TIME-DEPENDENT RATE PHENOMENON IN VIRUSES

4.1 ABSTRACT 168
4.2 INTRODUCTION 169
4.3 RESULTS 171
4.3.1 TDRP among Baltimore classification viral groups 171
4.3.2 TDRP within viral genera 175
4.3.3 Sensitivity analyses 177
4.3.4 Short-term TDRP 180
4.3.5 Case study: reestimating the time scale of evolution of lentiviruses 181
4.4 DISCUSSION 185
4.5 MATERIALS AND METHODS 190

4.5.1 DATA COLLECTION AND CONSTRUCTION OF PHYLOGENETICALLY INDEPENDENT DATA SETS OF RATE ESTIMATES 190
4.5.2 TDRP ANALYSES AT THE LEVEL OF BALTIMORE CLASSIFICATION VIRAL GROUPS 192
4.5.3 TDRP ANALYSES AT THE LEVEL OF VIRAL GENERA 193
4.5.4 SENSITIVITY ANALYSES 193
4.5.5 SHORT-TERM TDRP ANALYSES 194
4.5.6 ESTIMATING LENTIVIRAL EVOLUTIONARY TIMESCALES BY USING OUR TDRP MODEL 195
4.6 REFERENCES 197

CHAPTER FIVE: MARINE ORIGIN OF RETROVIRUSES IN THE EARLY PALAEOZOIC ERA
5.1 ABSTRACT 203
5.2 INTRODUCTION 204
5.3 RESULTS 206
5.3.1 DISCOVERY OF FLERVS IN AMPHIBIAN AND FISH GENOMES 206
5.3.2 CHARACTERISATION OF FULL LENGTH FLERVS 209
5.3.3 CHARACTERISATION OF FRAGMENTED FLERVS 214
5.3.4 PHYLOGENETIC ANALYSES 219
5.3.5 EVOLUTIONARY TIMESCALE OF FLERVS 222
5.4 DISCUSSION 227
5.5 MATERIALS AND METHODS 231
5.5.1 SCREENING FOR FV-LIKE SEQUENCES 231
5.5.2 CHARACTERISATION OF NOVEL FV-LIKE ELEMENTS 233
5.5.3 PHYLOGENETIC ANALYSES 234
5.5.4 CO-SPECIATION ANALYSES 235
5.5.5 ESTIMATING EVOLUTIONARY TIMESCALE OF FLERVS 236
5.5.6 PHYLOGENETIC ANALYSES BY USING ML FLERV ANCESTRAL SEQUENCES 236
5.6 REFERENCES 238

CHAPTER SIX: CONCLUSIONS AND FINAL THOUGHTS
6.1 CONCLUSIONS AND FINAL THOUGHTS 243
6.2 REFERENCES 251

APPENDIX
APPENDIX I 255
APPENDIX II 273

Chapter Zero: Introduction

Except from sections 0.2, 0.3, and 0.10, this chapter is modified from the following published paper:

Aiewsakun P, Katzourakis A. 2015. Endogenous viruses: Connecting recent and ancient viral evolution. Virology **479–480**:26–37.

PA wrote the paper. AK provided editorial guidance and supervisory advice.

0.1 Abstract

Viral molecular data have been growing at a phenomenal rate, and coupled with their high rates of evolution, this allows recent histories of viruses to be reconstructed in great detail. However, these rapid rates of evolution mean that the decay rate of evolutionary signal within viral molecular data is also very high, impeding reconstruction of the deep history of viruses from their molecular sequences. Furthermore, inferring complex histories of viral transmissions can be challenging, especially where extinct viral lineages are concerned. As a result, our knowledge of the long-term history of viral evolution has been highly fragmented and incomplete. These problems have been partially overcome by the discovery of viruses that integrated into host germ-line cells and have become a permanent part of host genomes, known as endogenous viral elements. Some endogenous viruses are derived from ancient relatives of modern-day viruses, allowing us to better examine their ancient host range, geographical distribution, and transmission routes, as well as their long-term evolutionary rate dynamics. This chapter introduces a number of concepts that underpin the content of this thesis, including viral classification and biology, as well as how paleovirology—the study of ancient, possibly extinct, viruses—can be achieved. The results from paleovirological studies have greatly improved and transformed our understanding of how viruses evolve in long term, connecting recent and ancient viral evolution.

0.2 Viral Classification

Viruses are arguably the most abundant biological entity on this planet (1). It has been estimated that the abundance of viruses exceeds that of bacteria by approximately 15-fold in the oceans (2) and between 5- and 1000-fold in the soil (3). They occupy a wide range of ecosystems, including soils (3–5), marine (2, 6–10), and fresh water (11–14), as well as extreme environments such as deserts (15), alkaline hypersaline lakes (16), hot acidic lakes (17), hot springs (18), and high-pressure deep-sea hydrothermal vents (19).

Not only are viruses extremely diverse in their ecology, but are also highly diverse in their morphology and biology. All viral particles contain nucleic acid, functioning as the viral genome, surrounded by a shell of proteins. Some viruses also possess a host-derived phospholipid bilayer enveloping the protein particle, playing an important role in cell attachment and entry (20) as well as host immune system evasion (21). Their genomes can either be single-stranded (ss) or double-stranded (ds), deoxyribonucleic acid (DNA) or ribonucleic acid (RNA). For ssRNA viruses, their genomes are either positive (+) strands, which can be directly translated into functional proteins, or negative (–) strands, which need to be transcribed into complementary strands before protein translation. Moreover, some viruses replicate via the reverse transcription process, where RNA is *retro*-transcribed into complementary DNA by an enzyme called reverse transcriptase (RT). These viruses are known as retroviruses. Based on these viral genomic features and replication strategies, David Baltimore invented the 'Baltimore classification system' in 1971 (22), cataloguing viruses into six groups:

Group I: dsDNA viruses

Group II: ssDNA viruses

Group III: dsRNA viruses

Group IV: (+)ssRNA viruses

Group V: (–)ssRNA viruses

Group VI: RT-ssRNA viruses

A decade after the invention of the Baltimore classification system, Jesse Summers and William S. Mason (23) discovered a virus that could not be accommodated by this system, namely duck hepatitis B virus. Duck hepatitis B virus has a dsDNA genome and employs a reverse transcription pathway for their replication (genome replication: dsDNA genome **—***transcription***—>** (+)ssRNA **—***reverse transcription***—>** (–)ssDNA intermediate **—***synthesis of (+) strand***—>** dsDNA genome). As a result, this virus could not be classified as either a group-I virus or a group-VI virus in the Baltimore classification system. Shortly thereafter, another virus with the same characteristic features was identified, namely cauliflower mosaic virus (24, 25). To accommodate these findings, Temin (26) grouped them together and classified them as 'pararetroviruses'. Today, these group of viruses is known as 'Group VII' of the Baltimore classification system.

0.3 Retrovirus

The main subject of this thesis is retroviruses—Group VI of the Baltimore classification system. According to the International Committee on Taxonomy of Viruses (27), this family (*Retroviridea*) has two subfamilies, namely *Orthoretrovirinae* and *Spumaretrovirinae*. The subfamily *Orthoretrovirinae* comprises six genera, including *Alpharetrovirus*, *Betaretrovirus*, *Gammaretrovirus*, *Deltaretrovirus*, *Epsilonretrovirus*, and *Lentivirus*. On the other hand, the subfamily *Spumaretrovirinae* only contains one genus, namely *Spumavirus*. Spumavirus, also known as foamy virus, is the primary focus of this thesis. This section provides some general background on retroviral genomic organisation and life cycle, with the emphasis on foamy viruses. It is worth noting that the life cycle of foamy virus is very unique, and different from those of other retroviruses in many aspects, and these differences are highlighted.

A retroviral particle is typically 80–100 nanometres in diameter, containing two identical linear (+)ssRNA molecules, functioning as its genome (28). Unlike other retroviruses however, foamy virus particles frequently (~10–20%) contain full-length viral genomic DNA (29–31). The viral genome is encapsulated within a protein shell, which is in turn enveloped by a membrane of

host-derived phospholipid bilayer (**Figure 0.1A**). The genome of retroviruses is typically ~7–10 kilobases (kb) in length, with that of foamy virus being the longest, ~12 kb (28). The retroviral genome consists of at least three genes, namely *gag*, *pol*, and *env* (28). The *gag* gene encodes internal structural proteins, including the matrix (MA), capsid (CA), and nucleocapsid (NC) proteins, forming the viral protein core (28). The *env* gene encodes the surface glycoproteins (SU), and transmembrane proteins (TM), linked together by disulphide bonds to form glycoproteins that display on the retroviral envelope membrane (28), responsible for cell attachment and entry (20) as well as host immune system evasion (21). The *pol* gene encodes two enzymes, namely reverse transcriptase (RT), which possesses (RNA- and DNA-dependent) DNA polymerase and ribonuclease H activities, and integrase (IN) (28). RT and IN are responsible for the synthesis of viral DNA, and the integration of the viral DNA into the host chromosome, respectively (28). Another protein coding domain that all retroviruses have is *pro,* coding for the protease (PR) enzyme locating between the *gag* and *pol* gene (28). In some viruses, for example foamy virus, PR coding domain is a part of the *pol* gene however (32). Some complex retroviruses possess accessory genes, situating between the *pol* and *env* gene, and/or after the *env* gene (33). Again, foamy virus is such virus, possessing two additional protein coding domains after the *env* gene, namely *bel1* and *bel2* (32, 33). These accessory genes have a variety of biological functions, including promoting viral infectivity (34, 35), regulating and coordinating viral gene expression (35–37), as well as modulating the activity of host immune system (38–42).

Flanking these genes are non-coding sequences, playing important roles in replication, gene expression, and genome packaging (28). The non-coding sequence on the 5’ end comprises four sub-regions, including (i) a repeated sequence region (R), (ii) a region that is unique to the 5’ end (U5), (iii) a primer binding site (PBS), which is complementary to the 3’ end of tRNA, and (iv) a leader region (L), containing the viral genome packaging signal (28). The 3’ end has 3 non-protein coding domains, including (i) a polypurine tract (PPT), which is a primer for (+)strand DNA synthesis during reverse transcription, (ii) a region unique to the 3’ end (U3), and (iii) a repeated sequence region (R), similar to that on the 5’ end (28).

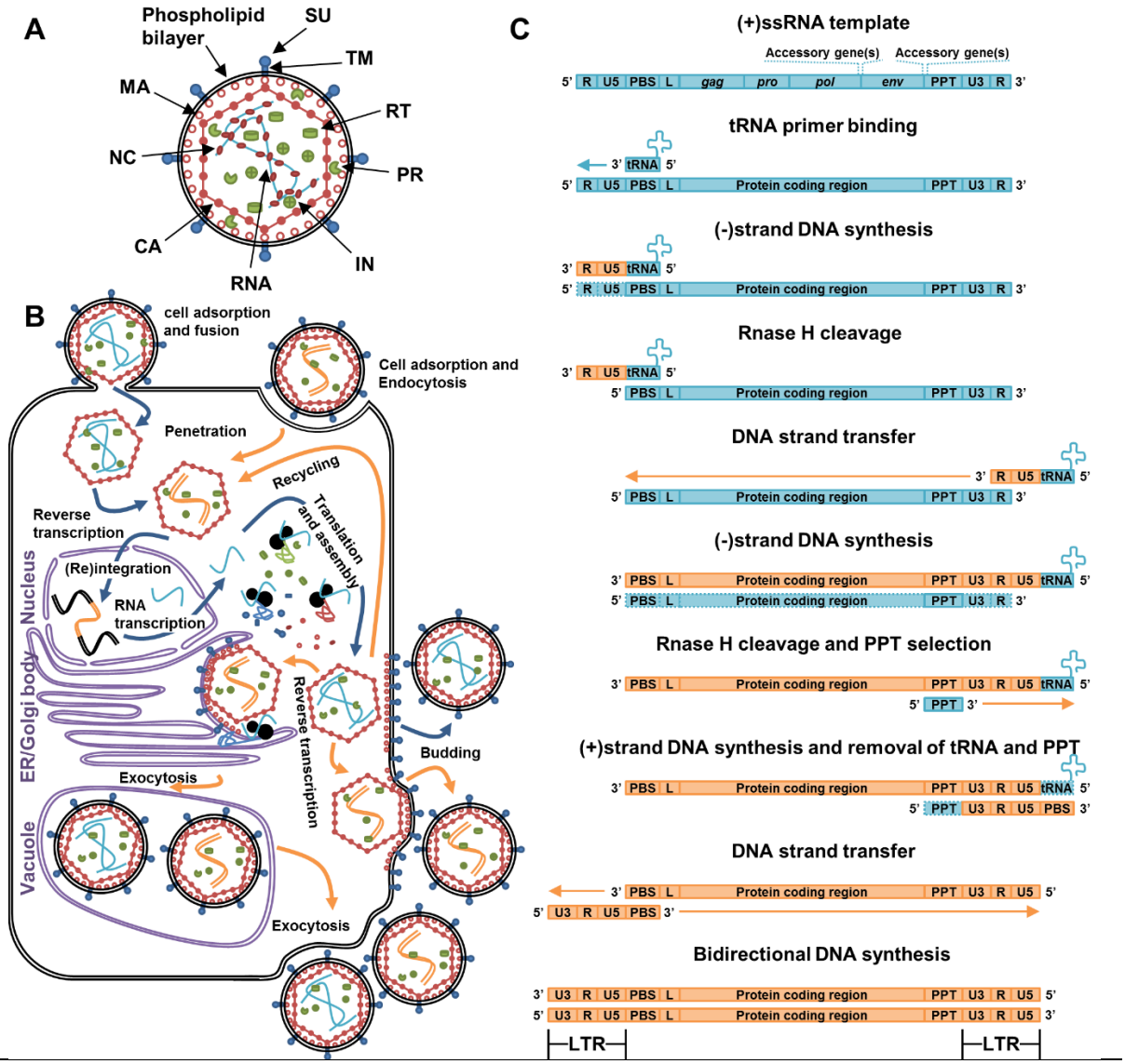


Figure 0.1 Retroviral particle, replication cycle, and reverse transcription process. (**A**) Schematic cross section through a retroviral particle. A retroviral particle typically contains two positive-sense single-stranded ribonucleic acid (RNA) molecules, functioning as its genome (light blue). The viral envelope (black circles) displays glycoproteins (blue), which are made up of two components, namely the transmembrane protein (TM) and surface glycoprotein (SU), linked together by disulphide bonds. The internal protein structure (red) is made up of three different types of proteins, including the matrix (MA), capsid (CA), and nucleocapsid (NC) proteins. The viral particle also contains several enzymatic proteins (green), including reverse transcriptase (RT), integrase (IN), and protease (PR). (**B**) Retroviral replication cycle. The replication pathway of a typical retrovirus is labelled by blue arrows, whereas that of a foamy virus is labelled by orange arrows. Note that, unlike typical retroviruses, foamy virus particles can have a double-stranded deoxyribonucleic acid (DNA) molecule as its genome (orange). The host organelles are in purple, including nucleus, Golgi body, endoplasmic reticulum (ER), and vacuole. (**C**) Schematic of the reverse transcription process. The top panel shows the organisation of a typical retroviral genome, containing at least four protein coding genes, including *gag* (which encodes MA, CA, and NC proteins), *pro* (which encodes PR proteins), *pol* (which encodes RT and IN proteins), and *env* (which encodes SU and TM proteins). Some retroviruses might have *pro* as a part of *pol*, however. Furthermore, some viruses might also have accessory genes located between *pol* and *env*, and/or after *env*. Flanking these protein coding genes are non-coding sequences. On the 5' end, there are four regions, including a repeated sequence region (R), a non-coding sequence region that is unique to the 5' end (U5), a primer biding site (PBS), and a leader sequence (L). On the 3' end, there are three regions. These include a polypurine track (PPT), a non-coding sequence region that is unique to the 3' end (U3), and a repeated sequence region (R) that is similar to the R region on the 5' end. Reverse transcription is a process that an RNA molecule (blue) is transcribed into a DNA molecule (orange). The blue boxes with dotted borders indicate that RNA is being degraded by ribonuclease H activities. By the end of the process, the DNA product contains long terminal repeats (LTRs), locating at both ends of the molecule.

The life cycle of a retrovirus is very complex, as shown in **Figure 0.1B** (28, 43). First, the viral particle recognises and binds specific receptors on the cell membrane. This typically triggers virus-host cell membrane fusion, releasing the viral core, i.e. capsid, into the host cell (44). Again, foamy virus is the exception to this, entering host cells via endocytosis, and the capsid is released to the cytoplasm via a pH-dependent membrane fusion process (45). Following the penetration, the (+)ssRNA genome is reverse transcribed into a dsDNA molecule, taking place within the cytoplasm inside the capsid (46–48). As shown in **Figure 0.1C**, the process of reverse transcription involves two 5'–to–3' DNA strand transfers, resulting in a duplication of terminal sequences, known as long terminal repeats (LTRs), locating at both ends of the dsDNA product (48). In the case of foamy virus, reverse transcription carries out in the producer cells rather than in newly infected cells (29–31), and thus a portion of their viral particles already contain genomic DNA; as a result, reverse transcription is not necessary at this step for foamy viruses. Disassembly of the capsid then proceeds, allowing the viral DNA to integrate with the host genomic DNA (49).

The integrated retrovirus is then transcribed into RNA molecules by the host machinery. Full-length RNA molecules serve as the next-generation retroviral genomes (28). Spliced RNA products, on the other hand, will be translated by ribosomes in the cytoplasm to produce viral accessory, capsid, and enzymatic proteins, as well as glycoproteins (28). For most retroviruses, their genomic RNAs and proteins assemble at the cell periphery to form viral particles, and are released from the cell via budding (50, 51). Foamy virus, in contrast, has a more complex route of particle releasing. Unlike other retroviruses, their glycoproteins are translated at the rough endoplasmic reticulum (ER) and targeted to the secretory pathway (43). Furthermore, while other retroviruses reverse transcribe their genomes after cell penetration, a foamy virus reverse transcribes its genome during or soon after the completion of the viral core assembly (29–31). Such particles, containing the DNA genomic molecule, might even be recycled, reintegrating into the host genome (52, 53). Foamy virus budding occurs predominantly at intracellular membranes (e.g. ER and Golgi body), and they might exit the

host cells via exocytosis (43). Foamy virus budding at the host cell membrane also occurs, but to a lesser extent (43).

This thesis investigates how retroviruses evolved and interacted with their hosts in deep time, with a particular focus on the long-term evolutionary history and dynamics of foamy viruses. This thesis also examines the long-term evolutionary rate dynamics of viruses across almost all genome types and replication strategies, with the aim to find a general evolutionary feature that is common to all viruses. These questions fall within the realm of paleovirology—the study of ancient, possibly extinct, viruses.

0.4 Paleovirology

Paleovirology is a recently established field of biology that involves studying ancient viruses, including, but not limited to, the study of molecular features of extinct viruses, the effects of pre-historic viral-host gene transfers, and the natural history of viruses over geological timescale (54). An ideal approach to answering paleovirological questions is to examine ancient viruses directly. However, this approach to studying viral evolution, commonly referred to as 'direct paleovirology', seemed to be impossible due to the simple fact that viruses are too small and too fragile to leave any observable geographical imprints and lack the ability to leave any geochemical signatures or to be fossilised. Indeed, physical fossils of viruses simply do not exist (54).

Traditionally, the deep histories of viruses were, and still are, investigated via analyses of the interactions between extant viruses and their hosts—known as 'indirect paleovirology'. For example, it has often been speculated that simian immunodeficiency viruses (SIVs) have a very long evolutionary relationship with their hosts based on their asymptomatic infection and wide host range (55–60). Indeed, indirect paleovirological analyses of natural selection footprints on simian retroviral restriction factors, including APOBEG3G (61), SAMHD1 (62, 63), Tetherin (64), and TRIMCyp (65), support this hypothesis, consistently suggesting that SIVs are at least 10 million years (Myr) old. Nevertheless, this specific approach to viral evolutionary

study is often complicated by the fact that viral restriction factors are frequently not specific to just one viral group (66) and that viruses that shaped the evolution of these factors might have already been extinct and/or are unknown (54). As a result, it is often the case that this method can provide us with only a limited and fragmented picture of how viruses evolve in deep time.

Another complication regarding using modern-day viruses to decipher their deep history is that their rates of evolution are extremely high. Viruses are arguably the fastest-evolving biological entity on this planet. Many of them evolve at a rate in the order of 10^{-3} substitutions per site per year (s/n/y) (67–69). This is approximately a million times faster than the rates of evolution of cellular host organisms, typically around 10^{-9} s/n/y (70). This extraordinarily high rate of evolution allows viruses to escape from host immunity and rapidly adapt to a new host when they cross species. This unique evolutionary feature causes viral genomic sequences to rapidly change at the molecular level, accumulating information that can be used to reconstruct their evolution (71). Combined with the large availability of molecular data of some extant viruses, this enables researchers to reconstruct the recent histories of viruses in great detail. For example, a reconstruction of the spread of the global pandemic human immunodeficiency virus (HIV) strain revealed that it originated in Kinshasa, Democratic Republic of Congo, in the 1920s, and shortly thereafter spread via three other highly-populated cities through the transportation network before emerging as a global pandemic (72). Such a detailed epidemiological history can be reconstructed only because HIV evolves extremely rapidly and there are many HIV molecular sequences available.

Although molecular data of extant viruses allow their recent history to be reconstructed to a high degree of resolution, such data are often insufficient to inform us how viruses evolved and interacted with their hosts in deep time. While their rapid rate of evolution causes viral genomic sequences to accumulate evolutionary signal rapidly, sequences can carry only a finite amount of information, and therefore the signal will also be overwritten at a rapid rate. This places a temporal limit on how far back in time we can accurately infer viral evolutionary history from molecular data of modern-day viruses, likely no more than a few million years into the past.

Furthermore, it is also a fact that extant viral sequences cannot shed light on the history of extinct viral relatives. The combination of these effects greatly hinders the investigation of deep viral evolution. Because of this, together with the high extinction rate of viruses and viral sampling biases (e.g. towards human- and livestock-infecting viruses), analyses of contemporary viral sequences are likely to miss many historical viral cross-species transmission events, for example. In order to obtain a more complete picture of how viruses evolved and interacted with their hosts in the distant past, it seems that more than just molecular data of extant viruses is needed.

More recently, researchers have begun to realise that direct paleovirology is actually achievable and not as farfetched as it seemed. Although viruses cannot be physically fossilised in the conventional sense, they occasionally leave long-lasting imprints within their host genomes, known as *endogenous viral elements* (EVEs), which means 'viruses from within'. The discovery of these viral imprints allows us to investigate the past history of viruses in more detail. Some viruses, i.e. retroviruses, enter host chromosomes as an obligate step during their life cycle, and other viruses do so less often by accident (73). EVEs result from the process called endogenisation, in which viral DNA copies integrate into host germ-line chromosomes, and in turn are vertically transmitted from parent to offspring. In cases where EVEs do not impose significant deleterious effects upon the hosts, their frequency might increase within the host population and lead to fixation. Once endogenised, most EVEs retain similarity to their ancestral exogenous virus (literally means 'viruses from the outside') for many millions of years, as they evolve at the host rate of evolution, which is typically several orders of magnitude slower than a virus. These EVEs give researchers unprecedented opportunities to directly observe ancient viruses in their host genomes, and examine how they might have evolved and interacted with these hosts. Direct and indirect paleovirological approaches are independent methods of studying ancient viruses, and hence can be used to validate one another. In cases where both provide congruent results, very robust and detailed viral evolutionary models can be made. The recent availability of a large number of cellular whole genomes and ongoing genomic screening has led to a steady increase in the number of

ancient viral sequences of all types (**Table 0.1**), enabling us to better examine the deep history of all viral groups.

The discovery of EVEs has markedly improved our knowledge of viral natural history; for example, how viruses were distributed across geographical space and among host species millions of years ago. By combining such information with the knowledge of contemporary viral host range and geographical distribution, the history of viral transmission sources and routes can be reconstructed like never before. Furthermore, EVEs also have profoundly impacted our understanding of viral evolutionary timescales and rate dynamics. Analyses of EVEs strongly suggest that all fast-evolving viruses are far older than previously thought and they actually evolve strikingly slowly in the long term. By exploring various detailed examples, this chapter summarises how EVEs have advanced and challenged our understanding of viral evolutionary dynamics, and improved our knowledge of the long-term viral natural history.

0.5 EVE discovery

A common approach researchers use to mine EVEs is to bioinformatically search for genomic elements in cellular organisms that exhibit similarity to known extant viral sequences. Owing to the advances in bioinformatics techniques and genomic sequencing technology, numerous EVEs have been uncovered to date. Retroviruses are the only known group of viruses that enter host chromosomes as an obligate part of their life cycle, and thus are predisposed to become endogenous. Surprisingly, although retroviral germ-line integration is expected to be rare, endogenous retroviruses (ERVs) are very common and make up a large portion of many eukaryotic genomes. For instance, it has been estimated that ERVs form ~7–8% of our genome (74, 75) with >98,000 retroviral fragments (76, 77), comprising >31 families (78). Nevertheless, it has been found that not only can retroviruses become endogenous, but every known group of viruses is capable of endogenisation (**Table 0.1**).

The route to endogenisation of retroviruses is clear due to their replication strategy. However, the means by which other viruses have endogenised are far less obvious. It has been

Table 0.1 Distribution of exogenous and endogenous viruses among various organisms.

Group/type*	Family (group)	Host range	
		Exogenous viruses	**Endogenous viruses (refs.)**
I/dsDNA virus	Herpesviridae	Eutherians, birds, and reptiles	Tarsier (79)
	Polydnaviridae	Wasps	Wasps (80–84)
	Phycodnaviridae	Marine filamentous brown algae	Marine filamentous brown algae (85)
II/ssDNA virus	Parvoviridae	Eutherians, birds, crustaceans and insects	Eutherians, marsupials, birds, fishes, tunicates, crustacean and flatworms (73, 86–90)
	Circoviridae	Eutherians, birds, fishes, and insects	Eutherians, marsupials, reptiles, amphibians, lancelets, insects, crustaceans, nematodes, gastropods, hydrozoans, and giardia (73, 86, 89, 91, 92)
	Geminiviridae	Plants, and fungi	Plants, fungi, and amoeba (92–95)
	Nanoviridae	Plants	Amphibians, crustaceans, molluscs, placozoans, algae, amoeba, diatoms, blastocystis, and giardia (86, 92)
III/dsRNA virus	Reoviridae	Mammals, birds, and insects	Insects (73)
	Partitiviridae	Plants, fungi, and protozoa	Plants, insects, ticks, and amoeba (96, 97)
	Totiviridae	Fungi, and protozoa	Insects, crustaceans, nematodes, fungi, plants, and diatoms (86, 97–100)

Table 0.1 (Continue).

Group/type*	Family (group)	Host range	
		Exogenous viruses	**Endogenous viruses (refs.)**
III/dsRNA virus	Chrysoviridae	Plants, and fungi	Plants (98)
	Endornaviridae	Plants, fungi, and protozoa	Crustaceans, plants, and protists (98)
IV/(+)ssRNA virus	Flaviviridae	Mammals, birds, and insects	Insects (73, 101, 102)
	Potyviridae	Plants	Plants (103)
V/(–)ssRNA virus	Orthomyxoviridae	Mammals, birds, and insects	Insects (73)
	Bunyaviridae	Vertebrates, and insects	Insects, and crustaceans (73, 86, 104)
	Bornaviridae	Mammals, and birds	Mammals, marsupials, reptile, fish, and insects (73, 91, 105–107)
	Filoviridae	Primates, bats, and pigs	Bats, rodents, srhews, tenrecs and marsupials (73, 107–110)
	Rhabdoviridae	Mammals, birds, and insects	Insects, ticks, and plants (73, 96)
	Nyamiviridae	Ticks and birds	Fish, and crustacean (86, 107)
VI/RT-RNA virus	Retroviridae (Alpharetrovirus)	Birds	Birds (111–115)
	Retroviridae (Betaretrovirus)	Boreoeutherians	Boreoeutherians, marsupials, and birds (113–125)
	Retroviridae (Gammaretrovirus)	Terrestrial boreoeutherians, marsupials, and birds	Marine and terrestrial boreoeutherians, marsupials, birds, reptiles, and amphibians (113, 119, 126–138)

Table 0.1 (Continue).

Group/type*	Family (group)	Host range	
		Exogenous viruses	**Endogenous viruses (refs.)**
VI/RT-RNA virus	Retroviridae (Epsilonretrovirus)	Fish	Primates, amphibians, and fish (139–142)
	Retroviridae (Lentivirus)	Simians, felines, cattle, equines, and small ruminants	Lagomorphs, prosimians and weasels (143–149)
	Retroviridae (Spumavirus)	Boreoeutherians	Xenarthrans, afrotherians, and prosimians. (150–152)
VII/RT-DNA virus	Hepadnaviridae	Mammals, birds, and reptiles	Birds, and reptile (73, 91, 153–157)
	Caulimoviridae	Plants	Plants (158–162)

* ‘ds’ = double-stranded;
‘ss’ = single-stranded;
‘+’ = positive-sense;
‘–’ = negative-sense;
‘RT’ = reverse transcribing

documented that various exogenous DNA viruses can occasionally undergo host chromosome integration through the process of non-homologous dsDNA end-joining (e.g. hepadnavirus, (163)), non-homologous DNA recombination (e.g. adeno-associated DNA virus, (164–166)), and telomeric homologous recombination (e.g. herpesviruses, (167)). Thus, finding that DNA viruses are capable of becoming endogenous is not entirely unforeseeable (168).

The discovery of endogenous non-retroviral RNA viruses came as a surprise to researchers however, as it requires three unusual steps that do not usually occur in their life cycle: (i) conversion of genomic RNA to DNA, (ii) nucleus entry and (iii) chromosomal integration. Structural analyses have suggested that many non-retroviral RNA viruses directly hijack RT and IN activity encoded by long interspersed nuclear elements for their endogenisation (73, 105, 107). Furthermore, it has been proposed that a non-retroviral RNA virus, namely potato virus Y, might have become endogenous in grapevine via the process of non-homologous recombination between an RNA of the virus and that of a retrotransposon, with the subsequent product inserted into the host genome by retrotransposition (103). It has been demonstrated that recombination between ERVs and exogenous non-retroviral RNA virus can result in viral chromosomal integration also (169). At present, the precise molecular details of how non-retroviral viruses become endogenous are still poorly characterised.

To date, numerous EVEs have been identified from diverse organisms (**Table 0.1**) and across various timescales (**Table 0.2**). They offer us opportunities to study and reconstruct the evolutionary history of the viruses that they are descended from, some of which are long extinct. EVEs have proved to be extremely valuable in expanding our understanding of deep viral natural history and long-term evolutionary dynamics, thereby closing the gap between recent and ancient viral evolution in our knowledge.

0.6 EVE discovery has vastly expanded viral host range

One of the most direct ways in which EVEs can inform us about viral natural history is that they can tell us about the diversity of the hosts with which viruses are capable of interacting. This

Table 0.2 Viral evolutionary timescales estimated from EVEs. Only the oldest age estimates available are shown.

Group/type*	Family (group)	Timescale of evolutionary (Myr old)	Method of timescale estimation*	Ref
I/dsDNA virus	*Herpesviridae* (*Betaherpesvirinae*)	>76	TMR-dating	(79)
II/ssDNA virus	*Parvoviridae*	>40	Ortholog-dating	(89)
	Circoviridae	>50	Ortholog-dating	(73, 89)
	*Geminiviridae (*Begomoviruses*)*	~20–30	Long-term rate date calibration	(95)
III/dsRNA virus	*Partitiviridae*	>10	Ortholog-dating	(96)
V/(–)ssRNA virus	*Bunyaviridae (*Phleboviruses*)*	>8	Ortholog-dating	(104)
	Bornaviridae	>93	Ortholog-dating	(73)
	Filoviridae	>30	Ortholog-dating	(73, 108)
	Rhabdoviridae	>8	Ortholog-dating	(96)
VI/RT-RNA virus	*Retroviridae (*Betaretroviruses*)*	>100	Host distribution	(113)
	*Retroviridae (*Gammaretroviruses*)*	>65	LTR-dating	(126)
	*Retroviridae (*Epsilonretroviruses*)*	>40	LTR-dating & ortholog-dating	(139, 140)
	*Retroviridae (*Lentivirus*)*	>14	Dating by using bio-geographical information	(146)
	*Retroviridae (*Spumavirus*)*	~100	Co-speciation	(150, 151)
VII/RT-DNA virus	*Hepadnaviridae*	>207	Ortholog-dating	(157)
	Caulimoviridae	>4.6	Ortholog-dating	(170)

* 'TMR' = telomeric repeats
'LTR' = Long terminal repeats

information has greatly broadened our knowledge of viral host range, in particular when they are identified in lineages that are not known to harbour exogenous, modern-day, representatives. The host range of extant viruses has been extensively surveyed among humans, livestock and companion animals, and this knowledge has aided the development of viral control and prevention strategies (171–174). Nevertheless, this information pertains largely to extant viral lineages that have survived to be sampled. Furthermore, there is also a tendency for non-agriculturally and non-medically important viruses to be overlooked. As a result, it could thus be that the actual viral host range is much broader than surveillance of existing viruses alone could reveal. This bias is less prominent for EVEs however, since searches for them are typically done across all cellular genomes that are publically available. Indeed, EVEs have revealed that many viruses have a much broader host range than previously appreciated.

By combining exogenous and endogenous viral host ranges together, our knowledge of the distribution of viruses among organisms has been vastly expanded. For example, all exogenous nanoviruses have only been found so far in plants (175). However, the identification of nanovirus EVEs in various animal and protist genomes (86, 92) provided conclusive evidence that, at least in the past, they could also infect animals and protists in addition to plants. Similarly, parvoviruses and densoviruses are known to be currently circulating only among eutherians, birds, crustaceans and insects (176). Hence, without the discoveries of their endogenous relatives in the genomes of marsupials, fish, tunicates, and flatworms (73, 88–90), we would not have known that they are (or at least were) capable of infecting these animals also. Likewise, the host range of protist- and fungus-infecting totiviruses has also been greatly expanded by the identifications of their endogenous counterparts in various animal genomes (86, 97–99). These examples (see **Table 0.1** for more) highlight how EVEs identified from diverse organisms across all these four eukaryotic kingdoms have greatly expanded the range of known hosts in a way that no extant virus surveillance could do. This information can be used to guide searches for potentially undescribed circulating exogenous viruses, which may allow us to better monitor and predict emergent viral infections.

0.7 Dating EVEs and generating robust viral evolutionary timescales

Besides expanding viral host range, EVEs have also greatly improved our knowledge of viral origins, providing robust evidence of the presence of viral lineages during particular periods. Because the formation of EVEs must have occurred after the origin of their respective viral families, their age can therefore be interpreted as a concrete lower-bound age estimate for the viral group that they are descended from. Several approaches can be used to date EVEs (**Figure 0.2** and **Table 0.2**). The simplest approach is to assume that the time to most recent common ancestor (tMRCA) of all exogenous and endogenous viruses is comparable to that of the hosts that carry them. For example, by examining the distribution of endogenous betaretroviruses across bird genomes, it has been proposed that they are likely >100 Myr old (113). Likewise, the timescale of endogenous begomoviruses has been estimated to be between 1.24 and 4.85 Myr old by this method (95). This approach, however, can be greatly biased by viral cross-species transmission, which can give the erroneous impression that the considered viruses are ancient, especially when transmissions take place between distantly related hosts.

A more robust approach, which is unaffected by viral cross-species transmission, is dating using EVE orthologs. Given the low probability of two independent EVE insertions occurring at the same site, the presence of EVEs in two or more host lineages at the same genomic location is a strong indication that they are orthologs, and thus the integration event must predate the host basal diversification. Subsequently, the minimum age of the EVEs can be directly inferred from the host basal split date. A way to confirm that two or more EVEs are indeed orthologous, resulting from a single integration (rather than multiple independent integrations at the exact same locus) is to calculate the distance between potentially orthologous EVEs and to ensure that it is consistent with what is known about the divergence time and substitution rate of the hosts (e.g. (157)). This method has been used to estimate the (minimum) timescales of many EVEs. For instance, analyses of orthologous bird hepadnavirus EVEs suggested that these

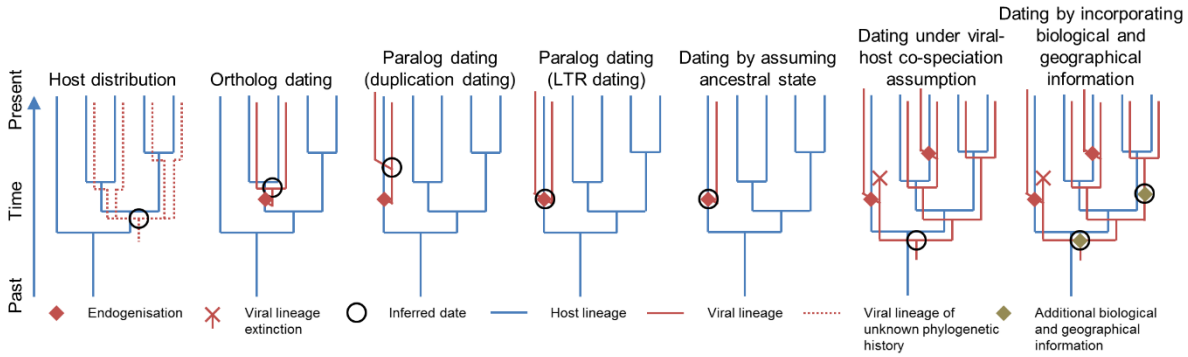


Figure 0.2 Inferring viral evolutionary history though endogenous viral element (EVE) analyses and dating. EVE timescales are generally inferred directly from their host timescales or based on the host neutral rate of evolution. Their age represents the lower-bound age estimate of the corresponding viral family. Several approaches can be used to date viral timescales through EVE analyses. The simplest approach is to infer that the timescale of the (combined exogenous and endogenous) viruses is comparable to that of the hosts carrying them, without explicit phylogenetic analyses (1st panel from left). If EVE orthologs are available, the basal diversification date of the hosts carrying orthologous EVEs can be interpreted as the minimum age of the EVEs, and thus the minimum age of the corresponding viral family (2nd panel). In the case that EVEs duplicated, the date of the event can be calculated from genetic difference observed between EVE paralogs, assuming that they accumulate genetic changes at the host neutral rate of evolution after the duplication. Likewise, this date can be interpreted as an EVE minimum age (3rd panel). In the specific case of endogenous retroviruses (ERVs), their integration event can be dated directly. During the process of retroviral endogenisation, two identical stretches of paralogous sequences are produced, known as long terminal repeats (LTRs; **Figure 0.1C**). Again, by assuming that once becoming endogenous the paired LTRs accumulate genetic changes independently from one another at the host neutral rate of evolution, the timescale required for them to be as different as observed can be calculated (4th panel). This is essentially the EVE timescale and thus the minimum age of the associated viral family. Alternatively, an EVE integration event can be dated by assuming an EVE ancestral state. This method is more general than LTR dating as it can be applied to any EVEs, and not just ERVs. By estimating the genetic difference between the current EVEs and the ancestral state, and assuming that the EVE evolutionary rate is equal to the host neutral rate of evolution, EVE integration can be dated (5th panel). Furthermore, if phylogenetic analyses suggest that viruses have been stably co-speciating with their hosts, the timescale of viruses can then be directly inferred from that of their hosts (6th panel). Lastly, by integrating multiple sources of information together, such as viral–host co-evolutionary history, selection imprints on antiviral genes, phylogenetic information, and bio-geographical distribution and timescale, our knowledge of viral evolutionary history can be improved even further (7th panel).

viruses must be >74 Myr old, originating in the Upper Cretaceous period of the Mesozoic Era (155). The identification of orthologous hepadnavirus EVEs in turtles pushed the origin of these viruses even further back, to the late Triassic period of the Mesozoic Era >207 Myr ago (Ma) (157). Similarly, analyses of orthologous rabbit-hare endogenous lentiviruses have shown that they are >12 Myr old (144), providing evidence for the origin of modern-day lentiviruses in the Miocene epoch or even earlier.

Another means to derive a minimum age for an EVE relies on cases where EVEs have been duplicated within the genome. This can be done by estimating the genetic divergence between duplicated EVE paralogs and converting it back to time by assuming the host neutral rate of evolution. This method has been used to calculate minimum ages of endogenous rabbit lentiviruses and sloth foamy viruses to be 7 Myr old (143), and 39 Myr old (150), respectively. A special case of duplication event can also be used to directly estimate the date of EVE integration. In the case of retroviruses, two identical stretches of paralogous LTR sequences flanking the viral genome are produced during the process of integration (**Figure 0.1C**). After endogenisation, these two sequences evolve independently from one another. By tracing the process backwards, and assuming host neutral evolution, it is possible to estimate the ERV integration date. This method has been used to date many ERV integration events, revealing that most retroviral families are many millions of years old (78, 125, 126, 139, 140, 177, 178). These methods are based on a reasonable assumption that EVE sequences evolve at the host neutral rate after the duplication process. Some studies applied the host neutral rate to calibrate phylogenies derived from multi-copy EVE families (95). However, such approaches may overestimate the true EVE timescale. This is because some of the variations observed among EVEs would have been generated under the viral rate of evolution, which is much faster than the host neutral rate. Thus, the application of the host neutral rate would lead to the dates of some nodes, in particular the deep ones, being overestimated.

Another approach to dating EVE integration events is to compare current EVEs against an ancestral viral state. This method is more general than the LTR-dating approach as it can be

applied to any EVE, and not just ERVs. For example, it has been used to estimate the integration date of a tarsier endogenous betaherpesvirus to be >76 Myr old (79). In brief, herpesviruses possess satellite telomeric repeats (TMRs). By knowing that the viral TMRs are identical to those of the host at the time of endogenisation, the endogenisation date of the virus can be calculated, assuming that any changes within the repeats occur at the host neutral rate of evolution. Indeed, this finding is consistent with the results from viral-host co-speciation analyses that suggested that, as a whole, betaherpesviruses are ~100 Myr old (179, 180).

A related approach that is based on knowledge of the ancestral viral state is to estimate EVE integration dates by using the frequency of stop codons contained in their sequences. Briefly, by assuming that ancestral viral sequences did not contain stop codons at the time of integration, and that the stop codons observed in EVEs are the result of neutral mutations, the timescale required to accumulate the observed number of stop codons can be estimated. To do so, nucleotide composition may be assumed (107), or extant viral sequences may be used as model sequences (151). This method has been used to estimate the integration dates of primate and rodent endogenous bornaviruses to be ~19–40 Myr ago (Ma) (107) and those of aye-aye and cape golden mole endogenous foamy viruses to be ~35 and ~65 Ma, respectively (151).

Apart from being minimum age estimates of their respective viral lineages as mentioned above, EVE integration dates can also be used as calibration points in downstream analysis in order to obtain timelines of viral evolution. For example, in a study of begomoviruses (95), an ancestral sequence of endogenous begomoviruses was inferred and its integration date was estimated. This information was then used to calibrate the timescale of the whole viral family to be ~20–30 Myr old. All together combined, results from paleovirological analyses seem to suggest that many, if not all, modern viral families have ancient origins, being many millions of years old (**Table 0.2**).

0.8 EVEs reveal ancient viral-host interaction and transmission routes

In addition to providing a means of estimating viral evolutionary timescales, EVEs have also proved instrumental in expanding our knowledge of ancient viral-host interaction. For example, phylogenetic analyses of extant simian foamy viruses suggested that they have an extremely stable history of co-speciation with their hosts for >30 Myr (181). The origin of this stable foamy virus-host co-speciation history was then pushed further back to the origin of eutherians, >100 Ma, by analyses of endogenous foamy viruses identified in xenarthran (sloth) and afrotherian (cape golden mole) genomes (150, 151). This example illustrates how EVEs can supplement and expand our knowledge of long-term viral-host interaction, and how this information can be used to date, and therefore improve our knowledge of, viral origin. Similarly, our knowledge of the interaction between hepadnaviruses and their hosts has also been greatly improved by the identification of ancient endogenous snake hepadnaviruses. Analyses suggested that they are basal to the clade of bird and mammal hepadnaviral relatives (91). Unlike in the case of foamy viruses however, this phylogenetic relationship is inconsistent with that of their hosts (where birds are more closely related to reptiles than mammals), indicative of deep viral cross-species transfers between vertebrate hosts of different classes (91). Since no extant xenarthran and afrotherian foamy viruses and reptile hepadnaviruses have been identified to date, it would have been impossible to obtain such information through analyses of currently available extant viral sequences alone.

In some cases, analyses of EVEs can reveal previously unknown reservoir host species and/or ancient viral geographical transmission routes, particularly where the hosts are limited geographically. Filoviruses consist of three viral subgroups, namely Ebola, Cueva, and Marburg viruses, and are known to be currently circulating primarily among bats and primates in Africa and East Asia. They are characterised by frequent cross-species transmissions, and occasionally jump from animals to humans, causing haemorrhagic fever outbreaks. Very little is known about their origin and reservoirs, with some evidence suggesting that bats are the

most likely candidates (182–187). Molecular analyses showed that South American and Australian marsupial endogenous filoviruses form a paraphyletic clade with, and are basal to, the clade of extant filoviruses (73, 108). Not only did these findings reveal that filoviruses are more geographically dispersed than previously thought, but also suggested that the Old World extant filoviruses might have originated in the New World (108). Further insight into the natural history and reservoirs of extant filoviruses came from the identification of filovirus EVEs in small mammals (73, 108, 109). Analyses have shown that the clade of rodent filovirus EVEs, which are >18 Myr old, is a sister of the clade of Ebola/Cueva viruses, but not Marburg viruses, indicating that modern filoviruses originated >18 Ma, in the early Miocene Epoch (109). Furthermore, more basal than endogenous marsupial filoviruses is the clade of filovirus EVEs identified in small bats, rodents, insectivores, and tenrecs, which are >30 Myr old (73, 108). Combined, it is also possible that, rather than marsupials, it was these small mammals that had been acting as filovirus reservoirs, repeatedly transmitting viruses to the New World marsupials and Old World primates. Indeed, this hypothesis is consistent with the prediction that, in addition to bats, insectivores and rodents are one of the most probable filovirus host reservoirs, based on a series of assumptions about the natural history of a likely filovirus host (188). This detailed example clearly demonstrates the power of EVEs to further elucidate the deep history of their modern-day viral relatives.

By combining viral and host phylogenetic information together with bio-geographical distribution and timescales, even more detailed hypotheses of viral transmission sources and routes can be formed. For instance, although not strongly supported, analyses have shown that an endogenous Malagasy lemur lentivirus is most closely related to simian lentiviruses (146, 147). This finding is consistent with a viral-host co-speciation scenario, which, if indeed is true, would posit that primate lentiviruses are >85 Myr old (146). Nevertheless, this pattern could also be a result of lentiviral cross-species transmission between prosimians and simians subsequent to their speciation (146). Under a terrestrial mode of viral transmission, it would require that the transmission was dated back >14 Ma (146), the most recent time terrestrial mammal invasion occurred (189). Alternatively, if the transmission occurred via an aerial vector

species, then the transmission could have happened more recently (146). Although it is still unclear which hypothesis is correct, this example nevertheless illustrates how tying together the ancient and recent viral geographical distribution and timescale allows us to form detailed hypotheses of past viral transmission routes.

Additionally, not only can EVEs inform us about their past history, but the models derived from their analyses may have some implications for the history of their hosts as well. For example, amidst the stable interaction between foamy viruses and their hosts, a few viral cross-species transmissions were identified, involving exogenous Chinese bat and endogenous Madagascan aye-aye foamy viruses (151). Analyses have shown that the two form their own deep clade, branching very deeply towards the root of the tree (151). Based on these findings and assuming the most parsimonious scenario (i.e. the least number of cross-species transmission events), it was postulated that the ancestor of these two viruses jumped from the African continent to the Madagascar-India landmass ~100 Ma, and were split into two groups ~80–90 Ma by the Madagascar-India landmass separation. The virus that remained on Madagascar later gave rise to the aye-aye endogenous foamy virus, and the other that was transferred to Asia then gave rise to the Chinese bat foamy virus. To date, this is the only documented bat (Yinpterochiroptera) foamy virus. However, if further analyses show that Yinpterochiroptera foamy viruses co-speciate with their hosts, it would then suggest that the cross-species transmission giving rise to the bat foamy virus likely occurred before the radiation of Yinpterochiroptera bats, ~63–53 Ma (190). However, the Indian landmass did not come into contact with Asia until ~55–35 Ma (191). Such observations would thus posit that the radiation of the bats did not occur in Asia, but rather on the Indian landmass while drifting in the Indian Ocean. This is another clear example highlighting the value of geographical history in elucidating viral evolution, and how this knowledge might affect our understanding of host natural history.

0.9 Discrepancy between viral evolutionary timescales and rates inferred from EVEs and modern-day viral sequences

Before the emergence of paleovirology, viral evolutionary timescales have generally been estimated from extant viral sequences, using phylogenetic techniques (**Figure 0.3**). In fact, most of the viral evolutionary timescale estimates available today have been calculated based on molecular sequences of modern-day viruses that were collected at different times (i.e. heterochronous molecular datasets). In brief, this method uses the differences in sampling dates and terminal branch lengths to infer the rate of viral evolution and extrapolates the rate to other parts of the tree to convert the units of the remaining branch lengths from molecular substitutions to time (192–195). This process allows us to infer the split-dates of all nodes, and thus the tMRCA of all viruses under investigation. Absolute divergence times of two or more viral lineages (i.e. node heights), usually inferred from the host geographical split or speciation dates, are another source of rate-calibration information (57, 179, 181, 196–200). Alternatively, if the rate of viral evolution has already been established elsewhere, instead of inferring the rate from the investigated sequences, such rates can be used to directly calibrate the timescale of the tree (201, 202). This is particularly useful when rate-calibration information is not available. Since viral evolutionary timescales estimated from these methods are independent from those derived from EVEs, they can be used to validate one another. Surprisingly, it appears that many EVE-derived age estimates are in conflict with those calculated from extant viral sequences.

There are many examples of such timescale estimate discrepancies. For example, analyses of positive selection footprints on anti-retroviral genes consistently supported ages of >10 Myr for SIVs (61–65), in line with an analysis of an endogenous prosimian lentivirus, suggesting that lentiviruses likely have been infecting primates for >14 Myr (146). In contrast, analyses of extant viral sequences suggested that they are only a few hundred years (55, 201) to tens of thousands of years old (57). The timescales of mammalian and bird hepadnaviruses estimated from EVEs and extant viral sequences are also very different. While analyses of EVEs

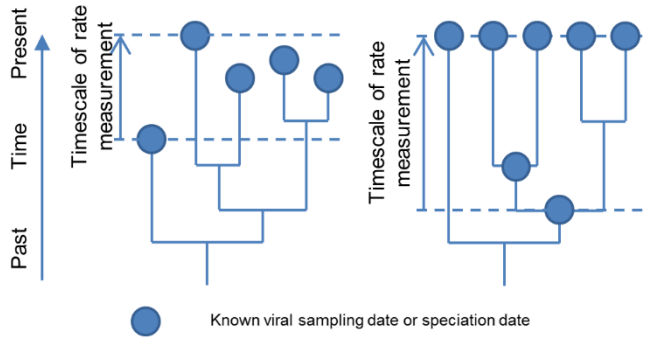


Figure 0.3 Estimating viral evolutionary timescale from extant viral sequences. Several phylogenetic techniques can be used to infer evolutionary timescales of viruses from their molecular sequences. The first method (**left**) is to use the differences in sampling dates and terminal branch lengths to infer the rate of viral evolution and extrapolates the rate to other parts of the tree to convert the units of the remaining branch lengths from molecular substitutions to time (192–195). If viral speciation dates are available, they can also be used to calibrate the rate and subsequently the timescale of viral evolution (**right**) (179, 181, 196–200). In the case that the rates of evolution of viruses have already been established elsewhere, instead of inferring the rate from the investigated sequences, such rates can be directly used to calibrate the timescale of the tree (201, 202). The rates calibrated from serially sampled viral sequences are generally short-term rate estimates, and those estimated from viral speciation dates are normally long-term rate estimates.

suggested that they are >74 Myr old (155), their origin was dated to be only ~30,000 (203) to ~125,000 (204) years ago by analyses of modern-day viruses. Likewise, while analyses of extant filoviruses suggested that the tMRCA for modern filoviruses ranges between 7,000 and 155,000 years (205–207), filovirus EVE analyses, identified in hamster and vole genomes, suggested that they are >18 Myr old (109). Similar observations have also been made for circoviruses and phleboviruses. Their timescales of evolution were estimated to be <500 years (208–210) whereas the discovery of their EVEs suggested that they are >50 Myr (73, 89) and >8 Myr (104), respectively. Overall, analyses of extant viruses seem to suggest that some modern viral families are young, whereas paleovirological analyses strongly support that almost all of these viral groups are of great antiquity, being millions of years old (**Table 0.2**).

So, which timescale estimates are closer to the true dates: the ones estimated from EVEs or the ones derived from the rate estimates of modern-day viruses? Since none of them are self-evidently wrong, it is difficult to answer the question with certainty. However, EVE timescales have generally been estimated based on the host neutral rate of evolution or directly inferred from the host timescales. Therefore, if the deep viral origins derived from EVEs were to be wrong and the shallow dates were to be closer to the truth, it would posit that (i) the host neutral rate must be several orders of magnitude higher than what we currently know, and (ii) the hosts have come to existence very recently (e.g. <1 Myr). These are extremely unlikely given that the host timescales and neutral rate of evolution are very well-established. Consequently, it is much more likely that the EVE-derived viral evolutionary timescales represent more accurate date estimates of viral origin. These observations have contributed to a radical shift in our knowledge of viral natural history, strongly suggesting that all viruses have a long co-evolutionary history with their hosts.

These findings have also challenged our understanding of viral evolutionary dynamics. Analyses of heterochronous extant viral sequences (collected over a time period of <500 years) often suggested that ssDNA viruses and many RNA viruses evolve extremely quickly. Their rates are often reported to be in the order of 10^{-2} to 10^{-5} s/n/y (68, 69, 201, 211–215).

As a result, these viruses are often said to be 'fast-evolving' viruses. Surprisingly, many ancient EVEs related to these 'fast-evolving' viruses appear to exhibit similarity higher than expected given their known rate of evolution (216). For instance, analyses of heterochronous extant lentiviral sequences have firmly established that their evolutionary rates are in the order of 10^{-4} s/n/y (217–220). Assuming that the rate is homogenous and constant at a value of 1×10^{-4} s/n/y, it is expected that, say, after a million years of evolution, each nucleotide position would have changed a hundred times on average, being fully saturated by nucleotide substitutions. This should render the ancestral and the evolved sequences to be no more similar than two random sequences. Despite this however, ancient endogenous lentiviruses that are millions of years old can still be easily recognised as relatives of modern lentiviruses (144–149). These findings mean that, in contrast to the conventional wisdom, lentiviruses evolve strikingly slowly in the long term. The same story can also be told for other 'fast-evolving' RNA and ssDNA viruses.

At the same time, viruses that appear to evolve slowly in the long term seem to evolve extremely rapidly over short timescales. For example, the evolutionary rates of classic 'slow-evolving' dsDNA viruses, foamy viruses and deltaretroviruses (221) are often reported to be in the range of $\sim 10^{-7}$–10^{-9} s/n/y, typically estimated over timescales of millions of years under the viral-host co-speciation assumption (179, 181, 196–200). However, many studies have shown that, over short timescales, these viruses evolve extremely rapidly. For example, analyses of serially-sampled dsDNA herpesviruses (collected over a timescale of 27 years) suggested that their rate is as high as 8.96×10^{-5} s/n/y (222), comparable to many of those of RNA viruses. Likewise, based on a 9-year longitudinal study of foamy viruses in an African green monkey population, their short-term rate estimate was calculated to be 3.74×10^{-4} s/n/y (223). This rate estimate is 3–4 orders of magnitudes higher than the typical reported rates of evolution of 'slow-evolving' foamy viruses. Deltaretroviruses have also been observed to evolve rapidly over short timescales. For instance, by following a population of bovine leukaemia viruses in a single sheep for 1.5 years, Willems et al. (224) estimated their rate of evolution to be as high as $\sim 3.43 \times 10^{-4}$ s/n/y. Analyses of human T-lymphotropic viruses type

II circulating among intravenous drug users also suggested that their rate is $\sim 2.7 \times 10^{-4}$ s/n/y (225, 226). These rate estimates are 2–4 orders of magnitude higher than their conventional long-term evolutionary rate estimates. Combined with the results from paleovirological analyses, it seems that all documented viruses so far appear to evolve extremely rapidly in the short-term but much slower in the long-term.

One possibility is that the observed rate change genuinely reflects changes in viral biology. Indeed, it has been found that the error rates of viral replication vary greatly among viruses (68, 227). Therefore, it is conceivable that the fidelity of viral polymerases might have varied with time, and thus might explain some portion of the discovered short-term and long-term rate discrepancy (153). Changes in viral transmission mode and/or replication speed have also been put forward as candidate explanations for the observed rate variation in viruses; the higher the replication and transmission speed, the higher the chance of mutations being accumulated, and thus the higher the rate of evolution (67, 225, 226). Nevertheless, for these factors to account for the observed rate change, not only must they alter the rate of viral evolution, but the effect must be systematically associated with time as well. Furthermore, given the generality of the observations, this would also posit that the biology of all viruses has changed in a similar way. Combined, although it is still currently unknown to what extent changes in viral biology contribute to the observed changes in the rates of viral evolution, we believe that they are unlikely to be the main contributors.

Large-scale meta-analyses have shown that, in fact, viral evolutionary rate estimates are negatively correlated with the timescale of rate measurement (228). This relationship stands both within RNA and DNA viral groups, by combining rates across a range of measurement techniques (228). This '*time-dependent rate phenomenon*' (TDRP) has been observed to be very common in nature, and was first noticed in cellular organisms (see review in (229)). We suggest that all of the observed viral rate discrepancies are in fact parts of this phenomenon, where rate estimates have been sampled from two extreme ends of the measurement timescale spectrum. That is, evolutionary rates of extant exogenous viruses tend to be

calculated over extremely short timescales (<100 years) and thus are extremely high, whereas EVE analyses typically span much longer timescales (>1 Myr) and hence yield rate estimates that are much lower.

Besides biological changes, many explanations have been proposed to explain the TDRP (**Figure 0.4**, see (229) for review). One explanation is that short-term rate estimates tend to include transient deleterious variations that have not yet been purged by purifying selection, and thus appear to be very high (69, 229, 230). Viral adaptive changes resulting from positive selection or viral-host evolutionary arm races are also lineage-specific, and hence will tend to manifest themselves as recent changes, inevitably elevating the short-term rate of evolution. Similarly, molecular variations that result from sequencing errors will also appear as recent changes, and because of this, may contribute to the observed high short-term rates to some degree (231–234). These processes could all lead to an elevation of short-term rates. In contrast, over longer timescales, failing to properly account for superimposed substitutions and substitution saturation can cause underestimation of rates (228, 229, 235–237). Indeed, using inappropriate substitution models can lead to sequence divergence, and thereby evolutionary rate, underestimation (237). This bias can be alleviated by statistical model-selection procedures, in other words choosing the best model available. Nevertheless, given model simplifications and complex interactions between different parts of viral genomes, such as gene overlapping and epistasis, it is likely that even the best model currently available will fail to properly account for these complex patterns of substitution.

Failing to account for rate heterogeneity among sites and the tendency for underestimating the true dates of calibrating nodes are other sources of errors that have the potential to contribute to the TDRP (229, 241). Alternatively, it is also possible that the observed TDRP can be caused by our tendency to use fast-evolving molecules for short-term evolutionary studies and slowly-evolving molecules for long-term evolutionary studies (229). At present, the extent to which each factor contributes to the TDRP is unclear. To bridge the gap between long-term and short-

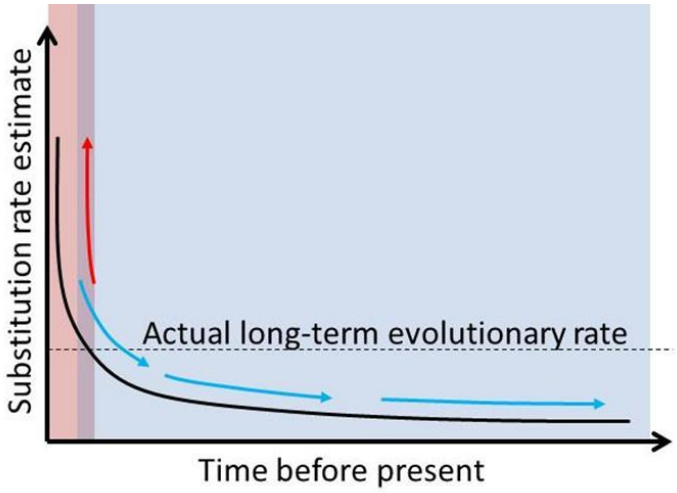


Figure 0.4 What cause evolutionary rate estimates to appear to be time-dependent? Numerous factors have the potential to cause evolutionary rate estimates to appear to be time-dependent. It has been proposed that evolutionary rates estimated over short-timescales (within the red region) tend to be overestimated (indicated by an upward red arrow) due to the inclusion of transient deleterious variations (69, 229), lineage-specific recent adaptive changes (238, 239), and sequencing errors (231–234). The tendency to underestimate the true dates of calibrating nodes can also result in rate overestimation (234). In contrast, misspecification of substitution model (237, 240), substitution saturations (229, 235–237), and rate heterogeneity among sites (241) have the potential to cause 'long-term' rate estimates (within the blue region) to be underestimated (indicated by downward blue arrows). Furthermore, it is also common for fast-evolving molecules to be used in short-term evolutionary studies and for slowly-evolving molecules to be used in long-term evolutionary studies. This itself can cause rates to appear to be time-dependent (229). As a result, a 'constant' actual rate (dotted line) may appear to be decreasing towards the past (black solid curve) (see (229) for review).

term viral natural history and evolutionary dynamics, understanding how the TDRP is generated will be absolutely crucial.

0.10 Thesis overview

The overarching aim of this thesis is to connect, reconcile, and fill in the gap between our knowledge of the recent and ancient viral evolution by using retroviruses as our study subject. Leading to this goal, this thesis investigates the long-term evolutionary history of retroviruses, with a particular focus on foamy viruses, and examines the general pattern of evolutionary rate dynamics that are shared by all viruses. This thesis is composed of five data chapters, exploring different specific paleovirological questions.

Chapter one reports the first complete genome of an exogenous prosimian galago foamy virus and two endogenous foamy viruses embedded within the genome of the prosimian aye-aye, and afrotherian Cape golden mole. Analyses reveal that, overall, foamy viruses and their mammalian hosts have been stably co-diverging with one another since the origin of eutherians, dating back >100 Ma. However, a small number of ancient viral cross-species transmissions could also be observed.

Foamy virus evolution was investigated further in Chapter two, focusing on that of New World monkey foamy viruses. The results show that New World monkey foamy viruses co-diverge broadly with their hosts at the level of host species and sub-population. However, analyses cannot detect a strong co-speciation pattern at the level of host genera, but instead reveal several ancient foamy virus cross-genus transmissions. This suggests that transmissions of foamy viruses among different New World monkey species were not rare in the past.

Chapter three presents analyses of the long-term evolutionary rate dynamics of foamy viruses under the assumption of viral-host co-speciation. The results show that the rate estimates of foamy virus evolution appear to be strongly negatively correlated with their measurement

timescales, following a power-law decay curve. We term this phenomenon the '*time-dependent rate phenomenon*' (TDRP). The work also demonstrates that currently available evolutionary inference tools cannot accommodate the TDRP, severely biasing evolutionary date estimates of foamy viruses. Indeed, the analyses presented in this chapter show that the power-law TDRP model outperforms all currently available evolutionary rate models.

The TDRP was further examined in Chapter four. Chapter four investigates the TDRP across almost all viral genome types and replication strategies. By using the power-law TDRP model as the basis of investigation, the analyses show that the TDRP can be found in all viruses, existing across multiple levels of viral taxonomy, from the Baltimore classification viral groups to genera. Furthermore, by taking the TDRP into account, a clear division between the rates of various types of viruses can be recovered, a distinction that has become less and less clear as more and more viral evolutionary rate estimates become available. Surprisingly, the results show that the rate-decay speed is universal across all viruses despite their large differences in biology, and therefore, might be useful for better estimation of the evolutionary timescale of any virus. To demonstrate this, the TDRP model was used to reestimate evolutionary timescales of extant lentiviruses—a group of viruses that have always been suggested to be very young by standard phylogenetic analyses despite conclusive evidence from paleovirological analyses showing that they are very old. The results suggest that modern-day lentiviruses originated millions of years ago, consistent with paleovirological evidence, and thus, for the first time, reconciling molecular analyses of extant and ancient viruses.

The last data chapter, Chapter five, integrates all the work in previous chapters to investigate the origin of retroviruses as a whole. This chapter reports the discovery of foamy-like ERVs in a wide range of amphibians and fish, including salamanders, a frog, lobe-finned fish, ray-finned fish, and shark. Examination of genomic organisation showed that the genome of ray-finned fish foamy-like ERVs differ significantly from those of typical mammalian foamy viruses, containing two additional 5' accessory genes, and is extremely large. Indeed, they are the largest known retroviral genome, being ~17kb in length. By using the TDRP model and co-

phylogenetic analyses, the work reveals that this major group of retroviruses originated together with their jawed vertebrate hosts in the Ordovician ocean, co-diverging and evolving together for >450 Myr. This is the first direct temporal evidence concretely supporting the ancient origin of retroviruses in the early Paleozoic era, if not earlier.

0.11 References

1. **Edwards RA**, **Rohwer F**. 2005. Opinion: viral metagenomics. Nat. Rev. Microbiol. **3**:504–510.
2. **Suttle CA**. 2007. Marine viruses--major players in the global ecosystem. Nat. Rev. Microbiol. **5**:801–812.
3. **Williamson KE**, **Corzo KA**, **Drissi CL**, **Buckingham JM**, **Thompson CP**, **Helton RR**. 2013. Estimates of viral abundance in soils are strongly influenced by extraction and enumeration methods. Biol. Fertil. Soils **49**:857–869.
4. **Williamson KE**, **Radosevich M**, **Wommack KE**. 2005. Abundance and diversity of viruses in six Delaware soils. Appl. Environ. Microbiol. **71**:3119–3125.
5. **Williamson KE**, **Radosevich M**, **Smith DW**, **Wommack KE**. 2007. Incidence of lysogeny within temperate and extreme soil environments. Environ. Microbiol. **9**:2563–2574.
6. **Hara S**, **Terauchi K**, **Koike I**. 1991. Abundance of viruses in marine waters: assessment by epifluorescence and transmission electron microscopy. Appl. Environ. Microbiol. **57**:2731–2734.
7. **Weinbauer MG**, **Fuks D**, **Puskaric S**, **Peduzzi P**. 1995. Diel, seasonal, and depth-related variability of viruses and dissolved DNA in the northern Adriatic Sea. Microb. Ecol. **30**:25–41.
8. **Guixa-Boixereu N**, **Vaque D**, **Gasol J**, **Pedros-Alio C**. 1999. Distribution of viruses and their potential effect on bacterioplankton in an oligotrophic marine system. Aquat. Microb. Ecol. **19**:205–213.
9. **Weinbauer MG**, **Suttle CA**. 1997. Comparison of epifluorescence and transmission electron microscopy for counting viruses in natural marine waters. Aquat. Microb. Ecol. **13**:225–232.
10. **Parada V**, **Sintes E**, **van Aken HM**, **Weinbauer MG**, **Herndl GJ**. 2007. Viral abundance, decay, and diversity in the meso- and bathypelagic waters of the north atlantic. Appl. Environ. Microbiol. **73**:4429–4438.
11. **Kepner RL**, **Wharton RA**, **Suttle CA**. 1998. Viruses in Antarctic lakes. Limnol.

Oceanogr. **43**:1754–1761.

12. **Hennes KP**, **Suttle CA**. 1995. Direct counts of viruses in natural waters and laboratory cultures by epifluorescence microscopy. Limnol. Oceanogr. **40**:1050–1055.
13. **Weinbauer M**, **Höfle M**. 1998. Size-specific mortality of lake bacterioplankton by natural virus communities. Aquat. Microb. Ecol. **15**:103–113.
14. **Nakayama N**, **Okumura M**, **Inoue K**, **Asakawa S**, **Kimura M**. 2007. Seasonal variations in the abundance of virus-like particles and bacteria in the floodwater of a Japanese paddy field. Soil Sci. Plant Nutr. **53**:420–429.
15. **Prigent M**, **Leroy M**, **Confalonieri F**, **Dutertre M**, **DuBow MS**. 2005. A diversity of bacteriophage forms and genomes can be isolated from the surface sands of the Sahara Desert. Extremophiles **9**:289–296.
16. **Jiang S**, **Steward G**, **Jellison R**, **Chu W**, **Choi S**. 2004. Abundance, distribution, and diversity of viruses in alkaline, hypersaline Mono Lake, California. Microb. Ecol. **47**:9–17.
17. **Diemer GS**, **Stedman KM**. 2012. A novel virus genome discovered in an extreme environment suggests recombination between unrelated groups of RNA and DNA viruses. Biol. Direct **7**:13.
18. **Breitbart M**, **Wegley L**, **Leeds S**, **Schoenfeld T**, **Rohwer F**. 2004. Phage community dynamics in hot springs. Appl. Environ. Microbiol. **70**:1633–1640.
19. **Wommack KE**, **Williamson SJ**, **Sundbergh A**, **Helton RR**, **Glazer BT**, **Portune K**, **Craig Cary S**. 2004. An instrument for collecting discrete large-volume water samples suitable for ecological studies of microorganisms. Deep. Res. Part I Oceanogr. Res. Pap. **51**:1781–1792.
20. **Lucas W**. 2010. Viral capsids and envelopes: structure and function. eLS. John Wiley & Sons Ltd, Chichester.
21. **Finlay BB**, **McFadden G**. 2006. Anti-immunology: evasion of the host immune system by bacterial and viral pathogens. Cell **124**:767–782.
22. **Baltimore D**. 1971. Expression of animal virus genomes. Bacteriol. Rev. **35**:235–241.
23. **Summers J**, **Mason WS**. 1982. Replication of the genome of a hepatitis B-like virus by

reverse transcription of an RNA intermediate. Cell **29**:403–415.

24. **Hull R**, **Covey SN**. 1983. Does cauliflower mosaic virus replicate by reverse transcription? Trends Biochem. Sci. **8**:119–121.

25. **Marco Y**, **Howell SH**. 1984. Intracellular forms of viral DNA consistent with a model of reverse transcriptional replication of the cauliflower mosaic virus genome. Nucleic Acids Res. **12**:1517–1528.

26. **Temin HM**. 1985. Reverse transcription in the eukaryotic genome: retroviruses, pararetroviruses, retrotransposons, and retrotranscripts. Mol Biol Evol **2**:455–468.

27. **International Committee on Taxonomy of Viruses**. 2012. Family – *Retroviridae*, p. 477–495. *In* King, AMQ, Lefkowitz, E, Adams, MJ, Carstens, EB (eds.), Virus Taxonomy: Ninth Report of the International Committee on Taxonomy of Viruses. Elsevier Academic Press, San Diego.

28. **Coffin JM**, **Hughes SH**, **Varmus HE**. 1997. Retroviruses. Cold Spring Harbor Laboratory Press.

29. **Moebes A**, **Enssle J**, **Bieniasz PD**, **Heinkelein M**, **Lindemann D**, **Bock M**, **McClure MO**, **Rethwilm A**. 1997. Human foamy virus reverse transcription that occurs late in the viral replication cycle. J. Virol. **71**:7305–7311.

30. **Yu SF**, **Sullivan MD**, **Linial ML**. 1999. Evidence that the human foamy virus genome is DNA. J. Virol. **73**:1565–1572.

31. **Roy J**, **Rudolph W**, **Juretzek T**, **Gartner K**, **Bock M**, **Herchenroder O**, **Lindemann D**, **Heinkelein M**, **Rethwilm A**. 2003. Feline foamy virus genome and replication strategy. J. Virol. **77**:11324–11331.

32. **Linial ML**. 1999. Foamy viruses are unconventional retroviruses. J. Virol. **73**:1747–1755.

33. **Cullen BR**. 1992. Mechanism of action of regulatory proteins encoded by complex retroviruses. Microbiol. Rev. **56**:375–394.

34. **Neil SJD**, **Eastman SW**, **Jouvenet N**, **Bieniasz PD**. 2006. HIV-1 Vpu promotes release and prevents endocytosis of nascent retrovirus particles from the plasma membrane. PLoS Pathog. **2**:354–367.

35. **Löchelt M**, **Yu SF**, **Linial ML**, **Flügel RM**. 1995. The human foamy virus internal promoter is required for efficientgene expression and infectivity. Virology **206**:601–610.

36. **Kao SY**, **Calman a F**, **Luciw P a**, **Peterlin BM**. 1987. Anti-termination of transcription within the long terminal repeat of HIV-1 by *tat* gene product. Nature **330**:489–493.

37. **Hadzopoulou-Cladaras M**, **Felber BK**, **Cladaras C**, **Athanassopoulos A**, **Tse A**, **Pavlakis GN**. 1989. The *rev* (*trs*/*art*) protein of human immunodeficiency virus type 1 affects viral mRNA and protein expression via a *cis*-acting sequence in the *env* region. J. Virol. **63**:1265–1274.

38. **Goila-Gaur R**, **Strebel K**. 2008. HIV-1 Vif, APOBEC, and intrinsic immunity. Retrovirology **5**:51.

39. **Muthumani K**, **Choo AY**, **Zong W-X**, **Madesh M**, **Hwang DS**, **Premkumar A**, **Thieu KP**, **Emmanuel J**, **Kumar S**, **Thompson CB**, **Weiner DB**. 2006. The HIV-1 Vpr and glucocorticoid receptor complex is a gain-of-function interaction that prevents the nuclear localization of PARP-1. Nat. Cell Biol. **8**:170–179.

40. **Quaranta MG**, **Mattioli B**, **Giordani L**, **Viora M**. 2009. Immunoregulatory effects of HIV-1 Nef protein. BioFactors **35**:169–174.

41. **Lukic DS**, **Hotz-Wagenblatt A**, **Lei J**, **Räthe A-M**, **Mühle M**, **Denner J**, **Münk C**, **Löchelt M**. 2013. Identification of the feline foamy virus Bet domain essential for APOBEC3 counteraction. Retrovirology **10**:76.

42. **Russell RA**, **Wiegand HL**, **Moore MD**, **Schafer A**, **McClure MO**, **Cullen BR**. 2005. Foamy virus Bet proteins function as novel inhibitors of the APOBEC3 family of innate antiretroviral defense factors. J. Virol. **79**:8724–8731.

43. **Lindemann D**, **Rethwilm A**. 2011. Foamy virus biology and its application for vector development. Viruses **3**:561–585.

44. **Wilen CB**, **Tilton JC**, **Doms RW**. 2012. HIV: cell binding and entry. Cold Spring Harb. Perspect. Med. **2**:a006866.

45. **Picard-Maureau M**, **Jarmy G**, **Berg A**, **Rethwilm A**, **Lindemann D**. 2003. Foamy virus envelope glycoprotein-mediated entry involves a pH-dependent fusion process. J. Virol. **77**:4722–4730.

46. **Campbell EM**, **Hope TJ**. 2015. HIV-1 capsid: the multifaceted key player in HIV-1 infection. Nat. Rev. Microbiol. **13**:471–483.

47. **Hu W-S**, **Hughes SH**. 2012. HIV-1 reverse transcription. Cold Spring Harb. Perspect. Med. **2**:a006882.

48. **Hughes SH**. 2015. Reverse transcription of retroviruses and LTR retrotransposons. Microbiol. Spectr. **3**:MDNA3-0027-2014.

49. **Hindmarsh P**, **Leis J**. 1999. Retroviral DNA integration. Microbiol. Mol. Biol. Rev. **63**:836–843.

50. **Martin-Serrano J**, **Neil SJD**. 2011. Host factors involved in retroviral budding and release. Nat. Rev. Microbiol. **9**:519–531.

51. **Pincetic A**, **Leis J**. 2009. The mechanism of budding of retroviruses from cell membranes. Adv. Virol. **2009**:623969.

52. **Heinkelein M**, **Pietschmann T**, **Jármy G**, **Dressler M**, **Imrich H**, **Thurow J**, **Lindemann D**, **Bock M**, **Moebes A**, **Roy J**, **Herchenröder O**, **Rethwilm A**. 2000. Efficient intracellular retrotransposition of an exogenous primate retrovirus genome. EMBO J. **19**:3436–3445.

53. **Heinkelein M**, **Rammling M**, **Juretzek T**, **Lindemann D**, **Rethwilm A**. 2003. Retrotransposition and cell-to-cell transfer of foamy viruses. J. Virol. **77**:11855–11858.

54. **Katzourakis A**. 2013. Paleovirology: inferring viral evolution from host genome sequence data. Philos. Trans. R. Soc. Lond. B. Biol. Sci. **368**:20120493.

55. **Sharp PM**, **Bailes E**, **Gao F**, **Beer BE**, **Hirsch VM**, **Hahn BH**. 2000. Origins and evolution of AIDS viruses: estimating the time-scale. Biochem. Soc. Trans. **28**:275–282.

56. **VandeWoude S**, **Apetrei C**. 2006. Going wild: lessons from naturally occurring T-lymphotropic lentiviruses. Clin. Microbiol. Rev. **19**:728–762.

57. **Worobey M**, **Telfer P**, **Souquière S**, **Hunter M**, **Coleman CA**, **Metzger MJ**, **Reed P**, **Makuwa M**, **Hearn G**, **Honarvar S**, **Roques P**, **Apetrei C**, **Kazanji M**, **Marx PA**. 2010. Island biogeography reveals the deep history of SIV. Science **329**:1487.

58. **Sharp PM**, **Shaw GM**, **Hahn BH**. 2005. Simian immunodeficiency virus infection of chimpanzees. J. Virol. **79**:3891–3902.

59. **Silvestri G**. 2005. Naturally SIV-infected sooty mangabeys: are we closer to understanding why they do not develop AIDS? J. Med. Primatol. **34**:243–252.

60. **Silvestri G**, **Paiardini M**, **Pandrea I**, **Lederman MM**, **Sodora DL**. 2007. Understanding the benign nature of SIV infection in natural hosts. J. Clin. Invest. **117**:3148–3154.

61. **Compton AA**, **Emerman M**. 2013. Convergence and divergence in the evolution of the APOBEC3G-Vif interaction reveal ancient origins of simian immunodeficiency viruses. PLoS Pathog. **9**:e1003135.

62. **Laguette N**, **Rahm N**, **Sobhian B**, **Chable-Bessia C**, **Münch J**, **Snoeck J**, **Sauter D**, **Switzer WM**, **Heneine W**, **Kirchhoff F**, **Delsuc F**, **Telenti A**, **Benkirane M**. 2012. Evolutionary and functional analyses of the interaction between the myeloid restriction factor SAMHD1 and the lentiviral Vpx protein. Cell Host Microbe **11**:205–217.

63. **Lim ES**, **Fregoso OI**, **McCoy CO**, **Matsen FA**, **Malik HS**, **Emerman M**. 2012. The ability of primate lentiviruses to degrade the monocyte restriction factor SAMHD1 preceded the birth of the viral accessory protein Vpx. Cell Host Microbe **11**:194–204.

64. **Lim ES**, **Malik HS**, **Emerman M**. 2010. Ancient adaptive evolution of tetherin shaped the functions of Vpu and Nef in human immunodeficiency virus and primate lentiviruses. J. Virol. **84**:7124–7134.

65. **Malfavon-Borja R**, **Wu LI**, **Emerman M**, **Malik HS**. 2013. Birth, decay, and reconstruction of an ancient TRIMCyp gene fusion in primate genomes. Proc. Natl. Acad. Sci. U. S. A. **110**:E583–E592.

66. **Compton A a**, **Malik HS**, **Emerman M**. 2013. Host gene evolution traces the evolutionary history of ancient primate lentiviruses. Philos. Trans. R. Soc. Lond. B. Biol. Sci. **368**:20120496.

67. **Hanada K**, **Suzuki Y**, **Gojobori T**. 2004. A large variation in the rates of synonymous substitution for RNA viruses and its relationship to a diversity of viral infection and transmission modes. Mol. Biol. Evol. **21**:1074–1080.

68. **Sanjuán R**. 2012. From molecular genetics to phylodynamics: evolutionary relevance of mutation rates across viruses. PLoS Pathog. **8**:e1002685.

69. **Jenkins GM**, **Rambaut A**, **Pybus OG**, **Holmes EC**. 2002. Rates of molecular evolution

in RNA viruses: a quantitative phylogenetic analysis. J. Mol. Evol. **54**:156–165.

70. **Kumar S**, **Subramanian S**. 2002. Mutation rates in mammalian genomes. Proc. Natl. Acad. Sci. U. S. A. **99**:803–808.

71. **Holmes EC**. 2004. The phylogeography of human viruses. Mol. Ecol. **13**:745–756.

72. **Faria NR**, **Rambaut A**, **Suchard MA**, **Baele G**, **Bedford T**, **Ward MJ**, **Tatem AJ**, **Sousa JD**, **Arinaminpathy N**, **Pepin J**, **Posada D**, **Peeters M**, **Pybus OG**, **Lemey P**. 2014. The early spread and epidemic ignition of HIV-1 in human populations. Science **346**:56–61.

73. **Katzourakis A**, **Gifford RJ**. 2010. Endogenous viral elements in animal genomes. PLoS Genet. **6**:e1001191.

74. **International Human Genome Sequencing Consortium**. 2001. Initial sequencing and analysis of the human genome. Nature **409**:860–921.

75. **Smit AF**. 1999. Interspersed repeats and other mementos of transposable elements in mammalian genomes. Curr. Opin. Genet. Dev. **9**:657–663.

76. **Paces J**, **Pavlícek A**, **Paces V**. 2002. HERVd: database of human endogenous retroviruses. Nucleic Acids Res. **30**:205–206.

77. **Paces J**, **Pavlícek A**, **Zika R**, **Kapitonov V V**, **Jurka J**, **Paces V**. 2004. HERVd: the Human Endogenous RetroViruses database: update. Nucleic Acids Res. **32**:D50.

78. **Katzourakis A**, **Tristem M**. 2005. Phylogeny of human endogenous and exogenous retroviruses, p. 186–239. *In* Sverdlov, ED (ed.), Retroviruses and Primate Genome Evolution.

79. **Aswad A**, **Katzourakis A**. 2014. The first endogenous herpesvirus, identified in the tarsier genome, and novel sequences from primate rhadinoviruses and lymphocryptoviruses. PLoS Genet. **10**:e1004332.

80. **Belle E**, **Beckage NE**, **Rousselet J**, **Poirie M**, **Lemeunier F**, **Drezen J-M**. 2002. Visualization of polydnavirus sequences in a parasitoid wasp chromosome. J. Virol. **76**:5793–5796.

81. **Stoltz DB**. 1990. Evidence for chromosomal transmission of polydnavirus DNA. J. Gen. Virol. **71**:1051–1056.

82. **Deng L**, **Stoltz DB**, **Webb BA**. 2000. A gene encoding a polydnavirus structural polypeptide is not encapsidated. Virology **269**:440–450.

83. **Bézier A**, **Annaheim M**, **Herbinière J**, **Wetterwald C**, **Gyapay G**, **Bernard-Samain S**, **Wincker P**, **Roditi I**, **Heller M**, **Belghazi M**, **Pfister-Wilhem R**, **Periquet G**, **Dupuy C**, **Huguet E**, **Volkoff A-N**, **Lanzrein B**, **Drezen J-M**. 2009. Polydnaviruses of braconid wasps derive from an ancestral nudivirus. Science **323**:926–30.

84. **Bézier A**, **Herbinière J**, **Lanzrein B**, **Drezen J-M**. 2009. Polydnavirus hidden face: the genes producing virus particles of parasitic wasps. J. Invertebr. Pathol. **101**:194–203.

85. **Delaroque N**, **Maier I**, **Knippers R**, **Müller DG**. 1999. Persistent virus integration into the genome of its algal host, *Ectocarpus siliculosus* (Phaeophyceae). J. Gen. Virol. **80**:1367–1370.

86. **Thézé J**, **Leclercq S**, **Moumen B**, **Cordaux R**, **Gilbert C**. 2014. Remarkable diversity of endogenous viruses in a crustacean genome. Genome Biol. Evol. **6**:2129–2140.

87. **Tang KFJ**, **Lightner D V**. 2006. Infectious hypodermal and hematopoietic necrosis virus (IHHNV)-related sequences in the genome of the black tiger prawn Penaeus monodon from Africa and Australia. Virus Res. **118**:185–191.

88. **Liu H**, **Fu Y**, **Xie J**, **Cheng J**, **Ghabrial SA**, **Li G**, **Peng Y**, **Yi X**, **Jiang D**. 2011. Widespread endogenization of densoviruses and parvoviruses in animal and human genomes. J. Virol. **85**:9863–9876.

89. **Belyi VA**, **Levine AJ**, **Skalka AM**. 2010. Sequences from ancestral single-stranded DNA viruses in vertebrate genomes: the *Parvoviridae* and *Circoviridae* are more than 40 to 50 million years old. J. Virol. **84**:12458–12462.

90. **Kapoor A**, **Simmonds P**, **Lipkin WI**. 2010. Discovery and characterization of mammalian endogenous parvoviruses. J. Virol. **84**:12628–12635.

91. **Gilbert C**, **Meik JM**, **Dashevsky D**, **Card DC**, **Castoe TA**, **Schaack S**. 2014. Endogenous hepadnaviruses, bornaviruses and circoviruses in snakes. Proc. Biol. Sci. **281**:20141122.

92. **Liu H**, **Fu Y**, **Li B**, **Yu X**, **Xie J**, **Cheng J**, **Ghabrial SA**, **Li G**, **Yi X**, **Jiang D**. 2011. Widespread horizontal gene transfer from circular single-stranded DNA viruses to

eukaryotic genomes. BMC Evol. Biol. **11**:276.

93. **Bejarano ER**, **Khashoggi A**, **Witty M**, **Lichtenstein C**. 1996. Integration of multiple repeats of geminiviral DNA into the nuclear genome of tobacco during evolution. Proc. Natl. Acad. Sci. U. S. A. **93**:759–764.

94. **Ashby MK**, **WaZrry A**, **Bejarano ER**, **Khashoggi A**, **Burrell M**, **Lichtenstein CP**. 1997. Analysis of multiple copies of geminiviral DNA in the genome of four closely related *Nicotiana* species suggest a unique integration event. Plant Mol. Biol. **35**:313–321.

95. **Lefeuvre P**, **Harkins GW**, **Lett J-M**, **Briddon RW**, **Chase MW**, **Moury B**, **Martin DP**. 2011. Evolutionary time-scale of the begomoviruses: evidence from integrated sequences in the *Nicotiana* genome. PLoS One **6**:e19193.

96. **Chiba S**, **Kondo H**, **Tani A**, **Saisho D**, **Sakamoto W**, **Kanematsu S**, **Suzuki N**. 2011. Widespread endogenization of genome sequences of non-retroviral rna viruses into plant genomes. PLoS Pathog. **7**:e1002146.

97. **Liu H**, **Fu Y**, **Jiang D**, **Li G**, **Xie J**, **Cheng J**, **Peng Y**, **Ghabrial SA**, **Yi X**. 2010. Widespread horizontal gene transfer from double-stranded RNA viruses to eukaryotic nuclear genomes. J. Virol. **84**:11876–11887.

98. **Liu H**, **Fu Y**, **Xie J**, **Cheng J**, **Ghabrial SA**, **Li G**, **Yi X**, **Jiang D**. 2012. Discovery of novel dsRNA viral sequences by *in silico* cloning and implications for viral diversity, host range and evolution. PLoS One **7**:e42147.

99. **Taylor DJ**, **Bruenn J**. 2009. The evolution of novel fungal genes from non-retroviral RNA viruses. BMC Biol. **7**:88.

100. **Frank AC**, **Wolfe KH**. 2009. Evolutionary capture of viral and plasmid DNA by yeast nuclear chromosomes. Eukaryot. Cell **8**:1521–1531.

101. **Crochu S**, **Cook S**, **Attoui H**, **Charrel RN**, **De Chesse R**, **Belhouchet M**, **Lemasson JJ**, **de Micco P**, **de Lamballerie X**. 2004. Sequences of flavivirus-related RNA viruses persist in DNA form integrated in the genome of *Aedes* spp. mosquitoes. J. Gen. Virol. **85**:1971–1980.

102. **Roiz D**, **Vázquez A**, **Seco MPS**, **Tenorio A**, **Rizzoli A**. 2009. Detection of novel insect

flavivirus sequences integrated in *Aedes albopictus* (Diptera: Culicidae) in Northern Italy. Virol. J. **6**:93.

103. **Tanne E**, **Sela I**. 2005. Occurrence of a DNA sequence of a non-retro RNA virus in a host plant genome and its expression: evidence for recombination between viral and host RNAs. Virology **332**:614–622.

104. **Ballinger MJ**, **Bruenn JA**, **Kotov AA**, **Taylor DJ**. 2013. Selectively maintained paleoviruses in Holarctic water fleas reveal an ancient origin for phleboviruses. Virology **446**:276–282.

105. **Horie M**, **Honda T**, **Suzuki Y**, **Kobayashi Y**, **Daito T**, **Oshida T**, **Ikuta K**, **Jern P**, **Gojobori T**, **Coffin JM**, **Tomonaga K**. 2010. Endogenous non-retroviral RNA virus elements in mammalian genomes. Nature **463**:84–87.

106. **Horie M**, **Kobayashi Y**, **Suzuki Y**, **Tomonaga K**. 2013. Comprehensive analysis of endogenous bornavirus-like elements in eukaryote genomes. Philos. Trans. R. Soc. Lond. B. Biol. Sci. **368**:20120499.

107. **Belyi VA**, **Levine AJ**, **Skalka AM**. 2010. Unexpected inheritance: multiple integrations of ancient bornavirus and ebolavirus/marburgvirus sequences in vertebrate genomes. PLoS Pathog. **6**:e1001030.

108. **Taylor DJ**, **Leach RW**, **Bruenn J**. 2010. Filoviruses are ancient and integrated into mammalian genomes. BMC Evol. Biol. **10**:193.

109. **Taylor DJ**, **Ballinger MJ**, **Zhan JJ**, **Hanzly LE**, **Bruenn JA**. 2014. Evidence that ebolaviruses and cuevaviruses have been diverging from marburgviruses since the Miocene. PeerJ **2**:e556.

110. **Taylor DJ**, **Dittmar K**, **Ballinger MJ**, **Bruenn JA**. 2011. Evolutionary maintenance of filovirus-like genes in bat genomes. BMC Evol. Biol. **11**:336.

111. **Astrin SM**, **Robinson HL**, **Crittenden LB**, **Buss EG**, **Wyban J**, **Hayward WS**. 1980. Ten genetic loci in the chicken that contain structural genes for endogenous avian leukosis viruses. Cold Spring Harb. Symp. Quant. Biol. **44**:1105–1109.

112. **Frisby DP**, **Weiss RA**, **Roussel M**, **Stehelin D**. 1979. The distribution of endogenous chicken retrovirus sequences in the DNA of galliform birds does not coincide with avian

phylogenetic relationships. Cell **17**:623–634.

113. **Bolisetty M**, **Blomberg J**, **Benachenhou F**, **Sperber G**, **Beemon K**. 2012. Unexpected diversity and expression of avian endogenous retroviruses. MBio **3**:e00344-12.

114. **Herniou E**, **Martin J**, **Miller K**, **Cook J**, **Wilkinson M**, **Tristem M**. 1998. Retroviral diversity and distribution in vertebrates. J. Virol. **72**:5955–5966.

115. **Gifford R**, **Kabat P**, **Martin J**, **Lynch C**, **Tristem M**. 2005. Evolution and distribution of class II-related endogenous retroviruses. J. Virol. **79**:6478–6486.

116. **Baillie GJ**, **van de Lagemaat LN**, **Baust C**, **Mager DL**. 2004. Multiple groups of endogenous betaretroviruses in mice, rats, and other mammals. J. Virol. **78**:5784–5798.

117. **van der Kuyl AC**. 2011. Characterization of a full-length endogenous beta-retrovirus, EqERV-beta1, in the genome of the horse (*Equus caballus*). Viruses **3**:620–628.

118. **Ericsson T**, **Oldmixon B**, **Blomberg J**, **Rosa M**, **Patience C**, **Andersson G**. 2001. Identification of novel porcine endogenous betaretrovirus sequences in miniature swine. J. Virol. **75**:2765–2770.

119. **Patience C**, **Switzer WM**, **Takeuchi Y**, **Griffiths DJ**, **Goward ME**, **Heneine W**, **Stoye JP**, **Weiss RA**. 2001. Multiple groups of novel retroviral genomes in pigs and related species. J. Virol. **75**:2771–2775.

120. **Hayward J a**, **Tachedjian M**, **Cui J**, **Field H**, **Holmes EC**, **Wang L-F**, **Tachedjian G**. 2013. Identification of diverse full-length endogenous betaretroviruses in megabats and microbats. Retrovirology **10**:35.

121. **Baba K**, **Nakaya Y**, **Shojima T**, **Muroi Y**, **Kizaki K**, **Hashizume K**, **Imakawa K**, **Miyazawa T**. 2011. Identification of novel endogenous betaretroviruses which are transcribed in the bovine placenta. J. Virol. **85**:1237–1245.

122. **Arnaud F**, **Varela M**, **Spencer TE**, **Palmarini M**. 2008. Endogenous retroviruses: coevolution of endogenous betaretroviruses of sheep and their host. Cell. Mol. Life Sci. **65**:3422–3432.

123. **Mayer J**, **Tsangaras K**, **Heeger F**, **Ávila-Arcos M**, **Stenglein MD**, **Chen W**, **Sun W**, **Mazzoni CJ**, **Osterrieder N**, **Greenwood AD**. 2013. A novel endogenous betaretrovirus group characterized from polar bears (*Ursus maritimus*) and giant pandas (*Ailuropoda*

melanoleuca). Virology **443**:1–10.

124. **Baillie GJ**, **Wilkins RJ**. 2001. Endogenous type D retrovirus in a marsupial, the common brushtail possum (*Trichosurus vulpecula*). J. Virol. **75**:2499–2507.

125. **Lavie L**, **Medstrand P**, **Schempp W**, **Meese E**, **Mayer J**. 2004. Human endogenous retrovirus family HERV-K(HML-5): status, evolution, and reconstruction of an ancient betaretrovirus in the human genome. J. Virol. **78**:8788–8798.

126. **Cui J**, **Tachedjian G**, **Tachedjian M**, **Holmes EC**, **Zhang S**, **Wang L-F**. 2012. Identification of diverse groups of endogenous gammaretroviruses in mega- and microbats. J. Gen. Virol. **93**:2037–2045.

127. **Tarlinton RE**, **Barfoot HKR**, **Allen CE**, **Brown K**, **Gifford RJ**, **Emes RD**. 2013. Characterisation of a group of endogenous gammaretroviruses in the canine genome. Vet. J. **196**:28–33.

128. **Wang L**, **Yin Q**, **He G**, **Rossiter SJ**, **Holmes EC**, **Cui J**. 2013. Ancient invasion of an extinct gammaretrovirus in cetaceans. Virology **441**:66–69.

129. **Cui J**, **Tachedjian M**, **Wang L**, **Tachedjian G**, **Wang L-F**, **Zhang S**. 2012. Discovery of retroviral homologs in bats: implications for the origin of mammalian gammaretroviruses. J. Virol. **86**:4288–4293.

130. **Lamere SA**, **St Leger JA**, **Schrenzel MD**, **Anthony SJ**, **Rideout BA**, **Salomon DR**. 2009. Molecular characterization of a novel gammaretrovirus in killer whales (*Orcinus orca*). J. Virol. **83**:12956–12967.

131. **Elleder D**, **Kim O**, **Padhi A**, **Bankert JG**, **Simeonov I**, **Schuster SC**, **Wittekindt NE**, **Motameny S**, **Poss M**. 2012. Polymorphic integrations of an endogenous gammaretrovirus in the mule deer genome. J. Virol. **86**:2787–2796.

132. **Anai Y**, **Ochi H**, **Watanabe S**, **Nakagawa S**, **Kawamura M**, **Gojobori T**, **Nishigaki K**. 2012. Infectious endogenous retroviruses in cats and emergence of recombinant viruses. J. Virol. **86**:8634–8644.

133. **Bamunusinghe D**, **Liu Q**, **Lu X**, **Oler A**, **Kozak C a**. 2013. Endogenous gammaretrovirus acquisition in *Mus musculus* subspecies carrying functional variants of the XPR1 virus receptor. J. Virol. **87**:9845–9855.

134. **Martin J**, **Herniou E**, **Cook J**, **O'Neill RW**, **Tristem M**. 1999. Interclass transmission and phyletic host tracking in murine leukemia virus-related retroviruses. J. Virol. **73**:2442–2449.

135. **Tristem M**, **Kabat P**, **Lieberman L**, **Linde S**, **Karpas A**, **Hill F**. 1996. Characterization of a novel murine leukemia virus-related subgroup within mammals. J. Virol. **70**:8241–8246.

136. **Hanger JJ**, **Bromham LD**, **McKee JJ**, **O'Brien TM**, **Robinson WF**. 2000. The nucleotide sequence of koala (*Phascolarctos cinereus*) retrovirus: a novel type C endogenous virus related to gibbon ape leukemia virus. J. Virol. **74**:4264–4272.

137. **Fiebig U**, **Hartmann MG**, **Bannert N**, **Kurth R**, **Denner J**. 2006. Transspecies transmission of the endogenous koala retrovirus. J. Virol. **80**:5651–5654.

138. **Jaratlerdsiri W**, **Rodríguez-Zárate CJ**, **Isberg SR**, **Damayanti CS**, **Miles LG**, **Chansue N**, **Moran C**, **Melville L**, **Gongora J**. 2009. Distribution of endogenous retroviruses in crocodilians. J. Virol. **83**:10305–10308.

139. **Brown K**, **Emes RD**, **Tarlinton RE**. 2014. Multiple groups of endogenous epsilon-like retroviruses conserved across primates. J. Virol. **88**:12464–12471.

140. **Sinzelle L**, **Carradec Q**, **Paillard E**, **Bronchain OJ**, **Pollet N**. 2011. Characterization of a *Xenopus tropicalis* endogenous retrovirus with developmental and stress-dependent expression. J. Virol. **85**:2167–2179.

141. **Kambol R**, **Kabat P**, **Tristem M**. 2003. Complete nucleotide sequence of an endogenous retrovirus from the amphibian, *Xenopus laevis*. Virology **311**:1–6.

142. **Basta HA**, **Cleveland SB**, **Clinton RA**, **Dimitrov AG**, **McClure MA**. 2009. Evolution of teleost fish retroviruses: characterization of new retroviruses with cellular genes. J. Virol. **83**:10152–10162.

143. **Katzourakis A**, **Tristem M**, **Pybus OG**, **Gifford RJ**. 2007. Discovery and analysis of the first endogenous lentivirus. Proc. Natl. Acad. Sci. U. S. A. **104**:6261–6265.

144. **Keckesova Z**, **Ylinen LMJ**, **Towers GJ**, **Gifford RJ**, **Katzourakis A**. 2009. Identification of a RELIK orthologue in the European hare (*Lepus europaeus*) reveals a minimum age of 12 million years for the lagomorph lentiviruses. Virology **384**:7–11.

145. **van der Loo W**, **Abrantes J**, **Esteves P**. 2009. Sharing of endogenous lentiviral gene fragments among leporid lineages separated for more than 12 million years. J. Virol. **83**:2386–2388.

146. **Gifford RJ**, **Katzourakis A**, **Tristem M**, **Pybus OG**, **Winters M**, **Shafer RW**. 2008. A transitional endogenous lentivirus from the genome of a basal primate and implications for lentivirus evolution. Proc. Natl. Acad. Sci. U. S. A. **105**:20362–20367.

147. **Gilbert C**, **Maxfield DG**, **Goodman SM**, **Feschotte C**. 2009. Parallel germline infiltration of a lentivirus in two Malagasy lemurs. PLoS Genet. **5**:e1000425.

148. **Han G-Z**, **Worobey M**. 2012. Endogenous lentiviral elements in the weasel family (*Mustelidae*). Mol. Biol. Evol. **29**:2905–2908.

149. **Cui J**, **Holmes EC**. 2012. Endogenous lentiviruses in the ferret genome. J. Virol. **86**:3383–3385.

150. **Katzourakis A**, **Gifford RJ**, **Tristem M**, **Gilbert MTP**, **Pybus OG**. 2009. Macroevolution of complex retroviruses. Science **325**:1512.

151. **Katzourakis A**, **Aiewsakun P**, **Jia H**, **Wolfe ND**, **LeBreton M**, **Yoder AD**, **Switzer WM**. 2014. Discovery of prosimian and afrotherian foamy viruses and potential cross species transmissions amidst stable and ancient mammalian co-evolution. Retrovirology **11**:61.

152. **Han G-Z**, **Worobey M**. 2012. An endogenous foamy virus in the aye-aye (*Daubentonia madagascariensis*). J. Virol. **86**:7696–7698.

153. **Gilbert C**, **Feschotte C**. 2010. Genomic fossils calibrate the long-term evolution of hepadnaviruses. PLoS Biol. **8**:e1000495.

154. **Liu W**, **Pan S**, **Yang H**, **Bai W**, **Shen Z**, **Liu J**, **Xie Y**. 2012. The first full-length endogenous hepadnaviruses: identification and analysis. J. Virol. **86**:9510–9513.

155. **Suh A**, **Brosius J**, **Schmitz J**, **Kriegs JO**. 2013. The genome of a Mesozoic paleovirus reveals the evolution of hepatitis B viruses. Nat. Commun. **4**:1791.

156. **Cui J**, **Holmes EC**. 2012. Endogenous hepadnaviruses in the genome of the budgerigar (*Melopsittacus undulatus*) and the evolution of avian hepadnaviruses. J. Virol. **86**:7688–7691.

157. **Suh A**, **Weber CC**, **Kehlmaier C**, **Braun EL**, **Green RE**, **Fritz U**, **Ray DA**, **Ellegren H**.

2014. Early Mesozoic coexistence of amniotes and hepadnaviridae. PLoS Genet. **10**:e1004559.

158. **Staginnus C**, **Richert-Pöggeler KR**. 2006. Endogenous pararetroviruses: two-faced travelers in the plant genome. Trends Plant Sci. **11**:485–491.

159. **Iskra-Caruana M-L**, **Baurens F-C**, **Gayral P**, **Chabannes M**. 2010. A four-partner plant–virus interaction: enemies can also come from within. Mol. Plant. Microbe. Interact. **23**:1394–1402.

160. **Bertsch C**, **Beuve M**, **Dolja V V**, **Wirth M**, **Pelsy F**, **Herrbach E**, **Lemaire O**. 2009. Retention of the virus-derived sequences in the nuclear genome of grapevine as a potential pathway to virus resistance. Biol. Direct **4**:21.

161. **Gayral P**, **Noa-Carrazana J-C**, **Lescot M**, **Lheureux F**, **Lockhart BEL**, **Matsumoto T**, **Piffanelli P**, **Iskra-Caruana M-L**. 2008. A single Banana streak virus integration event in the banana genome as the origin of infectious endogenous pararetrovirus. J. Virol. **82**:6697–6710.

162. **Kunii M**, **Kanda M**, **Nagano H**, **Uyeda I**, **Kishima Y**, **Sano Y**. 2004. Reconstruction of putative DNA virus from endogenous rice tungro bacilliform virus-like sequences in the rice genome: implications for integration and evolution. BMC Genomics **5**:80.

163. **Bill CA**, **Summers J**. 2004. Genomic DNA double-strand breaks are targets for hepadnaviral DNA integration. Proc. Natl. Acad. Sci. U. S. A. **101**:11135–11140.

164. **Young SM**, **Samulski RJ**. 2001. Adeno-associated virus (AAV) site-specific recombination does not require a Rep-dependent origin of replication within the AAV terminal repeat. Proc. Natl. Acad. Sci. U. S. A. **98**:13525–13530.

165. **Kotin RM**, **Linden RM**, **Berns KI**. 1992. Characterization of a preferred site on human chromosome 19q for integration of adeno-associated virus DNA by non-homologous recombination. EMBO J. **11**:5071–5078.

166. **Urcelay E**, **Ward P**, **Wiener S**, **Safer B**, **Kotin R**. 1995. Asymmetric replication in vitro from a human sequence element is dependent on adeno-associated virus Rep protein. J. Virol. **69**:2038–2046.

167. **Morissette G**, **Flamand L**. 2010. Herpesviruses and chromosomal integration. J. Virol.

84:12100–12109.

168. **Holmes EC**. 2011. The evolution of endogenous viral elements. Cell Host Microbe **10**:368–377.

169. **Geuking MB**, **Weber J**, **Dewannieux M**, **Gorelik E**, **Heidmann T**, **Hengartner H**, **Zinkernagel RM**, **Hangartner L**. 2009. Recombination of retrotransposon and exogenous RNA virus results in nonretroviral cDNA integration. Science **323**:393–396.

170. **Gayral P**, **Iskra-Caruana ML**. 2009. Phylogeny of banana streak virus reveals recent and repetitive endogenization in the genome of its banana host (*Musa* sp.). J. Mol. Evol. **69**:65–80.

171. **Laminger F**, **Prinz A**. 2010. Bats and other reservoir hosts of *Filoviridae*. Danger of epidemic on the African continent?--a deductive literature analysis. Wien. Klin. Wochenschr. **122 Suppl**:19–30.

172. **Calisher CH**, **Childs JE**, **Field HE**, **Holmes K V.**, **Schountz T**. 2006. Bats: important reservoir hosts of emerging viruses. Clin. Microbiol. Rev. **19**:531–545.

173. **Chantrey J**, **Meyer H**, **Baxby D**, **Begon M**, **Bown KJ**, **Hazel SM**, **Jones T**, **Montgomery WI**, **Bennett M**. 1999. Cowpox: reservoir hosts and geographic range. Epidemiol. Infect. **122**:455–460.

174. **Daniels PW**, **Halpin K**, **Hyatt A**, **Middleton D**. 2007. Infection and disease in reservoir and spillover hosts: determinants of pathogen emergence. Curr. Top. Microbiol. Immunol. **315**:113–131.

175. **Vetten HJ**. 2008. Nanoviruses, p. 385–391. *In* Mahy, BWJ, Van Regenmortel, MHV (eds.), Encyclopedia of Virology, 3rd ed. Elsevier, Oxford.

176. **Cotmore SF**, **Agbandje-McKenna M**, **Chiorini JA**, **Mukha D V**, **Pintel DJ**, **Qiu J**, **Soderlund-Venermo M**, **Tattersall P**, **Tijssen P**, **Gatherer D**, **Davison AJ**. 2014. The family *Parvoviridae*. Arch. Virol. **159**:1239–1247.

177. **Martins H**, **Villesen P**. 2011. Improved integration time estimation of endogenous retroviruses with phylogenetic data. PLoS One **6**:e14745.

178. **Tristem M**. 2000. Identification and characterization of novel human endogenous retrovirus families by phylogenetic screening of the human genome mapping project

database. J. Virol. **74**:3715–3730.

179. **McGeoch DJ**, **Cook S**, **Dolan A**, **Jamieson FE**, **Telford EAR**. 1995. Molecular phylogeny and evolutionary timescale for the family of mammalian herpesviruses. J. Mol. Biol. **247**:443–458.

180. **McGeoch DJ**, **Rixon FJ**, **Davison AJ**. 2006. Topics in herpesvirus genomics and evolution. Virus Res. **117**:90–104.

181. **Switzer WM**, **Salemi M**, **Shanmugam V**, **Gao F**, **Cong M-E**, **Kuiken C**, **Bhullar V**, **Beer BE**, **Vallet D**, **Gautier-Hion A**, **Tooze Z**, **Villinger F**, **Holmes EC**, **Heneine W**. 2005. Ancient co-speciation of simian foamy viruses and primates. Nature **434**:376–380.

182. **Kuzmin I V.**, **Niezgoda M**, **Franka R**, **Agwanda B**, **Markotter W**, **Breiman RF**, **Shieh WJ**, **Zaki SR**, **Rupprecht CE**. 2010. Marburg virus in fruit bat, Kenya. Emerg. Infect. Dis. **16**:352–354.

183. **Leroy EM**, **Epelboin A**, **Mondonge V**, **Pourrut X**, **Gonzalez J-P**, **Muyembe-Tamfum J-J**, **Formenty P**. 2009. Human Ebola outbreak resulting from direct exposure to fruit bats in Luebo, Democratic Republic of Congo, 2007. Vector Borne Zoonotic Dis. **9**:723–728.

184. **Peterson AT**, **Bauer JT**, **Mills JN**. 2004. Ecologic and geographic gistribution of filovirus disease. Emerg. Infect. Dis. **10**:40–47.

185. **Towner JS**, **Amman BR**, **Sealy TK**, **Reeder Carroll SA**, **Comer JA**, **Kemp A**, **Swanepoel R**, **Paddock CD**, **Balinandi S**, **Khristova ML**, **Formenty PBH**, **Albarino CG**, **Miller DM**, **Reed ZD**, **Kayiwa JT**, **Mills JN**, **Cannon DL**, **Greer PW**, **Byaruhanga E**, **Farnon EC**, **Atimnedi P**, **Okware S**, **Katongole-Mbidde E**, **Downing R**, **Tappero JW**, **Zaki SR**, **Ksiazek TG**, **Nichol ST**, **Rollin PE**. 2009. Isolation of genetically diverse Marburg viruses from Egyptian fruit bats. PLoS Pathog. **5**:e1000536.

186. **Leroy EM**, **Kumulungui B**, **Pourrut X**, **Rouquet P**, **Hassanin A**, **Yaba P**, **Délicat A**, **Paweska JT**, **Gonzalez J-P**, **Swanepoel R**. 2005. Fruit bats as reservoirs of Ebola virus. Nature **438**:575–576.

187. **Pourrut X**, **Kumulungui B**, **Wittmann T**, **Moussavou G**, **Délicat A**, **Yaba P**, **Nkoghe D**, **Gonzalez JP**, **Leroy EM**. 2005. The natural history of Ebola virus in Africa. Microbes

Infect. **7**:1005–1014.

188. **Peterson AT**, **Carroll DS**, **Mills JN**, **Johnson KM**. 2004. Potential mammalian filovirus reservoirs. Emerg. Infect. Dis. **10**:2073–2081.

189. **Tattersall I**. 2006. Historical biogeography of the strepsirhine primates of Madagascar. Folia Primatol. (Basel). **77**:477–487.

190. **Teeling EC**, **Springer MS**, **Madsen O**, **Bates P**, **O'brien SJ**, **Murphy WJ**. 2005. A molecular phylogeny for bats illuminates biogeography and the fossil record. Science **307**:580–584.

191. **Aitchison JC**, **Ali JR**, **Davis AM**. 2007. When and where did India and Asia collide? J. Geophys. Res. **112**:B05423.

192. **Rambaut A**. 2000. Estimating the rate of molecular evolution: incorporating non-contemporaneous sequences into maximum likelihood phylogenies. Bioinformatics **16**:395–399.

193. **Drummond A**, **Rodrigo AG**. 2000. Reconstructing genealogies of serial samples under the assumption of a molecular clock using serial-sample UPGMA. Mol. Biol. Evol. **17**:1807–1815.

194. **Drummond A**, **Forsberg R**, **Rodrigo AG**. 2001. The inference of stepwise changes in substitution rates using serial sequence samples. Mol. Biol. Evol. **18**:1365–1371.

195. **Drummond AJ**, **Nicholls GK**, **Rodrigo AG**, **Solomon W**. 2002. Estimating mutation parameters, population history and genealogy simultaneously from temporally spaced sequence data. Genetics **161**:1307–1320.

196. **Bernard HU**. 1994. Coevolution of papillomaviruses with human populations. Trends Microbiol. **2**:140–143.

197. **McGeoch DJ**, **Gatherer D**. 2005. Integrating reptilian herpesviruses into the family *Herpesviridae*. J. Virol. **79**:725–731.

198. **McGeoch DJ**, **Cook S**. 1994. Molecular phylogeny of the *Alphaherpesvirinae* subfamily and a proposed evolutionary timescale. J. Mol. Biol. **238**:9–22.

199. **Sakaoka H**, **Kurita K**, **Iida Y**, **Takada S**, **Umene K**, **Kim YT**, **Ren CS**, **Nahmias AJ**. 1994. Quantitative analysis of genomic polymorphism of herpes simplex virus type 1

strains from six countries: studies of molecular evolution and molecular epidemiology of the virus. J. Gen. Virol. **75**:513–527.

200. **Lemey P**, **Pybus OG**, **Van Dooren S**, **Vandamme A-M**. 2005. A Bayesian statistical analysis of human T-cell lymphotropic virus evolutionary rates. Infect. Genet. Evol. **5**:291–298.

201. **Sharp PM**, **Li WH**. 1988. Understanding the origins of AIDS viruses. Nature **336**:315.

202. **Yokoyama S**, **Chung L**, **Gojobori T**. 1988. Molecular evolution of the human immunodeficiency and related viruses. Mol. Biol. Evol. **5**:237–251.

203. **Orito E**, **Mizokami M**, **Ina Y**, **Moriyama EN**, **Kameshima N**, **Yamamoto M**, **Gojobori T**. 1989. Host-independent evolution and a genetic classification of the hepadnavirus family based on nucleotide sequences. Proc. Natl. Acad. Sci. U. S. A. **86**:7059–7062.

204. **Van Hemert FJ**, **van de Klundert MAA**, **Lukashov V V.**, **Kootstra NA**, **Berkhout B**, **Zaaijer HL**. 2011. Protein X of hepatitis B virus: origin and structure similarity with the central domain of DNA glycosylase. PLoS One **6**:e23392.

205. **Suzuki Y**, **Gojobori T**. 1997. The origin and evolution of Ebola and Marburg viruses. Mol. Biol. Evol. **14**:800–806.

206. **Carroll SA**, **Towner JS**, **Sealy TK**, **McMullan LK**, **Khristova ML**, **Burt FJ**, **Swanepoel R**, **Rollin PE**, **Nichol ST**. 2013. Molecular evolution of viruses of the family *Filoviridae* based on 97 whole-genome sequences. J. Virol. **87**:2608–2616.

207. **Negredo A**, **Palacios G**, **Vázquez-Morón S**, **González F**, **Dopazo H**, **Molero F**, **Juste J**, **Quetglas J**, **Savji N**, **de la Cruz Martínez M**, **Herrera JE**, **Pizarro M**, **Hutchison SK**, **Echevarría JE**, **Lipkin WI**, **Tenorio A**. 2011. Discovery of an ebolavirus-like filovirus in Europe. PLoS Pathog. **7**:e1002304.

208. **Firth C**, **Charleston MA**, **Duffy S**, **Shapiro B**, **Holmes EC**. 2009. Insights into the evolutionary history of an emerging livestock pathogen: porcine circovirus 2. J. Virol. **83**:12813–12821.

209. **Lam TTY**, **Liu W**, **Bowden TA**, **Cui N**, **Zhuang L**, **Liu K**, **Zhang YY**, **Cao WC**, **Pybus OG**. 2013. Evolutionary and molecular analysis of the emergent severe fever with thrombocytopenia syndrome virus. Epidemics **5**:1–10.

210. **Bird BH**, **Khristova ML**, **Rollin PE**, **Ksiazek TG**, **Nichol ST**. 2007. Complete genome analysis of 33 ecologically and biologically diverse Rift Valley fever virus strains reveals widespread virus movement and low genetic diversity due to recent common ancestry. J. Virol. **81**:2805–2816.

211. **Lemey P**, **Rambaut A**, **Pybus OG**. 2006. HIV evolutionary dynamics within and among hosts. AIDS Rev. **8**:125–140.

212. **Davis PL**, **Holmes EC**, **Larrous F**, **Van der Poel WHM**, **Tjørnehøj K**, **Alonso WJ**, **Bourhy H**. 2005. Phylogeography, population dynamics, and molecular evolution of European bat lyssaviruses. J. Virol. **79**:10487–10497.

213. **Shackelton LA**, **Parrish CR**, **Truyen U**, **Holmes EC**. 2005. High rate of viral evolution associated with the emergence of carnivore parvovirus. Proc. Natl. Acad. Sci. U. S. A. **102**:379–384.

214. **Shackelton LA**, **Holmes EC**. 2006. Phylogenetic evidence for the rapid evolution of human B19 erythrovirus. J. Virol. **80**:3666–3669.

215. **Duffy S**, **Holmes EC**. 2008. Phylogenetic evidence for rapid rates of molecular evolution in the single-stranded DNA begomovirus tomato yellow leaf curl virus. J. Virol. **82**:957–965.

216. **Feschotte C**, **Gilbert C**. 2012. Endogenous viruses: insights into viral evolution and impact on host biology. Nat. Rev. Genet. **13**:283–296.

217. **Lemey P**, **Van Dooren S**, **Vandamme A-M**. 2005. Evolutionary dynamics of human retroviruses investigated through full-genome scanning. Mol. Biol. Evol. **22**:942–951.

218. **Aulicino PC**, **Holmes EC**, **Rocco C**, **Mangano A**, **Sen L**. 2007. Extremely rapid spread of human immunodeficiency virus type 1 BF recombinants in Argentina. J. Virol. **81**:427–429.

219. **Bello G**, **Aulicino PC**, **Ruchansky D**, **Guimarães ML**, **Lopez-Galindez C**, **Casado C**, **Chiparelli H**, **Rocco C**, **Mangano A**, **Sen L**, **Morgado MG**. 2010. Phylodynamics of HIV-1 circulating recombinant forms 12_BF and 38_BF in Argentina and Uruguay. Retrovirology **7**:22.

220. **Biek R**, **Rodrigo AG**, **Holley D**, **Drummond A**, **Anderson CR**, **Ross HA**, **Poss M**.

2003. Epidemiology, genetic diversity, and evolution of endemic feline immunodeficiency virus in a population of wild cougars. J. Virol. **77**:9578–9589.

221. **Duffy S**, **Shackelton LA**, **Holmes EC**. 2008. Rates of evolutionary change in viruses: patterns and determinants. Nat. Rev. Genet. **9**:267–276.

222. **Firth C**, **Kitchen A**, **Shapiro B**, **Suchard MA**, **Holmes EC**, **Rambaut A**. 2010. Using time-structured data to estimate evolutionary rates of double-stranded DNA viruses. Mol. Biol. Evol. **27**:2038–2051.

223. **Schweizer M**, **Schleer H**, **Pietrek M**, **Liegibel J**, **Falcone V**, **Neumann-Haefelin D**. 1999. Genetic stability of foamy viruses: long-term study in an African green monkey population. J. Virol. **73**:9256–9265.

224. **Willems L**, **Thienpont E**, **Kerkhofs P**, **Burny A**, **Mammerickx M**, **Kettmann R**. 1993. Bovine leukemia virus, an animal model for the study of intrastrain variability. J. Virol. **67**:1086–1089.

225. **Salemi M**, **Lewis M**, **Egan JF**, **Hall WW**, **Desmyter J**, **Vandamme AM**. 1999. Different population dynamics of human T cell lymphotropic virus type II in intravenous drug users compared with endemically infected tribes. Proc Natl Acad Sci U S A **96**:13253–13258.

226. **Vandamme AM**, **Bertazzoni U**, **Salemi M**. 2000. Evolutionary strategies of human T-cell lymphotropic virus type II. Gene **261**:171–180.

227. **Drake JW**, **Charlesworth B**, **Charlesworth D**, **Crow JF**. 1998. Rates of spontaneous mutation. Genetics **148**:1667–1686.

228. **Duchêne S**, **Holmes EC**, **Ho SYW**. 2014. Analyses of evolutionary dynamics in viruses are hindered by a time-dependent bias in rate estimates. Proc. Biol. Sci. **281**:20140732.

229. **Ho SYW**, **Lanfear R**, **Bromham L**, **Phillips MJ**, **Soubrier J**, **Rodrigo AG**, **Cooper A**. 2011. Time-dependent rates of molecular evolution. Mol. Ecol. **20**:3087–3101.

230. **Loogväli EL**, **Kivisild T**, **Margus T**, **Villems R**. 2009. Explaining the imperfection of the molecular clock of hominid mitochondria. PLoS One **4**:e8260.

231. **Clark AG**, **Whittam TS**. 1992. Sequencing errors and molecular evolutionary analysis. Mol. Biol. Evol. **9**:744–752.

232. **Ho SYW**, **Heupink TH**, **Rambaut A**, **Shapiro B**. 2007. Bayesian estimation of sequence

damage in ancient DNA. Mol. Biol. Evol. **24**:1416–1422.

233. **Johnson PLF**, **Slatkin M**. 2008. Accounting for bias from sequencing error in population genetic estimates. Mol. Biol. Evol. **25**:199–206.

234. **Ho SYW**, **Phillips MJ**, **Cooper A**, **Drummond AJ**. 2005. Time dependency of molecular rate estimates and systematic overestimation of recent divergence times. Mol. Biol. Evol. **22**:1561–1568.

235. **Garcia-Moreno J**. 2004. Is there a universal mtDNA clock for birds ? J. Avian Biol. **35**:465–468.

236. **Emerson BC**. 2007. Alarm bells for the molecular clock? No support for Ho et al.'s model of time-dependent molecular rate estimates. Syst. Biol. **56**:337–345.

237. **Sullivan J**, **Joyce P**. 2005. Model selection in phylogenetics. Annu. Rev. Ecol. Evol. Syst. **36**:445–466.

238. **Mishmar D**, **Ruiz-Pesini E**, **Golik P**, **Macaulay V**, **Clark AG**, **Hosseini S**, **Brandon M**, **Easley K**, **Chen E**, **Brown MD**, **Sukernik RI**, **Olckers A**, **Wallace DC**. 2003. Natural selection shaped regional mtDNA variation in humans. Proc. Natl. Acad. Sci. U. S. A. **100**:171–176.

239. **Ruiz-Pesini E**, **Mishmar D**, **Brandon M**, **Procaccio V**, **Wallace DC**. 2004. Effects of purifying and adaptive selection on regional variation in human mtDNA. Science **303**:223–226.

240. **Gatesy J**. 2007. A tenth crucial question regarding model use in phylogenetics. Trends Ecol. Evol. **22**:509–510.

241. **Soubrier J**, **Steel M**, **Lee MSY**, **Der Sarkissian C**, **Guindon S**, **Ho SYW**, **Cooper A**. 2012. The influence of rate heterogeneity among sites on the time dependence of molecular rates. Mol. Biol. Evol. **29**:3345–3358.

Chapter one: Ancient co-evolution of foamy viruses and their mammalian hosts

---

This chapter is modified from the following published paper (**Appendix I**):

Katzourakis A†, Aiewsakun P†, Jia H, Wolfe ND, LeBreton M, Yoder AD, Switzer WM. 2014. Discovery of prosimian and afrotherian foamy viruses and potential cross species transmissions amidst stable and ancient mammalian co-evolution. Retrovirology **11**:61.

†Contributed equally

The *Author contributions* in the publication state that "AK and WMS conceived the project, AK, PA carried out the analysis, HJ performed the laboratory work, and NDW, ML, ADY provided samples. AK, PA and BS wrote the paper. All authors read and commented on the manuscript."

For the material in this chapter, PA performed the genomic annotations, phylogenetic and evolutionary analyses. AK identified the endogenous foamy virus elements and their taxonomy based on preliminary analyses. WMS and his collaborators provided the exogenous galago sequence. AK provided supervision. PA wrote the chapter. AK and WMS provided editorial guidance.

1.1 Abstract

Foamy viruses (FVs) are an important retroviral macroevolutionary model, demonstrated to have a stable co-speciation history with their mammalian hosts. Nevertheless, little is known about how they co-evolve with prosimians and afrotherians, as their molecular data are not currently available. This study reports and describes the first complete genome of an extant prosimian FV (PSFV) isolated from a galago (PSFVgal), and a novel endogenous afrotherian FV present in the Cape golden mole genome (ChrEFV). The work presented here also further describes PSFVaye—the previously discovered aye-aye endogenous PSFV, which was only partially characterised. Phylogenetic analyses reveal an extended co-speciation history between mammals and FVs, dating back to the eutherian diversification ~100 million years ago. The analyses also infer a small number of ancient cross-species transmissions of FVs nevertheless, involving bat, aye-aye, and New World monkey FVs. The results from this study provide important insights into the diversity and evolutionary history of prehistoric FVs, as well as that of their hosts.

1.2 Introduction

Foamy viruses (FVs) are complex retroviruses (*Spumaretrovirinae* subfamily) that are extremely widespread among mammals, in particular laurasiatherian and euarchontoglires mammals, together known as boreoeutherian mammals. The host range of FVs includes apes (1–9), Old World monkeys (10–15), New World monkeys (NWMs) (16–24), prosimians (16), cats (25, 26), cattle (27, 28), horses (29), and bats (30) (**Table 1.1**). Moreover, similar to other retroviruses, FVs occasionally leave *viral genomic fossils* within their host genomes, known as endogenous FVs. However, unlike their retroviral relatives such as alpha- (31, 32), beta- (31–34) and gammaretroviruses (31, 32), FVs become endogenous at a much lower frequency in comparison. Despite the large availability of mammalian genomes and the broad range of FV mammalian hosts, only two endogenous FVs have been discovered thus far (**Table 1.1**), namely the sloth endogenous FV (SloEFV), which exhibits all characteristic features of extant mammalian FVs (35), and (ii) the aye-aye endogenous FV (PSFVaye), which has been only partially characterised (36). The discovery of SloEFV is of particular importance, as it expanded the host range of FVs from boreoeutherian mammals to also cover xenarthrans (35).

One of the unique evolutionary features of FVs that distinguishes them from other retroviruses is their extended and stable co-speciation history with their hosts. Analyses have shown that Old World monkey and ape FVs have been stably co-speciating with their hosts for >30 million years (Myr) since the origin of catarrhines (37). Analyses of SloEFV (35) supported and extended this co-speciation pattern further down to the boreoeutherian-xenarthran split, dating back >100 Myr ago (Ma) (38) (**Box 1.1**). Because of this stable and extended FV-host co-speciation pattern, coupled with the large availability of FV sequences from diverse mammals (**Table 1.1**), the history of FV evolution is arguably one of the best described retroviral evolutionary histories that we have. Nonetheless, sequence data from prosimian and afrotherian FVs are still missing from our record, and thus, how they evolve and interact with their hosts is still unclear at the present.

Table 1.1 Mammalian foamy viruses (FVs).

FVs†‡	Host species	Host group	Year first isolate (ref)	Year first full-genome available (ref)
Exogenous FV				
PFV*	Human (*Homo sapiens*)	Ape	1971 (1)	1987 (2), 1988 (3)
SFVcpz (SFV-6, SFV-7)	Chimpanzee (*Pan troglodytes verus*)		1967 (4)	1994 (5)
SFVgor	Gorilla (*Gorilla gorilla sp.*)		1995 (6)	2011 (7)
SFVora (SFV-11)	Orangutan (*Pongo pygmaeus*)		1994 (8)	2003 (9)
SFVmac (SFV-1, SFV-2)	Macaque (*Macaca mulatta, M. cyclopis*)	Old World monkey	1955 (10), 1961 (11)	1991 (12)
SFVagm (SFV-3)	African green monkey (*Cercopithecus aethiops*)		1964 (13)	1992 (14)
SFVbab (SFV-10)	Baboon (*Papio cynocephalus*)		1975 (15)	–
SFVsqu (SFV-4)	Squirrel monkey (*Saimiri sciureus*)	New World monkey	1971 (16)	2010 (17)
SFVspm (SFV-8)	Spider monkey (*Ateles sp.*)		1973 (18)	2007 (19)
SFVcap (SFV-9)	Capuchin (*Cebus sp.*)		1975 (20)	2015 (21)
SFVmar	Marmoset (*Callithrix jacchus*)		1981 (22)	2010 (17)
SFVowl	Owl monkey (*Aotus sp.*)		2013 (23)	-

Table 1.1 (Continue).

FVs†‡	Host species	Host group	Year first isolate (ref)	Year first full-genome available (ref)
SFVtam	Tamarin monkey (*Saguinus sp.*)	New World monkey	2013 (23)	–
SFVlio tam	Lion tamarin monkey (*Leontopithecus sp.*)		2013 (23)	–
SFVtit	Titi monkey (*Callicebus sp.*)		2013 (23)	–
SFVbea sak	Bearded saki monkey (*Chiropotes sp.*)		2013 (23)	–
SFVhow	Howler monkey (*Alouatta sp.*)		2013 (23)	–
SFVwoo	Woolly monkey (*Lagothrix sp.*)		2015 (24)	–
SFVuak	Uakari monkey (*Cacajao sp.*)		2015 (24)	–
SFVsak	Saki monkey (*Pithecia sp.*)		2015 (24)	–
PSFVgal (SFV-5)	Galago (*Otolemur crassicaudatus panganiensis*)	Prosimian	1971 (16)	–
FFV	Domestic cat (*Felis catus*)	Feline	1969 (25)	1997 (26)
BFV	Cow (*Bos taurus*)	Bovine	1969 (27)	1994 (28)
EFV	Horse (*Equus ferus caballus)*	Equine	2000 (29)	2000 (29)
RhiFV	Bat (*Rhinolophus affinis)*	Bat	2012 (30)	–

Table 1.1 (Continue).

FVs†‡	Host species	Host group	Year first isolate (ref)	Year first full-genome available (ref)
Endogenous FV				
PSFVaye	Aye-aye (*Daubentonia madagascariensis*)	Prosimian	2012 (36)	-
SloEFV	Two-toed sloth (*Choloepus hoffmanni*)	Xenarthran	2009 (35)	2009 (35)

† Acronyms used in FV names: PFV, prototype FV; SFVcpz, chimpanzee simian FV (SFV); SFVgor, gorilla SFV; SFVora, orangutan SFV; SFVmac, macaque SFV; SFVagm, African green money SFV; SFVbab, baboon SFV; SFVsqu, squirrel monkey SFV; SFVspm, spider monkey SFV; SFVcap, capuchin SFV; SFVmar, common marmoset SFV; SFVowl, owl monkey SFV; SFVtam, tamarin monkey SFV; SFVlio tam, lion tamarin monkey SFV; SFVtit, titi monkey SFV; SFVbea saki, bearded saki monkey SFV; SFVhow, howler monkey SFV; SFVwoo, woolly monkey SFV; SFVuak, uakari monkey SFV; SFVsak, saki monkey SFV; PSFVgal, galago prosimian FV; FFV, feline FV; BFV, bovine FV; EFV, equine FV; RhiFV, *Rhinolophus* FV; PSFVaye, aye-aye prosimian FV; SloEFV, sloth endogenous FV.

‡ Names within brackets are old nomenclatures based on serotyping, which are not used anymore.

* Although the prototype FV (PFV) was isolated from a human and was originally called a human FV (HFV), it has now been well established that it is in fact a variant FV of chimpanzee origin, most closely related to, and clustering well, within the clade of FVs from *Pan troglodytes schweinfurthii* (39–41). Phylogenetic analyses revealed evidence for coevolution of chimpanzee hosts and their specific FVs at the subspecies level (39, 41). Therefore, PFV is here treated as a *Pan troglodytes schweinfurthii* FV, which diverged from SFVcpz about 0.96 million years ago, inferred under the FV-host co-speciation assumption using the divergence date of *Pan troglodytes schweinfurthii* and *Pan troglodytes verus* (42).

Box 1.1 Early history of mammalian evolution.

Phylogenetic analyses of mammalian genes consistently agree on the division of eutherians into four principle clades, namely Afrotheria, Xenarthra, Euarchontogilres, and Laurasiatheria, with the last two forming the clade called Boreoeutheria (see (43)). However, the exact placement of the root among these clades is ambiguous. Three main competing hypotheses have been proposed:

i) Atlantogenata hypothesis: Afrotheria and Xenarthra are sister groups forming the Atlantogenata clade, with the most basal split being that of Boreoeutheria and Atlantogenata. This hypothesis is supported by limited genetic evidence (44, 45);

ii) Epitherian hypothesis: Boreoeutheria and Afrotheria are sister groups forming the Epitheria clade, where Xenarthra is at the base of the Eutherian tree. This hypothesis is supported by morphological features and retroposon analyses (46, 47);

iii) Exafroplacentalia hypothesis: Boreoeutheria and Xenarthra are sister groups forming the Exafroplacentalia clade, and the Exafroplacentalia-Afrotheria split is the most basal split of eutherians. This hypothesis is supported by molecular studies of large data sets (38, 48, 49).

The Exafroplacentalia hypothesis has received the most extensive molecular support, and thus we use this hypothesis in the interpretation of our phylogenetic analyses.

In this study, we report and describe the first complete genome of an extant prosimian FV (PSFV) isolated from a lorisiforme galago (*Otolemur crassicaudatus panganiensis*) (PSFVgal), and the first afrotherian endogenous FV present in the Cape golden mole (*Chrysochloris asiatica*) genome (ChrEFV). We also further describe the endogenous PSFVaye, which has been only partially characterised (36). Together with other mammalian FVs, we reexamine the coevolutionary history of FVs and their mammalian hosts.

1.3 Results

1.3.1 Characterisation of the complete genome of the extant PSFVgal from a galago

The complete genome of PSFVgal (GenBank accession number KM233624) was provided by William Switzer (Centers for Disease Control and Prevention, personal communication). Genomic inspection shows that PSFVgal exhibits all the typical structural features of mammalian FV genomes, including the presence of *gag*, *pol*, *env*, *bel1*, and *bel2* genes flanked by long terminal repeats (LTRs) (**Figure 1.1**), supporting that it is a FV. There are no in-frame stop codons identified in the PSFVgal coding regions, suggesting that PSFVgal is an exogenous FV. The length of the complete genome is 12,118 nucleotides (nt). A detailed annotation of the PSFVgal genome is provided in **Figure S1.1** in **Electronic Appendix**.

Sequence similarity analysis revealed that the LTR length is 1,267 nt. An asparagine-1,2 tRNA-utilizing primer binding site (PBS) was identified downstream of the 5'−LTR before *gag* (TGGCGTCCCTGGGTGGGC, nt 1,270−1,287), but which is $tRNA^{Lys}$ (TGGCGCCCAACGTG GGGG/C) in all other mammalian FVs. By comparison with the LTR of a macaque FV (12), the cap site defining the 5' boundary of the R region of the PSFVgal LTR was determined to be 20 nt downstream from the TATA box (TATATAA, nt 928−934) at nt 955 within the 5'−LTR, and the polyadenylation site within the 3'−LTR, which defines the 3' boundary of the R region, was located at the CA dinucleotide (nt 11,955) 20 nt downstream from the polyadenylation signal (AATAAA, nt 11,930−11,935). Thus, the PSFVgal LTR U3, R, and U5 regions are 954, 150,

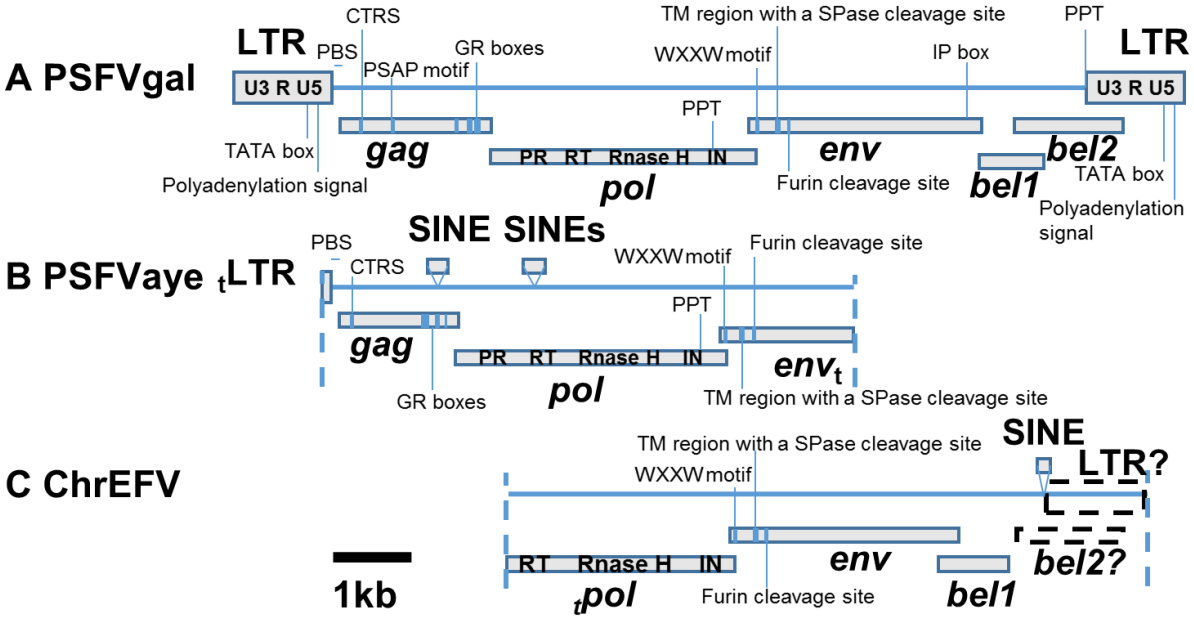


Figure 1.1 Complete and partial genomic organisations of PSFVgal, PSFVaye and ChrEFV. PSFVgal (**A**), PSFVaye (**B**), and ChrEFV (**C**) all exhibit at least some characteristic FV genomic features as indicated (see **Main text**). PSFVaye and ChrEFV are interrupted by a few short interspersed nuclear elements (SINEs). Dashed lines indicate where genomes are truncated. Dashed boxes represent hypothetical domains which may be present but could not be identified. 't'−subscription indicates that the domain is truncated (5'−truncated: preceding the domain name; 3'−truncated: following the domain name). The scale bar (black solid line) represents a nucleotide length of 1 kb. LTR, long-terminal repeat; PBS, primer binding site; CTRS, cytoplasmic retention signal; GR boxes, glycine-arginine rich boxes; PR, protease; RT, reverse transcriptase; IN, integrase; PPT, polypurine tract; WXXW, conserved WXXW site; TM, transmembrane; Spase, signal-peptide peptidase; IP, internal promoter.

and 163 nt long, respectively. Two polypurine tracts (PPT: AAGGAGAGGG) for the dual initiation of plus-stand DNA synthesis (50) were located at the 5' boundary of the 3'−LTR (nt 10,842−10,853) and at the centre of the genome toward the 3' terminus of *pol* (nt 6,068−6,077) as anticipated.

The *gag* gene was predicted to be 1,932 nt long (nt 1,363−3,294) to generate a 644 amino acid (aa) protein. The cytoplasmic retention signal (CTRS), important for particle assembly and viral budding (51), was located at aa 73−90 (GNWGPGDRFARIEVLLRD). The position of a late-assembly domain proline rich PSAP motif, which is essential for primate FV particle assembly and budding (52, 53), was located at aa positions 217−220. The PSAP motif is highly conserved in all simian FVs (SFVs) but its position within Gag is variable (53). The three FV-specific glycine-arginine rich (GR I−III) boxes within the C-terminal region of Gag (54) were also determined (aa 479−493, 541−563, 567−597, respectively).

The Pol protein is 1,156 aa long (nt 3,248−6,718) and protease (PR), reverse transcriptase (RT), Rnase H, and integrase (IN) were predicted to be located at aa positions 10−168, 154−355, 581−737, and 742−1,140, respectively. The catalytic centres of PR (DTG) and RT (YVDD) were located at aa 20−22 and 304−307, respectively. Rnase H active sites were found to be at D^{589} and D^{659} and IN catalytic motif (DD35E) within the IN core domain (aa 862−972) was at D^{926}-D^{938}-E^{962}.

The Env protein is 999 aa long (nt 6,603−9,602). We found a highly conserved WXXW motif (WLRW, aa 10−13) in the N-terminus, which is essential for Gag-Env protein interaction during the budding process, and is found in all known extant FVs (55, 56). The internal promoter (TATAAAA) within the *env* gene, which serves for the initiation of *bel1* transcription, is present at nt 9,299−9,305. We also identified a hydrophobic transmembrane region, which contains a putative signal-peptide peptidase cleavage site (TMGWCIGLFCLLLILLFS↓LVIVIL) at aa 81−104. The cleavage occurs within the endoplasmic reticulum to remove the signal peptide (57). A potential furin-cleavage site was also found (TRPNYTAARSRR↓SVE) at aa 131–145.

It is this site at which the FV Env protein is cleaved, either by furin or furin-like proteases, to produce an Env leader protein—a component that interacts with the N-terminus of Gag via its WXXW motif and is absolutely required for FV particle budding (57).

The *bel1* and *bel2* genes were predicted to be present at nt 9,569−10,402 and nt 10,026−11,414, respectively. The putative Bel1 protein (277 aa) exhibits a weak similarity (<15% identity) to extant FV Bel1 proteins, which could only be shown by comparison with Hidden Markov Model (HMM) of FV Bel1 proteins (HMM profile GyDB (http://www.gydb.org/): Bel1 spumaretroviridae; $E = 3.7 \times 10^{-6}$). In contrast, the putative Bel2 protein (511 aa) exhibits a relatively high similarity with other FV Bel2 proteins with the best BLASTp score to the bovine FV (BFV) Bel2 ($E = 4.0 \times 10^{-8}$; maximum identity = 26%).

1.3.2 Discovery and characterisation of PSFVaye and ChrEFV

To search for endogenous FVs, we screened all publically available GenBank whole genome shotgun (WGS) sequences with various FV proteins (see **Materials and Methods**) using tBLASTn. As previously reported (36), several matches were returned from the WGS assembled sequence of the aye-aye (*Daubentonia madagascariensis),* spanning four non-overlapping contigs (**Table 1.2**). Sequence analysis shows that these four contigs represent different parts of a single ERV 'PSFVaye', confirmed by genome walking (William Switzer, personal communication). Between the first and the second, and the second and the third contigs are short interspersed nuclear elements (SINEs), and between the third and the fourth contigs are 5 ambiguous nucleotides. We found that PSFVaye contains numerous frame-shift and in-frame stop codon mutations, and is interrupted by several transposable elements throughout the genome. This finding confirms that PSFVaye is indeed endogenous and not replication competent. This *in silico* screening process also returned one contig from the WGS assembled sequence from a Cape golden mole (*Chrysochloris asiatica*) (**Table 1.2**). We designated this element 'ChrEFV'. Like PSFVaye, ChrEFV contains numerous frame-shift and in-frame stop codon mutations, and is interrupted by transposable elements, confirming that it is not replication competent as is typical of ancient ERVs. Genomic inspection revealed that

Table 1.2 Identification of foamy virus (FV)-like sequences in the aye-aye and Cape golden mole genome.

Accession no.	FV genomic region	Position*	BLASTp best hit	Query Cover	Max Identity	E-value
Aye-aye (*Daubentonia madagascariensis*)						
AGTM011839603	(5' truncated) 5'−LTR	3208..3318	N/A	N/A	N/A	N/A
	PBS	3321..3338	N/A	N/A	N/A	N/A
	gag	3405..4654	CAA70074 Gag FFV	93%	37%	6×10^{-72}
AGTM012070104		131..361				
	pol	323..1247	AAF64414 Pol EFV	95%	48%	0.0
AGTM011519752		90..1996 (end)				
AGTM010361255		1..635				
	(3' truncated) *env*	520..2763	AFK85016 Env RhiFV	80%	42%	7×10^{-115}
Cape golden mole (*Chrysochloris asiatica*)						
AMDV01151999	(5' truncated) *pol*	<615..>3534	AFX98084 Pol SFV	95%	47%	0.0
	env	<3464..>6371	AFR79245 Env BFV	96%	30%	5×10^{-121}
	*bel1***	6338..6978	−	−	−	−

'N/A' = 'Not applicable'.
* The nucleotide positions are with respect to the contig length and do not take in-del mutations into account.
** The similarity between ChrEFV Bel1 and other FV Bel1 proteins was determined by comparing the ChrEFV putative Bel1 protein against the Hidden Markov Model (HMM) of FV Bel1 proteins available in the GyDB (http://www.gydb.org/) (HMM profile: Bel1 spumaretroviridae; E: 0.24)

the putative genes of both PSFVaye and ChrEFV are in the same order as those of typical exogenous mammalian FVs (**Figure 1.1**; see **Figure S1.1** in **Electronic Appendix** for detailed sequence annotations). Moreover, it is likely that there is only one copy of each virus present in either host genome, since both the aye-aye and Cape golden mole genomes were extensively sequenced (38× (58) and 66× (59) coverage, respectively) but paralogs of each were not found.

Within PSFVaye, via sequence similarity, we identified (i) a PBS (TGACACCCAATGTGGGG C**;** nt 114−131) with a broad tRNA usage predicted ($tRNA^{Lys, Asn, Thr, Pro}$), (ii) a complete putative *gag* gene (nt 198−1,971) which is interrupted by a SINE (AluJo (60), nt 1,449−1,748), (iii) a complete putative *pol* gene (nt 1,941−5,727; PR: nt 1,968−2,459; RT: nt 2,415−3,306; Rnase H: nt 3,985−4,467; IN: nt 4,480−5,676) which is also interrupted by SINEs (AluJ Mim (61) & ALU (62), nt 2,871−3,165), and (iv) a 3'-truncated putative *env* gene (nt 5,612−7,867). We also found ~113 nt on the 5' end of the PBS to be uniquely-mapping within the aye-aye genome. This might represent some portion of the 5' end of the 5' PSFVaye LTR (nt 1−111). The truncation of the 5'−LTR and *env* gene, as well as the absence of a 3' LTR in the PSFVaye genome were confirmed by using a genome walking of genomic DNA from two different aye-ayes (William Switzer, personal communication). We found that PSFVaye exhibits several characteristic FV features. For example, its Gag protein contains a degenerate, but still recognisable, putative CTRS (GPR?VGD*WQRICLAFQY, aa 37−54) and three GR I−III boxes (box I: nt 1,248−1,358; box II: nt 1,422−1,781 interrupted by a SINE (nt 1,449−1,748); box III nt 1,851−1,871); the putative *pol* gene harbours a PPT (AAGGACAAAG, nt 5,092−5,101); and the N-terminus of the putative Env contains a highly conserved WXXW motif (WLAW, aa 10−13). A hydrophobic transmembrane region containing a signal-peptide peptidase cleavage site (ILIWIMLFLILFSAILVS↓TLIAVF) was also located in the N-terminus of the putative Env at aa 63−86. We could not identify a furin-cleavage site in the Env of PSFVaye, most likely due to its defective nature. However, by comparing the Env sequence of PSFVaye to those of other non-defective exogenous FVs, we speculate that it should be located between aa 113 and 127 (YFSQAHV*KSRAIHF). Moreover, its putative Gag and Env proteins only exhibit similarities to

FV proteins (**Table 1.2**), and share no detectible similarities with other (non-FV) retroviral proteins. All these findings strongly suggest that the PSFVaye sequence is indeed an endogenous FV, and not merely a distantly related FV-like ERV.

ChrEFV contains (i) a 5'−truncated putative Pol coding gene (nt <1−<2,937) which only contains full RT (nt 1−597), Rnase H (nt 1,270−1,737), and IN coding domains (nt 1,750−2,931), (ii) a full putative Env coding gene (nt <2,867−>5,785) and (iii) a full putative Bel1 coding gene (nt 5,752−6,393). Following the putative *bel1* gene is a long non-repetitive sequence of ~1,700 nt containing a SINE (AFROSINE1B (63)). The *bel2* gene and the 3'−LTR of ChrEFV may be located within this region but cannot be identified, most likely due to divergence from the probes we used. Nonetheless, the N-terminus of the ChrEFV Env protein harbours the highly conserved WXXW motif (WMRW, aa 10−13), and only shows significant similarity to FV Env proteins (**Table 1.2**). A hydrophobic transmembrane region harboring a putative signal-peptide peptidase cleavage site (VMTSYVS?LILLGIIITA↓SFITIC, aa 74−97) was also found in the putative Env at the anticipated location, and a potential furin-cleavage site (GQIISNSSSNRRKI) could be located at aa 125−138 within the putative Env as well. Furthermore, the putative Bel1 protein also exhibits some similarity to FV Bel1 proteins, which could be demonstrated by using the HMM searching technique (HMM profile GyDB (http://www.gydb.org/): Bel1 spumaretroviridae; E = 0.24; note that this weak support is rather typical and expected given that this is a part of an endogenous FV and that FV Bel1 protein is not conserved among FVs (64)). Although a number of FV-specific genomic features as outlined in the PSFVgal characterisation section could not be located, possibly due to the lack of sequence data and/or divergence, this genomic information supports a FV progenitor of ChrEFV.

Aside from the aye-aye genome, currently available genomes of other lemurs (*Lemur catta* and *Microcebus murinus*) do not seem to contain PSFVaye-like ERVs. These results support the hypothesis that endogenisation of PSFVaye occurred in the ancestral aye-aye lineage after it diverged from the lineage of all other lemurs, dating back ~59−66 Ma (65–67). Similarly,

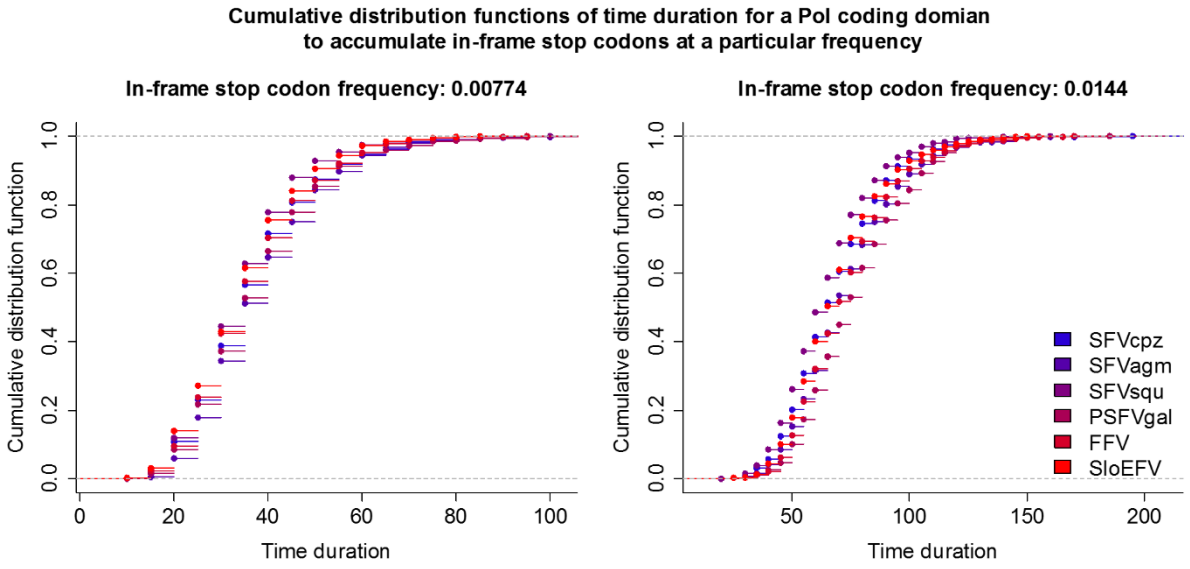


Figure 1.2 PSFVaye and ChrEFV integration dates, estimated from in-frame stop codon frequencies. Cumulative distribution functions (CDFs) of time duration for foamy virus Pol protein sequences to accumulate in-frame stop codons at the frequencies of 0.00774 (**right**, stop codon frequency of PSFVaye Pol) and 0.0144 (**left**, stop codon frequency of ChrEFV Pol). Based on these CDFs, the mean age estimate of PSFVaye is ~35.2−39.6 millions of years (Myr) old (95% confidence interval: 15−75 Myr old, mode: 30−35 Myr old, median: 35 Myr old). The mean age estimate of ChrEFV is ~65.6−78.2 Myr old (95% confidence interval: 35−125 Myr old, mode: 60−65 Myr old, median: 65−75 Myr old). SFVcpz, chimpanzee simian foamy virus (SFV); SFVagm, Arfican green monkey SFV; SFVsqu, squirrel monkey SFV; PSFVgal, galago prosimian FV; FFV, feline FV; and SloEFV, sloth endogenous FV.

ChrEFV-like ERVs could not be identified in elephant, hyrax, elephant shrew, and tenrec genomes, implying a maximum age of ~60−85 Myr based on the proposed Afrosoricida basal split date (38, 68). To further estimate the age of these elements, we analysed the frequencies of in-frame stop codons in their putative Pol proteins with a Monte Carlo simulation approach, using *pol*s of PSFVgal, SFVcpz, SFVagm, SFVsqu, FFV, and SloEFV (**Figure 1.2**) as model sequences. Consistently across all simulations, the mean age estimates of PSFVaye and ChrEFV are ~35.2−39.6 and ~65−78 Myr, respectively, congruent with (i.e. less than) their upper-bound age limit.

1.3.3 Phylogenetic reconstruction and reconciliation analysis

A retrovirus phylogeny (**Figure 1.3**) was estimated by using an alignment of RT protein sequences to establish the phylogenetic positions of PSFVgal, PSFVaye, and ChrEFV among retroviruses. We found that all three new FV sequences cluster together with the known extant mammalian FVs. Together with the genomic characterisation above, the phylogeny strongly supports the classification of PSFVgal as a true exogenous FV, and that both PSFVaye and ChrEFV are ERVs of FV origin. Our discovery of PSFVaye, and ChrEFV raises the possibility of the presence of undetected FVs circulating among lemurs and afrotherians.

To reconstruct the co-evolutionary history of FVs and their mammalian hosts, a phylogenetic reconciliation analysis was performed. A Bayesian phylogeny of FVs was constructed based on a Pol-Env protein alignment and was compared to the previously published host phylogeny (38) (**Figure 1.4**). Potential recombination among the sequences within the alignment was assessed by using a quartet-based recombination detection program VisRD3 (69), and the results showed no significant evidence for recombination, both at nucleotide and protein levels (nucleotide: *P* value = 0.621; protein: *P* value = 0.495). Furthermore, we also found that neither of the separate Pol nor Env alignments reject the topology of the best phylogeny inferred from the concatenated Pol-Env alignment (approximately unbiased test (AU): $P\ value_{Pol} = 0.987$; $P\ value_{Env} = 0.700$; **Figure 1.5**), congruent with the results from the recombination test.

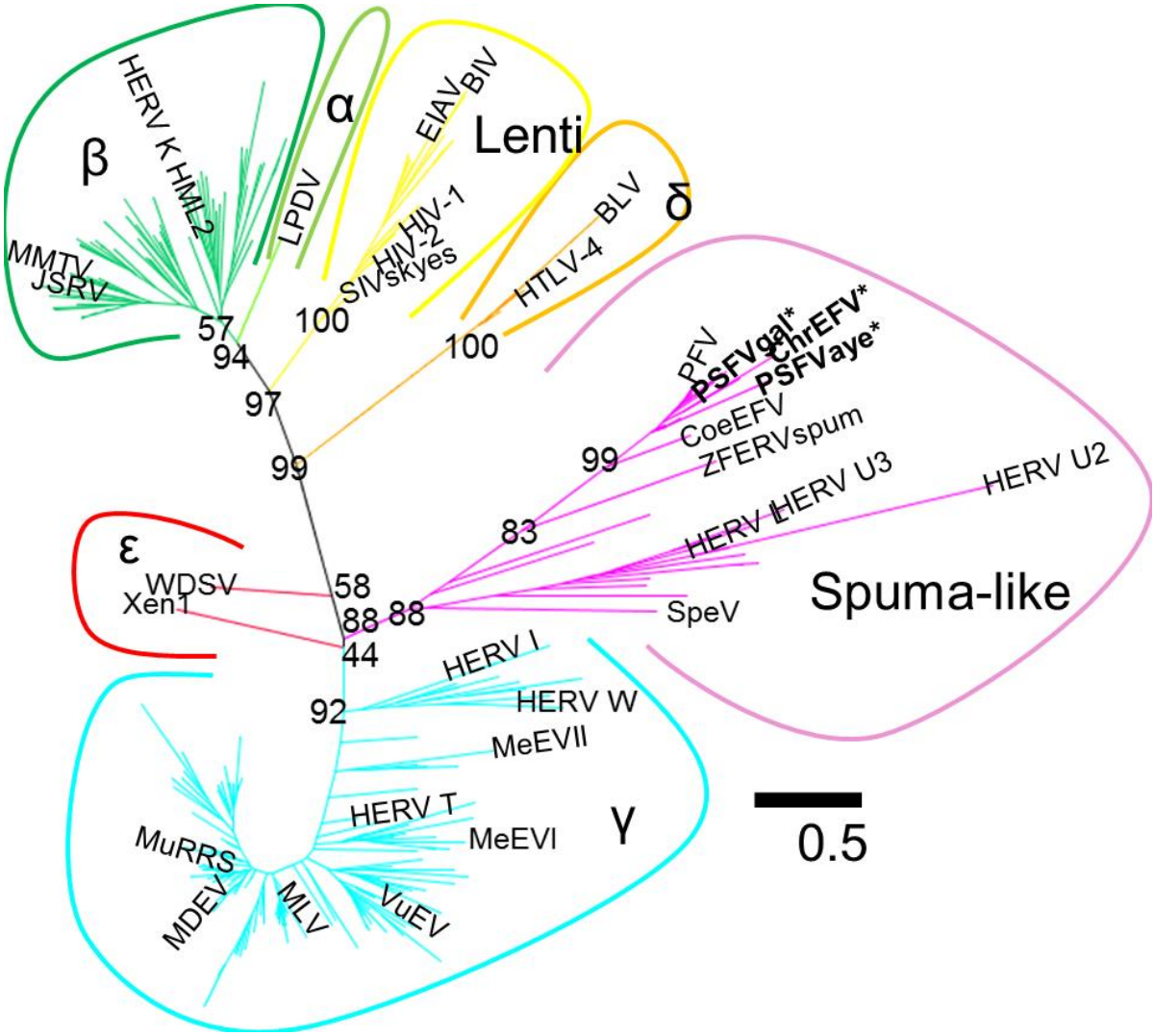


Figure 1.3 Retrovirus phylogeny. The maximum-likelihood phylogeny of retroviruses was built by using an alignment of reverse transcriptase proteins (162 aa) of several retrovirus genera (240 taxa): alpharetrovirus (α, light green), betaretrovirus (β, dark green), gammaretrovirus (γ, blue), deltaretrovirus (δ, orange), epsilonretrovirus (ε, red), lentiretrovirus (Lenti, yellow), and spuma-like retrovirus (Spuma-like, purple). The tree was constructed using RaxML 7.2.8-HPC2 on XSEDE (70). The best substitution model was determined by ProtTest 2.4 (71) to be the LG+Γ(4)+F model. The scale bar is in units of amino acid substitution per site. Numbers on nodes are bootstrap support values, estimated using 5,000 pseudoreplicates. PSFVgal, PSFVaye, and ChrEFV are indicated in bold type and with asterisks. The tree is not rooted.

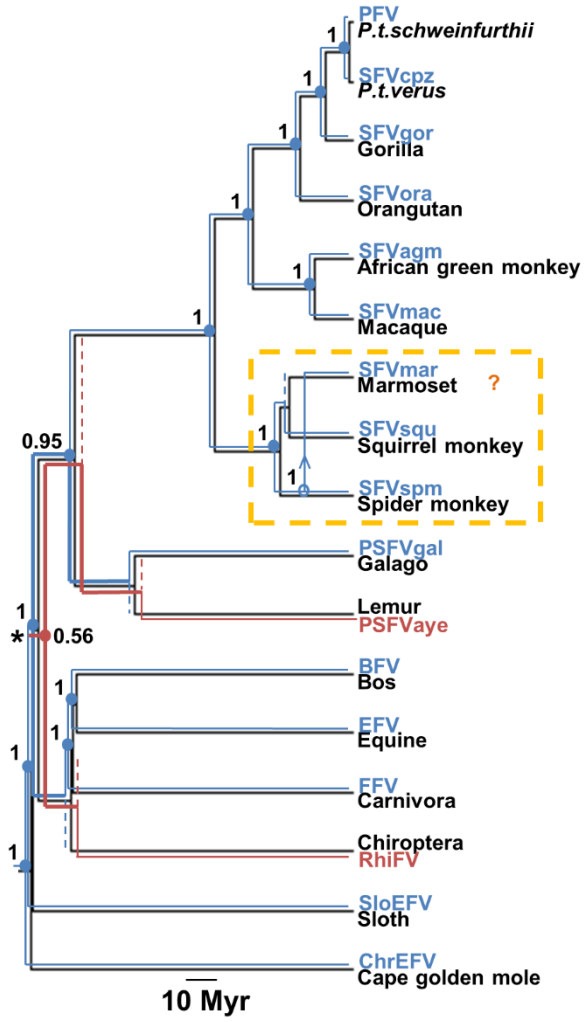
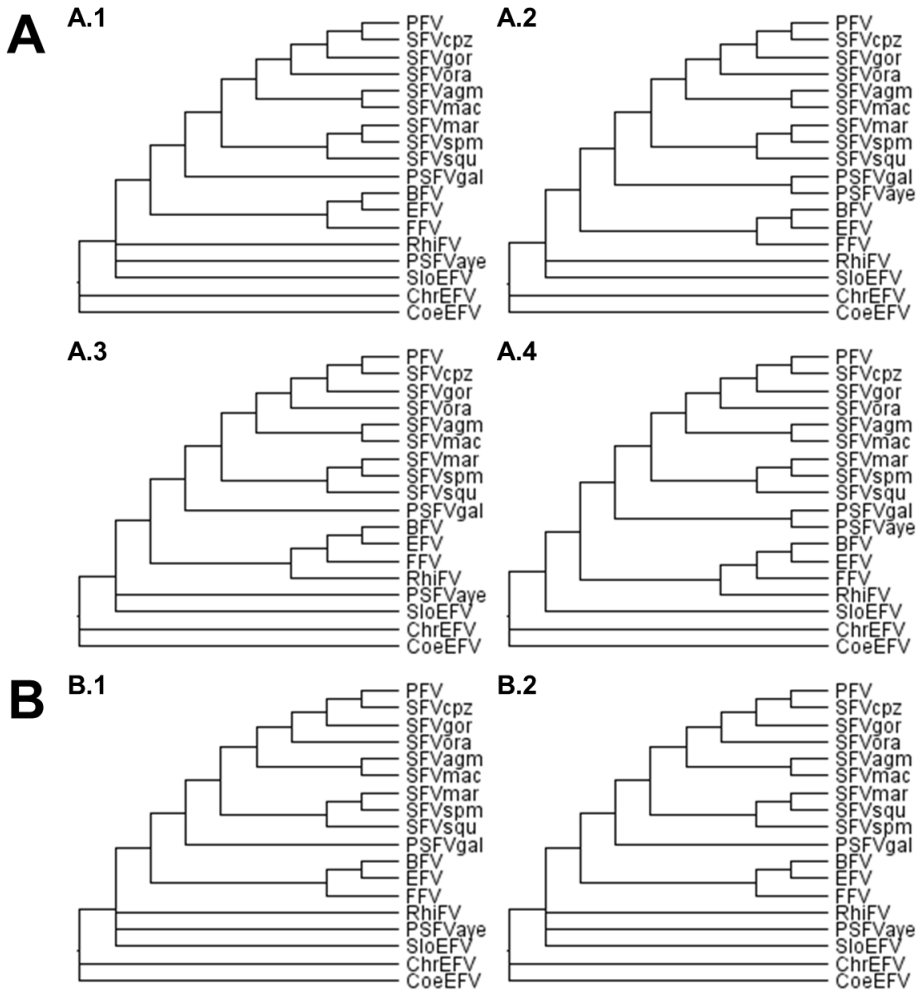


Figure 1.4 Phylogenetic reconciliation analysis. The foamy virus (FV) phylogeny (blue and red) was superimposed upon a previously published host phylogeny (black, published in (38)) and is scaled to host divergence times. The scale bar is in units of millions of years. Concatenated Pol-Env protein alignments (893 and 603 amino acids, respectively) were used to infer the FV trees. Numbers on nodes are Bayesian posterior FV clade supports. Solid circles represent inferred FV-host co-speciation events, open circles represent host switching events, and an asterisk indicates a lineage duplication event in the absence of host diversification, which gave rise to two FV lineages sharing host species. The thick lines in the FV phylogeny correspond to the evolutionary period, spanning about 30 million years, during which the two lineages shared their host species. A hypothetical scenario of host switching of New World monkey FVs (as indicated by 'an orange **?**') is shown in an orange box. An alternative scenario is shown in **Figure 1.8**.

Overall, the two phylogenies show remarkably similar topologies, reflecting the well-established history of stable FV-mammal co-speciation (35, 37). In total, 14/17 (82.4%) potential FV-host co-speciation events were inferred among the 17 FV and host sequences by using the co-phylogeny reconstruction software Jane v4.0 (72) (**Figure 1.4**). The deepest co-speciation event occurred early on in eutherian diversification, corresponding to the Exafroplacentalia-Afrotheria split (**Figure 1.4**, see also **Box 1.1**).

The reconstruction that maximises the number of co-speciation events also suggests that the most recent common ancestor (MRCA) of PSFVaye and the Chinese bat (*Rhinolophus affinis*) FV (RhiFV) was present in the MRCA of their hosts, and requires a duplication of the virus lineage at the base of the exafroplacentalian mammal clade (**Figure 1.4**). However, based on what we know about FV biology, it is unlikely that a FV will diversify into two lineages within a single natural host, i.e. in the absence of host diversification. We postulate that this viral lineage duplication represents an ancient host speciation event, where one of the resulting species is not sampled in our data set. We therefore adopt a conservative approach and do not count the PSFVaye/RhiFV co-speciation event, reducing the number of co-speciation events to 13 (76.5%). Nevertheless, 13 co-speciation events are still greater than expected by chance (random tip mapping: sample size = 1000, *P* value < 0.001), indicating that the co-speciation history of FVs and their hosts is stable.

Amidst this extended pattern of FV-host co-speciation, our analysis revealed several FV cross-species transmissions. One involves PSFVaye and RhiFV as aforementioned, both of which are placed robustly outside the clade of boreoeutherian FVs (posterior probability = 1; **Figure 1.4**), and the other one involves NWM FVs (**Figure 1.4**). AU tests were performed to evaluate the robustness of these inferred host switches. The results show that the co-speciation histories of PSFVaye, and RhiFV (**Figure 1.5A**), and that of NWM FVs (**Figure 1.5B**) are all rejected by our data, supporting the inferred FV cross-species transmissions. Interestingly, our analysis also suggested a sister group relationship between PSFVaye and RhiFV, although statistical support for this relationship is weak (posterior probabilities = 0.56, **Figure 1.4**).



Tree		Pol-Env		Pol		Env	
		LLH	P-AU	LLH	P-AU	LLH	P-AU
A	A.1	-32032.17	0.997	-17465.40	0.987	-14566.77	0.700
	A.2	-32070.13	0.003	-17487.43	0.028	-14582.70	0.006
	A.3	-32057.71	0.008	-17488.36	0.003	-14569.36	0.396
	A.4	-32083.78	4×10^{-4}	-17501.92	0.001	-14581.85	0.027
B	B.1	-32032.17	0.997	-17465.40	0.993	-14566.77	0.961
	B.2	-32071.27	0.003	-17487.90	0.007	-14583.37	0.039

Figure 1.5 Foamy virus (FV)-host evolutionary conflicts. Our analyses suggested two FV-host evolutionary conflicts: one is of PSFVaye and RhiFV (**A**), and the other is of New World monkey (NWM) FVs (**B**). **A.1** shows our best estimated phylogeny, in which both PSFVaye and RhiFV are placed robustly outside the Boreoeutherian FV clade but still remain together with exafroplacentalian FVs. **A.2-4** show alternative phylogenies in which either PSFVaye or RhiFV (**A.2** and **A.3**) or both (**A.4**) co-speciate with their hosts. We used approximately unbiased (AU) tests (73) to compare these alternative phylogenies against one another given Pol-Env/Pol/Env alignments, based on comparisons of site-wise log-likelihood scores (LLH) computed using PAML 4.7a (74). The AU tests were performed in Consel (75). The same analyses were performed to compare our best estimate of the phylogenetic placement of NWM FVs, which show FV-host evolutionary conflicts (**B.1**) against topologies that do not (**B.2**). Results are summarised in the table; only total LLH scores and AU probabilities (P-AUs) are shown.

1.3.4 The effects of neutral genetic changes on phylogenetic inference

Our finding that PSFVaye does not group together with other SFVs is consistent with the result from a previous study (36). However, it has been suggested that this inferred placement of PSFVaye may be erroneous, resulting from neutral genetic changes that have accumulated since its endogenisation (36). To examine whether or not neutral evolution alone is enough to explain this observation, a phylogeny of a Pol-Env protein alignment containing *in silico* mutated SFVagm, SFVsqu, and PSFVgal sequences were inferred. These sequences were mutated with an overestimate of the mammalian neutral substitution rate of 1×10^{-8} substitutions per site per year (s/n/y) over a time period of 40 Myr, a mean age estimate of PSFVaye (**Figure 1.2**). The substitution process was assumed to be neutral, homogenous, and independent across sites and base types.

Our results suggest that, even with this biased substitution rate setting of 1×10^{-8} s/n/y, the accumulation of neutral genetic changes alone is insufficient to explain the inferred phylogenetic position of PSFVaye; it only increased the branch length and decreased the branch support without changing the tree topology (**Figure 1.6**). It is thus unlikely that the placement of PSFVaye is an artefact.

1.3.5 FV-host divergence correlation analysis

To further evaluate the inferred FV-host co-speciation history, we performed a divergence correlation analysis. In this analysis, we identified codiverging branches based on the FV-host co-speciation pattern that we inferred, and examined if the genetic divergences of FVs and their hosts are correlated. Unlike the phylogenetic reconciliation analysis, this analysis takes both the topologies of the virus and host trees, as well as their branch lengths into account.

Overall, we found that there is a strong positive linear correlation between FV and host divergence, which extends from the present to the exafroplacentalian radiation (linear regression: N = 23, $R^2 = 0.823$, *P* value = 3.48×10^{-8}; **Figure 1.7**). This result suggests that,

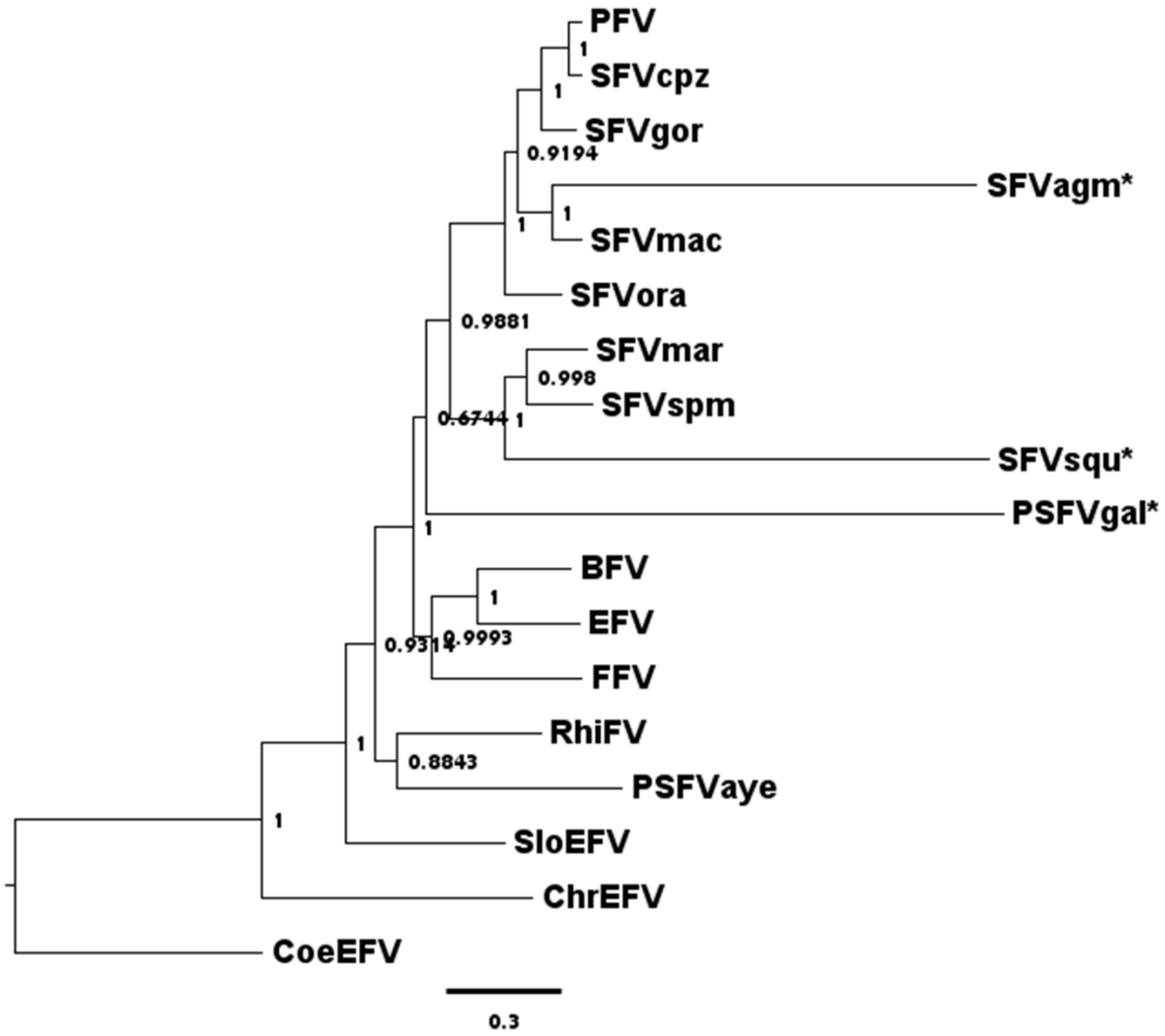


Figure 1.6 Investigation of the effects of neutral genetic changes on the inference of foamy virus (FV) phylogenetic relationships. We mutated the *pol* and *env* nucleotide sequences of SFVagm, SFVsqu, and PSFVgal (indicated with asterisks) *in silico* with a substitution rate of 1×10^{-8} substitutions per site per year over a hypothetical period of 40 million years. The mutated gene sequences were then translated into protein sequences and the FV phylogeny was re-built by using MrBayes 3.2.1 (76). The substitution process was assumed to be neutral, homogenous, and independent across all sites and base types. Numbers at branch nodes are posterior probabilities and the scale bar is in units of amino acid substitutions per site.

across this protracted time period, the accumulation of FV divergence occurred in proportion to that of their hosts. Together with the results from the phylogenetic reconciliation analysis, these findings strongly support the notion that FVs have been stably codiverging in concert with their mammalian hosts for >100 Myr throughout the Cretaceous and Cenozoic eras, as previously proposed (35).

We found that some of the results from this analysis are inconsistent with those from the phylogenetic reconciliation analysis however. The phylogenetic reconciliation analysis inferred the branching of ChrEFV to be a co-speciation event (posterior probability = 1; **Figure 1.4**), but the FV-host divergence correlation analysis showed that the branch of ChrEFV is too long, and is placed too deep in the tree to be consistent with the FV-host co-divergence history (**Figure 1.7**; identified as outliers by the Cook's distance inspection, green and blue dots, respectively). Another conflicting result is that while the split of BFV and equine FV (EFV) was inferred to be a co-speciation event by the phylogenetic reconciliation analysis (**Figure 1.4**), the BFV branch is too short to fit the FV-host co-speciation pattern (**Figure 1.7**, identified as an outlier by the Cook's distance inspection, red dot). The implications of these findings are discussed below.

1.4 Discussion

This study reports and describes the first complete genome of an exogenous PSFV isolated from a galago, namely PSFVgal, and describes in more detail the partial genome of PSFVaye obtained from the aye-aye genomic DNA. Phylogenetic analyses and genome inspection suggest that both of them are FVs. The defective nature of PSFVaye indicates that it is an endogenous FV, while PSFVgal does not contain any in-frame stop codons and/or frame-shift mutations, consistent with it being an exogenous FV. We also report and describe a novel afrotherian ERV found within the Cape golden mole genome, ChrEFV, which exhibits characteristic features of mammalian FVs.

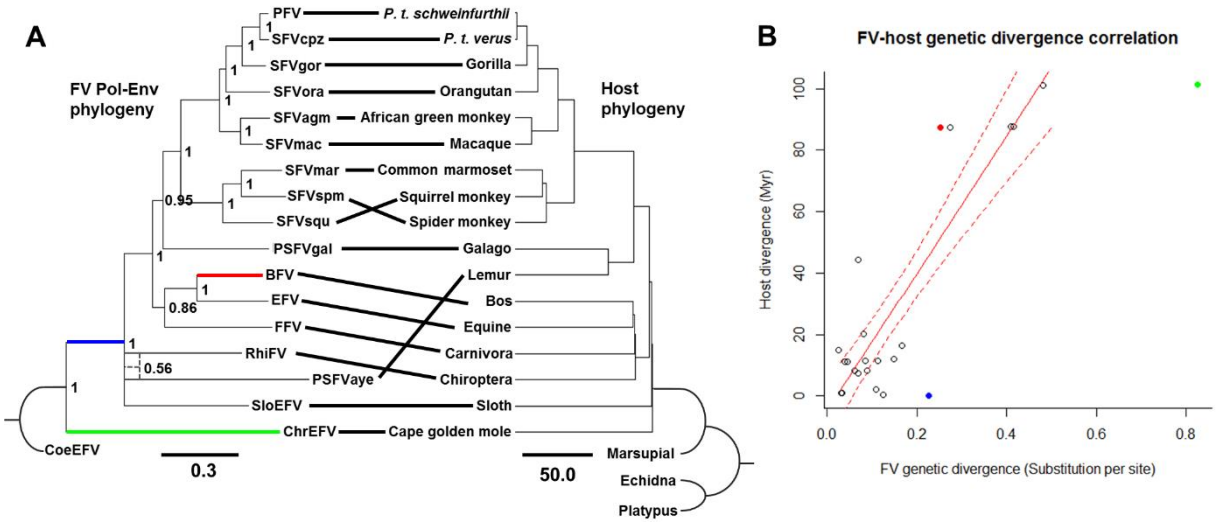


Figure 1.7 Co-evolution of foamy viruses (FVs) and their mammalian hosts. (**A**) FV Bayesian phylogeny (**left**) is compared to that of their hosts (right, previously published in (38)). Curved branches indicate outgroups. Numbers on nodes are Bayesian posterior clade supports. Solid lines between the two trees indicate FV-host associations. FV and host phylogeny scale bars are in units of amino-acid substitutions per site and million years, respectively. Weak support for the sister-taxon relationship of RhiFV and PSFVaye (dotted branches, posterior probability = 0.56) was inferred. Among 17 FV and host sequences, 3 apparent mismatches have been inferred by using co-phylogenetic reconstruction. Branches corresponding to FV-host co-evolution were identified and used in FV-host divergence correlation analysis (**B**). Branch lengths of FV tree and host divergence time were determined to have a linear correlation with a statistically strong support for its coefficient (linear regression: N = 23, R^2 = 0.823, *P* value = 3.48 $\times 10^{-8}$), represented by the solid red line. Dotted lines represent the 95% confidence interval of the estimated linear relationship. Outliers determined by Cook's distances (solid dots), including the bovine foamy virus (BFV) branch (red), the ChrEFV branch (green), and the ancestral branch leading to the exafroplacentalian FVs (blue), were not included in the linear regression. The colours of the outliers correspond to the colours of the branches in **A**.

Consistent with previous findings (35, 37), our analyses suggest an extended FV-host co-speciation history across a timescale spanning millions of years. Interestingly, this is in contrast to what is suggested by various *in vitro* experiments. For example, an *in vitro* study has shown that prototype FV (PFV) is capable of infecting not only primate, but also, rodent, laurasiatheria, avian and reptile cells, with an extremely broad range of susceptible cell types (77). Furthermore, *in vitro* investigations of FV cell-attachment and entry also suggested that FVs of different species likely utilise the same, perhaps promiscuous, receptor molecule(s) for cell-attachment and/or entry (78–80). Although these findings might in part help to explain FV cross-species transmissions observed in nature (see below), one would expect to see more FV cross-species transmissions happening given these findings. Analyses of anti-retroviral restriction factors have shown that they are specific to particular FVs. For example, while the tripartite motif protein 5αs (TRIM5αs) of most NWMs have been shown to be able to restrict some NWM FVs, PFV and a SFV from macaque, ape TRIM5αs cannot, but instead can restrict feline FV (17, 81). Inhibition of apolipoprotein B mRNA-editing, enzyme-catalytic, polypeptide-like 3C (APOBEC3C) by FVs has also been shown to be species-specific (82). Whether this species-specific FV-host antagonistic interaction is one of the factors that underline the stable co-speciation history or the result of it is still unclear. To date, the factors that determine this stable pattern of FV-host co-speciation are still very poorly understood.

Two analyses were performed to reconstruct the history of the co-speciation between FVs and their mammalian hosts: (i) a phylogenetic reconciliation analysis (**Figure 1.4**), and (ii) a FV-host divergence correlation analysis (**Figure 1.7**). However, we found that some of the results from these two analyses are incompatible. Although the phylogenetic reconciliation analysis suggests that the branching of ChrEFV is a co-speciation event (posterior probability = 1; **Figure 1.4A**), the ChrEFV branch is too long and is placed too deep in the tree (**Figure 1.7**). Coupled with the fact that ChrEFV is basal to all known extant FVs, this finding challenges the notion that the progenitor of ChrEFV co-speciated with its host, and also raises a possibility that it might not even be a FV. Nevertheless, ChrEFV is a long terminal branch that has accumulated numerous neutral genetic changes since its ancient endogenisation ~65−78 Ma.

Together with long-branch attraction, this 'pseudogene effect' (83) could cause artificial inflation of the branch length with a concomitant deep placement of ChrEFV within the FV phylogeny. Thus, given the congruent ChrEFV/host phylogenies and the FV-like features of ChrEFV outlined above, it is still likely that ChrEFV is a genuine endogenous FV, and that its progenitor codiverged with its host. To the best of our knowledge, this is the first (endogenous) afrotherian FV to be discovered, expanding the FV host range to cover the whole eutherian clade.

Resolution of ChrEFV evolutionary history requires further analysis with other afrotherian FVs should they be identified. For example, finding that other afrotherian FVs form a monophyletic clade with ChrEFV will strongly support a FV origin of ChrEFV. On the other hand, if they are paraphyletic, with ChrEFV being basal to all known FVs, this would raise a possibility that ChrEFV might not be a FV, but a class III non-FV ERV. The discovery of marsupial and/or monotreme FVs can also elucidate the origin of ChrEFV; finding them being more basal than ChrEFV would imply that ChrEFV is indeed an endogenous FV, and at the same time, would suggest that FVs have been co-diversifying with their hosts since the emergence of mammals.

Another conflicting result is that while the phylogenetic reconciliation analysis identified the BFV/EFV split as a co-speciation event (**Figure 1.4**), the BFV branch is too short to fit the FV-host co-speciation pattern (**Figure 1.7**). This finding raises a possibility that the BFV/EFV split might not be a co-speciation event as initially inferred, but instead represents a FV cross-species transmission between equine and bovine. An alternative explanation to this is that it is in fact a co-speciation event, but the rate of evolution of BFV and EFV is different from those of other FVs. Under this circumstance, one would expect the root-to-tip divergence for BFV and EFV to be different from those of other FVs. Our analyses however do not show such results (**Figure 1.4**). This in turn indicates that this scenario is perhaps unlikely. Additional analyses of other BFVs and EFVs are required to discriminate between these two possibilities.

Although FVs and mammals have an extended co-speciation pattern, it is not absolute. Our analyses inferred a small number of FV cross-species transmissions. One involves NWM FVs. NWM FVs form a monophyletic clade that is sister to catarrhine FVs, reflecting the branching order of their hosts; however, the branching orders within the clade clearly do not parallel that of NWMs (65, 84, 85) (**Figure 1.4** and **Figure 1.8, orange boxes**), confirmed by the AU test (**Figure 1.5B**). These results implied at least one ancestral NWM FV cross-species transmission. Marmosets, squirrel monkeys, and spider monkeys are closely related, and occupy large overlapping geographic ranges (17). It is thus plausible that heterologous NWM FV infections might have occurred in the past. This is unsurprising given that FV cross-species transmissions between closely related primate species has already been documented (41).

Our analyses also revealed viral cross-species transmissions involving PSFVaye and RhiFV. Both FVs are placed robustly outside the boreoeutherian FV clade (posterior probability = 1, **Figure 1.4**), suggesting that their origins might involve host switches across different mammalian orders. Indeed, the results from the AU tests showed that the co-speciation history of PSFVaye and RhiFV was rejected by our sequence data (**Figure 1.5A**). We propose the following scenario to explain the phylogenetic relationships of PSFVaye and RhiFV within the context of the evolutionary and geographical history of their hosts (**Figure 1.8**). ~180−160 Ma, Gondwana split into two separate landmasses, namely the Madagascar-India-Antarctica landmass and the Africa-South America landmass (86, 87), and the former further split into the Madagascar-India landmass and the Antarctica landmass ~135 Ma (86, 87). Thereafter, early eutherians arrived onto the Madagascar-India landmass upon which their FVs established stable populations (**Figure 1.8A**). Since then, these FVs evolved independently from, and hence are only distantly related to, FVs on the African landmass, which later would give rise to all extant simian, lorisiforme, and fereungulata FVs. The India landmass was then split from the Madagascar landmass ~80−90 Ma (86, 87), separating the FV populations into two groups (**Figure 1.8B**). One group remained on Madagascar, which later would give rise to the progenitor of PSFVaye via at least one viral cross mammalian order transmission. The second group was transferred to Laurasia via the India landmass continental drift, which independently

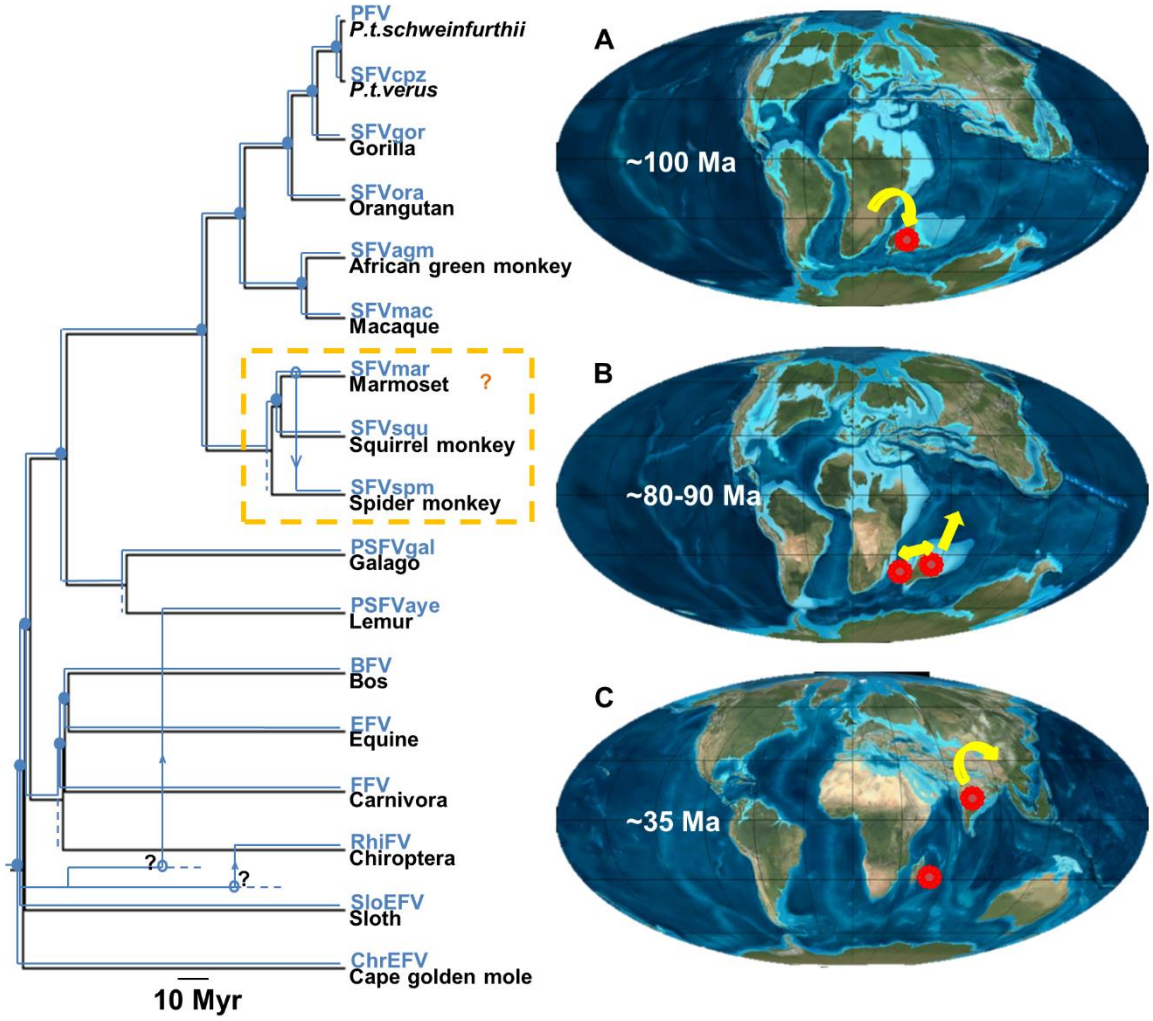


Figure 1.8 Hypothesis explaining the mismatches in foamy virus (FV)-host evolutionary history. FV phylogeny (blue) was superimposed upon host phylogeny (black) and is scaled to host divergence times. The scale bar is in units of millions of years. Solid circles represent FV-host co-speciation events. Open-circles represent possible host switching events of which the inferred switching directions are indicated by arrows. A hypothetical scenario of host switching (as indicated by an orange '**?**') within the New World monkey FV clade is shown in an orange box. An alternative scenario is shown in **Figure 1.4**. For PSFVaye and RhiFV, we speculate that, in the very early history of mammalian evolution, eutherians came to the Madagascar-India landmass with their FVs (red circles) upon which they established stable populations (**A**). The landmass then split into the Madagascar and India landmasses resulting in the FV population splitting into two groups ~80−90 Ma (**B**); one FV variant remained on Madagascar and the other was transported to Laurasia via the Indian landmass continental drift (**C**). The former gave rise to the PSFVaye progenitor while the latter gave rise to RhiFV. Directions of infected FV host population movements are indicated by yellow arrows. However, it is unknown which ancestral species introduced FV to the lemur and bat lineages and when the transmissions occurred (as indicated by a black '**?**'). Additional sequence data from other mammals, especially bats and lemurs, are required to resolve these aspects of FV evolutionary history.

gave rise to RhiFV, involving at least one viral cross mammalian order transmission also (**Figure 1.8C**).

Our hypothesis helps to explain the deep phylogenetic placement and dispersed geographical distribution of PSFVaye and RhiFV. In addition, this hypothesis also gives an explicit prediction that PSFVaye and RhiFV should share a MRCA about 80−90 Ma based on the estimated Madagascar-India landmass split date (86, 87). Indeed, although not well-supported, our analysis suggests that PSFVaye and RhiFV have a sister-group relationship (posterior probability = 0.56; **Figure 1.4**). Assuming that this is correct, we estimated their time to MRCA to be ~93 (95% confidence interval = 85−101) Myr based on the linear relationship of FV-host genetic divergence (**Figure 1.7B**). Although slightly high, our estimate range largely overlaps with the date range derived from this hypothesis. The accumulated neutral changes within the PSFVaye sequence may explain both the weak branch support and the slightly overestimated divergence date.

Our model does not suggest which ancestral species introduced FVs to the lemur and bat lineages and/or when the transmissions occurred. Nonetheless, since both PSFVaye and RhiFV are placed robustly outside the boreoeutherian FV clade but are still within the clade of exafroplacentalian FVs (posterior probability = 1 and 1, respectively, **Figure 1.4** and **1.7A**), it is likely that their original hosts were exafroplacentalians, but not boreoeutherians, and that the host switches occurred between species across deep clades that long diverged from each other. More sequence data from other bat and lemuriforme FVs are required to further elucidate the history of PSFVaye and RhiFV. Below, we outline various possible scenarios how the inclusion of more sequence data might offer us some insights into the origins of RhiFV and PSFVaye. We only discuss scenarios in which the number of viral cross-species transmissions is minimal and FV-host co-speciation is assumed to be the main background.

Regarding PSFVaye, we may find that other lemuriforme FVs disclose similarity to PSFVaye and together exhibit a pattern of co-speciation with their hosts. This would strongly suggest

that FVs were introduced into lemurs before the lemur diversification, and have been co-diversifying with their hosts for >59 Myr. On the other hand, finding that other lemuriforme FVs clustering with PSFVgal would imply that the cross-species transmission that gave rise to PSFVaye occurred after the lemur diversification, and that lemurs have their own specific FVs, which were not sampled in our study.

Regarding RhiFV, its natural host is *Rhinolophus affinis,* which is a Yinpterochiroptera bat endemic to China. If the cross-species transmission that gave rise to RhiFV happened after the radiation of Yinpterochiroptera bats, other Yinpterochiroptera FVs should disclose similarity to fereuungulata FVs, if they indeed exist. Conversely, finding other Yinpterochiroptera FVs clustering with RhiFV and exhibiting a co-divergence pattern with their hosts would imply that the cross-species transmission occurred before the radiation of Yinpterochiroptera bats. Observing such pattern would have a profound implication regarding the origin of bats. The basal diversification of Yinpterochiroptera bats was estimated to occur ∼63−53 Ma (88), but the India landmass did not come into contact with Laurasia until ∼50−35 Ma (**Figure 1.8C**). This would imply that the cross-species transmission that gave rise to RhiFV did not occur in Laurasia, but happened on the India landmass while it was drifting across the Indian Ocean. The most parsimonious scenario that would explain this pattern (i.e. having the least number of viral cross-species transmissions) is that the ancestor of Yinpterochiroptera bats (whose lineage originated ∼88Ma (38)) migrated from Africa to the India landmass, were infected by indigenous FVs, and thereafter were carried to Laurasia via the Indian landmass continental drift upon which the bats and their FVs diversified. This in turn would posit the origin of bats to be in Gondwana and not Laurasia. Such results would demonstrate how virus evolution can be used to elucidate host evolution under specific circumstances. Our results highlight the value of integrating multiple sources of information to elucidate the evolutionary history of mammals and their viruses, including continental and geographical histories, ancestral host locations, in addition to the natural history of host and virus.

1.5 Materials and Methods

1.5.1 Discovery and characterisation of ChrEFV and PSFVaye

Publically available GenBank whole genome shotgun (WGS) sequences were screened using tBLASTn and various FV protein sequences (**Table 1.3**). As previously reported by Han and Worobey (36), several matches were returned from the WGS assembled sequence of the aye-aye, *Daubentonia madagascariensis,* spanning four non-overlapping contigs. Genomic sequencing confirmed that these four contigs represent different parts of a single ERV (William Switzer, personal communication). Furthermore, this *in silico* screening process also returned one contig from the WGS assembled sequence of a Cape golden mole (*Chrysochloris asiatica*), designated 'ChrEFV' here. Both were annotated via sequence homology. The presence of simple repetitive elements was determined by CENSOR (http://www.girinst.org/censor/).

1.5.2 Estimating the integration dates of PSFVaye and ChrEFV with Monte Carlo simulation

To estimate the age of PSFVaye and ChrEFV, we analysed the frequencies of in-frame stop-codons in their Pol protein sequences. Excluding a hypothetical last stop codon, PSFVaye (1,163 codons) and ChrEFV (975 codons) Pols contain 9 and 14 in-frame stop codon mutations, which are equivalent to stop codon frequencies of 0.00774 and 0.0144, respectively. Using a Monte Carlo simulation approach and *pol*s of PSFVgal, SFVcpz, SFVagm, SFVsqu, FFV, and SloEFV as model sequences, we estimated two cumulative distribution functions (CDFs) of time duration for in-frame stop codons to be accumulated at these two particular frequencies for each model sequence. Each *pol* model sequence was mutated *in silico* over various hypothetical time durations from 5 Myr to 200 Myr with an increment of 5 Myr. The mutation (substitution) process was assumed to be neutral, homogenous, and independent across all sites and base types with the rate of 2.2×10^{-9} s/n/y, an average substitution rate of mammalian genomes (89). 1,000 simulations were performed for each period of time and the frequency of in-frame stop codons was then calculated for each

Table 1.3 GenBank accession numbers of protein sequences used as probes to search for integrated mammalian foamy viruses (FVs) as well as for phylogenetic analyses.

FV† (Accession number)	Gag	Pol	Env	Bel-1	Bel-2
PFV (Y07725)	CAA69002	CAA69003	CAA69004	CAA69005	NA‡
SFVcpz (U04327)	AAA19977	AAA19978	AAA19979	AAA19980	AAA19981
SFVgor (HM245790)	ADN65590	ADN65591	ADN65592	ADN65594	ADN65593
SFVora (AJ544579)	CAD67561	CAD67562	CAD67563	CAD67564	CAD67565
SFVagm (M74895)	AAA47795	AAA47796	AAA47798	AAA47799	AAA47800
SFVmac (NC_010819)	YP_001961121	YP_001961122	YP_001961123	YP_001961124	NA
SFVmar (GU356395)	ADE05999	ADE06000	ADE06001	ADE06002	ADE06003
SFVsqu (GU356394)	ADE05994	ADE05995	ADE05996	ADE05997	ADE05998
SFVspm (EU010385)	ABV59398	ABV59399	ABV59400	ABV59401	ABV59402
BFV (U94514)	AAB68769	AAB68770	AAB68771	AAB68772	AAB68773
EFV (AF201902)	AAF64413	AAF64414	AAF64415	AAF64416	AAF64417
FFV (Y08851)	CAA70074	CAA70075	CAA70076	CAA70077	CAA70078

'NA' = 'Not available'.
†Acronyms used in FV names: PFV, prototype FV; SFVcpz, chimpanzee simian FV (SFV); SFVgor, gorilla SFV; SFVora, orangutan SFV; SFVagm, African green monkey SFV; SFVmac, macaque SFV; SFVmar, common marmoset SFV; SFVsqu, squirrel SFV; SFVspm, spider monkey SFV; BFV, bovine FV; EFV, equine FV; FFV, feline FV.

simulation. For each sequence model, two separate CDFs of time duration were constructed using stop-codon frequencies of 0.00774 and 0.0144. In total, we constructed six CDFs for each of the two stop-codon frequencies. Means, medians, modes and 95% confidence intervals of the time duration for in-frame stop codons to accumulate at the two frequencies were calculated.

1.5.3 Phylogenetic reconstruction

To investigate the phylogenetic relationships of PSFVgal, PSFVaye, and ChrEFV with other retroviruses, a consensus unrooted retroviral phylogeny was reconstructed using a manually-curated alignment of 240 RT protein sequences (162 aa; **Data set S1.1** in **Electronic Appendix**) from alpha-, beta-, gamma-, delta-, epsilon-, lenti-, and spuma-retroviruses using RAxML 7.2.8-HPC2 on XSEDE (70). The LG+Γ(4)+F model was determined to fit the data best by using ProtTest 2.4 (71). Bootstrap support values for the branching order were calculated using 5,000 pseudoreplicates.

To further investigate the phylogenetic relationships of PSFVgal, PSFVaye, and ChrEFV with other FVs, a Bayesian FV phylogeny was estimated by using MrBayes 3.2.1 (76) and a manually-curated alignment of concatenated Pol-Env sequences (893 and 603 amino acids, respectively; **Data set S1.2** in **Electronic Appendix**). We used CoeEFV (90) to root our mammalian FV tree since it is the most immediate outgroup of mammalian FVs known to date. Potential recombination among the sequences within the alignment were checked by using VisRD3 (69) both at nucleotide and protein levels. For both analyses, the null distribution was built based on 1,000 sets of randomly-shuffled sequences and the extended statistical geometry (Hamming) weighting option was selected. In the nucleotide analysis, the window size and step size was 300 and 60 nt, respectively; while in the protein analysis, the window size and step size was 100 and 20 aa, respectively. The results showed no significant evidence for recombination. The phylogeny reconstruction was performed by using best available protein-specific model partitions of rtREV+Γ(4)+F (Pol) and WAG+I+Γ(4)+F (Env) as determined by ProtTest 2.4 (71) under the AIC criterion. The MCMC was run for 5,000,000

steps with the initial 25% discarded as burn-in. Trees and parameters were logged every 100 steps. Parameter value convergence and sampling independency were manually inspected and had effective sample sizes >1,000, indicating that they were all well sampled and had converged.

1.5.4 Phylogenetic reconciliation analysis

To reconstruct possible FV-mammal co-speciation histories, the topology of the FV tree was compared to that of their mammalian hosts previously published in (38) by using the co-phylogeny reconstruction software Jane v4.0 (72) (Genetic algorithm: number of generations = 100, population size = 100). The vertex-based cost mode was used, and the costs were set as follows: co-speciation = −1, duplication = 0, duplication & host switch = 0, loss = 0, and failure to diverge = 0. This setting was adopted in order to maximise the number of co-speciation events. The probability of observing the inferred number of co-speciation events by chance was calculated by using random tip mapping method implemented in Jane v4.0 (72) (Genetic algorithm: number of generation = 100, population size = 100, sample size = 1,000). This method of analysing phylogenetic incongruence cannot formally differentially weight transmission at different taxonomic levels however, and the derivation of *P* values relies upon the interchangeability of branches. Thus, it does not discriminate between distant and close interspecies transmissions.

1.5.5 Approximate unbiased (AU) test

Our inferred phylogeny suggests two possible cases of FV cross–species transmissions: one involves PSFVaye and RhiFV and the other involves NWM FVs. AU tests using multi-scale bootstrapping (73) were performed to investigate whether or not our Pol-Env/Pol/Env alignments reject the co-speciation model for PSFVaye and RhiFV, and also for NWM FVs (**Figure 1.5**). Four completing alternative phylogenetic placements of PSFVaye and RhiFV (**Figure 1.5A**), and two alternative placements of NWM FVs (**Figure 1.5B**) are compared to each other. The site-wise log likelihoods used in the tests were computed by using PAML 4.7a

(74). The best available amino acid substitution model, as determined by ProtTest 2.4 (71) under the AIC criterion, was used (Pol: rtREV+Γ(4)+F; Env: WAG+Γ(4)+F). A molecular clock was not imposed, and ambiguous sites were included. AU tests were performed in Consel (75) with default settings.

1.5.6 Examination of the effects of host neutral genetic changes on phylogenetic inference

It has been suggested that the phylogenetic placement of PSFVaye outside the clade of Boreoeutherian FVs might be erroneous, due to the host neutral changes accumulated in the sequence. To examine this possibility, an additional Bayesian phylogeny containing *in silico* mutated SFVagm, SFVsqu, and PSFVgal sequences was reconstructed by using the same settings as described above. The sequences were mutated with an overestimate of the mammalian neutral substitution rate of 1×10^{-8} s/n/y over a time period of 40 Myr, which is a mean age estimate of PSFVaye. The substitution process was assumed to be neutral, homogenous, and independent across sites and base types. The alignment is available in **Electronic Appendix**, **Data set S1.3**.

1.5.7 FV-host divergence correlation analysis

To evaluate the FV-host co-speciation history inferred by the phylogenetic reconciliation analysis, we identified branches in the FV and host trees that represent their co-evolutionary histories and investigated whether or not the lengths of these corresponding branches were linearly correlated, i.e. whether or not the accumulation of FV sequence divergence occurred in proportion to host divergence, by using linear regression implemented in MATLAB (91). Outliers were determined by using Cook's distance inspection. Outliers were removed one at a time using a threshold value of 3× mean of the observed Cook's distances and the model was re-fitted until no outliers were found. The model was then used to estimate how long the parental branch of RhiFV and PSFVaye (0.0564 amino-acid substitutions per site) is in units of time. Assuming that the radiation of exafroplacentalians was 101.1 Ma (38), we calculated

that PSFVaye and RhiFV shared a MRCA ~93.34 (95% confidence interval = 85.10−101.59) Ma.

1.6 References

1. **Achong BG**, **Mansell PW**, **Epstein MA**, **Clifford P**. 1971. An unusual virus in cultures from a human nasopharyngeal carcinoma. J. Natl. Cancer Inst. **46**:299–307.
2. **Flügel RM**, **Rethwilm A**, **Maurer B**, **Darai G**. 1987. Nucleotide sequence analysis of the env gene and its flanking regions of the human spumaretrovirus reveals two novel genes. EMBO J. **6**:2077–2084.
3. **Maurer B**, **Bannert H**, **Darai G**, **Flügel RM**. 1988. Analysis of the primary structure of the long terminal repeat and the gag and pol genes of the human spumaretrovirus. J. Virol. **62**:1590–1597.
4. **Rogers NG**, **Basnight M**, **Gibbs CJ**, **Gajdusek DC**. 1967. Latent viruses in chimpanzees with experimental Kuru. Nature **216**:446–449.
5. **Herchenröder O**, **Renne R**, **Loncar D**, **Cobb EK**, **Murthy KK**, **Schneider J**, **Mergia A**, **Luciw PA**. 1994. Isolation, cloning, and sequencing of simian foamy viruses from chimpanzees (SFVcpz): high homology to human foamy virus (HFV). Virology **201**:187–199.
6. **Bieniasz PD**, **Rethwilm A**, **Pitman R**, **Daniel MD**, **Chrystie I**, **McClure MO**. 1995. A comparative study of higher primate foamy viruses, including a new virus from a gorilla. Virology **207**:217–228.
7. **Schulze A**, **Lemey P**, **Schubert J**, **McClure MO**, **Rethwilm A**, **Bodem J**. 2011. Complete nucleotide sequence and evolutionary analysis of a gorilla foamy virus. J. Gen. Virol. **92**:582–586.
8. **McClure MO**, **Bieniasz PD**, **Schulz TF**, **Chrystie IL**, **Simpson G**, **Aguzzi A**, **Hoad JG**, **Cunningham A**, **Kirkwood J**, **Weiss RA**. 1994. Isolation of a new foamy retrovirus from orangutans. J. Virol. **68**:7124–7130.
9. **Verschoor EJ**, **Langenhuijzen S**, **van den Engel S**, **Niphuis H**, **Warren KS**, **Heeney JL**. 2003. Structural and evolutionary analysis of an orangutan foamy virus. J. Virol. **77**:8584–8587.
10. **Rustigian R**, **Johnston P**, **Reihart H**. 1955. Infection of monkey kidney tissue cultures with virus-like agents. Proc. Soc. Exp. Biol. Med. **88**:8–16.

11. **Johnston PB**. 1961. A second immunologic type of simian foamy virus: monkey throat infections and unmasking by both types. J. Infect. Dis. **109**:1–9.

12. **Kupiec JJ**, **Kay A**, **Hayat M**, **Ravier R**, **Périès J**, **Galibert F**. 1991. Sequence analysis of the simian foamy virus type 1 genome. Gene **101**:185–194.

13. **Stiles GE**, **Bittle JL**, **Cabasso VJ**. 1964. Comparison of simian foamy virus strains including a new serological type. Nature **201**:1350–1351.

14. **Renne R**, **Friedl E**, **Schweizer M**, **Fleps U**, **Turek R**, **Neumann-Haefelin D**. 1992. Genomic organization and expression of simian foamy virus type 3 (SFV-3). Virology **186**:597–608.

15. **Heberling RL**, **Kalter SS**. 1975. Isolation of foamy viruses from baboon (*Papio cynocephalus*) tissues. Am. J. Epidemiol. **102**:25–29.

16. **Johnston PB**. 1971. Taxonomic features of seven serotypes of simian and ape foamy viruses. Infect. Immun. **3**:793–799.

17. **Pacheco B**, **Finzi A**, **McGee-Estrada K**, **Sodroski J**. 2010. Species-specific inhibition of foamy viruses from South American monkeys by New World Monkey TRIM5{alpha} proteins. J. Virol. **84**:4095–4099.

18. **Hooks JJ**, **Gibbs CJ**, **Chou S**, **Howk R**, **Lewis M**, **Gajdusek DC**. 1973. Isolation of a new simian foamy virus from a spider monkey brain culture. Infect. Immun. **8**:804–813.

19. **Thümer L**, **Rethwilm A**, **Holmes EC**, **Bodem J**. 2007. The complete nucleotide sequence of a New World simian foamy virus. Virology **369**:191–197.

20. **Hooks JJ**, **Gibbs CJ**. 1975. The foamy viruses. Bacteriol. Rev. **39**:169–185.

21. **Troncoso LL**, **Muniz CP**, **Siqueira JD**, **Curty G**, **Schrago CG**, **Augusto A**, **Fedullo L**, **Soares MA**, **Santos AF**. 2015. Characterization and comparative analysis of a simian foamy virus complete genome isolated from Brazilian capuchin monkeys. Virus Res. **208**:1–6.

22. **Marczynska B**, **Jones CJ**, **Wolfe LG**. 1981. Syncytium-forming virus of common marmosets (*Callithrix jacchus jacchus*). Infect. Immun. **31**:1261–1269.

23. **Muniz CP**, **Troncoso LL**, **Moreira M a**, **Soares E a**, **Pissinatti A**, **Bonvicino CR**, **Seuánez HN**, **Sharma B**, **Jia H**, **Shankar A**, **Switzer WM**, **Santos AF**, **Soares M a**.

2013. Identification and characterization of highly divergent simian foamy viruses in a wide range of new world primates from Brazil. PLoS One **8**:e67568.

24. **Ghersi BM**, **Jia H**, **Aiewsakun P**, **Katzourakis A**, **Mendoza P**, **Bausch DG**, **Kasper MR**, **Montgomery JM**, **Switzer WM**. 2015. Wide distribution and ancient evolutionary history of simian foamy viruses in New World primates. Retrovirology **12**:89.

25. **Riggs JL**, **Oshirls**, **Taylor DO**, **Lennette EH**. 1969. Syncytium-forming agent isolated from domestic cats. Nature **222**:1190–1191.

26. **Winkler I**, **Bodem J**, **Haas L**, **Zemba M**, **Delius H**, **Flower R**, **Flügel RM**, **Löchelt M**. 1997. Characterization of the genome of feline foamy virus and its proteins shows distinct features different from those of primate spumaviruses. J. Virol. **71**:6727–6741.

27. **Malmquist WA**, **Van der Maaten MJ**, **Boothe AD**. 1969. Isolation, immunodiffusion, immunofluorescence, and electron microscopy of a syncytial virus of lymphosarcomatous and apparently normal cattle. Cancer Res. **29**:188–200.

28. **Renshaw RW**, **Casey JW**. 1994. Transcriptional mapping of the 3' end of the bovine syncytial virus genome. J. Virol. **68**:1021–1028.

29. **Tobaly-Tapiero J**, **Bittoun P**, **Neves M**, **Guillemin MC**, **Lecellier CH**, **Puvion-Dutilleul F**, **Gicquel B**, **Zientara S**, **Giron ML**, **de Thé H**, **Saïb A**. 2000. Isolation and characterization of an equine foamy virus. J. Virol. **74**:4064–4073.

30. **Wu Z**, **Ren X**, **Yang L**, **Hu Y**, **Yang J**, **He G**, **Zhang J**, **Dong J**, **Sun L**, **Du J**, **Liu L**, **Xue Y**, **Wang J**, **Yang F**, **Zhang S**, **Jin Q**. 2012. Virome analysis for identification of novel mammalian viruses in bat species from Chinese provinces. J. Virol. **86**:10999–11012.

31. **Bolisetty M**, **Blomberg J**, **Benachenhou F**, **Sperber G**, **Beemon K**. 2012. Unexpected diversity and expression of avian endogenous retroviruses. MBio **3**:e00344-12.

32. **Hayward A**, **Cornwallis CK**, **Jern P**. 2015. Pan-vertebrate comparative genomics unmasks retrovirus macroevolution. Proc. Natl. Acad. Sci. U. S. A. **112**:464–469.

33. **Gifford R**, **Kabat P**, **Martin J**, **Lynch C**, **Tristem M**. 2005. Evolution and distribution of class II-related endogenous retroviruses. J. Virol. **79**:6478–6486.

34. **Baillie GJ**, **van de Lagemaat LN**, **Baust C**, **Mager DL**. 2004. Multiple groups of endogenous betaretroviruses in mice, rats, and other mammals. J. Virol. **78**:5784–5798.

35. **Katzourakis A**, **Gifford RJ**, **Tristem M**, **Gilbert MTP**, **Pybus OG**. 2009. Macroevolution of complex retroviruses. Science **325**:1512.

36. **Han G-Z**, **Worobey M**. 2012. An endogenous foamy virus in the aye-aye (*Daubentonia madagascariensis*). J. Virol. **86**:7696–7698.

37. **Switzer WM**, **Salemi M**, **Shanmugam V**, **Gao F**, **Cong M-E**, **Kuiken C**, **Bhullar V**, **Beer BE**, **Vallet D**, **Gautier-Hion A**, **Tooze Z**, **Villinger F**, **Holmes EC**, **Heneine W**. 2005. Ancient co-speciation of simian foamy viruses and primates. Nature **434**:376–380.

38. **Bininda-Emonds ORP**, **Cardillo M**, **Jones KE**, **MacPhee RDE**, **Beck RMD**, **Grenyer R**, **Price SA**, **Vos RA**, **Gittleman JL**, **Purvis A**. 2007. The delayed rise of present-day mammals. Nature **446**:507–512.

39. **Switzer WM**, **Bhullar V**, **Shanmugam V**, **Cong M-E**, **Parekh B**, **Lerche NW**, **Yee JL**, **Ely JJ**, **Boneva R**, **Chapman LE**, **Folks TM**, **Heneine W**. 2004. Frequent simian foamy virus infection in persons occupationally exposed to nonhuman primates. J. Virol. **78**:2780–2789.

40. **Meiering CD**, **Linial ML**. 2001. Historical perspective of foamy virus epidemiology and infection. Clin. Microbiol. Rev. **14**:165–176.

41. **Liu W**, **Worobey M**, **Li Y**, **Keele BF**, **Bibollet-Ruche F**, **Guo Y**, **Goepfert PA**, **Santiago ML**, **Ndjango J-BN**, **Neel C**, **Clifford SL**, **Sanz C**, **Kamenya S**, **Wilson ML**, **Pusey AE**, **Gross-Camp N**, **Boesch C**, **Smith V**, **Zamma K**, **Huffman MA**, **Mitani JC**, **Watts DP**, **Peeters M**, **Shaw GM**, **Switzer WM**, **Sharp PM**, **Hahn BH**. 2008. Molecular ecology and natural history of simian foamy virus infection in wild-living chimpanzees. PLoS Pathog. **4**:e1000097.

42. **Stone AC**, **Battistuzzi FU**, **Kubatko LS**, **Perry GH**, **Trudeau E**, **Lin H**, **Kumar S**. 2010. More reliable estimates of divergence times in *Pan* using complete mtDNA sequences and accounting for population structure. Philos. Trans. R. Soc. Lond. B. Biol. Sci. **365**:3277–3288.

43. **Springer MS**, **Stanhope MJ**, **Madsen O**, **de Jong WW**. 2004. Molecules consolidate the placental mammal tree. Trends Ecol. Evol. **19**:430–438.

44. **Murphy WJ**, **Pringle TH**, **Crider TA**, **Springer MS**, **Miller W**. 2007. Using genomic data

to unravel the root of the placental mammal phylogeny. Genome Res. **17**:413–421.

45. **Waddell PJ**, **Cao Y**, **Hasegawa M**, **Mindell DP**. 1999. Assessing the Cretaceous superordinal divergence times within birds and placental mammals by using whole mitochondrial protein sequences and an extended statistical framework. Syst. Biol. **48**:119–137.

46. **Shoshani J**, **McKenna MC**. 1998. Higher taxonomic relationships among extant mammals based on morphology, with selected comparisons of results from molecular data. Mol. Phylogenet. Evol. **9**:572–584.

47. **Kriegs JO**, **Churakov G**, **Kiefmann M**, **Jordan U**, **Brosius J**, **Schmitz J**. 2006. Retroposed elements as archives for the evolutionary history of placental mammals. PLoS Biol. **4**:e91.

48. **Waddell PJ**, **Shelley S**. 2003. Evaluating placental inter-ordinal phylogenies with novel sequences including RAG1, gamma-fibrinogen, ND6, and mt-tRNA, plus MCMC-driven nucleotide, amino acid, and codon models. Mol. Phylogenet. Evol. **28**:197–224.

49. **Nikolaev S**, **Montoya-Burgos JI**, **Margulies EH**, **Rougemont J**, **Nyffeler B**, **Antonarakis SE**. 2007. Early history of mammals is elucidated with the ENCODE multiple species sequencing data. PLoS Genet. **3**:e2.

50. **Kupiec JJ**, **Sonigo P**. 1996. Reverse transcriptase jumps and gaps. J. Gen. Virol. **77**:1987–1991.

51. **Eastman SW**, **Linial ML**. 2001. Identification of a conserved residue of foamy virus Gag required for intracellular capsid assembly. J. Virol. **75**:6857–6864.

52. **Patton GS**, **Morris SA**, **Chung W**, **Bieniasz PD**, **McClure MO**. 2005. Identification of domains in gag important for prototypic foamy virus egress. J. Virol. **79**:6392–6399.

53. **Stange A**, **Mannigel I**, **Peters K**, **Heinkelein M**, **Stanke N**, **Cartellieri M**, **Göttlinger H**, **Rethwilm A**, **Zentgraf H**, **Lindemann D**. 2005. Characterization of prototype foamy virus gag late assembly domain motifs and their role in particle egress and infectivity. J. Virol. **79**:5466–5476.

54. **Schliephake AW**, **Rethwilm A**. 1994. Nuclear localization of foamy virus Gag precursor protein. J. Virol. **68**:4946–4954.

55. **Lindemann D**, **Pietschmann T**, **Picard-Maureau M**, **Berg A**, **Heinkelein M**, **Thurow J**, **Knaus P**, **Zentgraf H**, **Rethwilm A**. 2001. A particle-associated glycoprotein signal peptide essential for virus maturation and infectivity. J. Virol. **75**:5762–5771.

56. **Wilk T**, **Geiselhart V**, **Frech M**, **Fuller SD**, **Flügel RM**, **Löchelt M**. 2001. Specific interaction of a novel foamy virus Env leader protein with the N-terminal Gag domain. J. Virol. **75**:7995–8007.

57. **Geiselhart V**, **Bastone P**, **Kempf T**, **Schnölzer M**, **Löchelt M**. 2004. Furin-mediated cleavage of the feline foamy virus Env leader protein. J. Virol. **78**:13573–13581.

58. **Perry GH**, **Reeves D**, **Melsted P**, **Ratan A**, **Miller W**, **Michelini K**, **Louis EE**, **Pritchard JK**, **Mason CE**, **Gilad Y**. 2012. A genome sequence resource for the aye-aye (*Daubentonia madagascariensis*), a nocturnal lemur from Madagascar. Genome Biol. Evol. **4**:126–135.

59. **Di Palma F**, **Alfoldi J**, **Johnson J**, **Berlin A**, **Gnerre S**, **Jaffe D**, **MacCallum I**, **Young S**, **Walker BJ**, **Lindblad-Toh K**. 2012. The draft genome of *Chrysochloris asiatica*.

60. **Jurka J**. 1995. Origin and evolution of Alu repetitive elements, p. 25–41. *In* Maraia, RJ (ed.), Molecular Biology Intelligence Unit: The impact of short interspersed elements (SINEs) on the host genome. R.G. Landes Company, Austin.

61. **Jurka J.** 2010. SINE elements from the mouse lemur. Repbase Reports **10**:2171.

62. **Jurka J**, **Smith T**. 1988. A fundamental division in the Alu family of repeated sequences. Proc. Natl. Acad. Sci. U. S. A. **85**:4775–4778.

63. **Jurka J**. 2010. SINE elements from African elephants. Repbase Reports **10**:491.

64. **Bodem J**. 2011. Regulation of foamy viral transcription and RNA export. Adv. Virus Res. **81**:1–31.

65. **Perelman P**, **Johnson WE**, **Roos C**, **Seuánez HN**, **Horvath JE**, **Moreira MAM**, **Kessing B**, **Pontius J**, **Roelke M**, **Rumpler Y**, **Schneider MPC**, **Silva A**, **O'Brien SJ**, **Pecon-Slattery J**. 2011. A molecular phylogeny of living primates. PLoS Genet. **7**:e1001342.

66. **Horvath JE**, **Weisrock DW**, **Embry SL**, **Fiorentino I**, **Balhoff JP**, **Kappeler P**, **Wray GA**, **Willard HF**, **Yoder AD**. 2008. Development and application of a phylogenomic

toolkit: resolving the evolutionary history of Madagascar's lemurs. Genome Res. **18**:489–499.

67. **Yoder AD**, **Yang Z**. 2004. Divergence dates for Malagasy lemurs estimated from multiple gene loci: geological and evolutionary context. Mol. Ecol. **13**:757–773.

68. **Springer MS**, **Murphy WJ**, **Eizirik E**, **O'Brien SJ**. 2003. Placental mammal diversification and the Cretaceous-Tertiary boundary. Proc. Natl. Acad. Sci. U. S. A. **100**:1056–1061.

69. **Lemey P**, **Lott M**, **Martin DP**, **Moulton V**. 2009. Identifying recombinants in human and primate immunodeficiency virus sequence alignments using quartet scanning. BMC Bioinformatics **10**:126.

70. **Stamatakis A**, **Hoover P**, **Rougemont J**. 2008. A rapid bootstrap algorithm for the RAxML Web servers. Syst. Biol. **57**:758–771.

71. **Abascal F**, **Zardoya R**, **Posada D**. 2005. ProtTest: selection of best-fit models of protein evolution. Bioinformatics **21**:2104–2105.

72. **Conow C**, **Fielder D**, **Ovadia Y**, **Libeskind-Hadas R**. 2010. Jane: a new tool for the cophylogeny reconstruction problem. Algorithms Mol. Biol. **5**:16.

73. **Shimodaira H**. 2002. An approximately unbiased test of phylogenetic tree selection. Syst. Biol. **51**:492–508.

74. **Yang Z**. 2007. PAML 4: phylogenetic analysis by maximum likelihood. Mol. Biol. Evol. **24**:1586–1591.

75. **Shimodaira H**, **Hasegawa M**. 2001. CONSEL: for assessing the confidence of phylogenetic tree selection. Bioinformatics **17**:1246–1247.

76. **Ronquist F**, **Huelsenbeck JP**. 2003. MrBayes 3: Bayesian phylogenetic inference under mixed models. Bioinformatics **19**:1572–1574.

77. **Hill CL**, **Bieniasz PD**, **McClure MO**. 1999. Properties of human foamy virus relevant to its development as a vector for gene therapy. J. Gen. Virol. **80 (Pt 8)**:2003–2009.

78. **Berg A**, **Pietschmann T**, **Rethwilm A**, **Lindemann D**. 2003. Determinants of foamy virus envelope glycoprotein mediated resistance to superinfection. Virology **314**:243–252.

79. **Plochmann K**, **Horn A**, **Gschmack E**, **Armbruster N**, **Krieg J**, **Wiktorowicz T**, **Weber C**, **Stirnnagel K**, **Lindemann D**, **Rethwilm A**, **Scheller C**. 2012. Heparan sulfate is an attachment factor for foamy virus entry. J. Virol. **86**:10028–10035.

80. **Herchenröder O**, **Moosmayer D**, **Bock M**, **Pietschmann T**, **Rethwilm A**, **Bieniasz PD**, **McClure MO**, **Weis R**, **Schneider J**. 1999. Specific binding of recombinant foamy virus envelope protein to host cells correlates with susceptibility to infection. Virology **255**:228–236.

81. **Yap MW**, **Lindemann D**, **Stanke N**, **Reh J**, **Westphal D**, **Hanenberg H**, **Ohkura S**, **Stoye JP**. 2008. Restriction of foamy viruses by primate Trim5alpha. J. Virol. **82**:5429–5439.

82. **Perkovic M**, **Schmidt S**, **Marino D**, **Russell RA**, **Stauch B**, **Hofmann H**, **Kopietz F**, **Kloke B-P**, **Zielonka J**, **Ströver H**, **Hermle J**, **Lindemann D**, **Pathak VK**, **Schneider G**, **Löchelt M**, **Cichutek K**, **Münk C**. 2009. Species-specific inhibition of APOBEC3C by the prototype foamy virus protein Bet. J. Biol. Chem. **284**:5819–5826.

83. **McAllister BF**, **Werren JH**. 1997. Phylogenetic analysis of a retrotransposon with implications for strong evolutionary constraints on reverse transcriptase. Mol. Biol. Evol. **14**:69–80.

84. **Opazo JC**, **Wildman DE**, **Prychitko T**, **Johnson RM**, **Goodman M**. 2006. Phylogenetic relationships and divergence times among New World monkeys (Platyrrhini, Primates). Mol. Phylogenet. Evol. **40**:274–280.

85. **Schneider H**. 2000. The current status of the New World monkey phylogeny. An. Acad. Bras. Cienc. **72**:165–172.

86. **Hawkesworth C**, **Kelley S**, **Turner S**, **Le Roex A**, **Storey B**. 1999. Mantle processes during Gondwana break-up and dispersal. J. African Earth Sci. **28**:239–261.

87. **Wit MJ de**. 2003. Madagascar: heads it's a continent, tails it's an Island. Annu. Rev. Earth Planet. Sci. **31**:213–248.

88. **Teeling EC**, **Springer MS**, **Madsen O**, **Bates P**, **O'brien SJ**, **Murphy WJ**. 2005. A molecular phylogeny for bats illuminates biogeography and the fossil record. Science **307**:580–584.

89. **Kumar S**, **Subramanian S**. 2002. Mutation rates in mammalian genomes. Proc. Natl. Acad. Sci. U. S. A. **99**:803–808.

90. **Han G-Z**, **Worobey M**. 2012. An endogenous foamy-like viral element in the coelacanth genome. PLoS Pathog. **8**:e1002790.

91. **MATLAB**. 2010. MATLAB and Statistics Toolbox Release R2012a. The MathWorks Inc., Massachusetts.

Chapter two: Codivergence and cross-species transmissions of New World monkey foamy viruses

---

This chapter is modified from the following published paper (**Appendix II**):

Ghersi BM, Jia H, Aiewsakun P, Katzourakis A, Mendoza P, Bausch DG, Kasper MR, Montgomery JM, Switzer WM. 2015. Wide distribution and ancient evolutionary history of simian foamy viruses in New World primates. Retrovirology **12**:89.

The *Author contributions* in the publication state that “BMG, JMM, DGB, and WMS conceived the project; PA, AK, HJ, and WMS performed the sequence analyses; HJ performed the laboratory work and BMG, PM, DGB, MRK, JMM, and WMS provided samples. BMG, PA, AK and WMS wrote the paper. All authors read and commented on the manuscript. All authors read and approved the final manuscript.”

For the material in this chapter, PA performed the analysis and wrote the chapter. AK provided editorial guidance and supervisory advice. SWM and his collaborators provided the sequence data and advice.

2.1 Abstract

Foamy viruses (FVs) are complex retroviruses well known for their extended co-speciation history that they share with their hosts. Their evolutionary history has been studied extensively, and they infect a wide range of mammals, including New World monkeys (NWMs); however, little is known about this specific part of FV evolution. The study presented here reconstructs a NWM-FV co-evolutionary model, and shows that the two co-diverge with one another at the level of species and sub-populations. Phylogenetic reconciliation analyses inferred 12 co-speciation events, supported by highly similar NWM and FV divergence dates, and positively correlated NWM-FV genetic divergence. However, the analyses could not detect a virus-host co-speciation pattern at the level of host genera, inferring FV cross-genus transmissions from *Pitheciinae* to *Atelidae*, *Cacajao* to *Callithrix* or *Cebus*, between *Callithrix* and *Cebus*, and from *Lagothrix* to *Alouatta*. The work shown here highlights the importance of using temporal evidence to validate virus-host co-evolutionary models, and shows that FV cross-species transmissions might not be as rare as previously thought.

2.2 Introduction

Foamy viruses (FV) are complex retroviruses, comprising the only genus of the *Spumaretrovirinae* subfamily. They have been identified in a wide range of mammals, including apes (1–9), Old World monkeys (OWM) (10–15), New World monkeys (NWMs) (16–24), prosimians (16), cats (25, 26), cattle (27, 28), horses (29), and bats (30). The discovery and analyses of *FV genomic fossils*, known as endogenous FVs, revealed that FVs were, and perhaps still are, capable of infecting xenarthran (31) and afrotherian mammals (32, 33) as well, making the host range of FVs cover the entirety of eutherian mammals.

The history of FV evolution has been investigated extensively. Broad scale evolutionary analyses have consistently shown that FVs have an extended co-evolutionary history with their mammalian hosts, co-diversifying with one another for >100 million years (Myr) since the basal radiation of eutherians (31–33). Facilitated by the wealth of molecular data, the history of simian FVs is perhaps the best described part of FV evolution, in particular those of OWM and ape (OWMA) FVs, which have an extended co-speciation pattern with their hosts, dating back >30 Myr ago (Ma) to the origin of OWMAs (34). In contrast, relatively little is known about NWM FV evolution, with the majority of studies performed using captive animals in the US, and evidence of infection was demonstrated by using only serology. Because of this, only a limited number of molecular sequences of NWM FVs are available (16, 18, 20, 22).

To date, only four complete genomes of NWM FVs have been reported, all isolated from captive monkeys, including a squirrel monkey (*Saimiri sp.*) (SFVsqu) (17), a common marmoset (*Callithrix jacchus*) (SFVmar) (17), a spider monkey (*Ateles sp.*) (SFVspm) (19), and a Brazilian yellow-breasted capuchin monkey (*Cebus xanthosternos*) (SFVcap) (21). Unlike in the case of OWMA FVs, while all four clustered together phylogenetically, studies consistently demonstrated that the branching pattern within the clade of NWM FVs does not mirror with their hosts (17, 21, 32), indicating that their evolution is dominated by frequent cross-species transmissions. This co-evolutionary pattern, however, could not be verified, and might not represent the ‘true’ long-term co-evolutionary relationship of NWMs and their FVs. This is

because all four sequences were isolated from captive monkeys, and thus sporadic, recent cross-species transmissions from other captive distantly related monkeys could not be excluded.

In a recent study, a large-scale survey of NWM FVs in Brazil has been conducted, identifying FVs in 8 NWM genera, including capuchin (*Cebus* species), owl (*Aotus sp.*), marmoset (*Callithrix sp.*), tamarin (*Saguinus sp.*), squirrel (*Saimiri sp.*), titi (*Callicebus sp.*), saki (*Chiropotes sp.*), and howler (*Alouatta sp.*) monkeys (23). However, phylogenetic analyses were performed only on NWM FVs isolated from captive *Alouatta,* and *Cebus* monkeys, with an addition of SFVsqu, SFVmar, and SFVspm sequences available from GenBank. Although these results are informative, our knowledge of NWM FV evolution is still far from complete.

Here, we re-evaluate the long-term co-evolutionary relationship of NWM FVs and their hosts, by using FVs obtained from captive and wild-caught NWMs of 8 different genera, including *Alouatta*, *Ateles*, *Cacajao*, *Callithrix*, *Cebus*, *Lagothrix*, *Pithecia*, and *Saimiri*. Our results provide new insights into NWM FV evolution, and highlight the value of temporal evidence in reconstructing viral-host co-evolutionary models.

2.3 Results

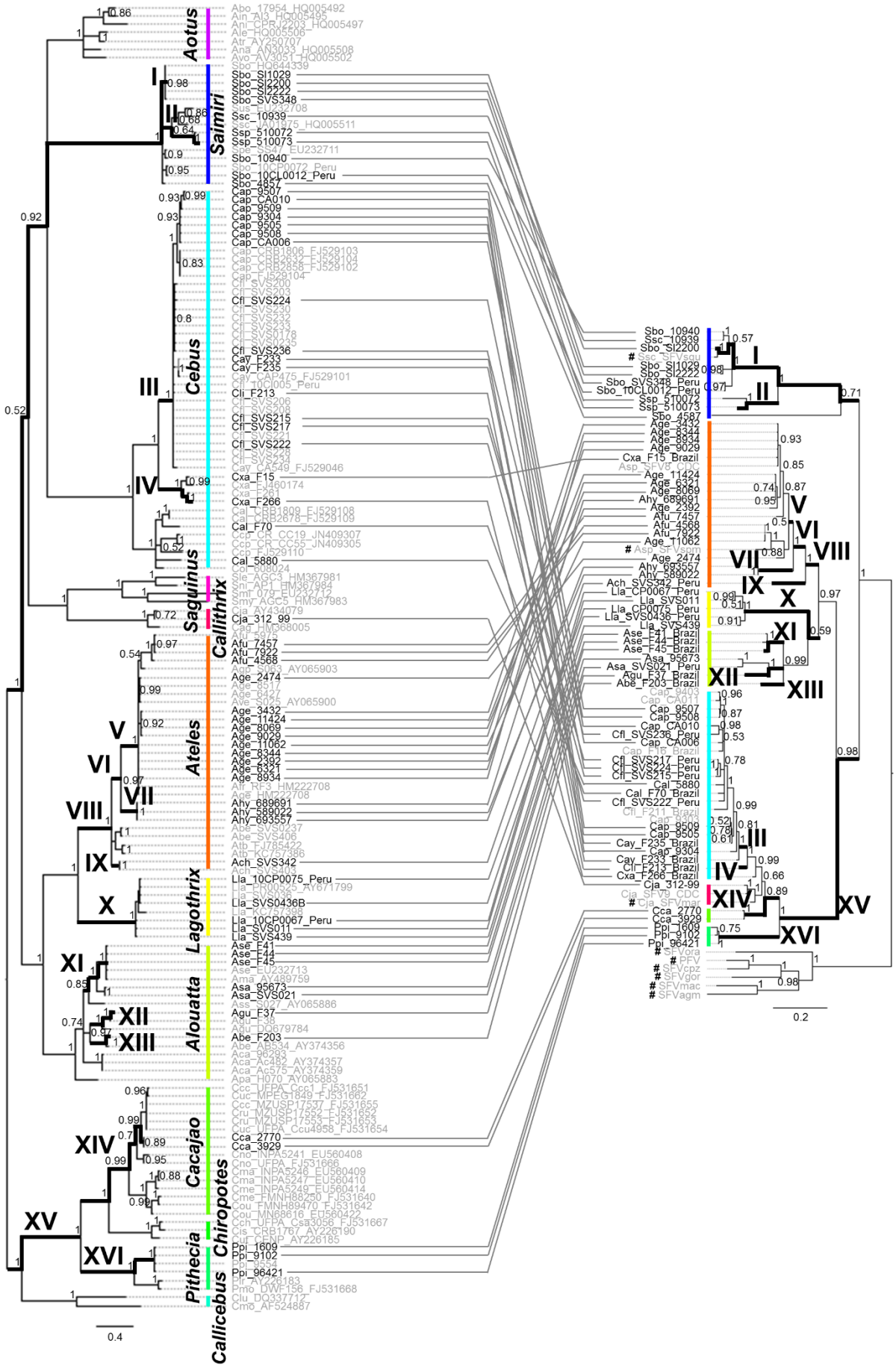
2.3.1 Phylogenetic reconciliation analysis

To better understand how FVs co-evolve with their NWM hosts, we perform evolutionary analyses on 74 FVs collected from 8 NWM genera, including *Alouatta*, *Ateles*, *Cacajao*, *Callithrix*, *Cebus*, *Lagothrix*, *Pithecia*, and *Saimiri* monkeys (**Table S2.1** in **Electronic Appendix**). The sequences were provided by William Switzer (Centers for Disease Control and Prevention, personal communication), obtained from wild-caught and captive Brazilian NWMs, and NWMs kept at Peruvian zoos, rescue centres, and illegal trade markets, as well as US zoological gardens, and research centres. To reconstruct possible co-evolutionary

histories between FVs and NWMs, we estimated their phylogenies, and compared the two together.

The host tree was reconstructed under the Bayesian phylogenetic framework by using *CytB* sequences, including 156 NWM individuals (**Table S2.1** in **Electronic Appendix**), and was rooted according to the phylogeny in (35). Their sequences were also provided by William Switzer (Centers for Disease Control and Prevention, personal communication). We found that *CytB* sequences of the same host genus cluster together to form monophyletic clades, and that the topology of the estimated tree is generally similar to the one obtained by Perelman et al. (35) (**Figure 2.1**, left**,** and **Figure 2.2A**) with one exception; while we estimated *Aotus* to be a sister clade to the clade of *Saimiri-Cebus* clade, Perelman et al. found that *Aotus* is sister to the *Callithrix-Saguinus* clade. This, however, would not affect our analyses since our FV tree does not contain any *Aotus* FVs. Others also have shown that the placement of *Aotus* in NWM phylogeny is not completely resolved (36).

A Bayesian tree of NWM FVs was estimated by using *pol* nucleotide sequences, and was rooted with OWMA FVs. The tree is shown in **Figure 2.1** (right) and **Figure 2.2B**. We noticed that there are some inconsistencies between our NWM FV tree and the one that was estimated in the previous study of Brazilian NWM FVs however (23). In the previous study, based on an analysis of short *pol* sequences (17 sequences, 276 nt), 'Cxa F266' isolated from a *Cebus xanthosternos* was found to cluster with a *Callithrix* FV, suggestive of a cross-genus transmission (23). However, in this work, we found Cxa F266 to be more closely related to *Cebus* FVs than to *Callithrix* FVs, consistent with the results from an analysis of long terminal repeat-*gag* sequences (18 sequences, 265 nt) which was also conducted in the previous study (23). Combined, it is thus more likely that Cxa F266 is in fact a *Cebus* FV, and does not represent a cross-genus transmission. Another inconsistency we observed is that while our study showed that SFVspm is more closely related to SFVmar than to SFVsqu (which is consistent with the results from previous Gag and Pol protein analyses (17, 37)), the previous study investigating evolution of NWM FVs in Brazil (23) reported that SFVspm is more closely

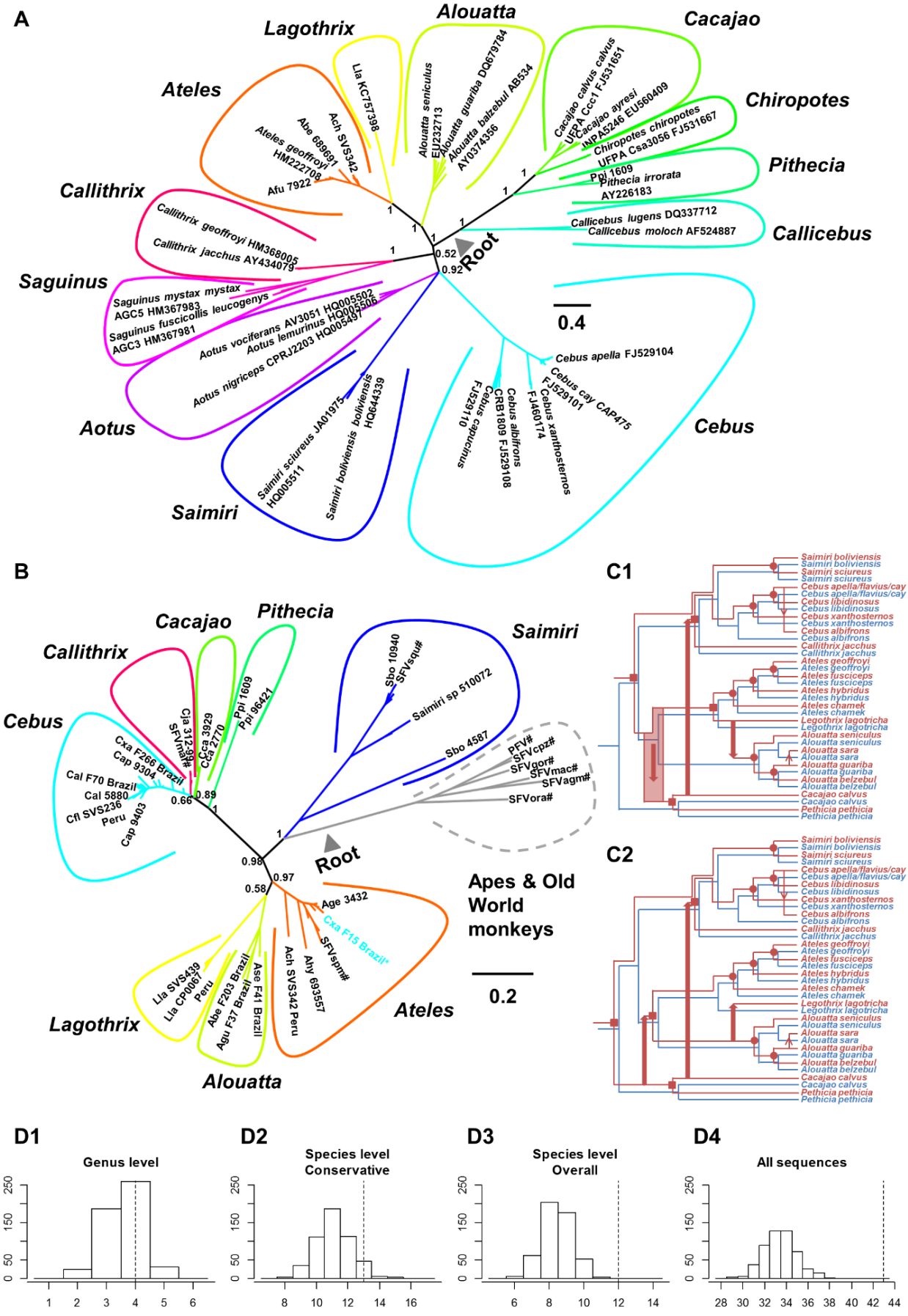

Aotus
Saimiri
Cebus
Saguinus
Callithrix
Ateles
Lagothrix
Alouatta
Cacajao
Chiropotes
Pithecia
Callicebus
I
II
III
IV
V
VI
VII
VIII
IX
X
XI
XII
XIII
XIV
XV
XVI
0.4
0.2

(See figure on previous page.)

Figure 2.1 Phylogenies of New World monkeys (NWM) and their foamy viruses (FVs). (**Left**) A NWM phylogeny estimated from a cytochrome-B nucleotide alignment (156 sequences, 618 nt). (**Right**) a NWM FV phylogeny estimated from a polymerase nucleotide alignment (80 sequences, 412 nt). Both trees were constructed under the Bayesian phylogenetic framework, and molecular clocks were not imposed. The host tree was rooted according to the tree in Perelman et al. (35). The FV tree was rooted with ape and Old World monkey FVs. The 'backbone' polymerase sequences in the FV tree are labelled with 'hash'. The scale bars are in the units of substitutions per site, and numbers on nodes are posterior probabilities. Grey lines indicate FV-host associations. FVs without associating hosts are labelled in grey, and were excluded from the phylogenetic reconciliation analysis. Thick branches are co-diverging branches used in the FV-host divergence correlation analysis, labelled with roman numerals (**I**-**XVI**), referring to dots in **Figure 2.3C**. See **Table S2.1** in **Electronic Appendix** for a complete list of species codes used in the study.

related to SFVsqu than SFVmar. These disparities are likely because the *pol* sequences used in the previous study were too short and/or the number of SFV sequences used in the analysis was too low.

Overall, we found that NWM FVs isolated from the same host genus tend to cluster together forming monophyletic clades. 'Cxa F15 Brazil' is the only exception to this; as shown in the previous study (23), this virus was isolated from a *Cebus xanthosternos* monkey but instead clustered within the *Ateles* FV clade. This likely represents a recent cross-genus FV transmission, which has been reported previously in captive and wild OWMAs (34, 38, 39). Our phylogenetic reconciliation analysis inferred another three cross-genus FV transmissions: (i) from an ancestral *Atelidae* monkey to an ancestral *Pitheciinae* monkey (**Figure 2.2C1**) or vice versa (**Figure 2.2C2**), (ii) from an ancestral *Cacajao* monkey to the lineage giving rise to *Cebus* and *Callithrix* monkeys, and (iii) from an ancestral *Lagothrix* monkey to an ancestral *Alouatta* monkey (**Figure 2.2C1**) or vice versa (**Figure 2.2C2**). Four potential co-speciation events at the level of host genera were also inferred: (i) the divergence of the *Pithecia* FV from *Cacajao* FV, (ii) the separation of the *Ateles* FV from *Lagothrix* FV (**Figure 2.2C1**) or *Alouatta* FV (**Figure 2.2C2**) FV, (iii) the split of an ancestral *Cebus* FV from an ancestral *Callithrix* FV, and (iv) the divergence of the *Saimiri* FV from *Atelidae* FV (**Figure 2.2C1**) or *Pitheciidae* FV (**Figure 2.2C2**). Nevertheless, the clade of *Lagothrix* and *Alouatta* FVs is weakly supported (posterior probability = 0.58, **Figure 2.2B**). An alternative scenario could imply that *Atelidae* FVs co- diverged with their hosts, decreasing the number of cross-genus transmissions by one and thereby increasing the number of co-divergence events by one. These four (or five) co-


A
Alouatta
Lagothrix
Cacajao
Ateles
Chiropotes
Callithrix
Pithecia
Callicebus
Saguinus
Aotus
Cebus
Saimiri
Root
0.4
B
Cacajao
Pithecia
Saimiri
Callithrix
Cebus
Root
Apes & Old World monkeys
0.2
Lagothrix
Ateles
Alouatta
C1
C2
D1
Genus level
D2
Species level Conservative
D3
Species level Overall
D4
All sequences

(See figure on previous page.)

Figure 2.2 Co-evolutionary history of New World monkeys (NWM) and their foamy viruses (FVs). (**A**) A Bayesian phylogeny of NWM hosts, rooted according to the phylogeny in (35). (**B**) A Bayesian phylogeny of NWM FVs, rooted with six ape and Old World monkey FVs. 'Backbone' sequences are indicated with '*hash*'. 'Cxa F15 Brazil' (written in *blue* and indicated with an *asterisk* '*') was isolated from a *Cebus xanthosternos* monkey in Brazil, but was found to be placed well within the clade of *Ateles* FVs. The roots of the trees are indicated by *grey triangles*. *Numbers* on nodes are posterior probability node supports. The *scale bars* are in the units of substitutions per site. The comprehensive FV and host trees are shown in **Figure 2.1**. (**C**) Competing models for the co-evolutionary histories of NWM FVs (*red*) and their hosts (*blue*) at the level of species. The directions of transmissions are indicated by *red arrows*. *Small arrows* indicate cross-species transmissions, and *large arrows* indicate cross-genus transmissions. The *red transparent bars* show the uncertainty of the cross-species transmission timing. Four co-speciation events at the genus level are indicated by *solid red squares*, and those at the species level are indicated by *solid red circles*. The trees are not scaled to time. (**D**) The distributions of the number of co-speciation events expected to occur by chance; **D1** genus level; **D2** species level, conservative tree collapsing method; **D3** species level, overall; and **D4** all sequences. The *dotted line* indicates the actual observed number of co-speciation events.

divergence events at the genus level, however, are not greater than expected to occur by chance (random tip mapping: sample size = 500, *P* value (4 co-speciation events) = 0.516, *P* value (5 co-speciation events) = 0.062, **Figure 2.2D1**).

To further examine whether or not NWM FVs co-speciate with their hosts at the species level, we collapsed the clades comprising host and viral sequences of the same species into one representative sequence and compared the trees. In total, 13 co-speciation events were inferred. Again, this is not greater than expected to occur by chance (random tip mapping: sample size = 500, *P* value = 0.108, **Figure 2.2D2**). Nonetheless, this method is highly conservative, retaining all cross-species transmissions in the trees. The FV samples we used in our study were collected from markets, zoos, and rescue centres where cross-species transmissions can readily occur due to extreme close proximity. It is therefore possible that this method might be too conservative and the noise from spurious recent cross-species transmissions might overwhelm the signal of the FV-host co-speciation.

To examine this possibility, we pruned the FV tree further such that it contains only one FV per one host species, of which the phylogenetic placement represents the majority of its kind. We then compared this pruned FV tree to the host species tree. 12 potential co-speciation events were inferred in total (**Figure 2.2C**), which are greater than expected to occur by chance (random tip mapping: sample size = 500, *P* value < 0.002, **Figure 2.2D3**). Combined with the

results above, this finding suggests that, overall, NWM SFVs stably and broadly co-diverge with their hosts at the species level over the long timescale; however, cross-species and genus transmissions are not rare. Lastly, we compared the entire SFV tree with the host tree (**Figure 2.1**). In total, 43 potential co-speciation events were inferred, which is greater than expected to occur by chance (Random tip mapping sample size = 500, *P* value < 0.002, **Figure 2.2D4**). These findings suggest that NWM FVs stably co-speciate with their hosts at the level of sub-population also.

2.3.2 Evaluating and refining the co-speciation model

We found that, based on the phylogenetic reconciliation analyses alone, many of the inferred NWM-FV co-speciation events cannot be placed conclusively onto the trees (**Figure 2.2C**). To determine which alternative co-evolutionary scenarios are more likely, we used the timescales of NWM FVs to evaluate the inferred co-speciation events. We time-calibrated the FV tree by using the dates of the co-speciation events that could be mapped unambiguously onto the tree, directly inferred from the host timescales, estimated in (35).

Four dates were used to time-calibrate the tree in total (**Table 2.1**): (i) the separation date of the NWM and OMWA FVs (~43.47 Ma (95% HPD = 38.55–48.36), (35)), (ii) the branching date of the *Cebus xanthosternos* FV (~1.95 Ma (95% HPD = 0.91–3.31), (35)), (iii) the branching date of the *Ateles chamek* FV (~5.07 Ma (95% HPD = 2.87–7.50), (35)), and (iv) the separation date between the *Alouatta belzebul* and *Alouatta senticulus* FV (~4.95 Ma (95% HPD = 2.93–7.26), (35)). The time-calibrated FV tree is shown in **Figure 2.3A**. Note that, although phylogenetic reconciliation analyses suggested that the separation of the *Pithecia* and *Cacajao* FVs and the separation of the *Cebus* and *Callithrix* FVs are virus-host co-divergence events (**Figure 2.2C**), we however did not use their dates to calibrate the FV tree. Based on the host divergence dates, the dates of the two FV separations were inferred to be ~13.29, and ~19.95 Ma, respectively. However, the split of the *Pithecia* FV was determined to be before that of the *Cebus* FV topologically with strong support (posterior probability = 0.89, **Figure 2.1**, and **Figure 2.2B**). This implies that at least one of these two FV splits was

Table 2.1 Time to most recent common ancestor (tMRCA) mean estimates for simians and their foamy viruses (FVs).

Branch node	tMRCA (95% HPD) (million years)		Fossil estimate†
	FV	Simian†	
Simiiformes	39.85 (34.90–44.89)	43.47 (38.55–48.36)*	43 ± 4.5
Platyrrhini	22.31 (14.84–31.04)	24.82 (20.55–29.25)	23.5 ± 3.0
Atelidae (*Ateles*/*Alouatta* split)	10.77 (7.45–15.04)	16.13 (10.52–21.35)	NA
Atelidae (*Ateles*/*Lagothrix* split)	10.77 (7.45–15.04)	11.25 (7.25-15.46)	NA
Ateles	6.28 (4.56–8.05)	5.07 (2.87–7.50)*	NA
Alouatta (excluding *Alouatta palliatta*)	5.54 (3.99–7.27)	4.95 (2.93–7.26)*	NA
Cebus (*C. apella*/*C. xanthosternos* split)	3.45 (2.61–4.35)	1.95 (0.91–3.31)*	NA
Cebus/*Callithrix*	4.36 (3.08–6.06)	19.95 (15.66–24.03)	NA
Pitheciiinae	8.11 (4.88–12.33)	13.69 (9.24–18.34)	NA
Saimirinae	16.89 (10.18–24.93)	2.24 (1.05–3.73)	NA

Using an alignment of 412 nt for 74 simian FV taxa. Medians inferred using Bayesian methods and a relaxed molecular clock
'NA' = 'not available'.
'HPD' = highest posterior density intervals
† Dating and fossil estimates from Perelman *et al.* (35)
* Dates used to calibrate the FV tree

incorrectly determined as a co-speciation event, and therefore they were not used to calibrate the tree.

Our analysis estimated the nucleotide substitution rate to be ~2.14×10^{-8} (95% HPD = 1.72×10^{-8}–2.62×10^{-8}) s/n/y, which falls within the range of previously estimated rates of FV evolution (~1.7×10^{-8} s/n/y (34) and ~7.79×10^{-9} s/n/y (23)). The separation date of the *Pithecia* and *Cacajao* FVs (~8.11 Ma (95% HPD = 4.88–12.33)) was estimated to be relatively more comparable to that of their hosts (~13.69 Ma (95% HPD = 9.24–18.34), (35)) than that of the *Callithrix* and *Cebus* FVs (~4.36 Ma (95% HPD = 3.08–6.06)), which was estimated to happen much later than the host split (~19.95 Ma (95% HPD = 15.66–24.03), (35)). These findings suggest that the *Pithecia-Cacajao* FV separation is likely a co-speciation event as initially inferred. However, the split of the *Cebus* and *Callithrix* FVs is not, likely representing a cross-genus transmission between the two host monkeys (**Figure 2.3B**, short blue or green large arrows). These results also imply that the initially inferred cross-genus transmission from an ancestral *Cacajao* monkey to the *Cebus* and *Callithrix* MRCA (**Figure 2.2C**) was incorrect; it is either a cross-genus transmission from an ancestral *Cacajao* to an ancestral *Callithrix* monkey (**Figure 2.3B**, long blue large arrow), or to an ancestral *Cebus* monkey (**Figure 2.3B**, long green large arrow). To distinguish between these two competing alternative scenarios would require additional sequence data from other NWM FVs, such as those of *Mico*, *Cebuella*, and *Callimico* monkeys.

In addition, our initial phylogenetic reconciliation analyses suggested that the branching of the *Ateles* FV is a co-speciation event, but it could be mapped either to the *Ateles-Lagothrix* split (**Figure 2.2C1**) or to the *Ateles-Alouatta* split (**Figure 2.2C2**). Our additional analyses estimated the branching of *Ateles* FVs to be ~10.77 Ma (95% HPD = 7.45–15.04), more comparable to the split date of the *Ateles* and *Lagothrix* monkeys (~11.25 Ma (95% HPD = 7.25–15.46), (35)) than that of the *Ateles* and *Alouatta* monkeys (~16.13 Ma (95% HPD = 10.52–21.35), (35)). This thus suggests that the former scenario is more likely.

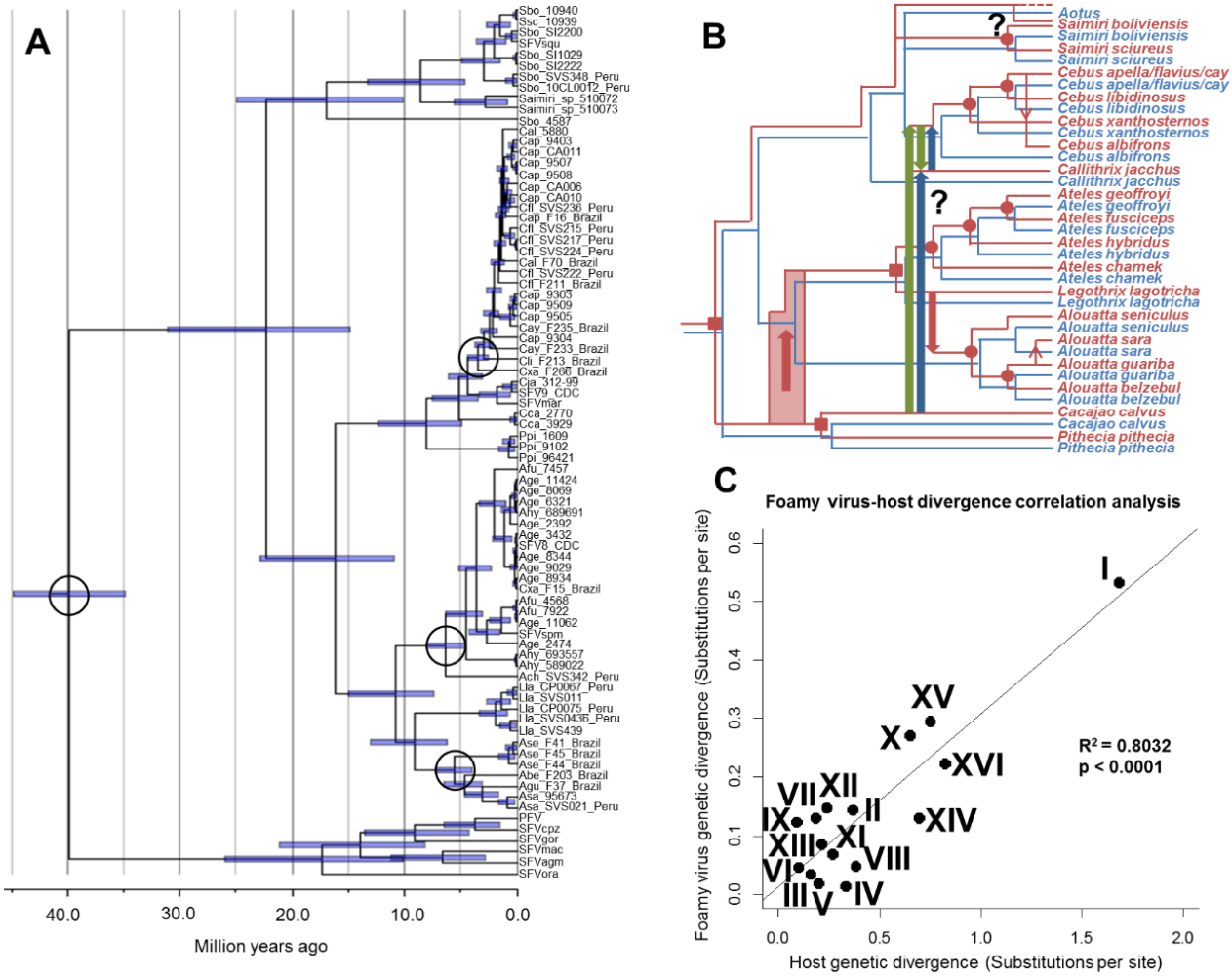


Figure 2.3 Evaluating and refining New World monkey foamy virus (NWM FV)-host co-evolutionary model. (**A**) Time-calibrated Bayesian NWM FV phylogeny, of which the topology was fixed to the one shown in **Figure 2.1** and **2.2B**. The calibrating nodes are encircled. The *bars* represent the uncertainty of the estimated node heights. The timescale is in millions of years. (**B**) Refined co-evolutionary history of NWM FVs (*red*) and their hosts (*blue*). *Red dotted branch* represents a possible *Aotus* ghost FV lineage that has not been sampled in our study. The directions of cross transmissions are indicated by *arrows*. *Small arrows* indicate cross-species transmission events, and *large arrows* indicate cross-genus transmission events. The *red transparent bar* shows the uncertainty of the cross-species transmission timing. Two alternative possible FV cross-species transmission scenarios involving *Cacajao*, *Cebus*, and *Callithrix* monkeys are shown in *green* and *blue*. '*Question mark*s' indicate ambiguous cross-species transmission pathways. Co-speciation events at the genus level are indicated by *solid red squares*, and those at the species level are indicated by *solid red circles*. The trees are not scaled to time. (**C**) NWM FV-host divergence correlation analysis. *Black dots* represent co-diverging branches identified under the refined NWM FV-host co-evolutionary model (see panel **B**). A well supported linear correlation was found (linear regression: N = 16, R^2 = 0.8032, p < 0.0001), represented by a *solid black line*. The *dots* are labelled with *roman numerals* (**I–XVI**), referring to branches in **Figure 2.1.**

Lastly, our analyses estimated the evolutionary timescale of *Saimiri* FV radiation (~16.89 Ma (95% HPD = 10.18–24.93) to be ~7.5× that of their hosts (~2.24 Ma (95% HPD = 1.05–3.73), (35)). This finding suggests that the *Saimiri* FV clade likely contains NWM FVs that arose from cross-species transmissions, but for which the ancestral FVs were not sampled. Assuming a stable FV host co-speciation history, we hypothesised that the ancestral FV 'ghost lineage' may have been an *Aotus* FV (**Figure 2.3B**), since the split between *Aotus* and *Saimiri* monkeys were inferred to occur ~19.95 Ma (95% HPD = 15.66–24.03) (35), comparable to our estimated *Saimiri* FV evolutionary timescale. To further examine this hypothesis, *Aotus* FV sequences are required. However, while *Aotus* species have been identified with antibodies to NWM FV antigens (23), *Aotus* FV sequences have not been reported to test this hypothesis. Furthermore, we also found that *Saimiri* FVs form a separate, ancestral lineage to all other NWM FVs, and that the two diverged from one another ~22.31 Ma (95% HPD = 14.84–31.04) (**Figure 2.3A**), comparable to the date of the basal NWM diversification (~24.82 Ma (95% HPD = 20.55–29.25), (35)). This date in turn supports that the split between *Saimiri* FVs and the rest is a co-speciation event that corresponds to the basal radiation of NWMs. The refined SFV-host co-evolutionary history is shown in **Figure 2.3B**.

2.3.3 NWM-FV divergence correlation analysis

A divergence correlation analysis was also performed to further evaluate our refined NWM-FV co-evolutionary model (**Figure 2.3C**). We found that, given the co-speciation model, the genetic divergence of NWM FVs is proportional to that of their hosts (linear regression: R^2 = 0.8032; $p < 0.0001$, **Figure 2.3C**). This indicates that the divergence of NWM FVs and their hosts are internally consistent, and simultaneously support our model, as well as the stable NWM-FV co-speciation hypothesis. We note that the number of co-speciation events is reduced by one, and the number of cross-species transmission events is increased by one under our refined co-evolutionary model. Nevertheless, these changes do not alter the conclusion about the stable NWM-FV co-divergence history at the species level, as 11 co-speciation events are still greater than expected to occur by chance (Random tip mapping sample size = 500, *P* value = 0.008).

2.4 Discussion

Our analyses showed that FV transmissions among closely related NWMs are relatively common. *Cebus* FVs are a clear example of this. Phylogenetic analyses show that *C. apella*, *C. flavius*, and *C. cay* FVs, as well as their hosts, cluster together without forming clear phylogenetic structures (**Figure 2.1**), indicative of frequent cross-species transmissions among these closely related NWMs with overlapping habitats. A few transmissions within the clades of *Ateles* monkeys and *Saimiri* monkeys were also observed (**Figure 2.1**). Despite these findings, our analyses nevertheless still inferred that NWM FVs and their hosts have a broad co-speciation history at the level of species and sub-population. These FV cross-species transmissions between closely related species are unsurprising however, given that they have already been documented in OWMAs (39).

Unlike their OWMA FV relatives, a strong co-speciation pattern between NWM FVs and their hosts could not be detected at the level of host genera, consistent with the results from previous analyses that used fewer samples (17, 21, 32). By comparing the virus and host tree topologies, and cross checking the consistency between virus and host divergence dates, we inferred four ancient NWM FV cross-genus transmissions, including (i) from *Pitheciinae* to *Atelidae*, (ii) from *Cacajao* to *Cebus* or *Callithrix*, (iii) between *Cebus* and *Callithrix*, and (iv) from *Lagothrix* to *Alouatta* (**Figure 2.3B**). It is unclear why NWM FVs have more successful onwards transmissions in the new host species than OWMA FVs. Analyses of antiretroviral restriction factors and host social behaviours might offer some insights into why the co-speciation pattern of NWM FVs is weaker than that of OWMA FVs. Notwithstanding this however, we were able to uncover that NWM FVs have a long co-evolutionary history with their hosts, co-evolving with one another since the origin of NWMs, dating back ~24 Ma.

2.5 Conclusions

Our study has filled in the gap in our knowledge of primate FV evolution, in particular that of NWM FVs. We found that, although FVs do have an extended co-evolutionary history with their mammalian hosts, FV cross-species transmissions might not be as rare as previously thought. Our analyses could detect NWM-FV co-speciation patterns at the level of host species and sub-population, but not at the host genus level. Several ancient FV cross-genus transmissions were found, including that from *Pitheciinae* to *Atelidae*, from *Cacajao* to *Cebus* or *Callithrix*, between *Cebus* or *Callithrix*, and from *Lagothrix* to *Alouatta*. Therefore, it is important that future investigation of FV evolution is performed with caution, and the FV-host co-speciation history must be validated when inferred and/or used for analyses. Our study highlights the value of temporal information in evaluating the validity of the virus-host co-evolutionary model. Supported by the similarity between the FV and host divergence dates, and the positive correlation of the FV and host genetic divergence, we inferred a number of NWM-FV co-speciation events, including the *Ateles/Leogotrix* FV separation, the split of the *Pithecia* FV and *Cacajao* FV, and the divergence of the *Saimiri* FV from *Pitheciidae* FV. Our results provide an important future reference for time-calibrating information essential for NWM FV evolutionary timescale inference, as well as evolutionary insight of NWM FVs. Further studies are needed to better understand the evolutionary history of NWM FVs and to clarify the possible “ghost” lineage identified by our analyses.

2.6 Materials and Methods

2.6.1 Phylogenetic reconstruction

A Bayesian phylogeny of 74 NWM FVs (**Table S2.1** in **Electronic Appendix**) was estimated from a manually-curated *pol* alignment (412 nt) by using MrBayes 3.2.1 (40). The sequences were provided by William Switzer (Centers for Disease Control and Prevention, personal communication). Genetic recombination within the *pol* alignments was checked by using Bootscan, Geneconv, MaxChi, Chimera, and RDP within the program RDP V3 (41) with the default setting, and no significant evidence for recombination was found. OWMA FVs (PFV;

Y07725, SFVcpz; U04327, SFVgor; HM245790, SFVora; AJ544579, SFVmac; NC_010819, and SFVagm; M784895) were used to root the tree. Nine sequences in the alignment were 3351 nt long (PFV, SFVcpz, SFVgor, SFVora, SFVmac, SFVagm, SFVmar; GU356395, SFVspm; EU010385, and SFVsqu; GU356394), included as 'backbones' to increase the power of deep node separation. The alignment is available in **Electronic Appendix**, **Data set S2.1**. The GTR+I+Γ(4) nucleotide substitution model was used without imposing a strict molecular clock. Two independent Markov chain Monte Carlo (MCMC) chains were run for 5 million steps with the initial 25% discarded as burn-in. Trees and parameters were logged every 2,500 steps thereafter. A metropolis coupling algorithm was applied to improve the MCMC samplings, using the setting of 3 hot and 1 cold chains. The convergence of estimated parameter values was diagnosed using potential scale reduction factors (PSRFs). PSRFs of all parameters were ~1.000, indicating that they were all well sampled from their posterior distributions and had converged.

A Bayesian phylogeny of 156 NWM hosts (**Table S2.1** in **Electronic Appendix**) was also estimated using the same protocol based on a manually-curated *CytB* alignment (618 nt). The sequences were provided by William Switzer (Centers for Disease Control and Prevention, personal communication). The alignment is available in **Electronic Appendix**, **Data set S2.2**. The phylogeny was rooted according to the tree in Perelman *et. al* (35). PSRFs of all parameters were ~1.000, indicating that they were all well sampled from their posterior distributions and had converged.

2.6.2 Phylogenetic reconciliation analysis

The topology of the tree of NWM FVs was compared to that of their hosts to infer possible co-speciation histories. The comparison was done by using the co-phylogeny reconstruction software Jane v4.0 (42) (Genetic algorithm: number of generations = 50 and population size = 100. Four phylogenetic reconciliation analyses were performed in total: (i) at the genus level, (ii) species level—conservative tree collapsing, (iii) species level—overall, and (iv) all sequences (see **Results** for details). In the last analysis, NWM FVs for which the

corresponding host sequences were not available were excluded from the analysis. The vertex-based cost mode was used, with costs set to maximise the number of co-speciation events (co-speciation = –1, duplication = 0, duplication & host switch = 0, loss = 0, and failure to diverge = 0). The probability of observing the inferred number of potential co-speciation events by chance was calculated by using random tip mapping method, implemented in Jane v4.0 (42) (Genetic algorithm: number of generation = 50, population size = 100, sample size = 500).

2.6.3 Estimating the timescales of NWM FVs

We found that many of the inferred co-divergence events cannot be placed conclusively onto the trees (see **Results**). To evaluate and refine the NWM-FV co-speciation model, we estimated the timescales of NWM FVs and compared them to those of their hosts, published in (35). The NWM FV tree was time-calibrated with the dates of the co-speciation events that could be mapped conclusively onto the tree, including the dates of (i) the NWM-OWMA FV separation, (ii) the branching of the *Cebus xanthosternos* FV, (iii) the branching of the *Ateles chamek* FV, and (iv) the separation between the *Alouatta belzebul* and *Alouatta senticulus* FVs. The dates of these events were directly inferred from those of their hosts, estimated in (35). The date estimation was performed under the Bayesian phylogenetic framework by using BEAST 1.8.2 (43) (**Table 2.1**). The SRD06 nucleotide substitution model, the Yule speciation process, a relaxed log-normal molecular clock and the default prior settings were employed. The tree topology was fixed to the one we obtained from the MrBayes analyses. The Bayesian MCMC was run for 50,000,000 steps with the initial 25% discarded as burn-in. Trees and parameters were logged every 2500 steps thereafter. Effective sample sizes of all parameters were >200, indicating that they were all well sampled and had converged. The timescales of NWM FV evolution were then used to evaluate which competing alternative scenarios of NWM-FV co-speciation were more likely.

2.6.4 NWM-FV divergence correlation analysis

To further evaluate the refined co-evolutionary model, a NWM-FV divergence correlation analysis was performed. A set of NWM-FV co-diverging branches were manually identified based on the refined model, and we examined if the divergences of NWMs and FVs are linearly correlated. The linear correlation, coefficient of determination (R^2), and *P* value were calculated by using the *lm* function implemented in R (https://cran.r-project.org).

2.7 References

1. **Achong BG**, **Mansell PW**, **Epstein MA**, **Clifford P**. 1971. An unusual virus in cultures from a human nasopharyngeal carcinoma. J. Natl. Cancer Inst. **46**:299–307.
2. **Flügel RM**, **Rethwilm A**, **Maurer B**, **Darai G**. 1987. Nucleotide sequence analysis of the env gene and its flanking regions of the human spumaretrovirus reveals two novel genes. EMBO J. **6**:2077–2084.
3. **Maurer B**, **Bannert H**, **Darai G**, **Flügel RM**. 1988. Analysis of the primary structure of the long terminal repeat and the gag and pol genes of the human spumaretrovirus. J. Virol. **62**:1590–1597.
4. **Rogers NG**, **Basnight M**, **Gibbs CJ**, **Gajdusek DC**. 1967. Latent viruses in chimpanzees with experimental Kuru. Nature **216**:446–449.
5. **Herchenröder O**, **Renne R**, **Loncar D**, **Cobb EK**, **Murthy KK**, **Schneider J**, **Mergia A**, **Luciw PA**. 1994. Isolation, cloning, and sequencing of simian foamy viruses from chimpanzees (SFVcpz): high homology to human foamy virus (HFV). Virology **201**:187–199.
6. **Bieniasz PD**, **Rethwilm A**, **Pitman R**, **Daniel MD**, **Chrystie I**, **McClure MO**. 1995. A comparative study of higher primate foamy viruses, including a new virus from a gorilla. Virology **207**:217–228.
7. **Schulze A**, **Lemey P**, **Schubert J**, **McClure MO**, **Rethwilm A**, **Bodem J**. 2011. Complete nucleotide sequence and evolutionary analysis of a gorilla foamy virus. J. Gen. Virol. **92**:582–586.
8. **McClure MO**, **Bieniasz PD**, **Schulz TF**, **Chrystie IL**, **Simpson G**, **Aguzzi A**, **Hoad JG**, **Cunningham A**, **Kirkwood J**, **Weiss RA**. 1994. Isolation of a new foamy retrovirus from orangutans. J. Virol. **68**:7124–7130.
9. **Verschoor EJ**, **Langenhuijzen S**, **van den Engel S**, **Niphuis H**, **Warren KS**, **Heeney JL**. 2003. Structural and evolutionary analysis of an orangutan foamy virus. J. Virol. **77**:8584–8587.
10. **Rustigian R**, **Johnston P**, **Reihart H**. 1955. Infection of monkey kidney tissue cultures with virus-like agents. Proc. Soc. Exp. Biol. Med. **88**:8–16.

11. **Johnston PB**. 1961. A second immunologic type of simian foamy virus: monkey throat infections and unmasking by both types. J. Infect. Dis. **109**:1–9.

12. **Kupiec JJ**, **Kay A**, **Hayat M**, **Ravier R**, **Périès J**, **Galibert F**. 1991. Sequence analysis of the simian foamy virus type 1 genome. Gene **101**:185–194.

13. **Stiles GE**, **Bittle JL**, **Cabasso VJ**. 1964. Comparison of simian foamy virus strains including a new serological type. Nature **201**:1350–1351.

14. **Renne R**, **Friedl E**, **Schweizer M**, **Fleps U**, **Turek R**, **Neumann-Haefelin D**. 1992. Genomic organization and expression of simian foamy virus type 3 (SFV-3). Virology **186**:597–608.

15. **Heberling RL**, **Kalter SS**. 1975. Isolation of foamy viruses from baboon (*Papio cynocephalus*) tissues. Am. J. Epidemiol. **102**:25–29.

16. **Johnston PB**. 1971. Taxonomic features of seven serotypes of simian and ape foamy viruses. Infect. Immun. **3**:793–799.

17. **Pacheco B**, **Finzi A**, **McGee-Estrada K**, **Sodroski J**. 2010. Species-specific inhibition of foamy viruses from South American monkeys by New World Monkey TRIM5{alpha} proteins. J. Virol. **84**:4095–4099.

18. **Hooks JJ**, **Gibbs CJ**, **Chou S**, **Howk R**, **Lewis M**, **Gajdusek DC**. 1973. Isolation of a new simian foamy virus from a spider monkey brain culture. Infect. Immun. **8**:804–813.

19. **Thümer L**, **Rethwilm A**, **Holmes EC**, **Bodem J**. 2007. The complete nucleotide sequence of a New World simian foamy virus. Virology **369**:191–197.

20. **Hooks JJ**, **Gibbs CJ**. 1975. The foamy viruses. Bacteriol. Rev. **39**:169–185.

21. **Troncoso LL**, **Muniz CP**, **Siqueira JD**, **Curty G**, **Schrago CG**, **Augusto A**, **Fedullo L**, **Soares MA**, **Santos AF**. 2015. Characterization and comparative analysis of a simian foamy virus complete genome isolated from Brazilian capuchin monkeys. Virus Res. **208**:1–6.

22. **Marczynska B**, **Jones CJ**, **Wolfe LG**. 1981. Syncytium-forming virus of common marmosets (*Callithrix jacchus jacchus*). Infect. Immun. **31**:1261–1269.

23. **Muniz CP**, **Troncoso LL**, **Moreira M a**, **Soares E a**, **Pissinatti A**, **Bonvicino CR**, **Seuánez HN**, **Sharma B**, **Jia H**, **Shankar A**, **Switzer WM**, **Santos AF**, **Soares M a**.

2013. Identification and characterization of highly divergent simian foamy viruses in a wide range of new world primates from Brazil. PLoS One **8**:e67568.

24. **Ghersi BM**, **Jia H**, **Aiewsakun P**, **Katzourakis A**, **Mendoza P**, **Bausch DG**, **Kasper MR**, **Montgomery JM**, **Switzer WM**. 2015. Wide distribution and ancient evolutionary history of simian foamy viruses in New World primates. Retrovirology **12**:89.

25. **Riggs JL**, **Oshirls**, **Taylor DO**, **Lennette EH**. 1969. Syncytium-forming agent isolated from domestic cats. Nature **222**:1190–1191.

26. **Winkler I**, **Bodem J**, **Haas L**, **Zemba M**, **Delius H**, **Flower R**, **Flügel RM**, **Löchelt M**. 1997. Characterization of the genome of feline foamy virus and its proteins shows distinct features different from those of primate spumaviruses. J. Virol. **71**:6727–6741.

27. **Malmquist WA**, **Van der Maaten MJ**, **Boothe AD**. 1969. Isolation, immunodiffusion, immunofluorescence, and electron microscopy of a syncytial virus of lymphosarcomatous and apparently normal cattle. Cancer Res. **29**:188–200.

28. **Renshaw RW**, **Casey JW**. 1994. Transcriptional mapping of the 3' end of the bovine syncytial virus genome. J. Virol. **68**:1021–1028.

29. **Tobaly-Tapiero J**, **Bittoun P**, **Neves M**, **Guillemin MC**, **Lecellier CH**, **Puvion-Dutilleul F**, **Gicquel B**, **Zientara S**, **Giron ML**, **de Thé H**, **Saïb A**. 2000. Isolation and characterization of an equine foamy virus. J. Virol. **74**:4064–4073.

30. **Wu Z**, **Ren X**, **Yang L**, **Hu Y**, **Yang J**, **He G**, **Zhang J**, **Dong J**, **Sun L**, **Du J**, **Liu L**, **Xue Y**, **Wang J**, **Yang F**, **Zhang S**, **Jin Q**. 2012. Virome analysis for identification of novel mammalian viruses in bat species from Chinese provinces. J. Virol. **86**:10999–11012.

31. **Katzourakis A**, **Gifford RJ**, **Tristem M**, **Gilbert MTP**, **Pybus OG**. 2009. Macroevolution of complex retroviruses. Science **325**:1512.

32. **Katzourakis A**, **Aiewsakun P**, **Jia H**, **Wolfe ND**, **LeBreton M**, **Yoder AD**, **Switzer WM**. 2014. Discovery of prosimian and afrotherian foamy viruses and potential cross species transmissions amidst stable and ancient mammalian co-evolution. Retrovirology **11**:61.

33. **Han G-Z**, **Worobey M**. 2014. Endogenous viral sequences from the Cape golden mole (*Chrysochloris asiatica*) reveal the presence of foamy viruses in all major placental mammal clades. PLoS One **9**:e97931.

34. **Switzer WM**, **Salemi M**, **Shanmugam V**, **Gao F**, **Cong M-E**, **Kuiken C**, **Bhullar V**, **Beer BE**, **Vallet D**, **Gautier-Hion A**, **Tooze Z**, **Villinger F**, **Holmes EC**, **Heneine W**. 2005. Ancient co-speciation of simian foamy viruses and primates. Nature **434**:376–380.

35. **Perelman P**, **Johnson WE**, **Roos C**, **Seuánez HN**, **Horvath JE**, **Moreira MAM**, **Kessing B**, **Pontius J**, **Roelke M**, **Rumpler Y**, **Schneider MPC**, **Silva A**, **O'Brien SJ**, **Pecon-Slattery J**. 2011. A molecular phylogeny of living primates. PLoS Genet. **7**:e1001342.

36. **Schneider H**, **Sampaio I**. 2015. The systematics and evolution of New World primates--a review. Mol Phylogenet Evol **82**:348–357.

37. **Kehl T**, **Tan J**, **Materniak M**. 2013. Non-simian foamy viruses: molecular virology, tropism and prevalence and zoonotic/interspecies transmission. Viruses **5**:2169–2209.

38. **Leendertz FH**, **Zirkel F**, **Couacy-Hymann E**, **Ellerbrok H**, **Morozov VA**, **Pauli G**, **Hedemann C**, **Formenty P**, **Jensen SA**, **Boesch C**, **Junglen S**. 2008. Interspecies transmission of simian foamy virus in a natural predator--prey system. J Virol **82**:7741–7744.

39. **Liu W**, **Worobey M**, **Li Y**, **Keele BF**, **Bibollet-Ruche F**, **Guo Y**, **Goepfert PA**, **Santiago ML**, **Ndjango J-BN**, **Neel C**, **Clifford SL**, **Sanz C**, **Kamenya S**, **Wilson ML**, **Pusey AE**, **Gross-Camp N**, **Boesch C**, **Smith V**, **Zamma K**, **Huffman MA**, **Mitani JC**, **Watts DP**, **Peeters M**, **Shaw GM**, **Switzer WM**, **Sharp PM**, **Hahn BH**. 2008. Molecular ecology and natural history of simian foamy virus infection in wild-living chimpanzees. PLoS Pathog. **4**:e1000097.

40. **Ronquist F**, **Teslenko M**, **Mark P**, **Ayres DL**, **Darling A**, **Hohna S**, **Larget B**, **Liu L**, **Suchard MA**, **Huelsenbeck JP**. 2012. MrBayes 3.2: efficient Bayesian phylogenetic inference and model choice across a large model space. Syst Biol **61**:539–542.

41. **Martin DP**, **Lemey P**, **Lott M**, **Moulton V**, **Posada D**, **Lefeuvre P**. 2010. RDP3: a flexible and fast computer program for analyzing recombination. Bioinformatics **26**:2462–2463.

42. **Conow C**, **Fielder D**, **Ovadia Y**, **Libeskind-Hadas R**. 2010. Jane: a new tool for the cophylogeny reconstruction problem. Algorithms Mol. Biol. **5**:16.

43. **Drummond AJ**, **Rambaut A**. 2007. BEAST: Bayesian evolutionary analysis by sampling trees. BMC Evol. Biol. **7**:214.

Chapter three: Time dependency of foamy virus evolutionary rate estimates

---

This chapter was originally published as:

Aiewsakun P, Katzourakis A. 2015. Time dependency of foamy virus evolutionary rate estimates. BMC Evol. Biol. **15**:119.

PA performed the analysis and wrote the paper. AK provided editorial guidance and supervisory advice.

3.1 Abstract

It appears that viral substitution rate estimates co-vary very strongly with their timescale of measurement; the shorter the timescale, the higher the estimated value. Foamy viruses have a long history of cospeciation with their hosts, and one of the lowest estimated rates of evolution among viruses. However, when their rate of evolution is estimated over short timescales, it is more reminiscent of the rapid rates seen in other RNA viruses. This discrepancy between their short-term and long-term rates could be explained by the time-dependency of substitution rate estimates. Several empirical models have been proposed and used to correct for this so-called *time-dependent rate phenomenon* (TDRP), such as a vertically-translated exponential rate decay model and a power-law rate decay model. Nevertheless, at present, it is still unclear which model best describes the rate dynamics. Here, we use foamy viruses as a case study to empirically describe the TDRP and to determine how to correct rate estimates for its effects. Four empirical models were investigated: (i) a vertically-translated exponential rate decay model, (ii) a simple exponential rate decay model, (iii) a vertically-translated power-law rate decay model, and (iv) a simple power-law rate decay model. Our results suggest that the TDRP is likely responsible for the large discrepancy observed in foamy virus short-term and long-term rate estimates, and the simple power-law rate decay model is the best model for inferring evolutionary timescales. Furthermore, we demonstrated that, within the Bayesian phylogenetic framework, currently available molecular clocks can severely bias evolutionary date estimates, indicating that they are inadequate for correcting for the TDRP. Our analyses also suggest that different viral lineages may have different TDRP dynamics, and this may bias date estimates if it is unaccounted for. As evolutionary rate estimates are dependent on their measurement timescales, their values must be used and interpreted under the context of the timescale of rate estimation. Extrapolating rate estimates across large timescales for evolutionary inferences can severely bias the outcomes. Given that the TDRP is widespread in nature but has been noted only recently, the estimated timescales of many viruses may need to be reconsidered and re-estimated. Our models could be used as a guideline to further improve current phylogenetic inference tools.

3.2 Introduction

It has been noted that evolutionary rates calculated over short timescales, such as those calculated from population genetic data, are much greater than those calculated across geological time frames, such as species evolutionary rates. This discrepancy between short-term and long-term rates is very widespread in nature, noticed in both viral genes (1–6) and cellular genes, including bacterial genes (7, 8), mitochondrial genes of worms (9), insects (10, 11), fish (12, 13), birds (14–16), and primates (17–23), as well as nuclear genes of various multicellular organisms (24–27) (see (28) for a review, and references therein). Further investigation has revealed that, in fact, the value of the rate estimate does not vary discretely, but continuously decreases as the measurement timescale increases (28). This so-called '*time-dependent rate phenomenon*' (TDRP) was first demonstrated in cellular genes (20). Subsequently, by pooling substitution rate estimates of diverse viruses together, Duchêne et al. (29) showed that the rate estimates of RNA and DNA viruses also exhibit this pattern.

To date, the processes that lead to the TDRP are still very much unclear. Many hypotheses have been proposed to explain it, such as temporal changes in organismal biology and natural selection pressure (28). Numerous methodological artefacts also have the potential to systematically bias the rate estimates in such a way that short-term rates will appear to be much greater than the long-term ones (28). Since the rate of evolution is a central component of evolutionary study, an accurate evolutionary inference requires that the TDRP is integrated into the analysis. An ideal approach to this problem is to understand how the TDRP is generated, and improve evolutionary inference tools so that they can account for the factors underlying the TDRP. However, given a large number of potential underlying factors and our current incomplete understanding of their interactions, untangling and explicitly accounting for each of them individually would be impractical at present (20, 30).

One pragmatic approach to this problem is to infer evolutionary timescales by using an empirical model describing the relationship between rate estimates and their measurement timescales. This approach has been employed in several studies (e.g. (12, 20)). A number of

empirical models, such as the vertically-translated exponential rate decay model and the power-law rate decay model, have been proposed, and used, for TDRP correction in evolutionary inferences (12, 20, 31). Nevertheless, at present, it is still unclear which model best describes the phenomenon, and to answer this question we require substitution rate estimates that are computed over various timeframes. These can be obtained from a data set of molecular sequences for which several divergence dates are known; the more divergence dates available, the more suitable the data set for this purpose. Furthermore, it is also preferable that the dates are distributed relatively evenly across the entire evolutionary timescale that is being examined (32). In this work, we seek to empirically describe the TDRP in detail and explore the various patterns of rate decay over time by using foamy viruses (FVs) as a case study.

FVs are a group of complex retroviruses that have a very stable and long cospeciation history with their hosts, stretching back more than a hundred million years (33–35), and because of this, almost all of their divergence dates can be directly inferred from those of their hosts (36–38). Based on this cospeciation and the known divergence dates of their primate hosts, the long-term rate of evolution of FVs has been estimated to be $\sim 7.79 \times 10^{-9}$ to 1.7×10^{-8} nucleotide substitutions per site per year (s/n/y) (33, 39). This is much slower than rates of substitution of other RNA viruses, typically reported to be in range of 10^{-3} to 10^{-4} s/n/y (6, 40, 41). The high similarity observed between extant FVs and their ancient endogenous counterparts (34, 35, 42, 43) has also lent further support to this notion of slow-evolving FVs. Altogether, FVs are thereby widely regarded as one of the most slow-evolving RNA viruses currently known (44).

These slow long-term rates of FV evolution stand in sharp contrast to their high mutation rate. *In vitro* analyses have shown that the FV replication error rate (5.8×10^{-5} s/n/replication) is comparable to that of human immunodeficiency virus (HIV) (6.5×10^{-5} s/n/replication) (45), which is one of the fastest-evolving viruses ever documented. This fast FV mutation rate has also been confirmed in human embryonic cell lines, where the *in vivo* rate was calculated to

be at least 1.1×10^{-5} s/n/replication (46). Moreover, by following a population of African Green Monkey FVs for 9 years, it was estimated that the FV evolutionary rate is as high as $\sim 3.75 \times 10^{-4}$ s/n/y (47). This short-term rate is ~4 to 5 orders of magnitude higher than the long-term rate counterparts, estimated under the FV-host cospeciation assumption.

The fact that almost all of the divergence dates of FVs can be directly inferred from those of their hosts makes FVs an ideal system to study the TDRP. Here, we use 14 extant FVs (**Table 3.1**) as a case study to present direct evidence of a smooth decay of nucleotide substitution rate estimates as the measurement timescale increases. We also empirically describe the rate decay pattern, examine whether or not the TDRP can explain the discrepancy between FV short-term and long-term rates, and discuss the applications and limitations of our empirical rate decay models, as well as how the TDRP may bias evolutionary inference and rate estimate interpretation.

3.3 Results

3.3.1 FV nucleotide substitution rate estimate decreases with measurement timescale

To compute FV nucleotide substitution rates for various timescales, we first estimated FV phylogenies using Pol protein (1,116 aa, see **Data set S3.1** in **Electronic Appendix**) and *pol* nucleotide (3,351 nt, see **Data set S3.2 Electronic Appendix**) alignments under the Bayesian and maximum-likelihood frameworks. The aligned sequences were checked for recombination, but no significant evidence was found, both at the nucleotide (*P* value = 0.266) and protein (*P* value = 0.357) levels. (See Materials and Methods for details.) Our results show that all phylogenies are perfectly in agreement with one another topologically (**Figure 3.1**), and also consistent with the results from previous studies (33–35). We thus considered the estimated tree topology as our best working hypothesis, and used it to estimate FV nucleotide substitution rates.

Table 3.1 GenBank accession numbers of foamy virus (FV) nucleotide sequences.

FV†	Accession number	Host
PFV‡	Y07725	West chimpanzee (*Pan troglodytes schweinfurthii*)
SFVcpz	U04327	East chimpanzee (*Pan troglodytes verus*)
SFVbnb	EU527700 & EU527595	Bonobo (*Pan paniscus*)
SFVgor	HM245790	Gorilla (*Gorilla gorilla sp.*)
SFVora	AJ544579	Orangutan (*Pongo pygmaeus*)
SFVagm	M74895	African green monkey (*Cercopithecus aethiops*)
SFVmac	NC_010819	Macaque (*Macaca mulatta*)
SFVspm	EU010385	Spider monkey (*Ateles sp.*)
SFVmar	GU356395	Common marmoset (*Callithrix jacchus*)
SFVsqu	GU356394	Squirrel monkey (*Saimiri sciureus*)
PSFVgal	KM233624	Galago (*Otolemur crassicaudatus panganiensis*)
BFV	U94514	Cow (*Bos taurus*)
EFV	AF201902	Horse (*Equus ferus caballus)*
FFV	Y08851	Domestic cat (*Felis catus*)

†Acronyms used in FV names: PFV, prototype FV; SFVcpz, chimpanzee simian FV (SFV); SFVbnb, bonobo SFV; SFVgor, gorilla SFV; SFVora, orangutan SFV; SFVagm, African green monkey SFV; SFVmac, macaque SFV; SFVmar, common marmoset SFV; SFVspm, spider monkey SFV; SFVsqu, squirrel monkey SFV; PSFVgal, galago prosimian FV; BFV, bovine FV; EFV, equine FV; FFV, feline FV.
‡ Although PFV was isolated from a human (48) it is well-established that it in fact has a West chimpanzee origin (49–51). It is therefore treated as a West chimpanzee FV here, which diverged from SFVcpz about 0.96 (95% HPD = 0.70–1.22) million years ago, inferred under the FV-host co-speciation assumption, based on the divergence date of *Pan troglodytes schweinfurthii* and *Pan troglodytes verus* (37).

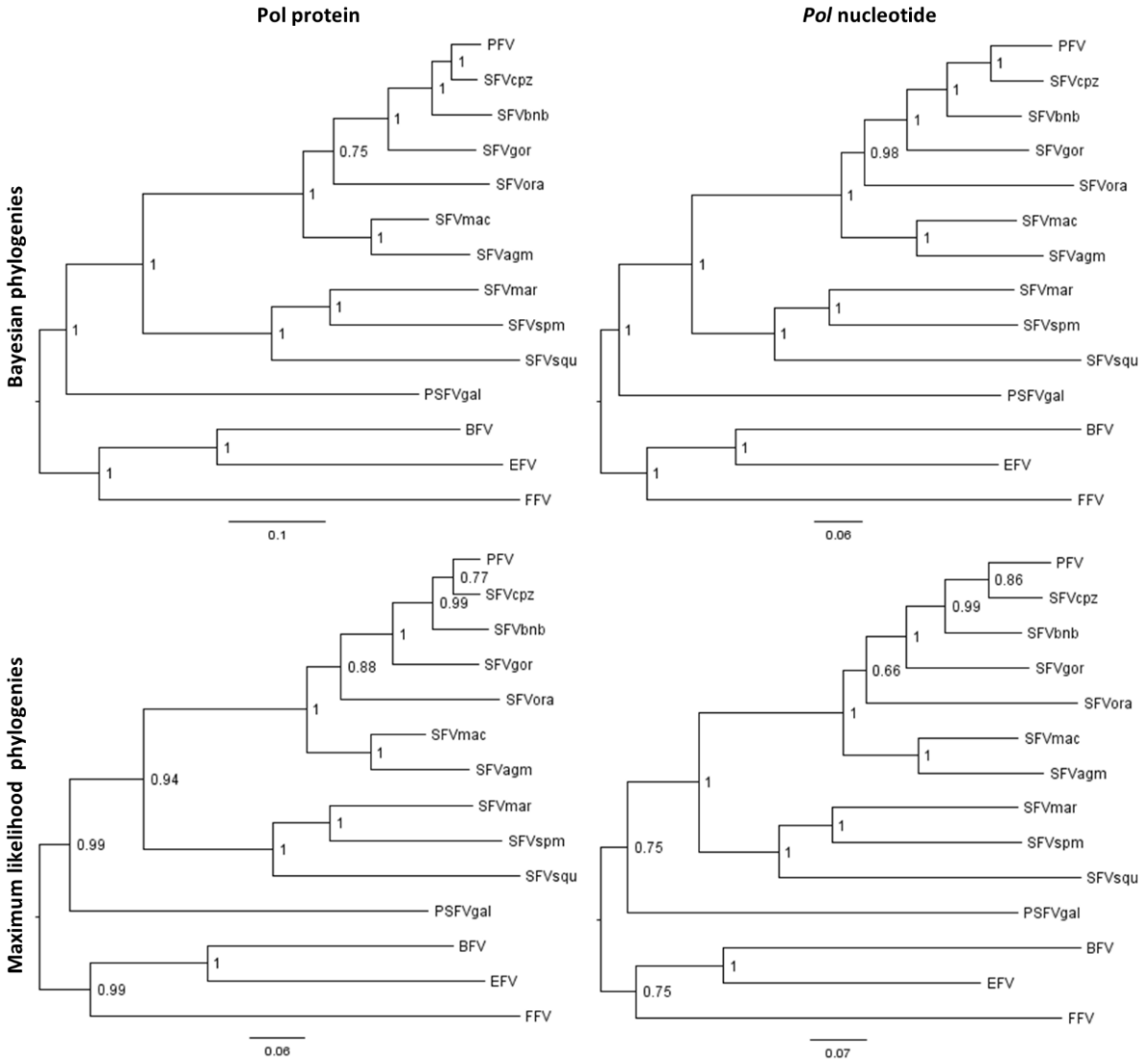


Figure 3.1 Foamy virus (FV) phylogenies. Four phylogenies of 14 extant FVs were constructed based on Pol protein (**left column**) and *pol* nucleotide (**right column**) alignments under the Bayesian (**top row**) and maximum-likelihood (**bottom row**) framework (See **Materials and Methods** for details). Numbers on nodes are node supports (**top row**: Bayesian posterior support; **bottom row**: bootstrap support), and scale bars are in the units of substitutions per site. See taxon definitions and GenBank accession numbers in **Table 3.1**.

We first estimated node-to-tip total per-lineage nucleotide substitutions (s estimates) from the *pol* nucleotide alignment under the fixed FV phylogeny, and the Bayesian phylogenetic framework. We employed a strict molecular clock with a fixed rate of 1 to obtain branch lengths in units of substitutions per site. A strict clock was applied (i.e., the tips were forced to align) under the assumption of consistency among rate estimates calculated using different nodes and tips. Thus, our study can also be viewed as an attempt to correct for the TDRP given the strict clock assumption. In total, 13 posterior distributions of s estimates were obtained (**Figure 3.2A**; **left**), one for each internal node, and we could assign timescales (t estimates) to 11 of them based on the FV-host cospeciation history (**Figure 3.2A**; **right**). This, in turn, allowed us to compute 11 distributions of node-to-tip average substitution rates ($\bar{r}$ estimates). (See Materials and Methods for details.) A summary of the results can be found in **Figure 3.2**, and **Table S3.1** in **Electronic Appendix**. Preliminary linear regression analyses suggest that log-transformed $\bar{r}$ is significantly negatively correlated with log-transformed t (linear regression: correlation coefficient = −0.577; 95% highest probability density (HPD) = −0.614–−0.542; randomisation test: number of randomization tests = 15,000, number of data points for null distribution construction in each test = 100, *P* value < 0.01 in all 15,000 tests; **Figure 3.2B**, see Materials and Methods for details).

3.3.2 Empirical description of FV nucleotide substitution rate decay

It has been proposed that a vertically-translated exponential decay function (**Eq. 1**) is a good empirical description for the apparent decay of the instantaneous substitution rate (r) (20). This model has been employed in several studies for TDRP correction (e.g. (31)). However, an examination of mitochondrial DNA control regions of cichlids showed that a simple power law function (**Eq. 4**) can also empirically describe the phenomenon reasonably well (12). To systematically explore these hypotheses, we examine another two empirical functions by varying whether the exponential and power law functions are vertically translated: a simple exponential rate decay function (**Eq. 2**) and a vertically-translated power-law function (**Eq. 3**). The four models can be expressed mathematically as follows:

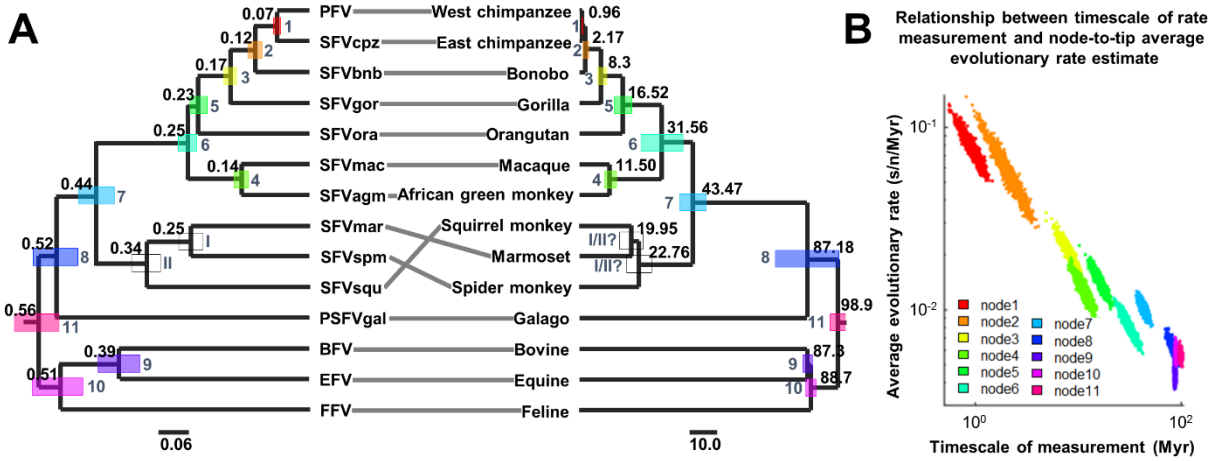


Figure 3.2 Foamy virus (FV) and corresponding host phylogenies, and the relationship between node-to-tip average evolutionary rate estimate and timescale of measurement. (**A**, left) FV phylogeny (taxon definitions and GenBank sequence accession numbers are in **Table 3.1**). Black numbers are estimated total per-lineage nucleotide substitutions in the units of substitutions per site. The node bars represent the uncertainties of the estimated nucleotide substitution divergences. The scale bar is in units of substitutions per site. (**A**, right) Host tree. Black numbers are estimated divergence dates in units of millions of years, for which the estimation uncertainties are shown by node bars. The scale bar is in units of millions of years. The topology of the host tree and the divergence dates were estimated elsewhere (see the references in **Table S3.1** in **Electronic Appendix**). Solid lines between the two trees indicate FV-host associations, and blue Arabic numbers (**1−11**) indicate matching FV-host nodes (i.e., FV-host cospeciation events). Nodes within the FV tree that could not be mapped conclusively onto the host tree are labelled by blue Roman numbers (**I** and **II**). The colours of the node bars on the FV tree correspond to the colours of the bars on the host tree. (**B**) Timescales of rate measurement and node-to-tip average evolutionary rate estimates on a log-log scale. The node numbers (**1−11**) refer to those in the FV and host trees (**A**). The colours correspond to the node bars' colours. The summary of the raw data can be found in **Table S3.1** in **Electronic Appendix**.

$$r_{VEX} = \alpha_{VEX} e^{-t\beta_{VEX}} + k_{VEX} \quad \text{(Eq.1)}$$

$$r_{EX} = \alpha_{EX} e^{-t\beta_{EX}} \quad \text{(Eq.2)}$$

$$r_{VPL} = \alpha_{VPL} t^{-\beta_{VPL}} + k_{VPL} \quad \text{(Eq.3)}$$

$$r_{PL} = \alpha_{PL} t^{-\beta_{PL}} \quad \text{(Eq.4)}$$

where t (unit time, ut) is the measurement/evolutionary timescale in the present-to-past direction, where $t = 0$ is the present; k (s/n/ut) is the stable long-term rate of evolution parameter; and α and β are arbitrary model parameters. 'VEX', 'EX', 'VPL', and 'PL' subscripts indicate the model to which the parameters and variables belong: vertically-translated exponential rate decay (VEX), simple exponential rate decay (EX), vertically-translated power-law rate decay (VPL) and simple power-law rate decay (PL) models, respectively. Note that, since a strict clock was applied and the tips were all aligned, the timescales of the rate measurement can be interpreted as node heights/divergence dates and vice versa.

To investigate how well each model describes the FV $\bar{r}$ dynamics, we first derived four equations depicting the relationship between s and t for the VEX (**Eq. 5**), EX (**Eq. 6**), VPL (**Eq. 7**), and PL (**Eq. 8**) models based on **Eq. 1−4**, respectively, as follows;

$$s_{VEX} = \int_{t=0}^{t=t} r_{VEX} dt = \frac{\alpha_{VEX}}{\beta_{VEX}} \left(1 - e^{-t\beta_{VEX}}\right) + tk_{VEX} \quad \text{(Eq.5)}$$

$$s_{EX} = \int_{t=0}^{t=t} r_{EX} dt = \frac{\alpha_{EX}}{\beta_{EX}} \left(1 - e^{-t\beta_{EX}}\right) \quad \text{(Eq.6)}$$

$$s_{VPL} = \int_{t=0}^{t=t} r_{VPL} dt = \frac{\alpha_{VPL} t^{1-\beta_{VPL}}}{1-\beta_{VPL}} + tk_{VPL} \quad \text{(Eq.7)}$$

$$s_{PL} = \int_{t=0}^{t=t} r_{PL} dt = \frac{\alpha_{PL} t^{1-\beta_{PL}}}{1-\beta_{PL}} \quad \text{(Eq.8)}$$

The curves were forced to go through the origin to conform the expectation that there are no substitutions at time equal to zero. We next simply divided both sides of the equations by t to derive equations describing the relationship between $\bar{r}$ and t for the four respective models (**Eq. 9–12**, respectively):

$$\frac{s_{VEX}}{t} = \bar{r}_{VEX} = \frac{\alpha_{VEX}}{t\beta_{VEX}}\left(1 - e^{-t\beta_{VEX}}\right) + k_{VEX} \quad \text{(Eq.9)}$$

$$\frac{s_{EX}}{t} = \bar{r}_{EX} = \frac{\alpha_{EX}}{t\beta_{EX}}\left(1 - e^{-t\beta_{EX}}\right) \quad \text{(Eq.10)}$$

$$\frac{s_{VPL}}{t} = \bar{r}_{VPL} = \frac{\alpha_{VPL}t^{-\beta_{VPL}}}{1-\beta_{VPL}} + k_{VPL} \quad \text{(Eq.11)}$$

$$\frac{s_{PL}}{t} = \bar{r}_{PL} = \frac{\alpha_{PL}t^{-\beta_{PL}}}{1-\beta_{PL}} \quad \text{(Eq.12)}$$

We then fitted all four models to the $\bar{r}$ and t estimates, and assessed how well the models describe the data by using adjusted R^2 ($\bar{R}^2$) scores. Although it is clear from visual inspection that the EX model tends to underestimate long-term rates (**Figure 3.3**), overall, all four models seem to describe the data well, indicated by their high $\bar{R}^2$ scores ($\bar{R}^2$ score: **Eq. 9**: 0.99; 95% HPD = 0.95–1.00; **Eq. 10**: 0.97; 95% HPD = 0.94–0.99; **Eq. 11**: 0.98; 95% HPD = 0.92–1.00; **Eq. 12**: 0.98; 95% HPD = 0.93–1.00). The results are shown in **Figure 3.3**, and **Table S3.2** in **Electronic Appendix**.

In addition, we also recovered FV short-term and long-term rate estimates (calculated over a timescale of 10 years and 30 million years (Myr), respectively) by using these four models. The VEX and EX models estimated the short-term rate of FVs to be ~10^{-7} to 10^{-8} s/n/y (median rate estimates: VEX: 1.06×10^{-7}; 95% HPD = 7.43×10^{-8}–1.51×10^{-7}; EX: 9.68×10^{-8}; 95% HPD = 6.92×10^{-8}–1.35×10^{-7}). This is much lower than the previously reported FV short-term rate estimate (~3.75×10^{-4} s/n/y (47)). On the other hand, the VPL and PL models estimated the rate to be in the order of 10^{-4} to 10^{-5} s/n/y (median rate estimates: VPL: 8.17×10^{-5}; 95% HPD = 3.24×10^{-5}–2.20×10^{-4}; PL: 7.53×10^{-5}; 95% HPD = 3.24×10^{-5}–1.45×10^{-4}), comparable to the previously reported FV short-term rate estimate. In contrast, all four models calculated the long-term rate to be ~10^{-8} to 10^{-9} s/n/y (median rate estimates: VEX: 8.75×10^{-9}; 95% HPD = 8.16×10^{-9}–9.39×10^{-9}; EX: 5.99×10^{-9}; 95% HPD = 4.58×10^{-9}–7.00×10^{-9}; VPL: 9.97×10^{-9}; 95% HPD = 9.02×10^{-9}–1.10×10^{-8}; PL: 9.93×10^{-9}; 95% HPD = 8.58×10^{-9}–1.09×10^{-8}), all comparable to the established long-term rates of FVs (~7.79×10^{-9} (39) to 1.7×10^{-8} (33) s/n/y).

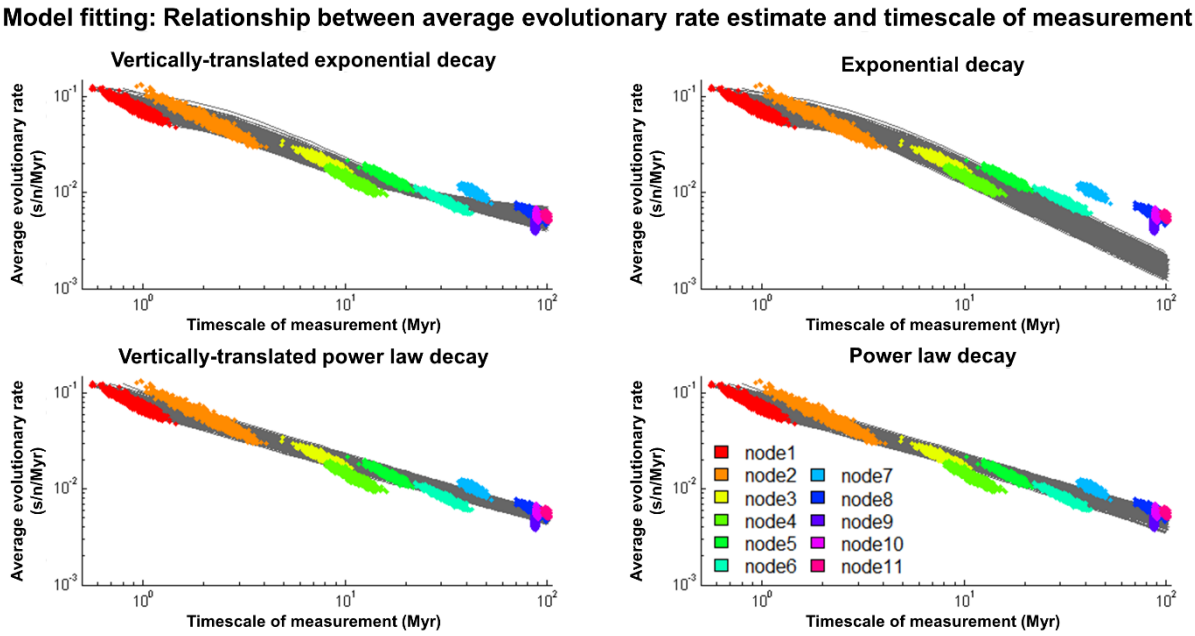


Figure 3.3 Empirical description of the relationship between rate measurement timescale (t) and average evolutionary rate estimate ($\bar{r}$). **Eq. 9** (**top-left**), **Eq. 10** (**top-right**), **Eq. 11** (**bottom-left**), and **Eq. 12** (**bottom-right**) were fitted to 1,500 randomly-sampled sets of corresponding $\bar{r}$ and t estimates under the criterion of the least sum of squared errors of $\bar{r}$ (grey lines). The colours correspond to those of the node bars in **Figure 3.2A**. The node numbers refer to those on the trees in **Figure 3.2A**. The summary of the results can be found in **Table S3.2** in **Electronic Appendix**.

3.3.3 Leave-one-out cross validation analyses

Although the high $\bar{R}^2$ scores suggest that all four models can describe the relationship between $\bar{r}$ and t estimates well (**Table S3.2** in **Electronic Appendix**), it has been noted that $\bar{R}^2$ scores are inappropriate for comparing the performance of nonlinear models, and can severely bias model selection in favour of models with more parameters (52). Corrected Akaike information criterion (AICc) and Bayesian information criterion (BIC) have been suggested to be more suitable for this purpose (52). However, the calculation of AICc and BIC scores requires likelihood functions of the models. Since our models are empirical, derived based on a top-down approach, we lack such information. Given these limitations and constrains, we thus used the leave-one-out cross validation (LOOCV) technique to compare our models, in the context of their ability to recover t values given the s values. **Eq.5−8** were used in these analyses as they depict how t relates to s. The t values inferred under the FV-host cospeciation assumption were used as references, and by comparing the recovered t values against them, we computed out-of-sample mean squared error (MSE_{OOS}) scores. These scores were used as a measurement of the overall predictability of the models. (See **Materials and Methods** for details.) The results are summarised in **Figure 3.4** and **Table S3.3** in **Electronic Appendix**.

Overall, we found that the t values recovered by all four models are largely comparable to the references (**Figure 3.4**), indicative of high predictability for all four models. This finding also implies that the cospeciation hypothesis is itself internally consistent. Two discrepancies were found between the reference and recovered t values however. For all four models, we found that (i) the recovered time to most recent common ancestor (tMRCA) of simian FVs (SFVs) (~68.87 to 74.50 Myr, node **7**, **Figure 3.2A** and **Table S3.3** in **Electronic Appendix**) are much greater than the corresponding host tMRCA (~43.47 Myr, (38)), and (ii) that the recovered tMRCA of bovine FV (BFV) and equine FV (EFV) (~43.06 to 48.85 Myr, node **9**, **Figure 3.2A** and **Table S3.3** in **Electronic Appendix**) are much lower than the respective host tMRCA (~87.3 Myr, (36)). To verify that these estimates are not correlated (i.e., that one did not cause the other), we re-calculated the tMRCAs for these two FV groups using all four models

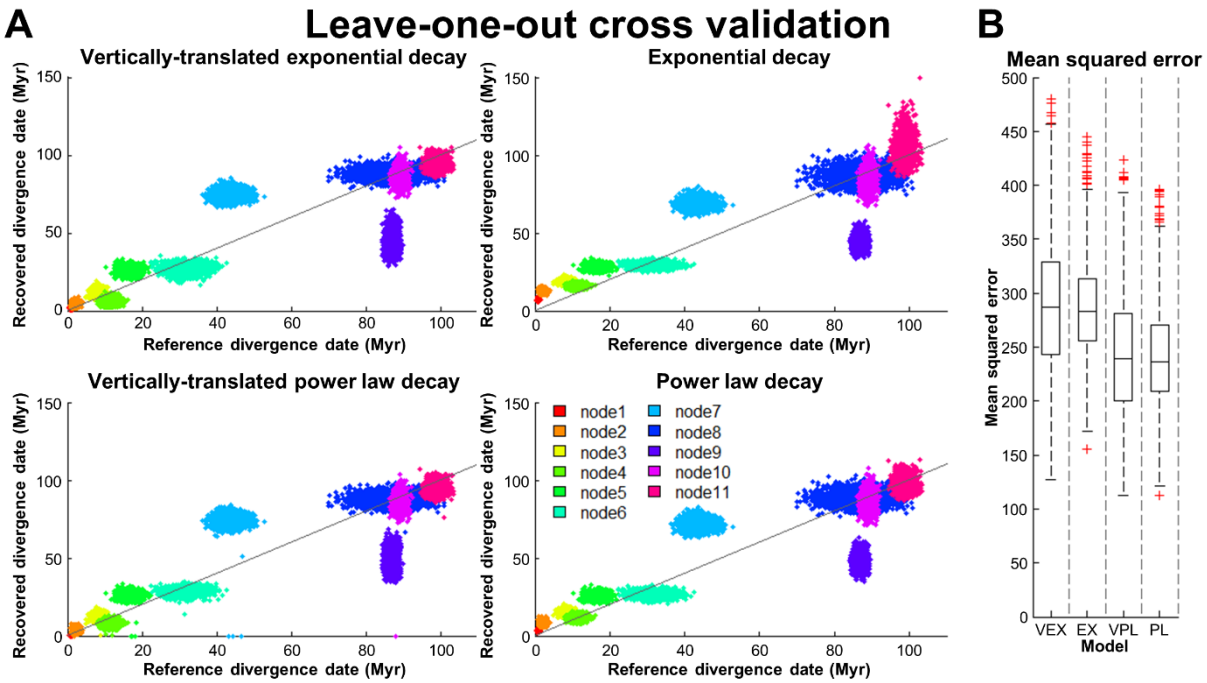


Figure 3.4 Leave-one-out cross validation (LOOCV) tests. The models were validated under the context of divergence date (t) inference, by using LOOCV tests (see Materials and Methods for details). In (**A**), the recovered t estimates are shown against the reference t values, inferred under the cospeciation assumption. The colours of the data points correspond to the colours of the node bars in **Figure 3.2A**. The node numbers refer to those in the trees in **Figure 3.2A**. The grey lines show where the recovered t values are exactly equal to the references. The out-of-sample mean squared errors are shown in (**B**). A summary of the results can be found in **Table S3.3** in **Electronic Appendix**. VEX: vertically-translated exponential rate decay model; EX: simple exponential rate decay model; VPL: vertically-translated power-law rate decay model; and PL: simple power-law rate decay models

estimated independently of nodes **7** and **9**. Both discrepancies could still be observed (data not shown), indicating that they are not artefacts.

Interestingly, unlike other models, the EX model in particular seems to have a tendency to overestimate shallow divergence dates (**Figure 3.4A**; **top right**). This is at odds with the observation that the EX model has a tendency to incorrectly describe the long-term rates (**Figure 3.3**; **top right**). Systematic differences in the distribution of the data points describing the $\bar{r}-t$ and $t-s$ relationships may explain this discrepancy. While the variance of $\bar{r}$ estimates is greatest when t is low, the variance of t estimates is greatest when s is high. Thus, the influence of the data points in parameter estimation differed when the EX model was fitted to the two data sets; that is, while **Eq. 10** parameter estimation (depicting the $\bar{r}-t$ relationship) was primarily influenced by the data points near the y axis, **Eq. 6** parameter estimation (depicting the $t-s$ relationship) was primarily influenced by the data points further away from the y axis. This problem is not as apparent in the other three models however, suggesting that they suffered to a much lesser extent from this effect.

By comparing the MSE_{OOS} scores, we found that the PL model (**Eq. 8**) has the least MSE_{OOS} overall (MSE_{OOS} = 236.25; 95% HPD = 146.21–325.40 Myr, mean rank = 1.62; **Figure 3.4B** and **Table S3.3** in **Electronic Appendix**), indicating that it is the best model for inferring t, and thus most preferable as a TDRP-correcting tool. Interestingly, the VPL model (**Eq. 7**) was found to be the second best model (MSE_{OOS} = 239.04; 95% HPD = 148.06–357.06 Myr, mean rank = 1.87; **Figure 3.4B** and **Table S3.3** in **Electronic Appendix**), suggesting that the extra parameter k_{VPL} in the VPL model does not significantly improve, but instead over-parameterises the model. Likewise, the VEX model (**Eq. 5**) was found to be the worst model (MSE_{OOS} = 287.52; 95% HPD = 169.43–400.58 Myr, mean rank = 3.32; **Figure 3.4B** and **Table S3.3** in **Electronic Appendix**), and the EX model (**Eq. 6**) was found to be the second worst model (MSE_{OOS} = 282.79; 95% HPD = 204.40–369.84 Myr, mean rank = 3.18; **Figure 3.4B** and **Table S3.3** in **Electronic Appendix**).

3.3.4 Examining the performance of currently available molecular clocks in TDRP correction

Several relaxed-clock models have been developed to address the problem of rate variation among lineages (53–55). These models allow rates to vary among branches, and thus over time; therefore, they have the potential to be used as a tool for correcting for the TDRP in evolutionary inferences. In this section, we explored how well currently available relaxed-clock models can accommodate the TDRP under the context of FV timescale inference. We chose to explore scenarios where only three nodes are available as calibrating nodes as they represent realistic circumstances where calibrating information is limited. Nodes **7** and **9** were excluded from this analysis. This is because the LOOCV analyses show that the reference and the recovered t values differ greatly, and it is not possible to determine which is closer to the true values.

Two schemes of date calibration were examined: (i) *aggregated-node calibration* scheme, where all three calibrating nodes are of similar timescales, and (ii) *dispersed-node calibration* scheme where calibrating nodes are of different timescales. In the former calibration scheme, we explored three different scenarios: (i) shallow- (t range: ~0.96−8.30 Myr; nodes **1**, **2**, and **3**), (ii) intermediate- (t range: ~11.50−31.56 Myr; nodes **4**, **5**, and **6**), and (iii) deep-timescale calibration scheme (t range: ~87.18−98.90 Myr; nodes **8**, **10**, and **11**). Similarly, we explored three (arbitrary) circumstances for the latter calibration scheme: (i) dispersed-I calibration scheme (t range: ~2.17−87.18 Myr; nodes **2**, **5**, and **8**), (ii) dispersed-II calibration scheme (t range: ~0.96−88.7 Myr; nodes **1**, **4** and **10**), and (iii) dispersed-III calibration scheme (t range: ~8.30−98.9 Myr; nodes **3**, **6**, and **11**). Three currently available molecular clocks were investigated, including (i) a strict molecular clock, (ii) a log-normal relaxed clock (55), and (iii) a random-local relaxed clock (54). The strict clock was included to examine how the TDRP would affect t inferences if it is ignored. We focused on Bayesian timescale estimates, and used the *pol* nucleotide alignment that was used in the phylogenetic reconstruction to estimate the timescales (see **Materials and Methods**). Again, the t estimates inferred under the FV-

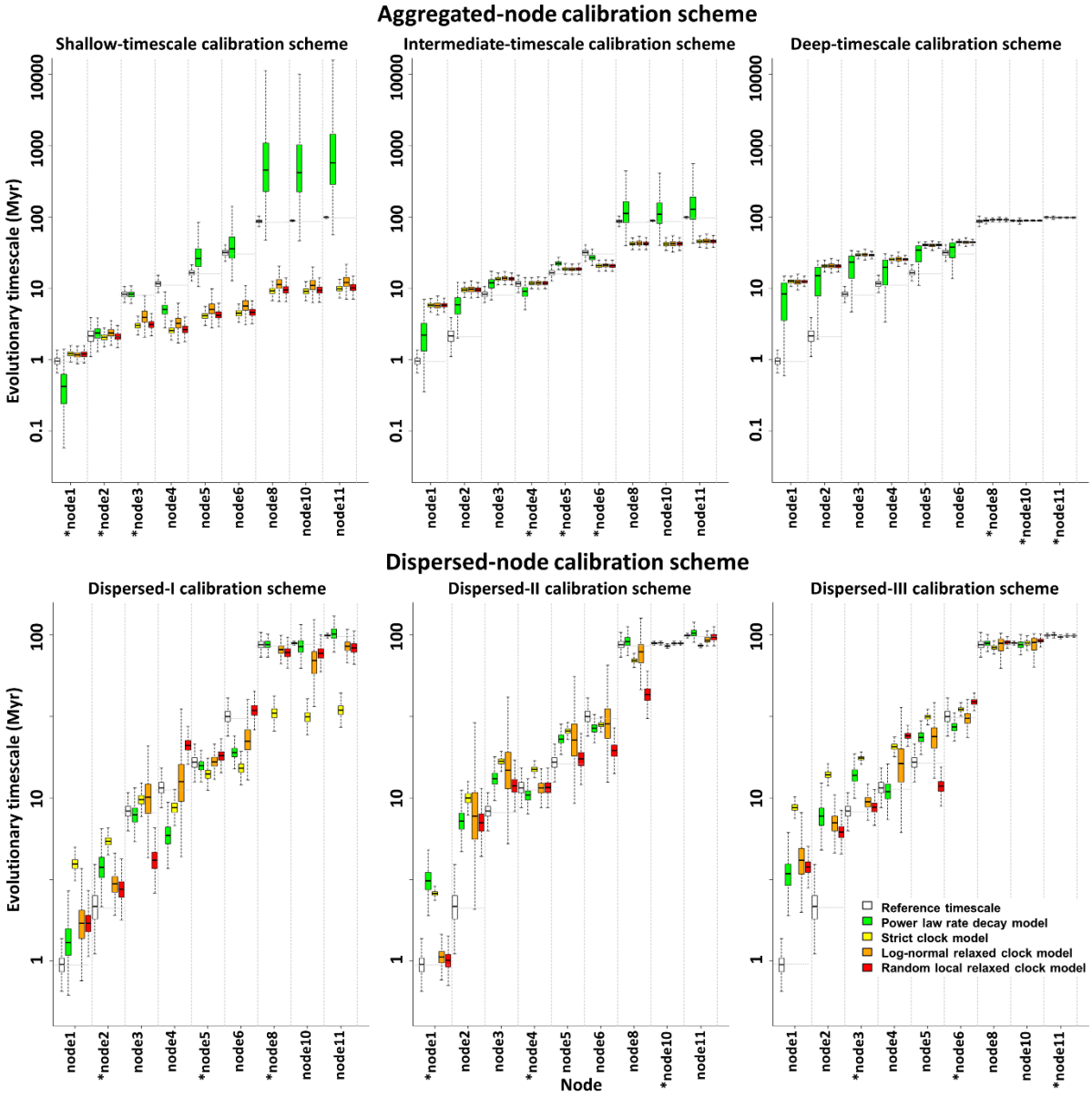


Figure 3.5 Performance of various molecular clocks in accommodating the time-dependent rate phenomenon (TDRP). We explored the performance of various clock models in accommodating the TDRP in the context of foamy virus (FV) evolutionary timescale inference under various hypothetical scenarios, where only three nodes are available as calibrating nodes. Two calibration schemes were examined: (i) aggregated-node calibration scheme (**top**), in which all three calibrating nodes are of similar timescales, and (ii) dispersed-node calibration scheme (**bottom**), in which calibrating nodes are of different timescales. The aggregated-node calibration scheme was sub-divided into three sub-schemes: (i) shallow-timescale calibration scheme (**top left**), (ii) intermediate-timescale calibration scheme (**top middle**), and (iii) deep-timescale calibration scheme (**top right**). The dispersed-node calibration scheme was also sub-divided into three arbitrary sub-schemes: (i) dispersed-I calibration scheme (**bottom left**), (ii) dispersed-II calibration scheme (**bottom middle**), and (iii) dispersed-III calibration scheme (**bottom right**). Three sets of dates from different clock implementations and one of our TDRP models were plotted: (i) simple power rate decay model (green), (ii) strict clock model (yellow), (iii) log-normal relaxed clock model (orange), and (iv) random-local relaxed clock model (red). FV evolutionary timescales inferred from the host timescales (reference timescales) are in white. Horizontal dotted grey lines are median reference timescales. The results are plotted on a log scale, and arranged in such a way that the evolutionary timescales of shallow nodes are on the left, and those of deep nodes are on the right. The node numbers refer to those in **Figure 3.2A**. Calibrating nodes are indicated by asterisks '*'. The summary of the results can be found in **Table S3.4** and **S3.5** in **Electronic Appendix**.

host cospeciation assumption were used as reference t estimates.

We also compared these currently available clocks to our PL model (**Eq. 8**). Unlike the above however, the calculation was not done under a full Bayesian phylogenetic framework. Rather, we used t and *s* estimates of the calibrating nodes to estimate the model, and then inferred the t values of other nodes based on their *s* estimates. The results are shown in **Figure 3.5**, and **Table S3.4** in **Electronic Appendix**.

3.3.4.1 Aggregated-node calibration scheme

Under the aggregated-node calibration scheme (**Figure 3.5**; **top**), our results show that the t values estimated under the strict clock, log-normal relaxed clock, and random-local relaxed clock are comparable to one another. This is likely because the calibrating nodes do not provide enough signal for the rate variation among lineages to be estimated. Indeed, we found that the coefficients of rate variation estimated under the two relaxed clock models do not deviate very far from zero (**Table S3.5** in **Electronic Appendix**).

Overall, we found that the t values of the nodes adjacent to the calibrating nodes are estimated relatively accurately, comparable to the reference values. However, as we move further away from the calibrating nodes, the t estimates become increasingly more inaccurate. Under the shallow-timescale calibration scheme, the t values of deep nodes are severely underestimated (**Figure 3.5**; **top left**). In contrast, the t estimates of shallow nodes are severely overestimated under the deep-timescale calibration scheme (**Figure 3.5**; **top right**). In the intermediate-timescale calibration scheme, the t values of deep nodes are underestimated while those of shallow nodes are overestimated (**Figure 3.5**; **top middle**). Our results are consistent with the findings in a previous study that solely focused on SFVs (56).

In contrast, our results showed that the PL model could accommodate the TDRP reasonably well under these examined scenarios, better than the currently available molecular clocks. Although we found that the uncertainties of t estimates are extremely large (thus providing

relatively little information about the actual dates themselves), the reference t values are almost always contained within the t distributions estimated under the PL model, whereas those estimated under the current clock models do not even overlap with the reference t distribution (**Figure 3.5**; **top**). The median t estimates from the PL model are also closer to the reference values, which is most apparent in the intermediate- and deep-timescale calibration schemes (**Figure 3.5**; **top middle** and **right**, respectively), but less so in the shallow-timescale calibration scheme (**Figure 3.5**; **top left**). These large uncertainties are unlikely an intrinsic property of the PL model however, but a shortcoming of our extrapolation approach in which the t and s values are not estimated in conjunction with one another. Coupled with the fact that these are extrapolations over large timescales and that the uncertainty of the s estimate increases down the tree (**Figure 3.2A**), it is unsurprising to find that this problem is most apparent in the shallow-timescale calibration scheme (**Figure 3.5**; **top left**), but less so in the intermediate-timescale calibration scheme (**Figure 3.5**; **top middle**), and least in the deep-timescale calibration scheme (**Figure 3.5**; **top right**). These findings thus indicate that deep nodes are preferable as calibrating nodes, consistent with the results from a previous simulation study that investigated the impact of calibrating node position on timescale estimates (56).

3.3.4.2 Dispersed-node calibration scheme

Under the dispersed-node calibration scheme (**Figure 3.5**; **bottom**), we found that the performances of the log-normal and random-local relaxed clocks differ from that of the strict clock. This is likely because the calibrating nodes are spread out in time and hence provide adequate information for the rate variation among branches, and thus over time, to be modelled more accurately. Indeed, the coefficients of rate variation estimated under the two relaxed clocks deviate far from zero (**Table S3.5** in **Electronic Appendix**).

Overall, we found that the strict clock model performs worst, and the t values obtained under the clock vary substantially between sub-schemes. For example, in the dispersed-I calibration scheme (**Figure 3.5**; **bottom left**), the t values of intermediate nodes were estimated relatively

accurately, but those of shallow nodes were overestimated, and those of deep nodes were underestimated. On the other hand, in the dispersed-II and -III calibration schemes (**Figure 3.5**; **bottom middle** and **left**, respectively), the t values of deep nodes were estimated relatively accurately whereas those of shallow nodes were severely overestimated. In contrast, the two relaxed clock models seem to perform reasonably well, returning t values that are comparable to the references, and also to those inferred under the PL model. This, in turn, indicates that the performance of these three models is comparable under these scenarios. Moreover, we also observed that the t distributions inferred under the PL model are tighter compared to those inferred under the aggregated-node calibration scheme. This is likely because the dispersed-node schemes involve interpolation rather than extrapolation.

3.3.5 Inferring missing FV divergence dates by using the PL model

Consistent with the results from previous studies (35, 57), amidst the stable FV-host cospeciation history, we found a few mismatches in the history of New World monkey (NWM) FVs and their hosts (**Figure 3.2A**). These involve (i) the split between marmoset and spider monkey FVs (SFVmar and SFVspm, respectively) (**Figure 3.2A**; node **I**), and (ii) the divergence leading to squirrel monkey FV (SFVsqu) lineage (**Figure 3.2A**; node **II**).

As a case study, we used the PL model to infer the t values for these two nodes. To do so, **Eq. 8** was fitted to the data set of t and s estimates, with those of nodes **7** and **9** excluded (as we could not conclusively determine the true t values for these two nodes; see above). The models were then used to infer the t values of node **I** and **II** from their s estimates. Our analyses inferred the tMRCA of SFVmar and SFVspm to be ~23.40 (95% HPD = 19.25–27.52) Myr, and the earlier branching of SFVsqu lineage to occur ~40.81 (95% HPD = 35.58–46.08) Myr ago.

3.4 Discussion

3.4.1 What are the causes of the TDRP?

Our analyses show that the estimated values of the FV substitution rates decrease continuously with measurement timescales (**Figure 3.2**), and the PL model is the best model for correcting for the TDRP. It is possible that the TDRP observed here is the result of changes in FV biology, such as polymerase fidelity, replication speed, and/or transmission modes. These ideas have been put forward as plausible explanations for the rate variations observed in several viruses, including hepadnaviruses (3), and human T-cell lymphotopic viruses (4, 5) as well as RNA viruses in general (6). These hypotheses would be reasonable and sensible if the values of the rate estimates changed discretely. However, as the rate estimate continuously decreases with the measurement timescale, this would posit that the biology of FVs has been continuously changing with time in such a way that that the viral rate of evolution gradually increases through time. Moreover, since the TDRP is not unique to FVs, but very common among viruses, these hypotheses would also imply that the biology of many viruses has changed in a similar manner. Although possible, this is extremely unlikely (58). We therefore believe that changes in viral biology likely do not play a major role in governing the overall decay trend of viral evolutionary rate estimates.

Changes in natural selection pressure have also been put forward as a potential underlying cause of the TDRP (27, 29, 59). Nevertheless, like viral biology, it is extremely difficult to imagine that environmental factors would have changed systematically with time so that viruses experience less and less purifying selection pressure as they evolve. It is important to note that we are by no means suggesting that changes in environment do not cause viral evolutionary rates to alter. Indeed, they can, and this has been observed. An analysis of bat rabies viruses has shown that their evolutionary rates are strongly correlated with the host local environments, suggesting that environmental changes can alter the rate of viral evolution (60). However, we argue that, for the alteration of environment and/or natural selection to play a major role in generating the TDRP, not only must they change the rate, but the effect must

also be time-correlated, which we consider unlikely. We thus believe that, rather than reflecting a genuine change in natural selection pressure, the observed TDRP is likely an artefact caused by other factors.

So, if viral biology, environmental factors, and selection pressure remain relatively constant through time, what could possibly cause the TDRP? It has been proposed that short-term rate estimates tend to be overestimated due to the inclusion of transient deleterious variations (28, 41, 61), recent adaptive changes (62–64), and sequencing errors (20, 65–67). Errors in coalescent calibration information have also been suggested as a sufficient explanation for elevated short-term rate estimates (68). In contrast, misspecification of substitution model (69, 70), saturation of nucleotide changes (14, 28, 69, 71), and improper accounting for rate heterogeneity among sequence positions (30) can lead to underestimation of long-term rate estimates. All of these factors could contribute to the TDRP, and at present, the importance and relevance of each factor is still poorly understood and continues to be debated (see (28) for a review).

3.4.2 The TDRP as a possible explanation of the short-term and long-term rate discrepancy

A large discrepancy of ~4 to 5 orders of magnitude has been observed between FV short-term and long-term rate estimates. Indeed, such discrepancies have been found not only in FVs, but many and diverse RNA and DNA viruses (3–5, 58, 64). Rather than viewing them as conflicting rate estimates, it is has been proposed that this discrepancy may result from estimating the rates over different extremes of the TDRP (58).

In order to investigate whether the TDRP can explain this discrepancy or not, we used our four rate models to recover the short-term and long-term rate estimates of FVs. Our analyses showed that all of the four models could recover the long-term rate relatively well. However, given that the models were parameterised on FV long-term rate estimates, this is expected

and unsurprising. In contrast, we found that only the VPL and PL models could recover the FV short-term rate accurately, but the VEX and EX models severely underestimated the rate by ~3 to 4 orders of magnitude. These results further support the use of the PL model as a tool for TDRP correction, and simultaneously indicate that, indeed, the short-term and long-term rate discrepancy in FVs can be explained by the TDRP.

3.4.3 Implications of the TDRP

There are a number of implications of the TDRP. One of them is that it is important to take the timescale of rate measurement into consideration when using or interpreting evolutionary rate estimates; otherwise, the outcomes could be severely biased. For example, it is inappropriate to use long-term rate estimates to infer or evaluate viral short-term epidemiological dynamics, as they could give an erroneous impression that the viruses being considered are of low adaptive and cross-species transmission potential. Similarly, in the specific case of FVs, which are candidate gene-carrying vectors for gene-therapy (72, 73), using their long-term rates to evaluate the risks of FV-derived gene vectors could be misleading, as they are in fact capable of evolving at a (short-term) rate as high as $\sim 3.75 \times 10^{-4}$ s/n/y (47), comparable to those of highly pathogenic and fast-adapting viruses like HIVs (41, 74–76) and influenzas (41, 77–79).

Another important, and perhaps more obvious, implication is that it will no longer be valid to naïvely extrapolate rate estimates across different time frames when inferring evolutionary timescales. The assumption of a single molecular clock can bias the timescale inference with the severity increasing with the timescale of rate extrapolation. As the value of the rate estimate continuously decreases with the measurement timescale, the TDRP should appear in, and is relevant to, every phylogenetic analysis. That is, if two or more evolutionary rate estimates are calculated over different timescales from a particular phylogeny (e.g. use different internal nodes to calibrate the rate), the TDRP should show up. Nevertheless, in practice, if the timescale of evolutionary investigation is short, e.g. hundreds of years (which are typical for infectious disease analyses), the uncertainty of the rate estimate may overwhelm the effect of the TDRP, and the phenomenon might not be observed. Indeed, the strict molecular clock has

sometimes been identified as appropriate for studies over short timescales such as epidemiological studies (41, 80, 81), but not for longer timescale analyses such as evolutionary investigations that compare viruses in different host species (4, 82, 83). Moreover, the TDRP can also bias demographic parameter estimations such as effective population sizes and migration rates if it is unaccounted for. This is simply because the calculation involves estimating or knowing substitution rates (28, 84).

Similar to the results from previous studies (56, 85), our results show that evolutionary timescales calibrated under currently available molecular clocks are highly sensitive to the choice of calibrating nodes. The effect is most pronounced when the calibrating nodes are of the same timescale (**Figure 3.5**; **top**). As short-term rate estimates are greater than the long-term ones, calibrating deep divergence dates with shallow nodes will underestimate them. Conversely, using deep nodes to calibrate shallow divergence dates will tend to overestimate them. Strikingly, we found that the current relaxed clocks do not perform any better than the strict clock under these circumstances, indicating that they are not an effective solution to the TDRP problem. Compared to the currently available relaxed-clock models, the PL model performs better and is more consistent across the calibration schemes, even when calibrating nodes are of similar timescales (**Figure 3.5**; **top**).

Our results suggest that the currently available relaxed-clock models suffer much less from the TDRP problem, and perhaps are equally good to the PL model, if calibrating nodes are dispersed in time (**Figure 3.5**; **bottom**), consistent with the results from a previous study (32). Nevertheless, in most realistic applications, the number and dispersal of calibrating nodes is serendipitous, dictated by their limited availability which often involves only one or a few nodes. Given this limited availability of time-calibrating information and the sensitivity of the current clocks to the choice of calibrating nodes, our results overall suggest that the current relaxed-clock models might not be an effective and practical solution to the TDRP problem yet. We, thus, believe that our PL model will be useful as a guideline to further improve our current evolutionary inference tools.

3.4.4 Possible evolutionary rate dynamics heterogeneity among viral lineages

Our analyses showed that the tMRCA estimates of SFVs recovered by the four models are considerably higher than the host tMRCA (**Figure 3.4A** and **Table S3.3** in **Electronic Appendix**; node **7**). Our analyses also estimated the tMRCA of SFVmar and SFVspm (~23.40 Myr) and that of NWM FVs (~40.81 Myr) to be greater than those of their hosts (~22.76 Myr, (38)). One possibility is that the tMRCAs inferred by our models are not artefacts, but resemble the real dates. This would however imply duplications of viral lineages in the absence of host diversification, which we consider to be unlikely. An alternative explanation is that the evolutionary rates of NWM FVs and the stem lineage are higher than average, and our analysis framework did not take this into account. Since we employed a strict clock to estimate s values, we thereby assumed that all taxa evolve under the same time-dependent rate dynamics. Thus, the incongruences between the predicted and inferred NWM FV divergence dates may be indicative of the heterogeneity of evolutionary rate dynamics among viral lineages, and this may bias evolutionary timescale inferences if it is unaccounted for. A solution to this problem would be to fit multiple time-dependent rate models to different parts of the tree.

Conversely, our results show that the recovered BFV/EFV tMRCAs are considerably lower than the host tMRCA (**Figure 3.4A** and **Table S3.3** in **Electronic Appendix**; node **9**), implying a cross-species FV transmission between equine and bovine hosts ~53 to 59 Ma. Unlike the scenarios discussed above, this is relatively reasonable and should not be ruled out. Alternatively, it could be that the substitution rate on the BFV/EFV stem branch is greater than that of other FVs, and our analysis with a strict clock failed to account for this rate difference. Further resolution of these questions would require the identification and analysis of FV genomes of other bovines and equines.

3.5 Conclusions

Our knowledge of viral natural history has been greatly advanced by molecular analyses. One of the key steps in viral evolutionary study involves estimating the rate of substitution. By using FVs as a case study, we show that their evolutionary rate estimates are negatively correlated with the timescale of rate measurement, and this is likely responsible for the short-term/long-term rate discrepancy observed in FVs, and perhaps other viruses as well (58). We also demonstrate that currently available relaxed-clock models are inadequate for accommodating the TDRP; using them to infer evolutionary timescales can severely bias the date estimates especially when rate-calibrating nodes are of similar timescales. We believe that the PL model developed here will be useful as a guideline for the further improvement of existing analytical tools. Our results also suggest that heterogeneity in rate dynamics among viral lineages may exist, and can affect evolutionary inference.

Combined, our work highlights the importance of the TDRP and heterogeneity in evolutionary rate dynamics among viral lineages. Given the potential impacts of the TDRP on evolutionary inference and rate estimate interpretation, the fact that it is so widespread in nature but has been noticed only recently could mean that the credibility of evolutionary timescale estimates of many viruses may need to be reconsidered.

3.6 Materials and Methods

3.6.1 Phylogenetic reconstruction

Four phylogenies of 14 extant foamy viruses (FVs) (**Table 3.1**) were estimated from manually-curated Pol protein (1,116 aa; see **Data set S3.1** in **Electronic Appendix**) and *pol* nucleotide (3,351 nt; see **Data set S3.2** in the **Electronic Appendix**) alignments. Potential recombination among aligned sequences was assessed with a quartet-based recombination detection program VisRD3 (86), both at nucleotide and protein levels. In both cases, the null distribution was built based on 1,000 data sets of randomly-shuffled sequences, and the extended statistical geometry (Hamming) weighting option was applied. In the nucleotide analysis, the

window and step size were 200 and 40 nt, respectively, and in the protein analysis, the window and step size were 100 and 20 aa, respectively. The results showed no significant evidence for recombination. The best-fit amino-acid and nucleotide substitution models used in the phylogenetic reconstructions were determined to be the RTREV + I + Γ(4) + F and GTR + I + Γ(4) by ProtTest 2.4 (87) and Jmodeltest 2.1.1 (88), respectively, under the AICc criterion.

The phylogenies were constructed under both the Bayesian and maximum-likelihood phylogenetic frameworks, by using MrBayes 3.2.1 (89) and MEGA 5.2 (90), respectively. In the Bayesian analyses, 2 independent MCMCs were run for 50,000,000 steps each, with the initial 12,500,000 steps discarded as burn-in. Parameters were thereafter logged every 2,500 steps. Metropolis coupling was applied, using the setting of 3 hot and 1 cold chains. Parameter estimate convergences were diagnosed using potential scale reduction factors (PSRFs). PSRFs of all parameters are ~1.000, indicating that they were all well sampled from their posterior distributions and had converged. In the maximum-likelihood analyses, bootstrap support values were calculated using 1,000 pseudoreplicates. A molecular clock was not imposed in either of the analyses. In total, four phylogenies were estimated: (i) a maximum-likelihood Pol protein tree, (ii) a Bayesian Pol protein maximum clade credibility (MCC) tree, (iii) a maximum-likelihood *pol* nucleotide tree, and (iv) a Bayesian *pol* nucleotide MCC tree (**Figure 3.1**), all of which show the same topology.

Note that we did not include endogenous mammalian FVs – SloEFV (34), PSFVaye (35, 42), and ChrEFV (or CaEFV) (35, 43) – in the analysis. The reason is that their evolutionary rate is a mixed rate, comprising the rate of viral evolution and the neutral rate of host evolution, which could bias the analysis. Even if we can decompose the rate into the two rate components, it is still unclear how the 'truncated' rates of viral evolution would fit into the dynamics of the evolutionary rate of extant viruses. As a result, we focus our study on the rate dynamics of extant FVs only.

3.6.2 Inferring FV node-to-tip total per-lineage substitutions (s estimates), evolutionary timescales (t estimates), and node-to-tip average evolutionary rates ($\bar{r}$ estimates) over various time frames, and assessing the correlation between $\bar{r}$ and t estimates

We estimated s values from the manually-curated *pol* nucleotide alignment under the Bayesian framework, by using BEAST 1.7.4 (91). The strict molecular clock assumption with a fixed rate of 1, and Yule speciation process were applied. The topology of the phylogeny was fixed according to the phylogeny obtained in the phylogenetic reconstruction. The MCMC was run for 50,000,000 steps, with the initial 12,500,000 steps discarded as burn-in. Parameters were logged every 2,500 steps. In total, 15,000 sets of parameter estimates were sampled. Parameter value convergence and sampling independency were manually inspected using Tracer v1.5 (92). We found all parameters had an effective sample size (ESS) of >350, indicating that all of them were well sampled and had converged.

In total, as there were 13 internal nodes, 13 posterior distributions of s estimates were obtained, 11 of which could be assigned to independently estimated evolutionary timescales of their hosts on the basis of the FV-host cospeciation assumption (**Figure 3.2A**, and **Table S3.1** in **Electronic Appendix**). We then divided s estimates by their evolutionary timescale (t estimates) to derive $\bar{r}$ estimates for various time frames. Note that, since a strict clock was applied and the tips were all aligned, the timescale of the rate estimation is equivalent to the node heights in units of time. To accommodate the uncertainty of t and $\bar{r}$ estimates, we simulated 15,000 sets of t values under the assumption that they are normally distributed, and randomly paired them to each of the data subsets of s estimates to compute 15,000 sets of $\bar{r}$ estimates. The means of the t distributions were assumed to be equal to the median t estimates reported in the literature, and their standard deviations were calculated from the reported upper- and lower-bounds of the corresponding 95% highest posterior density intervals: max ($\frac{Median-Lower\ 95\%\ HPD\ limit}{1.96}$, $\frac{Upper\ 95\%\ HPD\ limit-Median}{1.96}$). The t simulation was constrained by the estimated phylogeny in such a way that the t values of child nodes were

always lower than those of the corresponding parental nodes. This approach to model parameter estimation takes into account the uncertainty of s, t, and $\bar{r}$ estimates to the full extent, considering the whole space of their estimated Bayesian posterior distributions.

To preliminarily evaluate the correlation between $\bar{r}$ and t, we fitted a linear model to each of the 15,000 data subsets of the log-transformed $\bar{r}$ and t estimates, using the *LinearModel.fit* function implemented in MATLAB R2012a (93). However, it has been noted that a correlation analysis between a quotient and its denominator has a tendency to yield a seemingly significant but in fact spurious negative correlation (94). Randomisation tests have been recommended as a way to address this issue (94). Here, we randomly matched s and t estimates to compute $\bar{r}$ values under the null hypothesis that there is no correlation between s and t, and in turn computed a 'null' $\bar{r}$-t correlation coefficient. In each of the 15,000 sub correlation analyses, this process was repeated 100 times to construct a distribution of the null $\bar{r}$-t correlation coefficient, which was then used to compute the *P* value.

We also note that our data is not phylogenetically independent; for example, each of the terminal branches are included in several s, t, and $\bar{r}$ estimates. The results thus should be interpreted with an understanding that the data does not fully conform to the ordinary model-fitting assumption. It is also important to note that this problem is not unique to just the above analysis, but also applies to all subsequent model estimations in our study. In our model framework, this is inevitable, unfortunately, due to the fact that a phylogenetic independent data set, i.e., a data set of instantaneous rate estimates and their corresponding timescale, is extremely difficult, if not impossible, to obtain for FVs. However, validating the models under the context of t inference still shows that they work well despite this issue (see LOOCV analyses), indicating that our analyses suffer from this problem only to a low degree.

3.6.3 Describing the temporal dynamics of $\bar{r}$

Four equations depicting the $t-\bar{r}$ relationship were derived based on four empirical rate decay hypotheses: (i) vertically-translated exponential rate decay hypothesis (**Eq. 9**), (ii) simple exponential rate decay hypothesis (**Eq. 10**), (iii) vertically-translated power law rate decay hypothesis (**Eq. 11**), and (iv) simple power law rate decay hypothesis (**Eq. 12**). These equations were in turn fitted to 1,500 data sets of corresponding t and $\bar{r}$ estimates randomly sampled from their posterior distribution under the criterion of the least sum of squared errors (LSE) of $\bar{r}$, using the *lsqcurvefit* function implemented in MATLAB R2012a (93). All parameters were constrained to be greater than zero. $\bar{R}^2$ scores were used to preliminarily assess how well the models describe the data.

The estimated models were also used to recover previously reported short-term rate (47) and long-term rate (33, 39) estimates of FVs. In this study, the short-term and long-term rate estimates were defined and calculated over a timescale of 10 years, and 30 Myr, respectively.

3.6.4 Validating the performance of the models under the context of t inference by using the leave-one-out cross validation (LOOCV) technique

We first derived four equations describing the relationship of t and s from the four empirical rate decay hypotheses (**Eq. 5–8**, see Results), and fitted them to 1,500 sets of corresponding t and s estimates, sampled from their posterior distributions consisting of 15,000 data sets. For consistency, we constructed the t−s data set in such a way that it corresponded to the data set of t and $\bar{r}$ estimates used in the $\bar{r}$ dynamics analyses. In total, the models were validated for 1,500 rounds. In each round, the data set was partitioned into a testing set containing an estimate of s and t of one particular node, and a training set containing the rest of the data. The models were fitted to the training set under the criterion of the LSE of t, by using the *lsqcurvefit* function implemented in MATLAB R2012a (93). All parameters in the models were constrained to be greater than zero. The resultant models were then used to infer the t of the testing node from its s estimate. This process was repeated such that the t and s of every node

is used exactly once as the testing data to complete one round of LOOCV testing. The overall performance of the models was assessed based on the out-of-sample mean squared error (MSE_{OOS}), and the t values inferred under the FV-host cospeciation assumption were used as references. To compare the models, we applied Friedman's test to the MSE_{OOS}, and performed a post-hoc analysis (complete-pairwise MSE_{OOS} comparisons based on Wilcoxon signed-rank tests) with the Bonferroni multiple-testing correction. The significance was evaluated at $\alpha = 0.05$.

3.6.5 Examining the effect of the TDRP on t inference

We examined the effect of the TDRP on t inference under six hypothetical scenarios, in all of which only three nodes were used as calibrating nodes (see Results). Nodes **7** and **9** (**Figure 3.2A**) were excluded from the analysis. The t values inferred under the FV-host cospeciation assumption were used as references. Three currently available molecular clocks were examined: (i) a strict molecular clock, (ii) a log-normal relaxed clock (55), and (iii) a random-local relaxed clock (54). The t inference was performed under the Bayesian phylogenetic framework, implemented in BEAST 1.7.4 (91), by using the *pol* nucleotide alignment that was used in the phylogenetic reconstruction. The Yule speciation process and GTR + I + Γ(4) substitution model were applied. This model was determined to be the best for our alignment by Jmodeltest 2.1.1 (88) under the AICc criterion. The topology of the phylogeny obtained in the phylogenetic reconstruction was fixed. The MCMC was run for 50,000,000 steps, with the initial 12,500,000 steps discarded as burn-in. Parameters were logged every 2,500 steps. Parameter value convergence and sampling independency were manually inspected using Tracer v1.5 (92). For the runs that returned parameters with ESSs of <200 (a recommended cut-off (55)), we repeated the analyses and combined the results until the ESSs of all parameters were >200.

Moreover, we also examined the PL model for how well it addresses the TDRP under these various hypothetical scenarios. The t and s estimates of the calibrating nodes were used to

compute the model parameters. For consistency, the data set of t and s estimates used in this investigation was the same one that was used in the LOOCV analyses. The model was fitted to each of the 1,500 data subsets using the *lsqcurvefit* function implemented in MATLAB R2012a (93) under the criterion of the LSE of t, and the parameter estimation was constrained so that all parameters were greater than zero. The estimated model was then used to infer the t values of other nodes based on their s estimates.

3.6.6 Inferring missing FV divergence dates by using the PL model

The PL model was fitted to the data set of t and s estimates that was used in the LOOCV analyses under the criterion of the LSE of t, but without the data associated with nodes **7** and **9**. The fitting was performed using the *lsqcurvefit* function implemented in MATLAB R2012a (93), and was constrained so that all parameter values were greater than zero. The estimated model was then used to infer the missing t values for two nodes based on their s estimates: (i) node **I** (the split between marmoset and spider monkey FVs) and (ii) node **II** (the basal diversification of all New World monkey FVs).

3.7 References

1. **Alizon S**, **Fraser C**. 2013. Within-host and between-host evolutionary rates across the HIV-1 genome. Retrovirology **10**:49.
2. **Gray RR**, **Parker J**, **Lemey P**, **Salemi M**, **Katzourakis A**, **Pybus OG**. 2011. The mode and tempo of hepatitis C virus evolution within and among hosts. BMC Evol. Biol. **11**:131.
3. **Gilbert C**, **Feschotte C**. 2010. Genomic fossils calibrate the long-term evolution of hepadnaviruses. PLoS Biol. **8**:e1000495.
4. **Salemi M**, **Lewis M**, **Egan JF**, **Hall WW**, **Desmyter J**, **Vandamme AM**. 1999. Different population dynamics of human T cell lymphotropic virus type II in intravenous drug users compared with endemically infected tribes. Proc Natl Acad Sci U S A **96**:13253–13258.
5. **Vandamme AM**, **Bertazzoni U**, **Salemi M**. 2000. Evolutionary strategies of human T-cell lymphotropic virus type II. Gene **261**:171–180.
6. **Hanada K**, **Suzuki Y**, **Gojobori T**. 2004. A large variation in the rates of synonymous substitution for RNA viruses and its relationship to a diversity of viral infection and transmission modes. Mol. Biol. Evol. **21**:1074–1080.
7. **Rocha EPC**, **Smith JM**, **Hurst LD**, **Holden MTG**, **Cooper JE**, **Smith NH**, **Feil EJ**. 2006. Comparisons of dN/dS are time dependent for closely related bacterial genomes. J. Theor. Biol. **239**:226–235.
8. **Feng L**, **Reeves PR**, **Lan R**, **Ren Y**, **Gao C**, **Zhou Z**, **Ren Y**, **Cheng J**, **Wang W**, **Wang J**, **Qian W**, **Li D**, **Wang L**. 2008. A recalibrated molecular clock and independent origins for the cholera pandemic clones. PLoS One **3**:e4053.
9. **Denver DR**, **Morris K**, **Lynch M**, **Vassilieva LL**, **Thomas WK**. 2000. High direct estimate of the mutation rate in the mitochondrial genome of *Caenorhabditis elegans*. Science **289**:2342–2344.
10. **Haag-Liautard C**, **Coffey N**, **Houle D**, **Lynch M**, **Charlesworth B**, **Keightley PD**. 2008. Direct estimation of the mitochondrial DNA mutation rate in *Drosophila melanogaster*. PLoS Biol. **6**:1706–1714.
11. **Papadopoulou A**, **Anastasiou I**, **Vogler AP**. 2010. Revisiting the insect mitochondrial

molecular clock: the mid-Aegean trench calibration. Mol. Biol. Evol. **27**:1659–1672.

12. **Genner MJ**, **Seehausen O**, **Lunt DH**, **Joyce DA**, **Shaw PW**, **Carvalho GR**, **Turner GF**. 2007. Age of cichlids: new dates for ancient lake fish radiations. Mol. Biol. Evol. **24**:1269–1282.
13. **Burridge CP**, **Craw D**, **Fletcher D**, **Waters JM**. 2008. Geological dates and molecular rates: fish DNA sheds light on time dependency. Mol. Biol. Evol. **25**:624–633.
14. **Garcia-Moreno J**. 2004. Is there a universal mtDNA clock for birds ? J. Avian Biol. **35**:465–468.
15. **Millar CD**, **Dodd A**, **Anderson J**, **Gibb GC**, **Ritchie PA**, **Baroni C**, **Woodhams MD**, **Hendy MD**, **Lambert DM**. 2008. Mutation and evolutionary rates in Adélie penguins from the Antarctic. PLoS Genet. **4**:e1000209.
16. **Subramanian S**, **Denver DR**, **Millar CD**, **Heupink T**, **Aschrafi A**, **Emslie SD**, **Baroni C**, **Lambert DM**. 2009. High mitogenomic evolutionary rates and time dependency. Trends Genet. **25**:482–486.
17. **Parsons T**, **Muniec D**, **Sullivan K**, **Woodyatt N**, **Alliston-Greiner R**, **Wilson M**, **Berry D**, **Holland K**, **Weedn V**, **Gill P**, **Holland M**. 1997. A high observed substitution rate in the human mitochondrial DNA control region. Nat. Genet. **15**:363–368.
18. **Howell N**, **Smejkal CB**, **Mackey DA**, **Chinnery PF**, **Turnbull DM**, **Herrnstadt C**. 2003. The pedigree rate of sequence divergence in the human mitochondrial genome: there is a difference between phylogenetic and pedigree rates. Am. J. Hum. Genet. **72**:659–670.
19. **Santos C**, **Montiel R**, **Sierra B**, **Bettencourt C**, **Fernandez E**, **Alvarez L**, **Lima M**, **Abade A**, **Aluja MP**. 2005. Understanding differences between phylogenetic and pedigree-derived mtDNA mutation rate: a model using families from the Azores Islands (Portugal). Mol. Biol. Evol. **22**:1490–1505.
20. **Ho SYW**, **Phillips MJ**, **Cooper A**, **Drummond AJ**. 2005. Time dependency of molecular rate estimates and systematic overestimation of recent divergence times. Mol. Biol. Evol. **22**:1561–1568.
21. **Ho SYW**, **Shapiro B**, **Phillips MJ**, **Cooper A**, **Drummond AJ**. 2007. Evidence for time

dependency of molecular rate estimates. Syst. Biol. **56**:515–522.

22. **Subramanian S**, **Lambert DM**. 2011. Time dependency of molecular evolutionary rates? Yes and no. Genome Biol. Evol. **3**:1324–1328.

23. **Henn BM**, **Gignoux CR**, **Feldman MW**, **Mountain JL**. 2009. Characterizing the time dependency of human mitochondrial DNA mutation rate estimates. Mol. Biol. Evol. **26**:217–230.

24. **Denver DR**, **Morris K**, **Lynch M**, **Thomas WK**. 2004. High mutation rate and predominance of insertions in the *Caenorhabditis elegans* nuclear genome. Nature **430**:679–682.

25. **Houle D**, **Nuzhdin S V**. 2004. Mutation accumulation and the effect of copia insertions in *Drosophila melanogaster*. Genet. Res. **83**:7–18.

26. **Xue Y**, **Wang Q**, **Long Q**, **Ng BL**, **Swerdlow H**, **Burton J**, **Skuce C**, **Taylor R**, **Abdellah Z**, **Zhao Y**, **MacArthur DG**, **Quail MA**, **Carter NP**, **Yang H**, **Tyler-Smith C**. 2009. Human Y chromosome base-substitution mutation rate measured by direct sequencing in a deep-rooting pedigree. Curr. Biol. **19**:1453–1457.

27. **Subramanian S**, **Lambert DM**. 2012. Selective constraints determine the time dependency of molecular rates for human nuclear genomes. Genome Biol. Evol. **4**:1127–1132.

28. **Ho SYW**, **Lanfear R**, **Bromham L**, **Phillips MJ**, **Soubrier J**, **Rodrigo AG**, **Cooper A**. 2011. Time-dependent rates of molecular evolution. Mol. Ecol. **20**:3087–3101.

29. **Duchêne S**, **Holmes EC**, **Ho SYW**. 2014. Analyses of evolutionary dynamics in viruses are hindered by a time-dependent bias in rate estimates. Proc. Biol. Sci. **281**:20140732.

30. **Soubrier J**, **Steel M**, **Lee MSY**, **Der Sarkissian C**, **Guindon S**, **Ho SYW**, **Cooper A**. 2012. The influence of rate heterogeneity among sites on the time dependence of molecular rates. Mol. Biol. Evol. **29**:3345–3358.

31. **Gignoux CR**, **Henn BM**, **Mountain JL**. 2011. Rapid, global demographic expansions after the origins of agriculture. Proc. Natl. Acad. Sci. U. S. A. **108**:6044–6049.

32. **Seo T-K**, **Thorne JL**, **Hasegawa M**, **Kishino H**. 2002. A viral sampling design for testing the molecular clock and for estimating evolutionary rates and divergence times.

Bioinformatics **18**:115–123.

33. **Switzer WM**, **Salemi M**, **Shanmugam V**, **Gao F**, **Cong M-E**, **Kuiken C**, **Bhullar V**, **Beer BE**, **Vallet D**, **Gautier-Hion A**, **Tooze Z**, **Villinger F**, **Holmes EC**, **Heneine W**. 2005. Ancient co-speciation of simian foamy viruses and primates. Nature **434**:376–380.

34. **Katzourakis A**, **Gifford RJ**, **Tristem M**, **Gilbert MTP**, **Pybus OG**. 2009. Macroevolution of complex retroviruses. Science **325**:1512.

35. **Katzourakis A**, **Aiewsakun P**, **Jia H**, **Wolfe ND**, **LeBreton M**, **Yoder AD**, **Switzer WM**. 2014. Discovery of prosimian and afrotherian foamy viruses and potential cross species transmissions amidst stable and ancient mammalian co-evolution. Retrovirology **11**:61.

36. **Bininda-Emonds ORP**, **Cardillo M**, **Jones KE**, **MacPhee RDE**, **Beck RMD**, **Grenyer R**, **Price SA**, **Vos RA**, **Gittleman JL**, **Purvis A**. 2007. The delayed rise of present-day mammals. Nature **446**:507–512.

37. **Stone AC**, **Battistuzzi FU**, **Kubatko LS**, **Perry GH**, **Trudeau E**, **Lin H**, **Kumar S**. 2010. More reliable estimates of divergence times in *Pan* using complete mtDNA sequences and accounting for population structure. Philos. Trans. R. Soc. Lond. B. Biol. Sci. **365**:3277–3288.

38. **Perelman P**, **Johnson WE**, **Roos C**, **Seuánez HN**, **Horvath JE**, **Moreira MAM**, **Kessing B**, **Pontius J**, **Roelke M**, **Rumpler Y**, **Schneider MPC**, **Silva A**, **O'Brien SJ**, **Pecon-Slattery J**. 2011. A molecular phylogeny of living primates. PLoS Genet. **7**:e1001342.

39. **Muniz CP**, **Troncoso LL**, **Moreira M a**, **Soares E a**, **Pissinatti A**, **Bonvicino CR**, **Seuánez HN**, **Sharma B**, **Jia H**, **Shankar A**, **Switzer WM**, **Santos AF**, **Soares M a**. 2013. Identification and characterization of highly divergent simian foamy viruses in a wide range of new world primates from Brazil. PLoS One **8**:e67568.

40. **Sanjuán R**. 2012. From molecular genetics to phylodynamics: evolutionary relevance of mutation rates across viruses. PLoS Pathog. **8**:e1002685.

41. **Jenkins GM**, **Rambaut A**, **Pybus OG**, **Holmes EC**. 2002. Rates of molecular evolution in RNA viruses: a quantitative phylogenetic analysis. J. Mol. Evol. **54**:156–165.

42. **Han G-Z**, **Worobey M**. 2012. An endogenous foamy virus in the aye-aye (*Daubentonia*

madagascariensis). J. Virol. **86**:7696–7698.

43. **Han G-Z**, **Worobey M**. 2014. Endogenous viral sequences from the Cape golden mole (*Chrysochloris asiatica*) reveal the presence of foamy viruses in all major placental mammal clades. PLoS One **9**:e97931.

44. **Duffy S**, **Shackelton LA**, **Holmes EC**. 2008. Rates of evolutionary change in viruses: patterns and determinants. Nat. Rev. Genet. **9**:267–276.

45. **Boyer PL**, **Stenbak CR**, **Hoberman D**, **Linial ML**, **Hughes SH**. 2007. *In vitro* fidelity of the prototype primate foamy virus (PFV) RT compared to HIV-1 RT. Virology **367**:253–264.

46. **Gärtner K**, **Wiktorowicz T**, **Park J**, **Mergia A**, **Rethwilm A**, **Scheller C**. 2009. Accuracy estimation of foamy virus genome copying. Retrovirology **6**:32.

47. **Schweizer M**, **Schleer H**, **Pietrek M**, **Liegibel J**, **Falcone V**, **Neumann-Haefelin D**. 1999. Genetic stability of foamy viruses: long-term study in an African green monkey population. J. Virol. **73**:9256–9265.

48. **Achong BG**, **Mansell PW**, **Epstein MA**, **Clifford P**. 1971. An unusual virus in cultures from a human nasopharyngeal carcinoma. J. Natl. Cancer Inst. **46**:299–307.

49. **Switzer WM**, **Bhullar V**, **Shanmugam V**, **Cong M-E**, **Parekh B**, **Lerche NW**, **Yee JL**, **Ely JJ**, **Boneva R**, **Chapman LE**, **Folks TM**, **Heneine W**. 2004. Frequent simian foamy virus infection in persons occupationally exposed to nonhuman primates. J. Virol. **78**:2780–2789.

50. **Meiering CD**, **Linial ML**. 2001. Historical perspective of foamy virus epidemiology and infection. Clin. Microbiol. Rev. **14**:165–176.

51. **Liu W**, **Worobey M**, **Li Y**, **Keele BF**, **Bibollet-Ruche F**, **Guo Y**, **Goepfert PA**, **Santiago ML**, **Ndjango J-BN**, **Neel C**, **Clifford SL**, **Sanz C**, **Kamenya S**, **Wilson ML**, **Pusey AE**, **Gross-Camp N**, **Boesch C**, **Smith V**, **Zamma K**, **Huffman MA**, **Mitani JC**, **Watts DP**, **Peeters M**, **Shaw GM**, **Switzer WM**, **Sharp PM**, **Hahn BH**. 2008. Molecular ecology and natural history of simian foamy virus infection in wild-living chimpanzees. PLoS Pathog. **4**:e1000097.

52. **Spiess A-N**, **Neumeyer N**. 2010. An evaluation of R2 as an inadequate measure for

nonlinear models in pharmacological and biochemical research: a Monte Carlo approach. BMC Pharmacol. **10**:6.

53. **Yoder AD**, **Yang Z**. 2000. Estimation of primate speciation dates using local molecular clocks. Mol. Biol. Evol. **17**:1081–1090.

54. **Drummond AJ**, **Suchard MA**. 2010. Bayesian random local clocks, or one rate to rule them all. BMC Biol. **8**:114.

55. **Drummond AJ**, **Ho SYW**, **Phillips MJ**, **Rambaut A**. 2006. Relaxed phylogenetics and dating with confidence. PLoS Biol. **4**:699–710.

56. **Duchêne S**, **Lanfear R**, **Ho SYW**. 2014. The impact of calibration and clock-model choice on molecular estimates of divergence times. Mol. Phylogenet. Evol. **78**:277–289.

57. **Pacheco B**, **Finzi A**, **McGee-Estrada K**, **Sodroski J**. 2010. Species-specific inhibition of foamy viruses from South American monkeys by New World Monkey TRIM5{alpha} proteins. J. Virol. **84**:4095–4099.

58. **Aiewsakun P**, **Katzourakis A**. 2015. Endogenous viruses: connecting recent and ancient viral evolution. Virology **479**–**480**:26–37.

59. **Wertheim JO**, **Kosakovsky Pond SL**. 2011. Purifying selection can obscure the ancient age of viral lineages. Mol. Biol. Evol. **28**:3355–3365.

60. **Streicker DG**, **Lemey P**, **Velasco-Villa A**, **Rupprecht CE**. 2012. Rates of viral evolution are linked to host geography in bat rabies. PLoS Pathog. **8**:e1002720.

61. **Soares P**, **Ermini L**, **Thomson N**, **Mormina M**, **Rito T**, **Röhl A**, **Salas A**, **Oppenheimer S**, **Macaulay V**, **Richards MB**. 2009. Correcting for purifying selection: an improved human mitochondrial molecular clock. Am. J. Hum. Genet. **84**:740–759.

62. **Mishmar D**, **Ruiz-Pesini E**, **Golik P**, **Macaulay V**, **Clark AG**, **Hosseini S**, **Brandon M**, **Easley K**, **Chen E**, **Brown MD**, **Sukernik RI**, **Olckers A**, **Wallace DC**. 2003. Natural selection shaped regional mtDNA variation in humans. Proc. Natl. Acad. Sci. U. S. A. **100**:171–176.

63. **Ruiz-Pesini E**, **Mishmar D**, **Brandon M**, **Procaccio V**, **Wallace DC**. 2004. Effects of purifying and adaptive selection on regional variation in human mtDNA. Science **303**:223–226.

64. **Firth C**, **Kitchen A**, **Shapiro B**, **Suchard MA**, **Holmes EC**, **Rambaut A**. 2010. Using time-structured data to estimate evolutionary rates of double-stranded DNA viruses. Mol. Biol. Evol. **27**:2038–2051.

65. **Clark AG**, **Whittam TS**. 1992. Sequencing errors and molecular evolutionary analysis. Mol. Biol. Evol. **9**:744–752.

66. **Ho SYW**, **Heupink TH**, **Rambaut A**, **Shapiro B**. 2007. Bayesian estimation of sequence damage in ancient DNA. Mol. Biol. Evol. **24**:1416–1422.

67. **Johnson PLF**, **Slatkin M**. 2008. Accounting for bias from sequencing error in population genetic estimates. Mol. Biol. Evol. **25**:199–206.

68. **Peterson GI**, **Masel J**. 2009. Quantitative prediction of molecular clock and ka/ks at short timescales. Mol. Biol. Evol. **26**:2595–2603.

69. **Sullivan J**, **Joyce P**. 2005. Model selection in phylogenetics. Annu. Rev. Ecol. Evol. Syst. **36**:445–466.

70. **Gatesy J**. 2007. A tenth crucial question regarding model use in phylogenetics. Trends Ecol. Evol. **22**:509–510.

71. **Emerson BC**. 2007. Alarm bells for the molecular clock? No support for Ho et al.'s model of time-dependent molecular rate estimates. Syst. Biol. **56**:337–345.

72. **Trobridge GD**. 2009. Foamy virus vectors for gene transfer. Expert Opin. Biol. Ther. **9**:1427–1436.

73. **Williams DA**. 2008. Foamy virus vectors come of age. Mol. Ther. **16**:635–636.

74. **Sharp PM**, **Li WH**. 1988. Understanding the origins of AIDS viruses. Nature **336**:315.

75. **Lemey P**, **Rambaut A**, **Pybus OG**. 2006. HIV evolutionary dynamics within and among hosts. AIDS Rev. **8**:125–140.

76. **Lemey P**, **Van Dooren S**, **Vandamme A-M**. 2005. Evolutionary dynamics of human retroviruses investigated through full-genome scanning. Mol. Biol. Evol. **22**:942–951.

77. **Smith GJD**, **Vijaykrishna D**, **Bahl J**, **Lycett SJ**, **Worobey M**, **Pybus OG**, **Ma SK**, **Cheung CL**, **Raghwani J**, **Bhatt S**, **Peiris JSM**, **Guan Y**, **Rambaut A**. 2009. Origins and evolutionary genomics of the 2009 swine-origin H1N1 influenza A epidemic. Nature **459**:1122–1125.

78. **Fourment M**, **Wood JT**, **Gibbs AJ**, **Gibbs MJ**. 2010. Evolutionary dynamics of the N1 neuraminidases of the main lineages of influenza A viruses. Mol. Phylogenet. Evol. **56**:526–535.

79. **Buonagurio DA**, **Nakada S**, **Parvin JD**, **Krystal M**, **Palese P**, **Fitch WM**. 1986. Evolution of human influenza A viruses over 50 years: rapid, uniform rate of change in NS gene. Science **232**:980–982.

80. **Gojobori T**, **Moriyama EN**, **Kimura M**. 1990. Molecular clock of viral evolution, and the neutral theory. Proc. Natl. Acad. Sci. U. S. A. **87**:10015–10018.

81. **Leitner T**, **Albert J**. 1999. The molecular clock of HIV-1 unveiled through analysis of a known transmission history. Proc. Natl. Acad. Sci. U. S. A. **96**:10752–10757.

82. **Simmonds P**, **Smith DB**. 1999. Structural constraints on RNA virus evolution. J. Virol. **73**:5787–5794.

83. **Holmes EC**. 2003. Molecular clocks and the puzzle of RNA virus origins. J. Virol. **77**:3893–3897.

84. **Drummond AJ**, **Nicholls GK**, **Rodrigo AG**, **Solomon W**. 2002. Estimating mutation parameters, population history and genealogy simultaneously from temporally spaced sequence data. Genetics **161**:1307–1320.

85. **Sauquet H**, **Ho SYW**, **Gandolfo MA**, **Jordan GJ**, **Wilf P**, **Cantrill DJ**, **Bayly MJ**, **Bromham L**, **Brown GK**, **Carpenter RJ**, **Lee DM**, **Murphy DJ**, **Sniderman JMK**, **Udovicic F**. 2012. Testing the impact of calibration on molecular divergence times using a fossil-rich group: the case of *Nothofagus* (Fagales). Syst. Biol. **61**:289–313.

86. **Lemey P**, **Lott M**, **Martin DP**, **Moulton V**. 2009. Identifying recombinants in human and primate immunodeficiency virus sequence alignments using quartet scanning. BMC Bioinformatics **10**:126.

87. **Abascal F**, **Zardoya R**, **Posada D**. 2005. ProtTest: selection of best-fit models of protein evolution. Bioinformatics **21**:2104–2105.

88. **Darriba D**, **Taboada GL**, **Doallo R**, **Posada D**. 2012. jModelTest 2: more models, new heuristics and parallel computing. Nat. Methods **9**:772–772.

89. **Ronquist F**, **Huelsenbeck JP**. 2003. MrBayes 3: Bayesian phylogenetic inference

under mixed models. Bioinformatics **19**:1572–1574.

90. **Tamura K**, **Peterson D**, **Peterson N**, **Stecher G**, **Nei M**, **Kumar S**. 2011. MEGA5: Molecular evolutionary genetics analysis using maximum likelihood, evolutionary distance, and maximum parsimony methods. Mol. Bio. Evol. **28**:2731–2739.

91. **Drummond AJ**, **Rambaut A**. 2007. BEAST: Bayesian evolutionary analysis by sampling trees. BMC Evol. Biol. **7**:214.

92. **Rambaut A**, **Drummond AJ**. 2007. Tracer V1.5. Available from http//beast.bio.ed.ac.uk/Tracer.

93. **MATLAB**. 2010. MATLAB and Statistics Toolbox Release R2012a. The MathWorks Inc., Massachusetts.

94. **Jackson DA**, **Somers KM**. 1991. The spectre of “spurious” correlations. Oecologia **86**:147–151.

Chapter four: Time-dependent rate phenomenon in viruses

This chapter was originally published as:

Aiewsakun P, Katzourakis A. 2016. Time-dependent rate phenomenon in viruses. J. Virol. **90**: 7184-7195.

PA performed the analysis and wrote the paper. AK provided editorial guidance and supervisory advice.

4.1 Abstract

Among the most fundamental questions in viral evolutionary biology are how fast viruses evolve and how evolutionary rates differ among viruses and fluctuate through time. Traditionally, viruses are loosely classed into two groups: slow-evolving DNA viruses and fast-evolving RNA viruses. As viral evolutionary rate estimates become more available, it appears that the rates are negatively correlated with the measurement timescales and that the boundary between the rates of DNA and RNA viruses might not be as clear as previously thought. In this study, we collected 396 viral evolutionary rate estimates across almost all viral genome types and replication strategies, and we examined their rate dynamics. We showed that the time-dependent rate phenomenon exists across multiple levels of viral taxonomy, from the Baltimore classification viral groups to genera. We also showed that, by taking the rate decay dynamics into account, a clear division between the rates of DNA and RNA viruses as well as reverse-transcribing viruses could be recovered. Surprisingly, despite large differences in their biology, our analyses suggested that the rate decay speed is independent of viral types and thus might be useful for better estimation of the evolutionary time scale of any virus. To illustrate this, we used our model to reestimate the evolutionary timescales of extant lentiviruses, which were previously suggested to be very young by standard phylogenetic analyses. Our analyses suggested that these viruses are millions of years old, in agreement with paleovirological evidence, and therefore, for the first time, reconciled molecular analyses of ancient and extant viruses.

4.2 Introduction

An accurate and precise knowledge of the rate of viral evolution is central to the reconstruction of viral natural history, which is necessary for the calculation of many evolutionary parameters, from viral age estimates to population size. Generally, viruses are loosely classed into two groups according to their rates of evolution: “slow-evolving” and “fast-evolving” viruses. DNA viruses, especially double-stranded DNA (dsDNA) viruses, are traditionally thought of as slow-evolving viruses. To estimate molecular evolutionary rates, the absolute timescales for the observed genetic differences are required, and these can be derived from the divergence dates and/or sampling dates of the study subjects. Many dsDNA viruses have been shown to have an extremely stable cospeciation history with their hosts, and therefore their divergence dates can be directly inferred from those of their hosts. On the basis of these observations, their rates have been estimated to be of the order of 10^{-7} to 10^{-9} nucleotide substitutions per site per year (s/n/y) (1–5), comparable to those of their hosts (6, 7).

RNA viruses, on the other hand, are typically regarded as fast-evolving viruses. RNA viruses are generally characterised by frequent cross-species transmissions in nature; as a result, it is often difficult to calibrate their evolutionary rates using host evolutionary timescales. Their rates are thus often calculated by using molecular sequences collected at different time points (heterochronous molecular data sets). In this case, the differences among sampling times provide the timescales for the observed genetic divergence. Based on these analyses, their rates are commonly estimated to be between 10^{-2} and 10^{-5} s/n/y (8–11), 2 to 7 orders of magnitude higher than the typical rates of dsDNA viruses.

This conventional concept of a dichotomy between the rates of DNA and RNA viral evolution has recently been challenged, however, and it seems that the boundary between the rates of DNA and RNA viruses might not be as clear as previously thought (11–14). For example, analyses of heterochronous molecular data sets of dsDNA and single-stranded DNA (ssDNA) viruses have revealed that DNA viruses are in fact capable of evolving very rapidly over short timescales, with rates ranging between 10^{-3} and 10^{-6} s/n/y (15–19), comparable to the

established rates of many RNA viruses. On the other hand, when the cospeciation assumption is applicable to RNA viruses, such as deltaretroviruses, hantaviruses, and foamy viruses, analyses suggest that their long-term rates of evolution are extremely low, estimated to be in the range of 10^{-7} to 10^{-8} s/n/y (20–23), comparable to those of dsDNA viruses. Paleovirological analyses have also shown that many ancient endogenous viruses related to RNA viruses exhibit high similarity to their modern-day counterparts despite being millions of years old (24). This provides independent evidence indicating that RNA viruses can indeed evolve very slowly over geological timescales (24). If one naïvely combines all the rate estimates, the ranges of the rate estimates for both DNA and RNA viral evolution would appear to be extremely wide, spanning 10^{-3} to 10^{-9} s/n/y for DNA viruses and 10^{-2} to 10^{-8} s/n/y for RNA viruses, largely overlapping with one another. As a result, there is an emerging consensus that there is no strict division between the evolutionary rates of DNA and RNA viruses (13, 14).

As viral evolutionary rate estimates become more available, it is becoming increasingly clear that viral evolutionary rates appear to vary over time, continuously decreasing with the timescale of rate measurement (25–27). Many hypotheses have been proposed to explain this *time-dependent rate phenomenon* (TDRP), including temporal changes in selection pressure and/or viral biology, as well as the facts that short-term rates are methodologically prone to overestimation and long-term rates tend to be underestimated (for a review, see reference 25). This phenomenon may at least partly explain the observed large variation and overlap of viral evolutionary rate estimates.

Another consequence of the TDRP is that naïvely transferring the rate estimates over different timescales for evolutionary inference can severely bias the outcome (25–27). The best illustration of this is perhaps the severe underestimation of the evolutionary timescales of extant viruses by current standard phylogenetic tools. For example, while paleovirological analyses unequivocally show that simian immunodeficiency viruses (SIVs) are millions of years old (28, 29), all previous standard phylogenetic analyses, which do not account for the TDRP, have suggested that they are young, with the most recent common ancestor (MRCA) dating

to less than a million years ago (8, 30–32). One pragmatic approach to solving this problem is to use models describing the empirical relationship between the rate estimates and their measurement timescales to correct for the TDRP effects in evolutionary analyses (27, 33). Our study of foamy viruses has shown that a power law rate decay model can describe the TDRP very well empirically, and thus, it may be useful as a tool for correcting for the effects of the TDRP (27).

In this study, we collected 396 viral nucleotide substitution rates across almost all viral molecular features and replication strategies, and we examined their TDRP dynamics at various taxonomical levels by using the power law rate decay model as the basis of our investigation. We also examined how the rate dynamics differ among viruses and reexamined the concept of fast- and slow-evolving viruses. Last, we demonstrated the use of our TDRP model by estimating the evolutionary timescale of extant lentiviruses, which has always been severely underestimated by standard phylogenetic analyses.

4.3 Results

4.3.1 TDRP among Baltimore classification viral groups

A total of 396 viral nucleotide substitution rate estimates were collected from 133 studies for TDRP analyses (**Figure 4.1A**, and **Table S4.1** in **Electronic Appendix**). The rates were sampled from all Baltimore classification viral groups except group III double-stranded RNA viruses, since to our knowledge, long-term rate estimates (computed over timescales of >1,000 years) are not currently available for this group. We demonstrated previously that a simple power law curve can empirically describe the TDRP well for foamy viruses (27). We therefore use this model as the basis for our investigation. To do so, we log-transformed (base 10) the rate estimates($\bar{r}$) and their associated measurement timescales (t), and, in turn, fitted a linear model to them for statistical analyses; i.e., $\log(\bar{r}) = \alpha + \beta \log(t)$ (equivalent to the equation $\bar{r} = 10^{\alpha} t^{\beta}$), where α is an intercept and β is a slope.

We noted that these rate estimates are not phylogenetically independent, and this may bias our investigation. To address this issue, we constructed 1,000 pseudoreplicated data subsets, each of which contains between 84 and 95 phylogenetically independent rate estimates (see **Materials and Methods**), and used them in our analyses. The summary of the number of rate estimates for each data subset is in **Table 4.1.**

As shown previously (26, 27), our analyses suggested that viral evolutionary rate estimates are negatively correlated with their measurement timescales (*P* value < 0.05 in 1,000/1,000 subanalyses; combined *P* value < 0.001, **Figure 4.1B**). Surprisingly, the slopes of the rate decay (β estimates) do not differ significantly among viral groups (*P* value < 0.05 in 27/1,000 subanalyses; combined *P* value = 0.133) and are estimated to be ∼−0.65 (95% highest probability density (95% HPD) = −0.72–−0.57)); i.e., for every 3−fold increase in the measurement timescale, the value of the rate estimate decreases by approximately half (∼2.03×; 95% HPD = 1.86×–2.21×).

Our analyses showed that, overall, the power-law rate decay model can describe the data well (median R-squared values (R^2) = 0.89; 95% HPD = 0.85–0.93). We found that the rate estimates differ significantly among viral groups even when one adjusts for the rate decay dynamics (*P* value < 0.05 in 1,000/1,000 subanalyses; combined *P* value < 0.001). To determine the rate differences among viruses, the intercepts of the models (α estimates), which represent rate estimates that are controlled for the timescale of measurement (i.e., 10^{α} is equivalent to a rate that is estimated over a period of 1 year), were compared. The results show the following ranking of rates of viruses (expressed in units of s/n/y), from lowest to highest: group I dsDNA viruses (4.36×10^{-4}; 95% HPD = 1.24×10^{-4}–1.48×10^{-3}), group VII reverse-transcribing DNA (RT-DNA) viruses (8.12×10^{-4}; 95% HPD = 7.30×10^{-5}–9.82×10^{-3}) and group VI reverse-transcribing RNA (RT-RNA) viruses (1.12×10^{-3}; 95% HPD = 2.71×10^{-4}–4.50×10^{-3}) (approximately equal), group II ssDNA viruses (1.75×10^{-3}; 95% HPD = 7.37×10^{-4}–4.09×10^{-3}), group V negative ssRNA ((−)ssRNA) viruses (6.87×10^{-3}; 95% HPD =3.11×10^{-3}–1.46×10^{-2}) and group IV positive ssRNA ((+)ssRNA) viruses (8.63×10^{-3}; 95%

Table 4.1 Summary of the number of viral evolutionary rate estimates used in this study.

Viral group	Number of evolutionary rate estimates					
	Entire data set			Each data subset		
	Short-term rate	Long-term rate	Total	Short-term rate	Long-term rate	Total
Group I/dsDNA virus	9	12	21	3 to 8	1 to 5	7 to 10
Group II/ssDNA virus	45	2	47	9 to 12	1 to 2	11 to 13
Group IV/(+)ssRNA virus	122	1	123	45	1	46
Group V/(−)ssRNA virus	103	3	106	12 to 16	1	13 to 17
Group VI/RT-RNA virus	69	16	85	1 to 4	1 to 6	5 to 7
Group VII/RT-DNA virus	11	3	14	1	1	2
Total	359	37	396	71 to 85	6 to 15	84 to 95

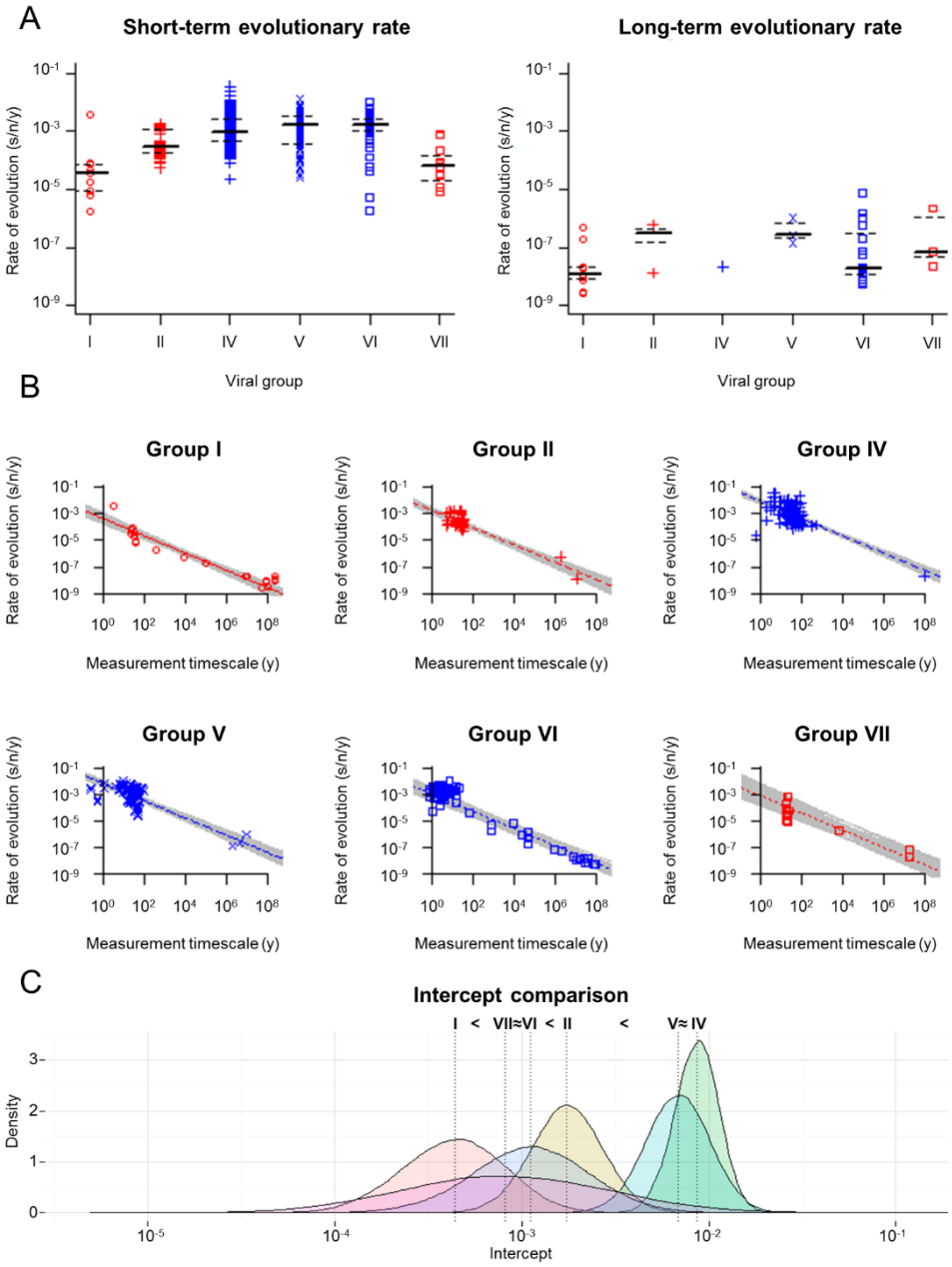


Figure 4.1 Time-dependent rate phenomenon among viral groups. (**A**) A total of 359 viral short-term rate estimates (calculated over timescales of 0.16 to 760 years) (left), and 37 long-term rate estimates (calculated over timescales of 6,600 to 2.28 × 10^8 years) (right) were collected from 133 publications, including 21 rate estimates for group I dsDNA viruses, 47 for group II ssDNA viruses, 123 for group IV (+)ssRNA viruses, 106 for group V (−)ssRNA viruses, 85 for group VI RT-RNA viruses, and 14 for group VII RT-DNA viruses (**Table S4.1** in **Electronic Appendix**). Lines indicating the upper quartile (top dashed line), median (central solid line), and lower quartile (bottom dashed line) were added to each viral group to aid visualising the point's density. (**B**) Viral evolutionary rate estimates are negatively correlated with their measurement timescales. Grey lines represent 1,000 individual best-fit models, where the slopes are the same across all viral groups. Red or blue lines represent models that are parameterised by median parameter estimates; open red circles and solid line, group I dsDNA viruses; red plus signs and dashed line, group II ssDNA viruses; blue plus signs and dashed line, group IV (+)ssRNA viruses; blue crosses and long-dashed line, group V (−)ssRNA viruses; open blue squares and dotted line, group VI RT-RNA viruses; open red squares and dotted line, group VII RT-DNA viruses. (**C**) Complete pairwise comparisons of the intercepts of the rate decay curves (representing rate estimates controlled for a 1-year timescale of rate measurement). Vertical dotted lines indicate median estimates. The differences among intercepts were evaluated at a significance level of 0.05.

HPD = 4.97×10^{-3}–1.45×10^{-2}) (approximately equal) (**Figure 4.1C**).

4.3.2 TDRP within viral genera

We noted that although the TDRP exists at the level of the Baltimore classification viral groups, it is possible that this might be merely an artefact, resulting from our tendency to use fast-evolving and slow-evolving viruses for short-term and long-term viral evolutionary studies, respectively. That is, it is possible that viral evolutionary rate estimates may in fact be independent of the rate measurement timescale at lower taxonomic levels, with some lineages evolving more slowly or more rapidly than others. However, due to our tendency to use fast- and slow-evolving viruses for short-term and long-term viral evolutionary studies, respectively, the TDRP may emerge as an artefact at the level of the Baltimore classification viral groups, in which all of these different rapidly and slowly evolving viruses were analysed together.

To examine this possibility, we investigated the TDRP at the level of viral genera to see whether the rate decay pattern still holds or not. We performed the same TDRP analyses on seven viral genera sampled across six viral groups, including the group I dsDNA virus genera *Simplexvirus* and *Varicellovirus*, the group II ssDNA virus genus *Mastrevirus*, the group IV (+)ssRNA virus genus *Tobamovirus*, the group V (−)ssRNA virus genus *Hantavirus*, the group VI RT−RNA virus *Deltaretrovirus*, and the group VII RT−DNA virus genus *Avihepadnavirus*. These genera were chosen because the timescales of their rate measurement cover the longest time span (in term of orders of magnitude). The rates of simplexviruses and varicelloviruses were combined for analysis because there were not enough data points to be analysed individually. Note that these rate estimates are not phylogenetically independent, and thus, the results should be interpreted with caution.

Our analyses show that the model can describe the data well (R^2 = 0.95), and that the TDRP still holds even at the level of viral genera (*P* value < 0.001). The slopes of the rate decay curves do not differ significantly among the genera (*P* value = 0.89) and are estimated to be∼−0.68 (95% HPD = −0.74–−0.62, **Figure 4.2A**). Remarkably, this slope estimate is

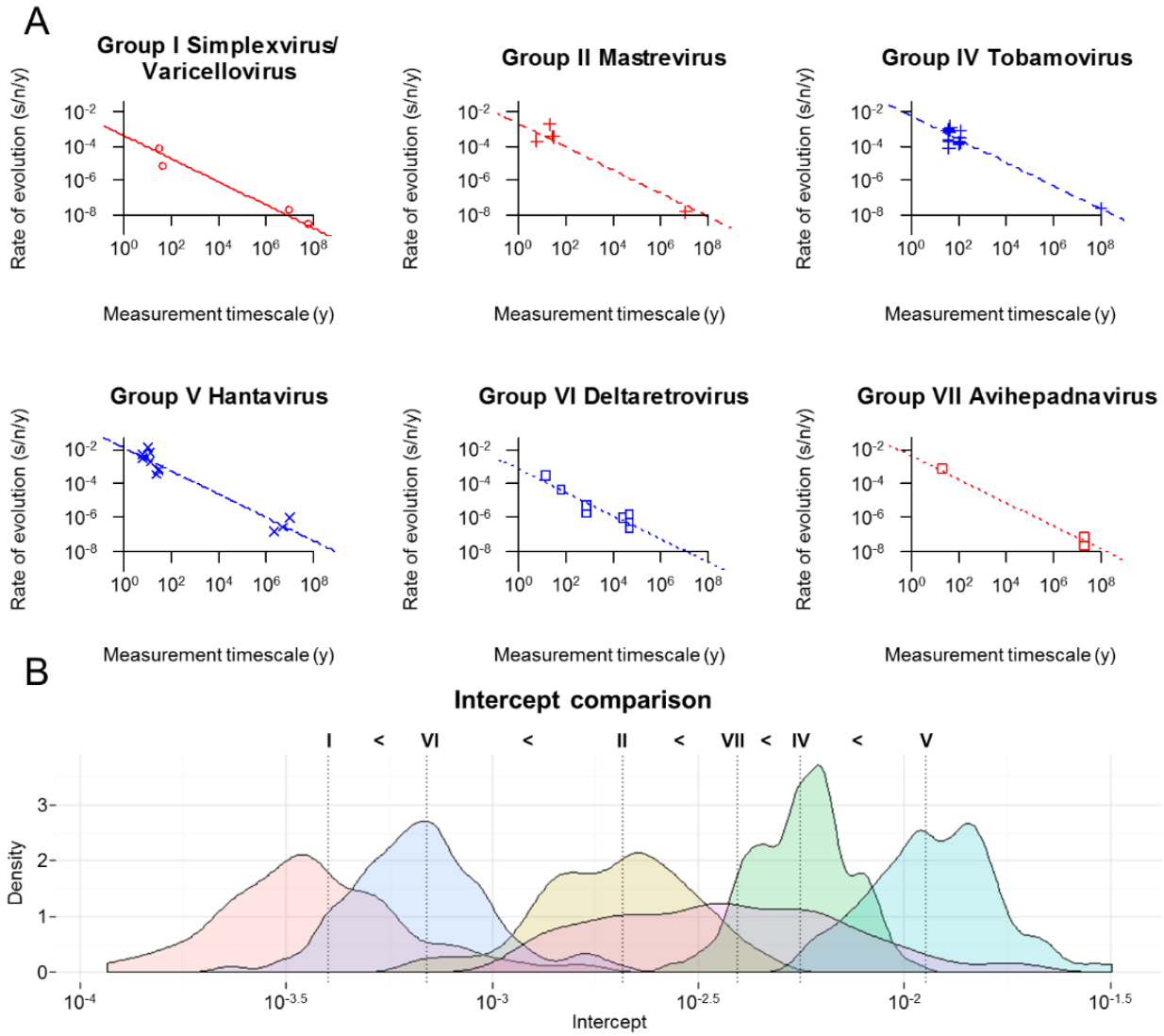


Figure 4.2 Time-dependent rate phenomenon among viral genera. (**A**) The lines represent the best fit models; open red circles and solid line, group I genera *Simplexvirus* and *Varicellovirus*; red plus signs and dashed line, group II genus *Mastrevirus*; blue plus signs and dashed line, group IV genus *Tobamovirus*; blue crosses and long-dashed line, group V genus *Hantavirus*; open blue squares and dotted line, group VI genus *Deltaretrovirus*; open red squares and dotted line, group VII genus *Avihepadnavirus*. (**B**) Complete pairwise comparisons of the intercepts of the rate decay curves (representing rate estimates controlled for a 1-year timescale of rate measurement). Vertical dotted lines indicate median estimates. The differences among intercepts were evaluated at a significance level of 0.05.

strikingly similar to the slope estimate we obtained from the TDRP analyses at the level of viral groups (∼−0.65; 95% HPD = −0.72–−0.57). These findings further support our observation that the rate decay slopes do not differ significantly among viruses and suggest, simultaneously, that the TDRP at the level of viral groups is not an artefact emerging from systematic biases in our tendency to use fast- and slow-evolving viruses for short-term and long-term evolutionary studies, respectively. Moreover, this finding also suggests that the problem of phylogenetic nonindependency, although present in our data set, does not greatly affect the analyses. By comparing the intercepts of the models, we found that different viral genera evolve at different rates (*P* value < 0.001), which may be ranked as follows (expressed in units of s/n/y), from lowest to highest: the group I genus *Simplexvirus*/*Varicellovirus* (3.99×10^{-4}; 95% HPD = 1.28×10^{-4}–1.24×10^{-3}), the group VI genus *Deltaretrovirus* (6.94×10^{-4}; 95% HPD = 3.05×10^{-4}–1.58×10^{-3}), the group II genus *Mastrevirus* (2.07×10^{-3}; 95% HPD = 8.93×10^{-4}–4.82×10^{-3}), the group VII genus *Avihepadnavirus* (3.92×10^{-3}; 95% HPD = 1.04×10^{-3}–1.47×10^{-2}), the group IV genus *Tobamovirus* (5.57×10^{-3}; 95% HPD = 3.26×10^{-3}–9.52×10^{-3}), and the group V genus *Hantavirus* (1.13×10^{-2}; 95% HPD = 5.88×10^{-3}–2.16×10^{-2}) (**Figure 4.2B**). These results are largely consistent with the results obtained from the TDRP analyses at the level of viral groups.

4.3.3 Sensitivity analyses

Despite the control of the rate measurement methods, it is likely that some of the rate estimates may still be erroneous, especially the short-term rate estimates, computed from data sets that do not contain sufficient temporal structure. In such a scenario, it is expected that the number of substitutions (s) and t are independent of one another (i.e., $s(t) = c$; where c is a constant), and thus, the relationship between $\bar{r}$ and t is expected to be as follows: $\bar{r} = s/t = c/t$; i.e., $\log(\bar{r}) = \log(c) - \log(t)$. Therefore, the presence of erroneous short-term rate estimates can systematically bias the rate-decay slope toward −1.

To evaluate the effects of these erroneous rate estimates on our analyses, we simulated 50 short-term and 5 long-term substitution rate estimates under the TDRP model and replaced

20%, 40%, 60%, 80%, 100% of the short-term rates with erroneous rates, calculated from noncorrelated substitution numbers and timescales (see **Materials and Methods**). The intercepts and slopes of the rate decay curves were compared to those obtained from the control simulation, in which all of the rates were simulated under the TDRP model.

Our analyses reveal that the intercepts obtained from all simulations largely overlap (**Figure 4.3, top right**). The intercept estimate (expressed in units of s/n/y) was 7.02×10^{-3} (95% HPD = 1.08×10^{-3}–2.41×10^{-2}) with 0% erroneous short-term rate estimates, 6.09×10^{-3} (95% HPD = 1.12×10^{-3}–3.25×10^{-2}) with 20%, 6.05×10^{-3} (95% HPD = 8.96×10^{-4}–2.89×10^{-2}) with 40%, 7.03×10^{-3} (95% HPD = 1.43×10^{-3}–4.01×10^{-2}) with 60%, 8.43×10^{-3} (95% HPD = 1.68×10^{-3}–4.19×10^{-2}) with 80%, and 7.65×10^{-3} (95% HPD = 1.61×10^{-3}–4.46×10^{-2}) with 100%. In fact, we could not detect biases in the intercept estimates even when all of the short-term rates were erroneous. When the intercept estimates were compared against the controlled intercept by the Wilcoxon test, *P* values were > 0.999 with 20% erroneous short-term rate estimates, > 0.999 with 40%, 0.632 with 60%, 0.705 with 80%, and 0.428 with 100%. We estimated the slopes to be −0.65 (95% HPD = −0.71–−0.55) with 0% erroneous short-term rate estimates, −0.66 (95% HPD = −0.74–−0.59) with 20%, −0.68 (95% HPD = −0.75–−0.60) with 40%, −0.69 (95% HPD = −0.76–−0.62) with 60%, −0.71 (95% HPD = −0.78–−0.63) with 80%, and −0.72 (95% HPD = −0.81–−0.64) with 100% (**Figure 4.3**, **top left**). The bias showed only when there were >20% erroneous rate estimates (*P* values were < 0.001 with 40% to 100% erroneous rate estimates); otherwise, it could not be detected (*P* value was 0.142 with 20% erroneous rate estimates). Furthermore, even when the bias could be detected, the effect was very small. For example, with 60% erroneous rate estimates, the slope only changes by ∼0.04 with respect to the slope obtained from the control simulation, changing from ∼−0.65 (95% HPD = −0.71–−0.55) to ∼−0.69 (95% HPD = −0.76–−0.62). Even in the case of 100% erroneous short-term rates, the biased slope distribution still largely overlaps with that obtained from the control simulation. Combined, these results suggest that our TDRP analyses are extremely robust against erroneous rates derived from data sets containing insufficient temporal structure.

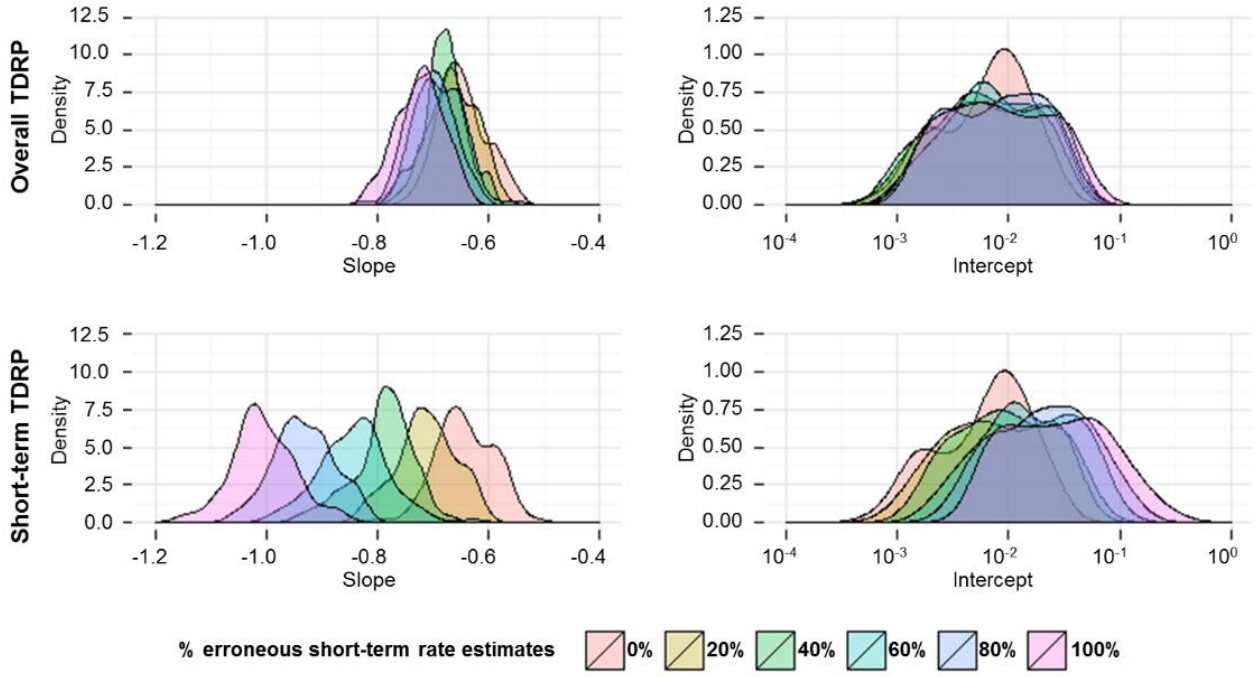


Figure 4.3 Effects of erroneous short-term rate estimates derived from noncorrelated substitution numbers and timescales of rate measurement on the rate decay curve inference. Simulations were used to examine how the presence of erroneous short-term rate estimates may bias the time-dependent rate phenomenon (TDRP) analyses, assuming a nonhomogeneous Poisson evolutionary process and a power law rate decay curve. Left and right graphs show slope and intercept estimates, respectively, of the rate decay curve, computed in the presence of 20%, 40%, 60%, 80%, and 100% erroneous short-term rate estimates. Top and bottom graphs show how the erroneous short-term rates affect overall and short-term TDRP curves, respectively. The intercept and slope of the rate decay curve obtained from control simulations (0% erroneous rate estimates) were used as controls.

We also performed sensitivity analyses exclusively on short-term rate estimates. We found that in this case, erroneous short-term rate estimates can severely bias the TDRP curve. As expected, as the number of erroneous rate estimates increased, the slope was increasingly biased toward −1 (**Figure 4.3**, **bottom left**). Slope estimates were −0.71 (95% HPD = −0.82–−0.62) with 20% erroneous short-term rate estimates, −0.78 (95% HPD = −0.92–−0.71) with 40%, −0.84 (95% HPD = −0.97–−0.73) with 60%, −0.94 (95% HPD = −1.02–−0.83) with 80%, and −1.01 (95% HPD = −1.10–−0.87) with 100%, all of which are significantly different from the controlled slope (−0.65; 95% HPD = −0.73–−0.55; *P* values < 0.001 for all comparisons). The intercept also became more and more overestimated (**Figure 4.3**, **bottom right**). The intercept estimate (expressed in units of s/n/y) was 6.88×10^{-3} (95% HPD = 9.56×10^{-4}–2.48×10^{-2}) with 0% erroneous short-term rates, 7.34×10^{-3} (95% HPD = 1.37×10^{-3}–4.10×10^{-2}) with 20%, 9.24×10^{-3} (95% HPD = 2.14×10^{-3}–6.23×10^{-2}) with 40%, 1.36×10^{-2} (95% HPD = 2.39×10^{-3}–6.98×10^{-2}) with 60%, 2.12×10^{-2} (95% HPD = 3.91×10^{-3}–1.14×10^{-1}) with 80%, and 2.86×10^{-2} (95% HPD = 5.37×10^{-3}–1.63×10^{-1}) with 100%. When the intercept estimates were compared with the controlled intercept, *P* values were 0.350 with 20% erroneous rate estimates, 0.015 with 40%, and <0.001 with 60% to 100%.

4.3.4 Short-term TDRP

We also performed TDRP analyses exclusively on the short-term rate estimates in order to examine whether or not the pattern we observed in the overall TDRP analyses still holds. Groups I, VI, and VII were not included, however, because there were only strictly less than 10 phylogenetically independent short-term rates for these groups in each data subset (**Table 4.1**). Again, our analyses showed that the short-term rates are negatively correlated with their measurement timescales (*P* value < 0.05 in 840/1,000 subanalyses; combined *P* value < 0.001). The slope was estimated to be −0.49, slightly greater than the slope obtained from the overall TDRP analyses (−0.65; 95% HPD = −0.72–−0.57), with a much greater uncertainty (95% HPD = −0.93–−0.06). Consistent with this is a low R^2 score of 0.26 (95% HPD = 0.10–0.41). This is expected, however, since the variance of the rate estimates is extremely high with respect to the narrow range of the rate measurement timescales (**Figure 4.4A**). In fact, it

is very striking that the rate decay pattern could still be observed even over a timescale of < 1,000 years. Our analyses show that the short-term rate decay slopes do not differ significantly among viral groups (*P* value < 0.05 in 24/1,000 subanalyses; combined *P* value = 0.656). The temporally adjusted rate estimates (expressed in units of s/n/y) are also clearly different among viruses (*P* value < 0.05 in 963/1,000 subanalyses; combined *P* value <0.001); those of group II ssDNA viruses (1.16×10^{-3}; 95% HPD = 2.35×10^{-4}–5.36×10^{-3}) are less than those of group V (−)ssRNA viruses (4.34×10^{-3}; 95% HPD = 1.02×10^{-3}–2.02×10^{-2}), which are approximately equal to those of group IV (+)ssRNA viruses (5.14×10^{-3}; 95% HPD = 1.03×10^{-3}–2.54×10^{-2}) (**Figure 4.4B**). These findings are consistent with the results obtained from the overall TDRP analyses.

4.3.5 Case study: reestimating the time scale of evolution of lentiviruses

Lentiviruses are a group of medically important viruses whose evolutionary history has been studied extensively. Nevertheless, there is still little consensus regarding when lentiviruses originated. In fact, estimating the origin of lentiviruses from extant viral molecular sequences has long been one of the most difficult challenges in lentiviral evolutionary biology. For example, while paleovirological analyses strongly support an age of millions of years for SIVs (28, 29), all previous standard molecular analyses of modern SIVs consistently underestimated their age to be at most tens of thousands of years old (8, 30–32). This is likely due to the fact that the TDRP was not accounted for in previous molecular analyses. Indeed, the TDRP has been noticed in lentiviruses(26), but their TDRP-corrected age estimate has not been inferred. Here we illustrate the use of our TDRP model by reestimating the time scale of lentiviruses. We first estimated a phylogeny of extant lentiviruses from their integrase coding region under the Bayesian phylogenetic framework and a strict molecular clock assumption (**Figure 4.5**). The rate was fixed to 1, and thus, the node heights were expressed in units of substitutions per site. We then simply converted the units of branch lengths from substitutions per site to times by using our TDRP model. The rate decay slope was sampled from the slope we obtained from the overall TDRP analyses. The lentivirus-specific intercept was calibrated by using the separation date of African mainland drill SIVs (SIVdrl) and Bioko drill SIVs (SIVdrl-

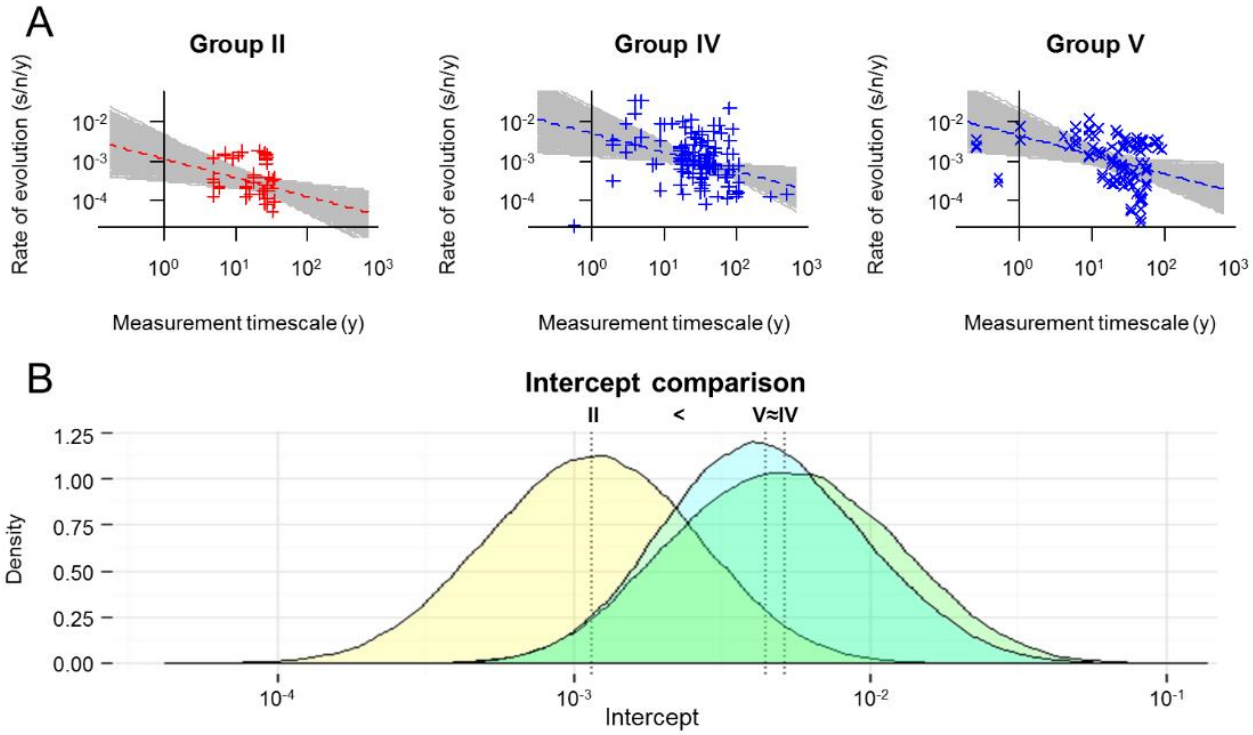


Figure 4.4 Short-term time-dependent rate phenomenon among viral groups. (**A**) Grey lines represent 1,000 individual best-fit models, where the slopes are the same across all viral groups. Red or blue lines represent models that are parameterised by median parameter estimates: red plus signs and dashed line, group II ssDNA viruses; blue plus signs and dashed line, group IV (+)ssRNA viruses; blue crosses and long-dashed line, group V (−)ssRNA viruses. (**B**) Complete pairwise comparisons of the intercepts of the rate decay curves (representing rate estimates controlled for a 1-year timescale of rate measurement). Vertical dotted lines indicate median estimates. The differences among intercepts were evaluated at a significance level of 0.05.

Bioko), which was >10,000 to 11,000 years ago (32) (see **Materials and Methods**).

The intercept of the lentivirus-specific TDRP curve was calculated to be ∼−2.20 (95% HPD: −2.54–−1.83). This result implies that the short-term rate of lentiviruses (calculated over a 1-year period) is $\sim 6.34 \times 10^{-3}$ (95% HPD = 2.86×10^{-3}–1.48×10^{-2}) s/n/y, comparable to the established short-term rate of lentiviral evolution (10, 34–37). We then used this model to compute the timescales of other nodes (**Figure 4.5**). Note that since the curve is calibrated by a geographical separation date, which likely postdates the actual SIVdrl/SIVdrl-Bioko speciation date (32), our lentiviral evolutionary time scale estimates should be interpreted as lower bound estimates.

Our analyses suggest that lentiviruses as a whole have an ancient origin and are >15.9 (95% HPD = 1.03–96.9) million years (Myr) old. We also estimated SIVs to be >2.57 (95% HPD = 0.35–10.9) Myr old, feline immunodeficiency viruses (FIVs) to be >1.63 (95% HPD = 0.27–6.50) Myr old, and the MRCA of SIVs and FIVs to be >10.1 (95% HPD = 1.07–55.8) Myr old. The uncertainties of the age estimates are large, however, spanning ∼2 orders of magnitude. We believe that this is caused partly by the fact that we used a short nucleotide alignment in our analyses (integrase (420 nucleotides)). Our alignment was limited to the integrase region, since this is the only region currently available for SIVdrl-Bioko—the only lentivirus that provides the timescale calibration information in our analyses. Note that, in our case, expanding the alignment would not alleviate the problem. This is because the SIVdrl-Bioko sequence would still be mostly blank, and thus, the uncertainty of its substitution estimate, with which the TDRP model was calibrated, would still be high. Furthermore, expanding the alignment would likely bias the mean estimates of the timescales as well. A longer sequence of SIVdrl-Bioko is required to overcome this particular problem and, in turn, to increase the precision of the date estimates.

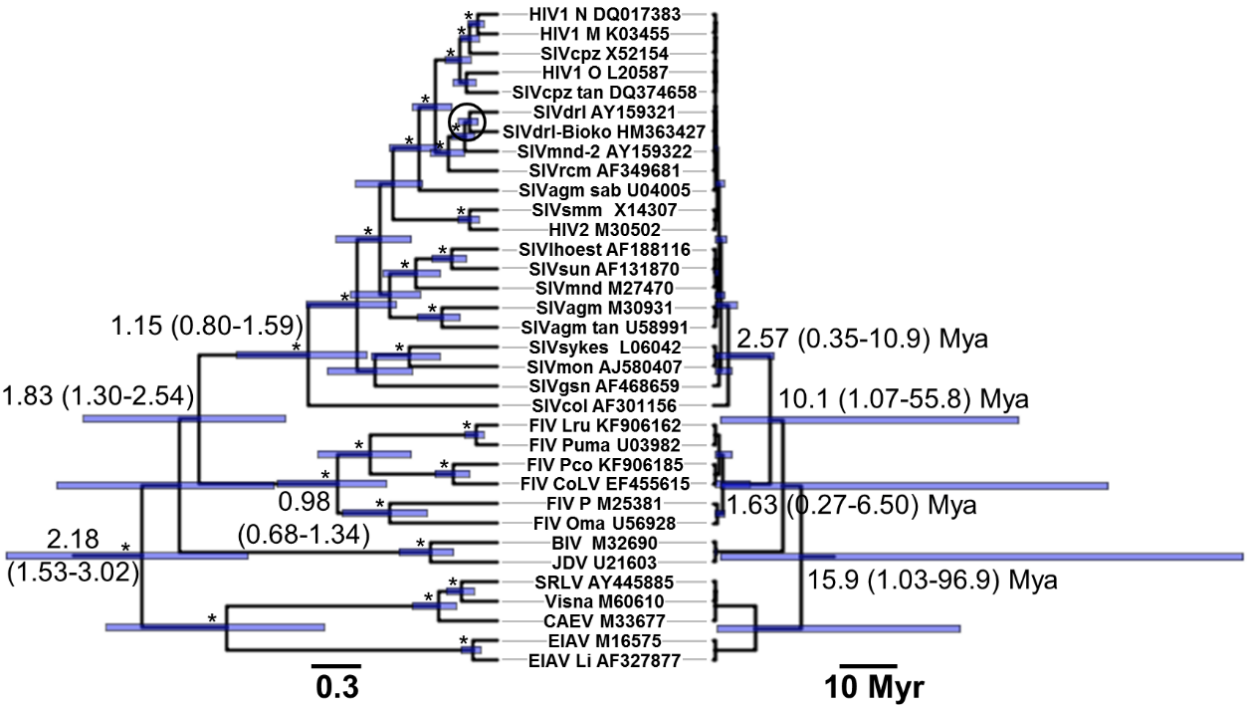


Figure 4.5 Lentivirus phylogeny and evolutionary timescale. (**Left**) Maximum clade credibility phylogeny of lentiviruses. The tree was estimated in the Bayesian phylogenetic framework under a strict clock assumption with a fixed rate of 1. The branch lengths and scale bar are in units of substitutions per site. The numbers on nodes are node heights in units of substitutions per site. The corresponding 95% highest posterior density intervals (HPDs) are given in parentheses. Asterisks indicate nodes with posterior support of >0.85. The split between SIVdrls and SIVdrl-Bioko (>10,000 to 11,000 years ago), which was used to calibrate the lentivirus-specific TDRP model, is circled. (**Right**) Time-calibrated lentivirus tree. The numbers on nodes are node heights in units of years before the present, inferred by using our TDRP model. Corresponding 95% HPDs are given in parentheses. The branch lengths and scale bar are also in units of millions of years (Myr).

4.4 Discussion

The last decade or so has seen several studies and reviews reporting and discussing the discrepancies among viral rate estimates that are calculated over different time frames (24, 26, 27, 32, 38, 39). The recurring theme is that, as can be seen in **Figure 4.1A**, viral short-term rate estimates (typically calculated from heterochronous molecular data sets) tend to be much greater than long-term rate estimates (usually inferred based on the virus-host cospeciation assumption or on geographical separation dates). Indeed, it has been observed that viral evolutionary rate estimates are systematically negatively correlated with the time scale of rate estimation, continuously decreasing as the measurement time scale increases (25–27). Duchêne *et al.* (26) demonstrated this phenomenon in both RNA and DNA viruses by fitting separate regressions onto them *a priori*, and analyses of foamy viruses indicate that this phenomenon can be empirically described very well by a power law curve (27). In this study, we used the power law model as our basis for further investigation of this *time-dependent rate phenomenon* (TDRP) at a finer taxonomic scale, across different Baltimore classification viral groups and genera. We also statistically compared how the TDRP dynamics of different viruses differ from one another.

We found that the TDRP holds both at the level of viral groups (**Figure 4.1**) and at the level of genera (**Figure 4.2**) and that the power law models can describe the phenomenon well, as indicated by high R^2 scores. Our sensitivity analyses show that overall, the inference of the rate decay curve is extremely robust against erroneous short-term rate estimates, derived from data sets with weak temporal signals (**Figure 4.3**, **top**). This robustness likely comes from the fact that the sampling times of the short-term rates are relatively short (<1,000 years) compared to the timescales of the overall TDRP examination (100 Myr). This result also suggests that the difference between short-term and long-term rate estimates is likely the main factor that drives the inference of the overall TDRP slope. This robustness, however, raises a particular concern, which is that the TDRP we observed here could be driven by only a few long-term rate estimates and, in the worst-case scenario, that the majority of the short-term

rates we used in our analyses are erroneous, calculated from data sets with no temporal signal. Nevertheless, our sensitivity analyses show that erroneous short-term rates can severely bias the short-term TDRP curve estimation (**Figure 4.3**, **bottom**). However, we found that the TDRP could still be detected even among short-term rate estimates and that the results from the short-term TDRP analyses are all consistent with those obtained from the overall TDRP analyses (**Figure 4.4**). This indicates a temporal structure for the short-term rate estimates and suggests that our analyses likely contain only a few erroneous short-term rate estimates, if any. This finding also shows that our overall TDRP analyses were not driven solely by a few long-term rate estimates and that the rate decay slope is indeed stable across all timescales.

Many factors have been put forward as underlying causes of the TDRP (for a review, see reference 25). Nevertheless, the fact that the TDRP can be found across all viruses suggests that it is a result of natural phenomena that all viruses encounter. There are a wide range of evolutionary scenarios that all viruses are expected to encounter, which can lead to depressed long-term substitution rate estimates and inflated short-term rate estimates, and hence have the potential to explain the TDRP. For example, despite their large diversity in biological features, life histories, and replication strategies, all viruses experience repeated bottlenecking transmission events, which can lead to the accumulation of deleterious mutations within viral genomes (40–42). It is also clear that all viruses have highly subdivided population structures. This means that slightly defective viruses within one subpopulation (e.g., within a host individual or a group of individuals within a transmission network) might not be immediately outcompeted by their slightly fitter relatives in different subpopulations, exacerbating the accumulation of deleterious mutations (43, 44). It is therefore expected that deleterious mutations can persist within viruses (as a species) for significant amounts of time, causing rates estimated over short timescales (i.e., toward the present) to be systematically overestimated. In contrast, the long-term rate of viral evolution is a product of coevolution between viruses and their slowly evolving hosts, constrained by many molecular features needed for proper biological functions and host interaction, such as overlapping reading frames, secondary structure, and regions with multiple functions (45). These strong

evolutionary constraints are expected to result in extensive site saturation and convergent evolution (45). This, in turn, can lead to systematically underestimated long-term rate estimates. Overall, given that the TDRP is a prevalent evolutionary feature of viruses, we hypothesise that the bottlenecking nature of viral transmission, highly subdivided population structure, and substitution saturation are the main drivers of the TDRP.

Surprisingly, all of our analyses reveal that the speeds at which the rates decay do not differ significantly among viruses despite their vast diversity in replication strategies, molecular features, and evolutionary histories. This uniformity of the rate decay slopes suggests that, irrespective of their biological and molecular features, all viruses may experience the same degrees of population subdivision and bottlenecking. Moreover, this finding may also suggest that the degree of substitution saturation might be the same across all viruses. The most parsimonious form of this hypothesis predicts that all viral genomes might have similar proportions of conserved, neutral, and adaptive sites. Indeed, it has been shown that poor modelling of rate heterogeneity among sites can cause rate estimates to appear to be time dependent (46). It is therefore possible that the TDRP might be caused partially by the proportions of conserved, neutral, and adaptive sites being poorly modelled by our current methods. Examination of these hypotheses will require experiments that elucidate how population subdivision, bottlenecking, and the proportions of conserved, neutral, and adaptive sites affect the measurement of evolutionary rates.

We want to emphasise that the TDRP is likely only an apparent phenomenon—that the actual rate of substitution fixation within viral species is time independent but *appears* to vary over time, likely due to the presence of transient deleterious mutations (which inflate short-term rate estimates) and substitution saturations (which deflates long-term rate estimates) (24, 25, 27). Lying between the two extreme ends of short-term/long-term rate estimates is an effective (apparent) rate that matches the true time-independent rate of viral evolution, resulting from a balance between the presence of site saturations and transient deleterious mutations within viral genomes. It is likely that these true rates of viral evolution differ between viral lineages;

after all, the replication strategies, molecular characteristics, and evolutionary histories of these highly distinct viruses are drastically different, and these should in some ways influence their (actual) substitution rates. On the other hand, it is expected that viruses with similar biology and molecular features should have similar evolutionary rates. It is therefore surprising to see emerging evidence showing that the difference in viral evolutionary rates is extremely large even among viruses of the same type (11, 12) and that the boundary between the rates of DNA and RNA viruses is very blurry to the point that there may be no strict division between the two (13, 14). Our analyses show, for the first time, that the TDRP can explain a large portion of this variance, and the boundary between the rates of different viruses becomes more apparent once the TDRP is accounted for.

Our analyses show that our TDRP model can account for ~90% of the variance within the rate estimates (R^2 = 0.89 for TDRP among viral groups and 0.95 for TDRP among viral genera), and after adjusting for the TDRP, we found that dsDNA viruses evolve more slowly than ssDNA viruses and reverse transcribing (RT) viruses, which, in turn, evolve more slowly than ssRNA viruses consistently across all analyses (**Figure 4.1C** and **2B**). These findings are consistent with the experimental observations that (i) DNA is more chemically stable than RNA (47), (ii) single-stranded nucleic acids have higher instability than double-stranded nucleic acids (48), and (iii) the replication fidelity of DNA polymerase is higher than that of reverse transcriptase, which, in turn, is higher than that of RNA polymerase (13, 49). Our results also greatly resemble the observations that the mutation rates of dsDNA viruses are lower than those of ssDNA viruses and RT viruses, which, in turn, are lower than those of RNA viruses (11, 50). Together, our findings support the notion that there indeed exists a clear basic division between the rates of evolution of dsDNA viruses, ssDNA/RT viruses, and ssRNA viruses and that this can be explained very well by differences in molecular biology among viruses

One of the most important implications of the TDRP is that naïvely extrapolating evolutionary rate estimates across different timescales for evolutionary inferences can severely bias the analyses (26, 27). One of the best-known cases of this is perhaps the severe underestimation

of lentiviral evolutionary timescales by standard phylogenetic analyses. Our TDRP model may offer a partial solution to this problem, allowing us to reconcile the long-term and short-term viral evolutionary studies. To illustrate this, we used our model to reestimate the ages of various lentiviral groups (**Figure 4.5**).

Our analyses estimate that lentiviruses are millions of years old, and the date estimates are highly consistent with paleovirological evidence and the knowledge of lentiviral host migration. SIVs were previously estimated to be at most tens of thousands of years old by phylogenetic analyses that did not account for the TDRP (32). Our TDRP model, on the other hand, estimated SIVs to be >2.57 (95% HPD = 0.35–10.9) Myr old. Compared to previous age estimates, our estimate fits much better with paleovirological evidence that suggests that SIVs are many millions of year old (28). We also estimated FIVs to be >1.63 (95% HPD = 0.27–6.50) Myr old. Based on the geographical distribution of FIVs, Troyer *et al.* (51) proposed that modern FIVs likely originated before the last time felids crossed the Bering Strait, in the late Pliocene (>2.58 Myr ago) to early Pleistocene (>1.80 Myr ago) epoch (52). Our estimate fits very well with this knowledge of felid migration. Furthermore, previous phylogenetic analyses revealed that a gray mouse lemur endogenous lentivirus (pSIVgml) is embedded within the clade of SIVs and FIVs (29), and it is >4 Myr old (53). This places a lower-bound estimate of the time to the MRCA (tMRCA) of SIVs and FIVs at 4 Myr. Again, our result is consistent with this finding, suggesting that the tMRCA of SIVs and FIVs is >10.1 (95% HPD = 1.07–55.8) Myr. Last, we estimated lentiviruses as a whole to be >15.9 (95% HPD = 1.03–96.9) Myr old. This is congruent with the results from analyses of orthologous rabbit-hare endogenous lentiviruses, which suggested that lentiviruses are >12 Myr old overall (54).

It is also noteworthy that, based on the observed basal phylogenetic placement of pSIVgml within SIVs, an ancient cospeciation between primates and lentiviruses has also been proposed, implying that primate lentiviruses are 85 Myr old (29). Our estimated tMRCA for SIVs and FIVs, at 10.1 Myr, is much more recent than that date. However, it should be noted that this date estimate is a lower-bound estimate, since our model was calibrated with a lower-

bound viral speciation date. Furthermore, given the limited number of basal viral lineages in our analyses, it is possible that the number of substitutions of basal branches, and therefore deep evolutionary timescales, might still be significantly underestimated due to site saturation. We therefore consider that it is too early to reject this hypothesis. Further investigation with more basal lentiviral sequences may resolve this conundrum.

Overall, our results support the predictive value of our TDRP model and suggest that our model is useful as a guideline for further improvement of current evolutionary analytical tools. To the best of our knowledge, this is the first demonstration that lentiviruses are millions of years old by phylogenetic analyses of extant viruses, closing the gap between ancient and recent viral evolution. Nonetheless, we note the large uncertainties of our lentiviral age estimates, spanning ~2 orders of magnitude (**Figure 4.5**). Rather than representing intrinsic problems of the model, we believe that, in addition to a short nucleotide alignment, this highlights the inherent difficulties in working with the small number of data points available for calibrating the model and extrapolating rate estimates across large timescales. This problem could be partially overcome by developing a TDRP model under a full Bayesian or maximum likelihood phylogenetic framework. Such a model would allow us to fully integrate and control the prior distributions of the rates, timescales, and substitutions, as well as to jointly estimate the parameters, thereby increasing precision.

4.5 Materials and Methods

4.5.1 Data collection and construction of phylogenetically independent data sets of rate estimates

A total of 396 viral nucleotide substitution rate estimates were collected from 133 pieces of published literature. The rates were arbitrarily divided into two groups: short-term rates (estimated over timescales of <1,000 years) and long-term rates (estimated over timescales of >1,000 years) (**Figure 4.1A**; see also **Table S4.1** in **Electronic Appendix**). We considered 1,000 years to be an appropriate cutoff because it corresponds to a large gap in the rate

measurement timescales (between 760 years and 6,600 years), clearly dividing rate estimates into two distinct rate categories. To control for the quality of the data, only evolutionary rates estimated under the Bayesian or maximum likelihood frameworks or from neighbour-joining trees were collected. All the rates estimated from neighbour-joining trees were calculated based on explicit nucleotide substitution models where the rates of transversion and transition substitutions were distinguished, at the very least. In two studies however, the models were not reported. In total, the data set comprises nucleotide substitution rate estimates of six major viral groups (according to the Baltimore classification), including 21 rate estimates for group I dsDNA viruses, 47 for group II ssDNA viruses, 123 for group IV (+)ssRNA viruses, 106 for group V (−)ssRNA viruses, 85 for group VI RT-RNA viruses and 14 for group VII RT-DNA viruses (**Table 4.1**).

These rate estimates are not phylogenetically independent, however, and this may bias our analyses if it is unaccounted for. To construct data subsets of phylogenetically independent rate estimates, we constrained the sampling criterion so that only one rate estimate was sampled per virus (defined by their common names (**Table S4.1** in **Electronic Appendix**)), and all of its rates had the same probability of being sampled.

We also noted that some of these rate estimates are still phylogenetically dependent despite being for different viruses. For example, the short-term rates of herpes simplex virus 1 and varicella-zoster virus are phylogenetically independent, but both are phylogenetically nested within the long-term rates of alphaherpesviruses, which are, in turn, nested within those of alpha- and betaherpesviruses. In such cases, we divided the rates into phylogenetically independent groups while maximising the number of rate estimates within each group. In this case, for instance, the rates were divided into three groups: (i) those of herpes simplex virus 1 and varicella-zoster virus, (ii) those of alphaherpesviruses, and (iii) those of alpha- and betaherpesviruses. For a particular data subset, only one of these groups was sampled with the same probability. Note that the number of rates may differ among groups, and thus, the number of rate estimates may differ across data subsets. Furthermore, the sampling was also

constrained such that in each data subset, there was at least one short-term and one long-term rate estimate for each viral group. This constraint was imposed so that the rate decay dynamics could be estimated reliably. In total, 1,000 pseudoreplicated data subsets of phylogenetically independent rate estimates were constructed. See the sampling count in **Table S4.1** in **Electronic Appendix** and the summary of the number of rate estimates for each data subset in **Table 4.1**.

4.5.2 TDRP analyses at the level of Baltimore classification viral groups

For each of the 1,000 pseudoreplicated data subsets, the rate estimates and measurement timescales were log-transformed (base 10) and were fitted to three linear models (equivalent to modelling a power law relationship between the rate estimates and measurement timescales): (i) rate estimate as a function of the timescale of rate measurement, (ii) rate estimate as a function of the timescale of rate measurement and viral group, and (iii) rate estimate as a function of the timescale of rate measurement and viral group, with separate rate decay slopes for each group. Model parameters were estimated by using the linear model (*lm*) function, implemented in R, version 3.1.2 (55). The significance levels of the estimated parameters were evaluated by *t* tests. Analyses of variance (ANOVA) and F tests were used to compare the models. Complete pairwise comparisons of the intercepts were performed by using the *glht* function, implemented in R, version 3.1.2 (55), to determine which viruses evolve faster or slower than others. In order to capture all of the parameter estimation uncertainties across 1,000 pseudoreplicated subanalyses, we constructed the distributions of parameter values by sampling 100 sets of parameter values from each of the 1,000 pseudoreplicated analyses and combined them as one to form an overall distribution (containing 100,000 sets of parameter estimates in total). Fisher's method was used to combine the *P* values from each subanalysis into a single overall *P* value. The results were evaluated at a significance level of 0.05.

4.5.3 TDRP analyses at the level of viral genera

A data set comprising rate estimates obtained from seven viral genera was assembled, including 4 rate estimates for the group I dsDNA virus genera *Simplexvirus* and *Varicellovirus*, 6 for the group II ssDNA virus genus *Mastrevirus*, 18 for the group IV (+)ssRNA virus genus *Tobamovirus*, 12 for the group V (−)ssRNA virus genus *Hantavirus*, 8 for the group VI RT-RNA virus genus *Deltaretrovirus*, and 3 for the group VII RT-DNA virus genus *Avihepadnavirus*. The *Simplexvirus* and *Varicellovirus* genera were combined as one data set, because there were not enough data points for separate analyses. The rate estimates and measurement timescales were log-transformed (base 10) and were fitted to three different linear models as described above. Parameter estimation, model comparison, and the construction of the distributions of estimated parameter values were also performed as described above.

4.5.4 Sensitivity analyses

Although we controlled for the rate measurement methods, some of the rate estimates, especially the short-term ones, might still be erroneous, since they are estimated from molecular data sets that do not contain sufficient temporal structure. We conducted simulations to evaluate how sensitive our analyses are to the presence of this specific type of erroneous rate estimate.

We simulated 55 molecular sequences, each of which is 1,000 bp long. The evolutionary process was assumed to be a nonhomogeneous Poisson process, with the instantaneous rate of substitution (r_t) being constant across sites but varying through time, governed by a power law decay function $r_t = r_0 t^{-\beta}$. Guided by the results obtained from the TDRP analyses, the distribution of β (the rate decay slope) was assumed to be normal, with a mean of 0.65 and a standard deviation of 0.041. r_0 (an apparent instantaneous evolutionary rate at one year before the present) was randomly sampled from a log-uniform distribution, ranging from 10^{-4} to 10^{-2} s/n/y. Fifty sequences were simulated to evolve over short timescales, spanning 10 to 1,000 years, and five sequences evolve over long timescales, ranging from 1,000 to 100 million

years. The timescales were sampled from log-uniform distributions. The total numbers of substitutions were counted and were divided by their corresponding timescales to derive average substitution rate estimates. If the simulation resulted in at least one zero rate estimate, the simulation was rerun with a new sampled set of parameter values. To model erroneous rate estimates, a portion of short-term substitution numbers (20%, 40%, 60%, 80%, 100%) was shuffled before dividing by time, breaking the correlation between the two. The rate estimates and measurement timescales were log-transformed (base 10) and were fitted to a linear function using the *lm* function, implemented in R, version 3.1.2 (55). One hundred simulations were performed for each setting (0%, 20%, 40%, 60%, 80%, 100% erroneous short-term rate estimates). We then compared the distributions of the intercepts and slopes of the rate decay curves obtained from simulations containing erroneous rates to those obtained from the control simulation (0% erroneous rate estimates) using Wilcoxon signed-rank tests, implemented in R, version 3.1.2 (55). The *P* values were corrected by using the Bonferroni correction for multiple testing. In addition, we also performed sensitivity analyses exclusively on short-term rate estimates to investigate the effects of erroneous rates on the short-term TDRP analyses (<1,000 years).

4.5.5 Short-term TDRP analyses

Only the short-term rates of groups II, IV, and V were used, since there were <10 phylogenetically independent short-term rate estimates for groups I, VI, and VII in each data subset (**Table 4.1**). The rate estimates and measurement timescales were log-transformed (base 10) and were fitted to three different linear models as described for the overall TDRP analyses. Parameter estimation, model comparison, and the construction of the distributions of estimated parameter values were also conducted as described above.

4.5.6 Estimating lentiviral evolutionary timescales by using our TDRP model

A phylogeny of lentiviruses was estimated from a manually curated integrase nucleotide alignment (420 nucleotides, 34 sequences) under the Bayesian phylogenetic framework by using BEAST, version 1.8.1 (56). We decided to use only the integrase region, because it is the only region currently available for SIVdrl-Bioko—the sole lentivirus that provides the timescale information we used to calibrate our TDRP model (GenBank accession no. HM363427; see below). See **Data set S4.1** in **Electronic Appendix** for the alignment. The Yule speciation process was applied, and the strict clock assumption with a fixed rate of 1 was used to estimate the total per lineage substitutions (s). Since we were interested only in the number of substitutions, and the phylogeny is not directly time calibrated, a strict molecular clock is sufficient. The best-fit nucleotide substitution models for each codon position, determined by jModelTest, version 2.1 (57), under the Akaike information criterion with a correction for finite sample sizes (AICc), were used (1st, GTR+I+Γ(4); 2nd, GTR+I+Γ(4); 3rd, TrN+Γ(4)). SIVdrl and SIVdrl-Bioko were constrained to be monophyletic. The Markov chain Monte Carlo (MCMC) was run for 50,000,000 steps. Parameters were logged every 5,000 steps, and the first 10% were discarded as burn-in. In total, 9,000 sets of posterior parameter estimates were obtained. Effective sample sizes of all parameters are >1,000, indicating that all parameters were well sampled and had converged.

We then simply converted the units of the branch lengths of each Bayesian posterior phylogeny from s to time (t) by using our TDRP model. The TDRP model states that $\log(\bar{r}) = \log(s/t) = \alpha + \beta \log(t)$, where $\bar{r}$ is the average evolutionary rate estimate (27); therefore, we can convert s to t simply by using the equation $t = (s/10^{\alpha})^{1/(\beta+1)}$. Nonetheless, our results show that although the rate decay slopes (β) do not differ significantly among viruses, the intercepts (α) do. Thus, an α that is specific to lentiviruses is required in order to conduct these analyses.

To calculate α, which is equal to $\log(s) - (\beta + 1)\log(t)$, we need β and at least a pair of corresponding s and t estimates. Based on the separation date of the African mainland and Bioko island, it has been proposed that SIVdrl and SIVdrl-Bioko separated >10,000 to 11,000 years ago (32), providing us with a timescale of the genetic divergence between SIVdrl and SIVdrl-Bioko. In this study, we assumed that the SIVdrl/SIVdrl-Bioko separation date is normally distributed, with a mean of 10,500 years and a standard error of (10,500–10,000) years/1.96, or (11,000–10,500) years/1.96, which equals 255.10 years. For each of the estimated Bayesian posterior trees, we first extracted the s estimate since the MRCA of SIVdrl and SIVdrl-Bioko to the present (s_{cal}), and sampled one SIVdrl/SIVdrl-Bioko split date (t_{cal}) from the assumed normal distribution. We then sampled one rate decay slope estimate (β) from the slope distribution, obtained from the overall TDRP analyses, and, in turn, computed the α of the lentivirus-specific TDRP model using the relationship $\alpha = \log(s_{cal}) - (\beta + 1)\log(t_{cal})$. The calibrated model, $t = (s/10^{\alpha})^{1/(\beta+1)}$, was then used to compute the t estimates of other nodes given their s estimates. This process was applied to all of the 9,000 posterior estimated trees in order to obtain the full posterior distributions of t estimates.

4.6 References

1. **Bernard HU**. 1994. Coevolution of papillomaviruses with human populations. Trends Microbiol. **2**:140–143.
2. **McGeoch DJ**, **Gatherer D**. 2005. Integrating reptilian herpesviruses into the family Herpesviridae. J. Virol. **79**:725–731.
3. **McGeoch DJ**, **Cook S**. 1994. Molecular phylogeny of the *Alphaherpesvirinae* subfamily and a proposed evolutionary timescale. J. Mol. Biol. **238**:9–22.
4. **McGeoch DJ**, **Cook S**, **Dolan A**, **Jamieson FE**, **Telford EAR**. 1995. Molecular phylogeny and evolutionary timescale for the family of mammalian herpesviruses. J. Mol. Biol. **247**:443–458.
5. **Sakaoka H**, **Kurita K**, **Iida Y**, **Takada S**, **Umene K**, **Kim YT**, **Ren CS**, **Nahmias AJ**. 1994. Quantitative analysis of genomic polymorphism of herpes simplex virus type 1 strains from six countries: studies of molecular evolution and molecular epidemiology of the virus. J. Gen. Virol. **75**:513–527.
6. **Perelman P**, **Johnson WE**, **Roos C**, **Seuánez HN**, **Horvath JE**, **Moreira MAM**, **Kessing B**, **Pontius J**, **Roelke M**, **Rumpler Y**, **Schneider MPC**, **Silva A**, **O'Brien SJ**, **Pecon-Slattery J**. 2011. A molecular phylogeny of living primates. PLoS Genet. **7**:e1001342.
7. **Kumar S**, **Subramanian S**. 2002. Mutation rates in mammalian genomes. Proc. Natl. Acad. Sci. U. S. A. **99**:803–808.
8. **Sharp PM**, **Li WH**. 1988. Understanding the origins of AIDS viruses. Nature **336**:315.
9. **Jenkins GM**, **Rambaut A**, **Pybus OG**, **Holmes EC**. 2002. Rates of molecular evolution in RNA viruses: a quantitative phylogenetic analysis. J. Mol. Evol. **54**:156–165.
10. **Lemey P**, **Rambaut A**, **Pybus OG**. 2006. HIV evolutionary dynamics within and among hosts. AIDS Rev. **8**:125–140.
11. **Sanjuán R**. 2012. From molecular genetics to phylodynamics: evolutionary relevance of mutation rates across viruses. PLoS Pathog. **8**:e1002685.
12. **Hanada K**, **Suzuki Y**, **Gojobori T**. 2004. A large variation in the rates of synonymous substitution for RNA viruses and its relationship to a diversity of viral infection and

transmission modes. Mol. Biol. Evol. **21**:1074–1080.

13. **Duffy S**, **Shackelton LA**, **Holmes EC**. 2008. Rates of evolutionary change in viruses: patterns and determinants. Nat. Rev. Genet. **9**:267–276.
14. **Biek R**, **Pybus OG**, **Lloyd-Smith JO**, **Didelot X**. 2015. Measurably evolving pathogens in the genomic era. Trends Ecol. Evol. **30**:306–313.
15. **Firth C**, **Kitchen A**, **Shapiro B**, **Suchard MA**, **Holmes EC**, **Rambaut A**. 2010. Using time-structured data to estimate evolutionary rates of double-stranded DNA viruses. Mol. Biol. Evol. **27**:2038–2051.
16. **Davis PL**, **Holmes EC**, **Larrous F**, **Van der Poel WHM**, **Tjørnehøj K**, **Alonso WJ**, **Bourhy H**. 2005. Phylogeography, population dynamics, and molecular evolution of European bat lyssaviruses. J. Virol. **79**:10487–10497.
17. **Shackelton LA**, **Parrish CR**, **Truyen U**, **Holmes EC**. 2005. High rate of viral evolution associated with the emergence of carnivore parvovirus. Proc. Natl. Acad. Sci. U. S. A. **102**:379–384.
18. **Shackelton LA**, **Holmes EC**. 2006. Phylogenetic evidence for the rapid evolution of human B19 erythrovirus. J. Virol. **80**:3666–3669.
19. **Duffy S**, **Holmes EC**. 2008. Phylogenetic evidence for rapid rates of molecular evolution in the single-stranded DNA begomovirus tomato yellow leaf curl virus. J. Virol. **82**:957–965.
20. **Switzer WM**, **Salemi M**, **Shanmugam V**, **Gao F**, **Cong M-E**, **Kuiken C**, **Bhullar V**, **Beer BE**, **Vallet D**, **Gautier-Hion A**, **Tooze Z**, **Villinger F**, **Holmes EC**, **Heneine W**. 2005. Ancient co-speciation of simian foamy viruses and primates. Nature **434**:376–380.
21. **Lemey P**, **Pybus OG**, **Van Dooren S**, **Vandamme A-M**. 2005. A Bayesian statistical analysis of human T-cell lymphotropic virus evolutionary rates. Infect. Genet. Evol. **5**:291–298.
22. **Sironen T**, **Vaheri A**, **Plyusnin A**. 2001. Molecular evolution of Puumala hantavirus. J. Virol. **75**:11803–11810.
23. **Hughes AL**, **Friedman R**. 2000. Evolutionary diversification of protein-coding genes of hantaviruses. Mol. Biol. Evol. **17**:1558–1568.

24. **Aiewsakun P**, **Katzourakis A**. 2015. Endogenous viruses: connecting recent and ancient viral evolution. Virology **479**–**480**:26–37.

25. **Ho SYW**, **Lanfear R**, **Bromham L**, **Phillips MJ**, **Soubrier J**, **Rodrigo AG**, **Cooper A**. 2011. Time-dependent rates of molecular evolution. Mol. Ecol. **20**:3087–3101.

26. **Duchêne S**, **Holmes EC**, **Ho SYW**. 2014. Analyses of evolutionary dynamics in viruses are hindered by a time-dependent bias in rate estimates. Proc. Biol. Sci. **281**:20140732.

27. **Aiewsakun P**, **Katzourakis A**. 2015. Time dependency of foamy virus evolutionary rate estimates. BMC Evol. Biol. **15**:119.

28. **Compton A a**, **Malik HS**, **Emerman M**. 2013. Host gene evolution traces the evolutionary history of ancient primate lentiviruses. Philos. Trans. R. Soc. Lond. B. Biol. Sci. **368**:20120496.

29. **Gifford RJ**, **Katzourakis A**, **Tristem M**, **Pybus OG**, **Winters M**, **Shafer RW**. 2008. A transitional endogenous lentivirus from the genome of a basal primate and implications for lentivirus evolution. Proc. Natl. Acad. Sci. U. S. A. **105**:20362–20367.

30. **Sharp PM**, **Bailes E**, **Gao F**, **Beer BE**, **Hirsch VM**, **Hahn BH**. 2000. Origins and evolution of AIDS viruses: estimating the time-scale. Biochem. Soc. Trans. **28**:275–282.

31. **Wertheim JO**, **Worobey M**. 2009. Dating the age of the SIV lineages that gave rise to HIV-1 and HIV-2. PLoS Comput. Biol. **5**:e1000377.

32. **Worobey M**, **Telfer P**, **Souquière S**, **Hunter M**, **Coleman CA**, **Metzger MJ**, **Reed P**, **Makuwa M**, **Hearn G**, **Honarvar S**, **Roques P**, **Apetrei C**, **Kazanji M**, **Marx PA**. 2010. Island biogeography reveals the deep history of SIV. Science **329**:1487.

33. **Ho SYW**, **Phillips MJ**, **Cooper A**, **Drummond AJ**. 2005. Time dependency of molecular rate estimates and systematic overestimation of recent divergence times. Mol. Biol. Evol. **22**:1561–1568.

34. **Lemey P**, **Van Dooren S**, **Vandamme A-M**. 2005. Evolutionary dynamics of human retroviruses investigated through full-genome scanning. Mol. Biol. Evol. **22**:942–951.

35. **Aulicino PC**, **Holmes EC**, **Rocco C**, **Mangano A**, **Sen L**. 2007. Extremely rapid spread of human immunodeficiency virus type 1 BF recombinants in Argentina. J. Virol. **81**:427–429.

36. **Bello G**, **Aulicino PC**, **Ruchansky D**, **Guimarães ML**, **Lopez-Galindez C**, **Casado C**, **Chiparelli H**, **Rocco C**, **Mangano A**, **Sen L**, **Morgado MG**. 2010. Phylodynamics of HIV-1 circulating recombinant forms 12_BF and 38_BF in Argentina and Uruguay. Retrovirology **7**:22.

37. **Biek R**, **Rodrigo AG**, **Holley D**, **Drummond A**, **Anderson CR**, **Ross HA**, **Poss M**. 2003. Epidemiology, genetic diversity, and evolution of endemic feline immunodeficiency virus in a population of wild cougars. J. Virol. **77**:9578–9589.

38. **Lefeuvre P**, **Harkins GW**, **Lett J-M**, **Briddon RW**, **Chase MW**, **Moury B**, **Martin DP**. 2011. Evolutionary time-scale of the begomoviruses: evidence from integrated sequences in the Nicotiana genome. PLoS One **6**:e19193.

39. **Gilbert C**, **Feschotte C**. 2010. Genomic fossils calibrate the long-term evolution of hepadnaviruses. PLoS Biol. **8**:e1000495.

40. **Campos PR a**, **Wahl LM**. 2010. The adaptation rate of asexuals: deleterious mutations, clonal interference and population bottlenecks. Evolution **64**:1973–1983.

41. **Clarke DK**, **Duarte E a**, **Moya a**, **Elena SF**, **Domingo E**, **Holland J**. 1993. Genetic bottlenecks and population passages cause profound fitness differences in RNA viruses. J. Virol. **67**:222–228.

42. **Duarte E**, **Clarke D**, **Moya a**, **Domingo E**, **Holland J**. 1992. Rapid fitness losses in mammalian RNA virus clones due to Muller's ratchet. Proc. Natl. Acad. Sci. U. S. A. **89**:6015–6019.

43. **Theodorou K**, **Couvet D**. 2006. Genetic load in subdivided populations: interactions between the migration rate, the size and the number of subpopulations. Heredity (Edinb). **96**:69–78.

44. **Whitlock MC**. 2003. Fixation probability and time in subdivided populations. Genetics **164**:767–779.

45. **Holmes EC**. 2003. Error thresholds and the constraints to RNA virus evolution. Trends Microbiol. **11**:543–546.

46. **Soubrier J**, **Steel M**, **Lee MSY**, **Der Sarkissian C**, **Guindon S**, **Ho SYW**, **Cooper A**. 2012. The influence of rate heterogeneity among sites on the time dependence of

molecular rates. Mol. Biol. Evol. **29**:3345–3358.

47. **Lesnik EA**, **Freier SM**. 1995. Relative thermodynamic stability of DNA, RNA, and DNA:RNA hybrid duplexes: relationship with base composition and structure. Biochemistry **34**:10807–10815.

48. **Frederico LA**, **Kunkel TA**, **Shaw BR**. 1990. A sensitive genetic assay for the detection of cytosine deamination: determination of rate constants and the activation energy. Biochemistry **29**:2532–2537.

49. **Steinhauer DA**, **Domingo E**, **Holland JJ**. 1992. Lack of evidence for proofreading mechanisms associated with an RNA virus polymerase. Gene **122**:281–288.

50. **Sanjuán R**, **Nebot MR**, **Chirico N**, **Mansky LM**, **Belshaw R**. 2010. Viral mutation rates. J. Virol. **84**:9733–9748.

51. **Troyer JL**, **Vandewoude S**, **Pecon-Slattery J**, **McIntosh C**, **Franklin S**, **Antunes A**, **Johnson W**, **O'Brien SJ**. 2008. FIV cross-species transmission: an evolutionary prospective. Vet. Immunol. Immunopathol. **123**:159–166.

52. **Johnson WE**, **Eizirik E**, **Pecon-Slattery J**, **Murphy WJ**, **Antunes A**, **Teeling E**, **O'Brien SJ**. 2006. The late Miocene radiation of modern *Felidae*: a genetic assessment. Science **311**:73–77.

53. **Gilbert C**, **Maxfield DG**, **Goodman SM**, **Feschotte C**. 2009. Parallel germline infiltration of a lentivirus in two Malagasy lemurs. PLoS Genet. **5**:e1000425.

54. **Keckesova Z**, **Ylinen LMJ**, **Towers GJ**, **Gifford RJ**, **Katzourakis A**. 2009. Identification of a RELIK orthologue in the European hare (*Lepus europaeus*) reveals a minimum age of 12 million years for the lagomorph lentiviruses. Virology **384**:7–11.

55. **R Core Team**. 2014. R: a Language and environment for statistical computing. R Found. Stat. Comput. **1**:409.

56. **Drummond AJ**, **Rambaut A**. 2007. BEAST: Bayesian evolutionary analysis by sampling trees. BMC Evol. Biol. **7**:214.

57. **Darriba D**, **Taboada GL**, **Doallo R**, **Posada D**. 2012. jModelTest 2: more models, new heuristics and parallel computing. Nat. Methods **9**:772–772.

Chapter five: Marine origin of retroviruses in the early Palaeozoic Era

---

This chapter was originally published as:

Aiewsakun P, Katzourakis A. 2016. Marine origin of retroviruses in the early Palaeozoic Era. Nat. Commun. **7**:13954.

PA performed the analysis and wrote the paper. AK provided editorial guidance and supervisory advice.

5.1 Abstract

Very little is known about the ancient origin of retroviruses, but owing to the discovery of their ancient endogenous viral counterparts, their early history is beginning to unfold. Here, we report 36 lineages of basal amphibian and fish foamy-like endogenous retroviruses (FLERVs). Phylogenetic analyses reveal that ray-finned fish FLERVs exhibit an overall co-speciation pattern with their hosts, while amphibian FLERVs might not. We also observe several possible ancient viral cross-class transmissions, involving lobe-finned fish, shark, and frog FLERVs. Sequence examination and analyses reveal two major lineages of ray-finned fish FLERVs, one of which had gained two novel accessory genes within their extraordinarily large genomes. Our phylogenetic analyses suggest that this major retroviral lineage, and therefore retroviruses as a whole, have an ancient marine origin and originated together with, if not before, their jawed vertebrate hosts >450 million years ago in the Ordovician period, early Palaeozoic Era.

5.2 Introduction

Retroviruses are a group of medically and economically important viruses (family *Retroviridae*) that infect a wide range of animals from fish to human (1, 2), and can occasionally leave *genomic fossils* within their host genome, known as endogenous retroviruses (ERVs). ERVs are relics of past infections, resulting from viral genomic integrations that occur in host germ-line cells, and are in turn passed down from parents to offspring as part of the host genome (3). Owing to the increasing availability of animal whole genomes, more and more ancient ERVs have been discovered, allowing the early history of retroviruses to be examined like never before (3). The oldest age estimates directly inferred for retroviruses are ~100 million years (Myr) old, derived from analyses of mammalian retroviruses and ERVs (4–6). Beyond this point in time, the origins of retroviruses remain unclear due to the difficulties of accurately identifying ancient retroviral integrations, and the limitations of extrapolating from extant retroviral sequence data.

Retroviruses are extremely widespread among vertebrates (1, 2), raising the possibility that they might be as old as their vertebrate hosts. However, retroviruses frequently cross species even across vertebrate classes, as evidenced by phylogenetic analyses incorporating distantly related ERVs (1, 2). Thus, retroviruses could be ~100 Myr old and have been transmitted across vertebrates, or may date back to the origins of vertebrates at ~460−550 Myr ago (Ma) (7–9). To better examine and estimate the date of origin of retroviruses necessitates analyses of retroviruses whose evolutionary dynamics are well understood, and because of this, we turn to foamy viruses (FVs).

FVs are a unique subgroup of retroviruses (genus *Spumaretrovirus*) that are characterised by an extremely stable history of co-speciation with their mammalian hosts, at least since the origin of eutherians ~100 Myr ago (5, 6). This unique evolutionary feature, together with the high availability of FV molecular data, allows their evolutionary dynamics to be described in unprecedented detail, making FVs one of the most important models of retroviral

macroevolution. Similar to other retroviruses, FVs occasionally leave viral genomic fossils in their host genomes. Three mammalian endogenous FVs have been discovered to date, including the prosimian aye-aye endogenous FV (*Daubentonia madagascariensis*; PSFVaye (6, 10)), sloth endogenous FV (*Choloepus hoffmanni*; SloEFV (5)), and Cape golden mole endogenous FV (*Chrysochloris asiatica*; ChrEFV (6, 11)). On the basis of protein sequence similarity, a number of ERVs have also been identified as FV-like in fish genomes, including coelacanth (*Latimeria chalumnae;* CoeEFV (12)), platyfish (*Xiphophorus maculatus*; platyfishEFV (13)), and zebrafish (*Danio rerio*; DrFV-1 (14)). Phylogenetic analyses showed that CoeEFV is positioned basal to mammalian FVs (12), and that DrFV-1 and platyfishEFV form a monophyletic clade, basal to CoeEFV and mammalian FVs (13). This phylogenetic pattern matches perfectly that of vertebrates, raising the possibility of an ancient marine origin of this major retroviral lineage (12, 13). However, unlike in the case of mammalian FVs, this inferred large-scale cospeciation history could be misleading, as limited data sampling may have obscured a history of cross-species transmissions. Outside of mammals, there is no statistical evidence of congruence of host and FV phylogenetic topologies, and therefore the two scenarios cannot be distinguished. Additional ERV data from multiple vertebrate classes could help resolve the ancient origins of FVs, and also shed light on the early history of retroviruses as a whole.

Here we present 36 novel lineages of amphibian and fish FV-like ERVs (FLERVs), some of which have the largest known retroviral genomes. We also re-examine the co-evolutionary history of this major retroviral lineage with their vertebrate hosts. To reconstruct the evolutionary history of these viruses, we incorporate a recently developed approach that can account for the fact that the rates of viral evolution are time-dependent (15–18), appearing to decay over time following a power-law pattern (16). By overcoming the limitations of extrapolating across different timescales, incorporating multiple endogenous FVs, and leveraging the stable evolutionary dynamics of FVs, we show that this major group of retroviruses emerged >450 million years ago in the early Palaeozoic Era, coinciding with the

origin of jawed vertebrates. To our knowledge, this is the first temporal evidence indicating that retroviruses are at least as old as their jawed vertebrate hosts.

5.3 Results

5.3.1 Discovery of FLERVs in amphibian and fish genomes

By using tBLASTn and the reverse transcriptase (RT) protein of CoeEFV as a screening probe, 1,752 RT sequences were retrieved from publically accessible nucleotide GenBank databases, including the database of GenBank non-redundant nucleotide sequences, expressed sequence tags, high throughput genomic sequences, whole genome shotgun sequences, and transcriptome shotgun assembly sequences. Together with 304 additional RT sequences from seven retroviral genera, phylogenetic analyses showed that 161 sequences from 28 distinct species (4 salamanders, 1 frog, 20 ray-finned fish, 2 lobe-finned fish, and 1 shark) cluster together with the known mammalian FVs and fish ERVs previously identified as FV-like. We therefore only focused on these 161 retroviruses, and defined them as FLERVs.

To identify potentially full-length FLERVs, we extended the sequences on both sides and BLASTed the sequences against themselves to search for long terminal repeats (LTRs)—the key characteristic feature of ERVs that defines the boundary of the element. By doing so, we discovered one potentially full-length novel FLERV in the eastern newt (*Notophthalmus viridescens*) and four in the midas cichlid (*Amphilophus citrinellus*) genomes. We designated them ‘NviFLERV’ and ‘AciFLERV’, respectively. Furthermore, upon the ERV insertion, the target site DNA is also duplicated, resulting in small target site duplications (TSDs) flanking the ERVs (19). We were able to identify NviFERLV’s TSDs as well as those of the four AciFLERVs (**Note S5.1** in **Electronic Appendix**), supporting that they are *bona fide* ERVs, and are not recombinants of multiple elements. Furthermore, we also found five potentially full-length FLERVs in the zebrafish (*Danio rerio*) and three in the West Indian Ocean coelacanth (*Latimeria chalumnae*) genomes. These ERVs have already been described as DrFV-1 (14) and CoeEFV (12), respectively.

To search for additional FLERVs, we screened the five publically accessible nucleotide databases again by using the RT protein sequences of SloEFV, NviFLERV, and AciFLERV. Thirty-two RT sequences from four additional ray-finned fish species were identified as FV-like, phylogenetically grouping with mammalian FVs and fish FLERVs. By expanding the sequences and based on the identification of their LTRs and TSDs, we discovered two more potentially full-length FLERVs in annual killifish (*Austrofundulus limnaeus*), and designated them 'AliFLERV'. We were able to identify TSDs of only one AliFLERV however, as the contig that harbours the other AliFLERV does not contain complete LTRs (**Note S5.1** in **Electronic Appendix**). In total, 193 sequences from 32 unique vertebrate species were identified as FLERVs (**Table S5.1** in **Electronic Appendix**). Potential recombination among these sequences was checked, but no significant evidence for recombination was found (*P* value = 0.604). Their phylogeny together with other retroviruses is shown in **Figure 5.1**. Our analyses show that they are indeed most closely related to the known FVs and FLERVs with strong support (bootstrap clade support = 95%, and Bayesian posterior probability = 1.00). Detailed phylogenetic relationships among FVs and FLERVs inferred from RT sequences are shown in **Figure S5.1**, in **Electronic Appendix**.

Our RT phylogeny strongly supports monophyletic clades of salamander, lobe-finned fish, and ray-finned fish FLERVs (bootstrap clade support > 84%, and Bayesian posterior probability = 1.00 for all clades). The clade of mammalian FVs was strongly supported by the Bayesian phylogenetic analysis (Bayesian posterior probability = 0.99), but not by the maximum-likelihood method (bootstrap clade support = 57%). On the other hand, we found that the frog FLERV clusters together with shark FLERVs (bootstrap clade support = 74%, and Bayesian posterior probability = 1.00), and the shark FLERVs appear to form two separate clades; however, the latter phylogenetic pattern is not strongly supported (bootstrap clade support = 38%, and Bayesian posterior probability = 0.73). We also observed that mammalian FVs, salamander FLERVs, and lobe-finned fish FLERVs cluster together (bootstrap clade support = 90%, and Bayesian posterior probability = 1.00), and ray-finned fish and shark FLERVs are basal to this clade, reflecting the host history. The phylogeny also shows that mammalian FVs

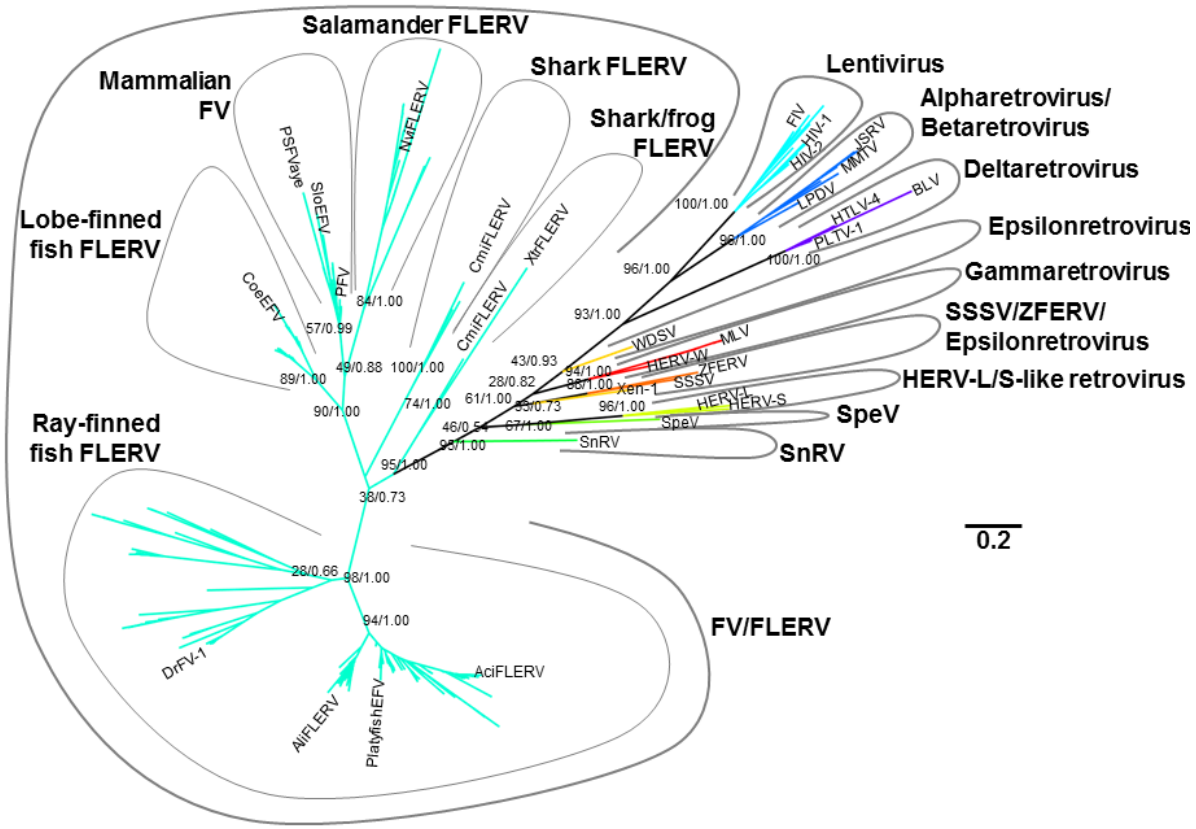


Figure 5.1 Retroviral phylogeny illustrating how FVs and FLERVs relate to other retroviruses. The un-rooted phylogenies were estimated from a reverse transcriptase protein alignment. Since both Bayesian maximum clade credibility and maximum likelihood trees have very similar topologies, only the Bayesian tree is shown. The scale bar is in units of amino acid substitutions per site. The numbers on nodes are bootstrap (first) and Bayesian posterior probability (second) clade support values.

are more closely related to salamander FLERVs than lobe-finned fish FLERVs, mirroring the host evolutionary relationship and thus consistent with a long-term co-speciation history between this retroviral lineage with their tetrapod hosts. However, again, the support for this relationship is low (bootstrap clade support = 49%, and Bayesian posterior probability = 0.88), likely because the RT sequences used to reconstruct the tree were short (130 aa), and thus, the topology of this tree should not be over-interpreted. Indeed, as noted by others (5, 20), RT trees are suitable only for broad classification of viruses. Below, we reconstructed a phylogeny of longer Pol protein sequences to determine the relationship among FVs and FLERVs more precisely (see **5.3 Results**: **5.3.4 Phylogenetic analyses**). Interestingly, our RT phylogenetic analysis shows that some fish and amphibian genomes harbour two FLERV lineages, including the genomes of Japanese fire belly newt (*Cynops pyrrhogaster*), fathead minnow (*Pimephales promelas)*, Atlantic cod (*Gadus morhua*), midas cichlid (*Amphilophus citrinellus*), and Amazon molly (*Poecilia Formosa*), as well as West Indian Ocean coelacanth (*Latimeria chalumnae*) (**Figure S5.1** in **Electronic Appendix**). To our knowledge, no known mammals have been shown to harbour two distinct lineages of FVs.

5.3.2 Characterisation of full length FLERVs

To characterise our novel amphibian and fish FLERVs, we first focused on the potentially full-length FLERVs, NviFLERV, AciFLERV, and AliFLERV, and used mammalian FVs as the main reference for comparison (**Figure 5.2**; see detailed genome annotation in **Figure S5.2** and **Note S5.1** in **Electronic Appendix**). We did not characterise DrFV-1 and CoeEFV as they have been previously described (12, 14). We found that many of the potentially full-length NviFLERV, AciFLERV, and AliFLERV contain large insertion/deletion mutations, in-frame stop codon and frame-shift mutations as well as transposable elements (TEs), making genome annotation difficult and potentially inaccurate. To address this problem, we used the full-length elements to retrieve additional (fragmented) elements from their respective genomes using BLASTn, and reconstructed the maximum-likelihood sequence of the basal node on the phylogeny. This inferred ancestral sequence was used for genome annotation. Unfortunately, only two NviFLERV elements were found in *N. viridescens*, designated ‘NviFLERV-1’, and

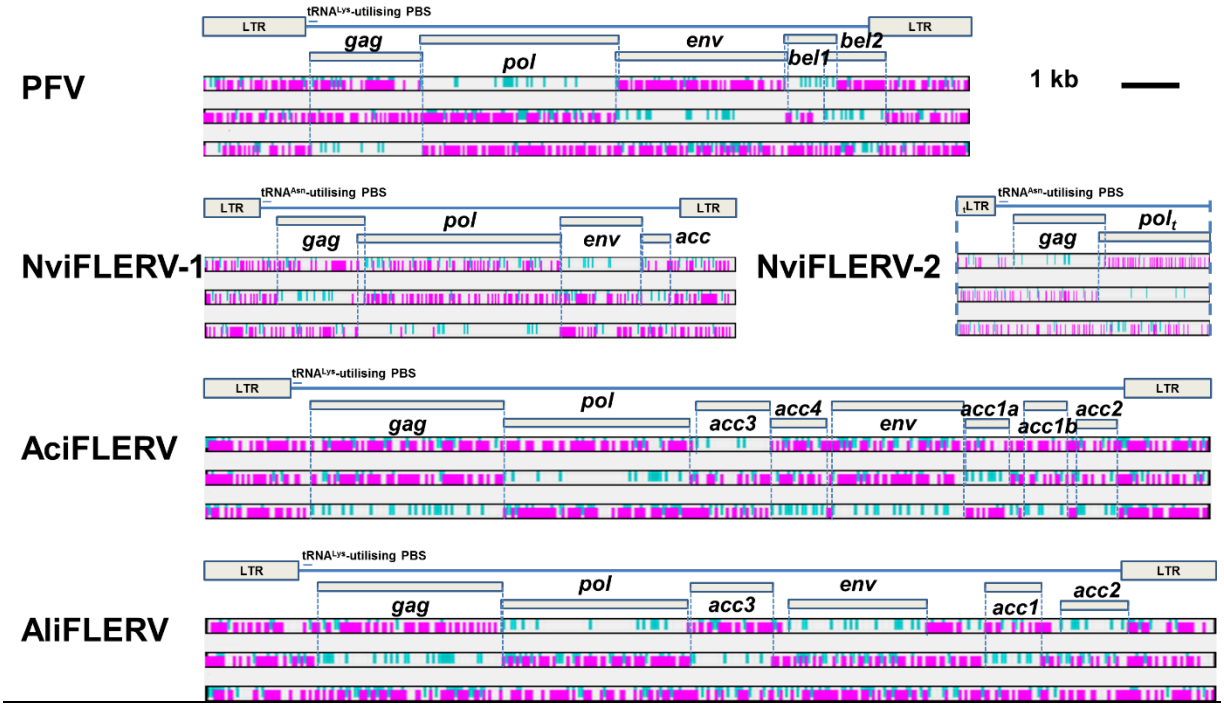


Figure 5.2 Genomic organisations of FVs and FLERVs. NviFLERV-1 and NviFLERV-2 are endogenous viral elements present within the genome of *N. viridescens*. This is in contrast to AciFLERV and AliFLERV, which were annotated by using their maximum-likelihood ancestral sequences, reconstructed from 9 and 23 elements, respectively. Under each element are the distributions of stop codons (TAA, TAG, and TGA; pink) and start codon (ATG; blue) in frame +1 (top), +2 (middle), and +3 (bottom), used to determine potential protein coding regions (blue thin vertical dashed lines). 't'-subscription indicates that the domain is truncated (5'-truncated: preceding the domain name; 3'-truncated: following the domain name; blue thick vertical dashed lines indicate where domains are truncated). Prototype foamy virus (PFV) was included as a reference. The scale bar (black solid line) represents a nucleotide length of 1 kilobase (kb).

'NviFLERV-2', and thus its ancestral sequence could not be reconstructed reliably.

NviFLERV-1 is a complete full-length FLERV (9,200 nt), containing putative *gag*, *pol*, and *env* genes that are similar to those of simian FVs (SFVs), flanked by 5'− and 3'−LTRs. Situated on the 3' end of the 5'−LTR is a $tRNA^{Asn}$-utilising putative primer binding site (PBS), identified via sequence similarity, similar to that of prosimian galago FV (6), but different from those of other mammalian FVs, which have $tRNA^{Lys}$-utilising PBSs. Several in-frame stop codons and frame-shift mutations were found in protein coding regions, indicating that it is defective, typical for an ERV. Between the *env* and 3'−LTR is a stretch of uniquely-mapping 492 nt, 164 aa, that does not exhibit similarity to any genes apart from a gene of Porcine reproductive and respiratory syndrome virus (BLASTx: AEQ61854; $E = 6 \times 10^{-4}$). This relatively high E-value of 6×10^{-4} makes it unlikely that they are homologues; nevertheless since mammalian FVs contain accessory genes in this region (*bel1* and *bel2*), and because it exhibits some similarity to a viral gene, we hypothesise that this nucleotide region might be an accessory gene of the progenitor of NviFLERV-1. Indeed, further analyses revealed that other salamander FLERVs also contain this gene (**Table 5.1**, see below), supporting that it has a viral origin and is not a host gene. We in turn annotated the gene as an '*acc'* gene.

The fact that NviFLERV-1 has paired 5'− and 3'−LTRs enables us to estimate its age. Due to the process of retroviral genomic integration, the paired LTRs are at first identical, but after becoming endogenous, the two gradually diverge from one another (3). Assuming that LTRs evolve neutrally after endogenisation, it is possible to calculate the integration date of the element, i.e. its age, from the genetic distance of the two paired LTRs. The synonymous substitution rate of amphibian evolution was estimated to be $\sim 0.924-1.53 \times 10^{-9}$ substitutions per site per year (21). Since synonymous substitutions are largely neutral, we considered it to be a reasonable approximation for a neutral rate of evolution of amphibians. We estimated the pairwise distance of the paired LTRs to be 4.4%. Given this estimated LTR pairwise distance and the synonymous substitution rate, we estimated NviFLERV-1 to be ~14 Myr old. The fact

that we could identify the TSDs of NviFLERV-1 indicates that NviFLERV-1 is likely a *bona fide* ERV and its ancient age estimate is not an artefact of recombination between multiple ERVs.

NviFLERV-2 is a truncated element (4,391 nt), containing a 5'-truncated 5' LTR, a complete *gag* gene, and a partial *pol* gene. A $tRNA^{Asn}$-utilising PBS similar to that of NviFLERV-1, and prosimian galago FV, was also identified by via sequence similarity. In contrast to NviFLERV-1, NviFLERV-2 *gag* and *pol* genes do not contain in-frame stop codon or frame-shift mutations. This raises two possibilities: either (i) that NviFLERV-2 is so young that it has not yet gained any in-frame stop codon or frame-shift mutations, or (ii) that it is a result of viral genomic contamination during the eastern newt genomic sequencing. To evaluate these possibilities, we screened short read archives for reads that span across the viral–host junction. Numerous reads mapping the junction were found, supporting that NviFLERV-2 is indeed a (very young) ERV, and not a result of viral genomic contamination.

Unlike NviFLERVs, we found 9 elements of AciFLERVs and 23 elements of AliFLERVs, allowing us to reconstruct their ancestral sequence for better genomic annotation. The original elements contain in-frame stop codon and frame-shift mutations as well as TEs and large insertion/deletion mutations, confirming that they are indeed ERVs. The inferred ancestral sequence of AciFLERV and AliFLERV are 17,409, and 17,490 nt long, respectively, much longer than the length of typical mammalian FV genomes, ~10kb. It is noteworthy that these lengths are not inflated by TEs and/or large insertions as they were removed during the reconstruction of ancestral sequences. By investigating the distribution of start and stop codons, we determined AciFLERV to have 8 open reading frames (ORFs; ORF-1 to -8, from the 5' end of the genome) flanked by 5'− and 3'−LTRs. Again, this is very different from mammalian FVs, which has only 5 ORFs, including (from the 5' end) *gag*, *pol*, and *env* genes followed by two accessory genes *bel1*, and *bel2*.

On the basis of sequence similarity and manual sequence inspection, we identified a $tRNA^{Lys}$-utilising PBS on the 3' end of the 5'−LTR, similar to those utilised by most mammalian FVs,

and determined that ORF-2, and ORF-5 are *pol*, and *env* genes, respectively. We in turn annotated ORF-1 as *gag* gene, and ORF-6, -7, and -8 as accessory genes based on our knowledge of retroviral genomic structure. Further inspection revealed that ORF-6 and -7 proteins are highly similar (BLASTp: $E = 1 \times 10^{-26}$), indicative of paralogy. We hence annotated ORF-6 (759 nt, 253 aa), -7 (753 nt, 251 aa), and -8 (702 nt, 234 aa) as an '*acc1a'*, an '*acc1b'*, and an '*acc2'* gene, respectively. *Acc1a* and *acc1b* are found in the same position as *bel1*, while *acc2* in the same position as *bel2*. The hypothetical proteins of these accessory genes, however, do not exhibit similarity to any known proteins, and/or contain any known conserved domains. Further studies of their functions and structures are required to shed more light on these genes. Similarly, ORF-3 (1,395 nt, 465 aa) and -4 (1,002 nt, 334 aa) proteins, which are much longer than those of ORF-6, -7, and -8, also do not contain any known conserved domains; therefore, it is unclear what they are. Nevertheless, it is known that some retroviruses possess accessory genes between *pol* and *env*. For example, the lentiviral human immunodeficiency virus contains accessory genes in this region, *vif*, *vpr*, and *vpu*, which are essential for the viral replication, assembly, and release (22). We thus believe that ORF-3 and -4 are viral accessory genes, and designated them '*acc3*' and '*acc4*', respectively. Again, our analyses revealed that other ray-finned fish FLERVs also contain these genes (**Table 5.1**, see below), supporting that they are indeed accessory genes of these FV-like viruses.

These gene annotations were then transferred to AliFLERV via protein sequence similarity. By examining the distribution of start and stop codons, we determined AliFLERV to have 6 ORFs (ORF-1 to -6) flanked by 5'− and 3'−LTRs. Our analyses suggest that ORF-1 to -6 are *gag*, *pol*, *acc3*, *env*, *acc1, and acc2* genes, respectively. A $tRNA^{Lys}$-utilising PBS was also identified after the 5'−LTR via sequence similarity as anticipated, similar to those of AciFLERV and most mammalian FVs. Detailed genomic annotations, such as PBS and TSD sequences, gene lengths, and locations of the genes on the contigs, are in **Note S5.1**, **Electronic Appendix**.

To estimate the age of AciFLERV and AliFLERV, we used the LTR-dating method as described above. Four pairs of AciFLERV paired LTRs were found, and their pairwise distances were

estimated to be between 0.2 and 1.1%. Likewise, we obtained two pairs of AliFLERV paired LTRs, and estimated their pairwise distance to be 3.1 and 4.9%. The neutral rate of fish genomic evolution has been reported to be $\sim 1.46 \times 10^{-8}$ substitutions per site per year (23). Given the estimated pairwise distances and the neutral rate of fish evolution, we inferred the oldest element of AciFLERV and AliFLERV to be 0.377 Myr old, and 1.68 Myr old, respectively, which are surprisingly young given the presence of TEs within their protein coding regions and their high copy numbers (at least 9 for AciFLERV and 23 for AliFLERV). This finding indicates that these FLERVs may still be active, and/or that they have been rapidly proliferating by helper infectious retroviruses via complementation *in trans*. Alternatively, it also could be that these groups of young FLERVs represent concomitant germ-line infections of closely related viruses. These young ages, however, should be interpreted with caution, as LTR-dating could severely underestimate the ages of ERVs if LTR gene conversion happens, reducing the divergence between paired LTRs (24). Nevertheless, this is unlikely as LTR gene conversion had to happen for all six elements to explain the observation. Further analyses revealed that the TEs within these FLERVs (none of which share the same genomic location) belong to TE groups that contain numerous highly similar elements, a strong indication of recent TE bursts. This observation might explain the presence of TEs in AciFLERV and AliFLERV despite their young age.

5.3.3 Characterisation of fragmented FLERVs

We used the annotated genomes of NviFLERV, AciFLERV, and AliFLERV, as well as those of CoeEFV and DrFV-1, to further annotate the genomic structure of other fragmented FLERVs in other vertebrate genomes via protein sequence similarity (**Table 5.1**). We were able to detect the presence of NviFLERV-like *gag*, *pol*, *env*, and *acc* genes in the genomes of the Japanese fire belly newt (*Cynops pyrrhogaster*), and Iberian ribbed newt (*Pleurodeles waltl*). Phylogenetic analyses showed that Gag, Env, and Acc proteins of *C. pyrrhogaster* FLERVs (CpyFLERVs) form two separate clades (**Figure 5.3**), confirming the existence of two lineages of CpyFLERVs as initially suggested by the RT phylogenetic analyses. The presence of NviFLERV-like *gag*, *pol*, and *env* genes were also detected in the genome of the Ezo

Table 5.1 Genomic structure of amphibian and fish FLERVs.

Class	Species	Gene*						
							Accessory gene (*acc*)	
		gag	***pol***	***acc3***	***acc4***	***env***	***acc1***	***acc2***
Amphibia	*Cynops pyrrhogaster*	○	○			○	○†	
		○	○			○	○†	
	Hynobius retardatus	○	○			○		
	Lissotriton vulgaris	○						
	Notophthalmus viridescens	○	○			○	○†	
	Pleurodeles waltl	○	○			○	○†	
	Xenopus tropicalis		○					
Chondrichthyes	*Callorhinchus milii*		○					
Osteichthyes	*Amphilophus citrinellus*		○					
		○	○	○	○	○	○‡	○
	Anoplopoma fimbria	○	○	○		○		○
			○					
	Austrofundulus limnaeus	○	○	○		○	○	○
	Cynoglossus semilaevis	○	○	○		○	○	
	Cyprinus carpio		○					
	Dicentrarchus labrax	○	○	○	○		○	○
	Esox lucius		○					
	Fundulus heteroclitus	○	○	○		○	○	
	Gadus morhua	○	○	○	○	○		
			○					
	Larimichthys crocea	○	○	○	○	○	○	○
	Lates calcarifer	○	○	○	○	○	○	○
	Nothobranchius furzeri		○			○		

Table 5.1 (Continue).

Class	Species	Gene*						
							Accessory gene (*acc*)	
		gag	***pol***	***acc3***	***acc4***	***env***	***acc1***	***acc2***
Osteichthyes	*Oreochromis niloticus*	○	○	○	○		○	
	Periophthalmodon schlosseri		○					
	Periophthalmus magnuspinnatus		○					
	Pimephales promelas		○					
		○	○			○		
	Poecilia formosa	○	○	○	○	○	○	
		○	○	○	○	○	○	
	Poecilia reticulate	○	○	○	○	○	○	
	Sebastes nigrocinctus		○				○	
	Sebastes rubrivinctus		○				○	
	Stegastes partitus	○	○	○	○	○	○	○
	Thunnus orientalis		○			○		
Sarcopterygii	*Latimeria menadoensis*	○	○			○		

* Open circles '○' indicate the presence of the gene. When there is more than one FLREV lineage in a genome, their genomic structures were annotated separately. NviFLERV, AciFLERV, AliFLERV, CoeEFV, and DrFV-1 protein sequences were used as probes.
† We identified only one accessory gene (*acc*) in salamander FLERVs (see also **Figure 5.2**); therefore it is unclear whether it is *acc1* or *acc2*.
‡ Two copies of *acc1* genes were found in the FLERV, annotated as *acc1a* and *acc1b* (see **Figure 5.2**).

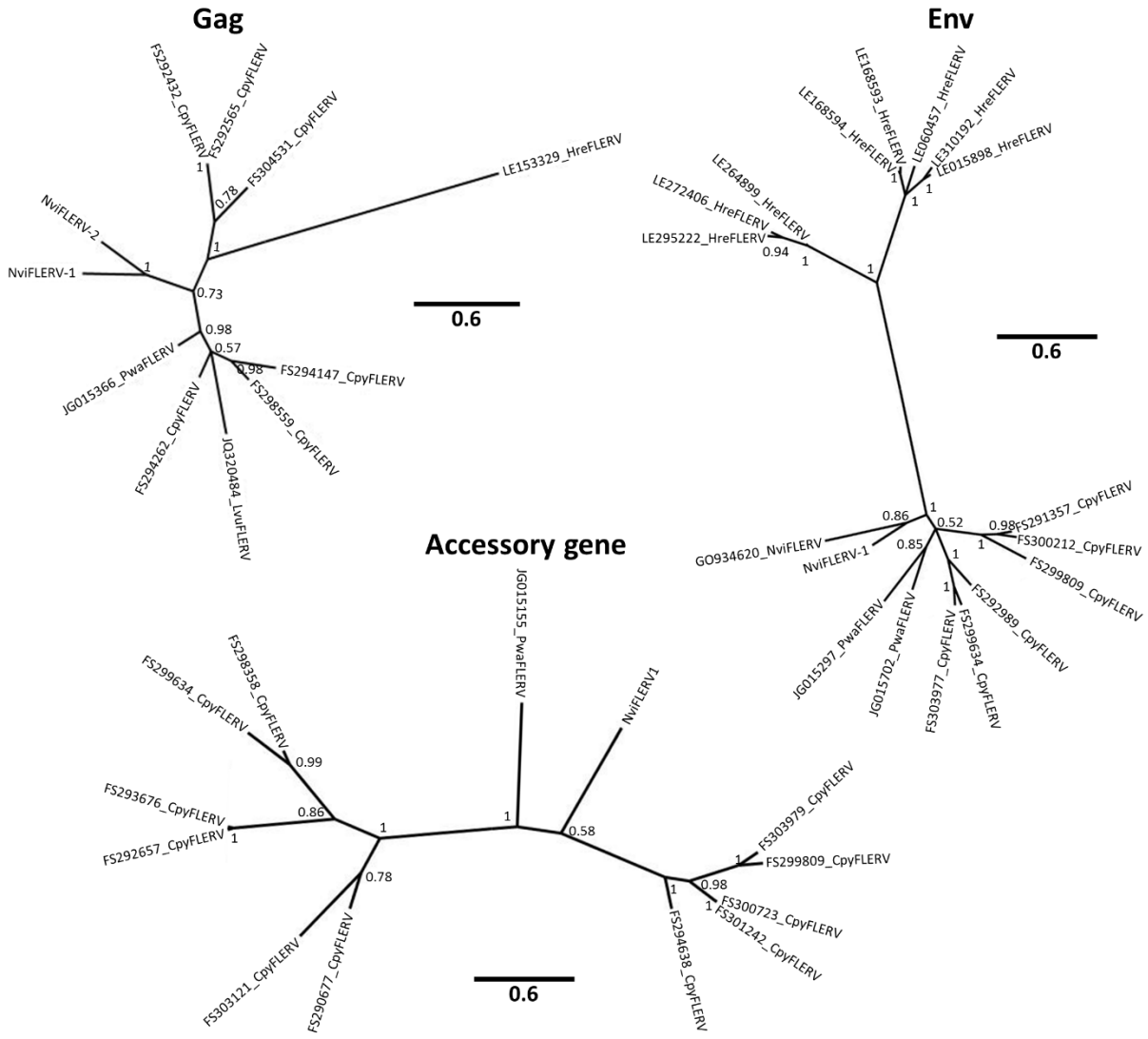


Figure 5.3 Phylogenies of salamander FLERVs. Unrooted Bayesian phylogenies of salamander FLERVs were estimated from Gag (left), Env (right) and accessory protein (middle) alignments. Preceding viral names are the contig accession numbers containing viral sequences used to reconstruct the phylogeny. The scale bar is in units of amino-acid substitutions per site, and the numbers on nodes are Bayesian posterior probability node support values.

salamander (*Hynobius retardatus*), but not the NviFLERV-like *acc* gene. This could be due to the lack of sequence data, genomic deletion, and/or high degree of sequence divergence. Our analyses also detected a NviFLERV-like *gag* gene in the genome of the smooth newt (*Lissotriton vulgaris*), a salamander that was not initially listed as a species containing FV-like RT sequences (**Table S5.1** in **Electronic Appendix**). Again, the absence of sequence data and/or genomic deletion might explain this result. Lastly, we could not detect any genes other than *pol* in the Western clawed frog (*Xenopus tropicalis*).

Regarding ray-finned fish FLERVs, while our analyses detected AciFLERV/AliFLERV-like *pol* genes in all of the FLERVs (*pol*: 27/27), AciFLERV/AliFLERV-like *gag, env* and *acc1* genes could only be detected in about half of the FLERVs (*gag*: 12/27; *env*: 12/27, *acc1*: 12/27), and *acc2* gene in about a fifth of the FLERVs (5/27). (Note that when there are multiple FLERV lineages present in a single genome, the genomic structure of each lineage was annotated separately by keeping track of which contigs the genes were found in.) Interestingly, in addition to the four fish species, our analyses revealed that sablefish (*Anoplopoma fimbria*) also contain two lineages of FLERVs; one with detectable *gag*, *pol*, *acc3*, *env*, and *acc2* genes, and another with only a detectable *pol* gene. Sablefish was not listed as a species containing two FLERV lineages in the initial RT analyses as the RT coding region for the former lineage was absent from the database. Another interesting observation was that, when the *acc1* gene was detected, only a single copy was found in all cases, except in AciFLERV. This suggests that the *acc1* duplication in the AciFLERV progenitor occurred very recently, and is lineage-specific. Our analyses confirmed the presence of *acc3* and *acc4* in 14 and 10 ray-finned fish FLERVs, respectively. The phylogenetic distribution of *acc3* and *acc4* was further examined to shed more light on their evolutionary history (see below). Finally, our analyses detected DrFV-like *gag, pol* and *env* genes in the genome of the fathead minnow (*P. promelas*), and CoeEFV-like *gag, pol* and *env* genes in the Indonesian coelacanth (*Latimeria menadoensis*), but could not detect any genes other than *pol* in Australian ghostshark (*Callorhinchus milii*). Again, we note that the absence of evidence is not the evidence of absence; the fact that we could not detect some genes in several fish FLERVs could be due to the lack of sequence data, genomic

deletion, genetic divergence, or indeed the genuine absence of the genes. To distinguish these possibilities requires genomic examination of their exogenous viral counterparts.

5.3.4 Phylogenetic analyses

To investigate the phylogenetic relationship between FLERVs and FVs in more detail, we estimated their Bayesian phylogeny based on a Pol protein alignment (**Figure 5.4**). *A. fimbria* FLERV (AfiFLERV) that contains detectable *gag*, *pol*, *acc3*, *env*, and *acc2* genes was not included in this analysis because its Pol sequence was extremely short (42 aa after alignment curation). Also, note that the phylogenetic placement of *L. vulgaris* FLERV (LvuFLERV) was determined separately by using a phylogenetic analysis of Gag proteins, as its Pol protein is not available. Potential recombination within both alignments was checked, but no significant evidence for recombination was found (FV/FLERV Pol protein alignment: *P* value = 0.089; salamander FLERV Gag protein alignment: *P* value = 0.649).

Overall, the results from phylogenetic analyses of RT and Pol protein sequences are largely consistent. We found that, as shown by the RT phylogenetic analysis, ray-finned fish, lobe-finned fish, and salamander FLERVs, as well as mammalian FVs all form well-supported monophyletic clades (Bayesian posterior probabilities > 0.99 for all clades), with tetrapod and lobe-finned fish FVs/FLERVs grouping together to the exclusion of ray-finned fish and shark FLERVs (Bayesian posterior probability = 0.98). Unlike the RT phylogeny however, the Pol phylogeny shows that shark FLERVs are paraphyletic instead of forming two separate clades; nevertheless, this phylogenetic pattern is not well-supported (Bayesian posterior probability = 0.72), similar to that in the RT phylogeny (**Figure 5.1**). Thus, it is still unclear how the progenitor of shark FLERVs interacted with their hosts.

As previously reported (5, 6, 25), we found strong evidence supporting the co-speciation history of mammals and their FVs (maximum number of co-speciation events inferred = 13/15 (86.67%); random tip mapping: N = 500, $p < 0.002$). Unlike in the case of mammals, several cross-species transmissions were found among ray finned fish (**Figure 5.4**). Despite this



Figure 5.4 Co-evolution of FVs and FLERVs and their vertebrate hosts. A Bayesian phylogeny of FVs and FLERVs (**left**) is compared with the published vertebrate host cladogram (26–29) (**right**). Preceding viral names are the contig accession numbers containing viral sequences. Class III retroviruses were used to root the viral tree. Solid grey lines between the two trees indicate viral–host associations. The scale bar is in units of amino acid substitutions per site, and Arabic numerals on nodes are Bayesian posterior probability node support values. Roman numerals indicate the nodes of which the total per-lineage substitutions to the chimpanzee simian FV (U14327 SFVcpz) were used to construct the time-dependent rate phenomenon model (**Figure 5.5**). The presence of *acc3* (orange squares) and *acc4* (red squares) is mapped onto the viral phylogeny. The most parsimonious timing of *acc3* and *acc4* acquisition is indicated (orange and pink arrows respectively).

observation however, our analyses still showed that ray-finned fish FLERVs co-diverge broadly with their hosts with strong support (maximum number of co-speciation events inferred = 18/27 (66.67%); random tip mapping: N = 500, $p < 0.002$). In the case of salamanders, no significant evidence was found for a history of cospeciation with their FV-like viruses (maximum number of co-speciation events inferred = 4/5 (80%); random tip mapping: N = 500, $p = 0.094$).

We also mapped the presence of *acc3* and *acc4* onto the phylogeny to determine the history of their acquisition. We found that they are exclusively limited to just one clade of ray-finned fish FLERVs (**Figure 5.4**), suggesting that they were acquired only once and perhaps both at the same time. This result further supports the phylogenetic distinctiveness of the two lineages of ray-finned fish FLERVs. Nonetheless, we noted that *acc3* and *acc4* genes could not be detected in some of the FLERVs in this clade. This could be due to the lack of sequence data or indeed multiple independent gene deletions as indicated by the non-monophyly of the absence of *acc3* and *acc4* genes (**Figure 5.4**).

Regarding the deeper evolutionary history, our analyses show that mammalian FVs are most closely related to lobe-finned fish FLERVs (posterior probability > 0.99), and basal to this clade are salamander FLERVs (Bayesian posterior probability > 0.99). This branching pattern, however, does not match that of the hosts, where mammals are more closely related to amphibians than lobe-finned fish. We note that this topology was not supported by the phylogeny estimated from a much shorter RT alignment (130 aa), which shows a sister-taxon relationship between mammalian FVs and salamander FLERVs (**Figure 5.1**), but with no statistical support. Indeed, when the tree estimation uncertainty is taken into account, this phylogenetic pattern falls within the RT tree estimation uncertainty. Combined, our results suggest that there were likely one or more ancient transmissions of FV-like viruses between tetrapods. We also found that, as shown by the RT phylogenetic analysis, instead of clustering with salamander FLERVs, the frog FLERV (XtrFLERV) was inferred to cluster with shark FLERVs (CmiFLERVs) with strong support (Bayesian posterior probability > 0.99), again indicative of viral cross-class transmission. Nevertheless, interpretation of these findings could

be complicated by '*incomplete lineage sorting*'. Incomplete lineage sorting is a phenomenon where multiple lineages of viruses continue to exist after the host basal diversification, but are sorted into different host lineages in the subsequent host divergence, and this could give raise to mismatches in virus/host evolutionary histories despite virus-host co-speciation. Temporal evidence can be used to differentiate between this phenomenon and viral cross-class transmissions (see **5.4 Discussion**). Another possibility is that these mismatches in virus/host evolutionary histories are a result of neutral genetic changes accumulating within FLERVs. However, a study has shown that neutral genetic changes only increase the branch length and decrease the clade support without altering the tree topology (6); thus it is unlikely that this is the case. It may nonetheless explain the extremely long branch of XtrFLERV (**Figure 5.4**, left).

5.3.5 Evolutionary timescale of FLERVs

Studies have shown that the rate estimates of mammalian FV evolution are time dependent (16, 18), and that this *time-dependent rate phenomenon* (TDRP) can be empirically described well by a power-law decay function (16). In fact, it has been demonstrated that virtually all viruses exhibit this evolutionary feature, and that the TDRP pattern is extremely stable across a very large timescale, spanning nine orders of magnitude, and across a wide range of host organisms from plants to animals (15). A consequence of this phenomenon is that the relationship between total per-lineage substitutions (S) and evolutionary timescales (T) will also be a power-law (16). To estimate the timescale of the FLERVs, we first built a TDRP model describing the relationship between S and T estimates of mammalian FVs by tracing the chimpanzee SFV lineage backward in time (U14327 SFVcpz; **Figure 5.4**). This was achievable due to the remarkably stable FV-mammal co-speciation history, allowing us to directly infer viral evolutionary timescales from those of mammals, and thereby obtaining a set of corresponding S and T estimates for model construction (16) (**Table 5.2**). We then extrapolated the model to estimate the timescale of FLERVs from their S estimates, under an explicit assumption that the TDRP dynamics of mammalian FVs are the same as those infecting their ancient ancestral (perhaps marine) vertebrates (**Figure 5.5** and **Table 5.2**).

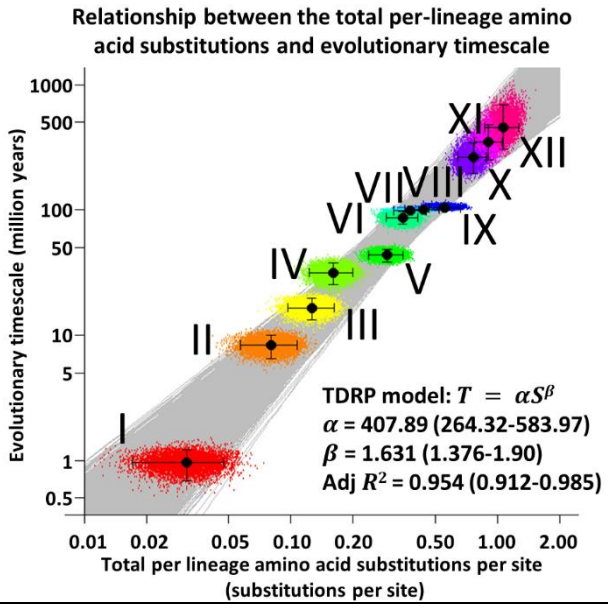


Figure 5.5 Evolutionary timescale of FLERVs estimated by using the power-law rate-decay model. The total per-lineage amino acid substitutions (S) from various nodes to the chimpanzee simian FV (U14327 SFVcpz) are plotted against corresponding evolutionary timescales (T). The S and T estimates are labelled with Roman numerals (**I–XI**), referring to nodes in **Figure 5.4**. Solid black dots are median estimates, and the associated 95% highest probability density (95% HPDs) intervals are indicated by error bars. The T estimates of node **I–IX** were directly inferred from those of their hosts (26, 30, 31). 7,500 power-law TDRP models were fitted to the posterior distributions of the S and T estimates of the nodes **I–IX** (grey lines). The median model parameter values, adjusted R^2 scores, and corresponding 95% HPDs (in the parentheses) are shown in the bottom right. The model was extrapolated to infer the branching date of lobe-finned fish FLERV lineage (node **X**), the separation date of salamander FLERV lineage (node **XI**), and the age of the entire FV/FLERV clade (node **XII**). See **Table 5.2** for the values.

Table 5.2 Evolutionary timescales of FVs and FLERVs.

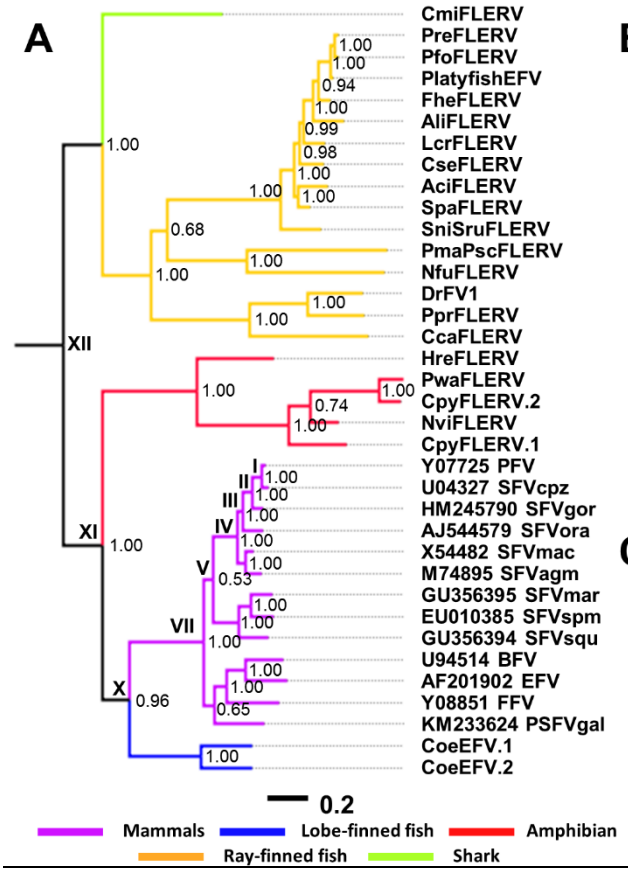
Node†	Total per-lineage substitution (substitutions per site)		Evolutionary timescale (million years) ‡		Ref
	Median	95% HPD*	Median	95% HPD	
I	0.031	0.017−0.048	0.96	0.69−1.21	(30)
II	0.080	0.057−0.107	8.29	6.52−10.05	(31)
III	0.126	0.097−0.163	16.52	13.33−19.61	
IV	0.160	0.123−0.199	31.60	25.36−37.85	
V	0.292	0.238−0.349	43.43	38.39−48.08	
VI	0.348	0.288−0.413	86.92	76.34−97.22	
VII	0.381	0.315−0.454	98.59	95.94−100.78	(26)
VIII	0.436	0.357−0.523	100.95	98.40−103.83	
IX	0.559	0.464−0.660	103.92	99.73−109.09	
X	0.764	0.641−0.889	262.76	195.00−342.08	NA
XI	0.908	0.771−1.056	347.89	250.62−479.97	
XII	1.071	0.869−1.274	454.65	303.69−683.66	

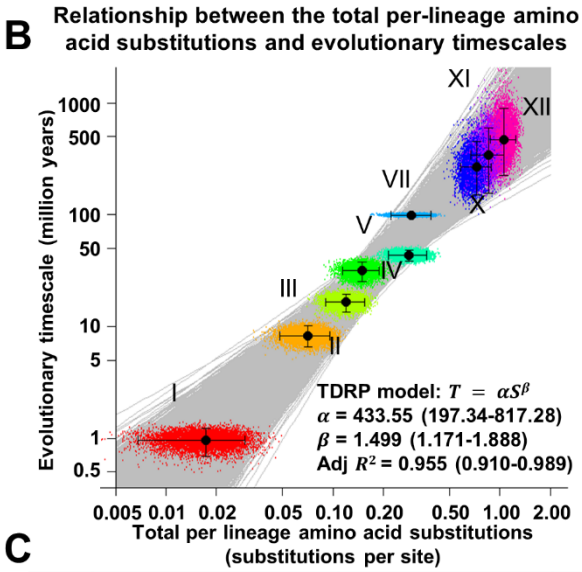
† Referring to the node numbers in the viral tree in **Figure 5.4** (left). Total per-lineage substitution and evolutionary timescale estimates of nodes **I–IX** were used to estimate the power-law time-dependent rate phenomenon model (**Figure 5.5**). The model was then extrapolated to infer the timescale of nodes **X**–**XII** from their total per-lineage substitution estimates.
‡ Evolutionary timescale of nodes **I**–**IX** were directly inferred from those of their hosts. The references refer to the host date references.
* '95% HPD' = 95% highest probability density interval.

Our analyses showed that the model can describe the dynamics of FV substitutions very well (adjusted R^2 = 0.954, 95 percent highest probability density interval (95% HPD) = 0.912–0.985). By assuming that viruses infecting modern-day mammals and their ancient vertebrate ancestors share the same TDRP dynamics, and based on the posterior distribution of the Bayesian Pol phylogeny, the clade of mammalian FVs and lobe-finned fish FLERVs was estimated to be ~263 (95% HPD = 195−342) Myr old. The separation of the salamander FLERV lineage was inferred to have happened ~348 (95% HPD = 251−478) Ma. The age of the entire clade of FVs/FLERVs was estimated to be ~455 (95% HPD = 304−684) Myr old.

Since our tree contains both extant retroviruses (which evolve under pure viral evolutionary rates) and ERVs (which have mixed rates of host and virus evolution) (**Figure 5.4**), we note that some branches in our tree have mixed rates of virus and host evolution, and because our model was built based on evolutionary dynamics of extant viruses, these mixed rates of ERV evolution could potentially bias our date estimates. Nevertheless, since we did not impose a molecular clock onto the tree (see **Materials and Methods**), each branch can therefore vary their length and height independently of time, allowing them to have different rates. Subsequently, those mixed rates of evolution will tend to limit to terminal branches that lead to ERVs. We avoided these branches with mixed rates of ERV evolution in our date estimation by measuring the node heights from SFVcpz tip down the tree, assuming that internal branches reflect only exogenous viral evolution and do not contain any evolutionary periods in which viruses might have been infectious ERVs.

As an independent verification that our date estimates are not biased by the mixed rates of ERV evolution, we performed phylogenetic analyses and date estimation by using FLERV Pol ancestral sequences, from which the host evolution has been removed. The alignment was checked for recombination, but no recombination was found (*P* value = 0.880). Indeed, we obtained similar results, estimating the divergence date of the lobe-finned fish FLERV lineage to be ~267 (95% HPD = 150−452) Ma, the separation date of the salamander FLERV lineage to be ~342 (95% HPD = 157−603) Ma, and the overall age of FVs/FLERVs to be ~473 (95%





C

Node	S (substitutions per site) Median	95% HPD	T (million years) Median	95% HPD	Ref
I	0.017	0.007-0.030	0.96	0.69-1.21	(30)
II	0.071	0.048-0.096	8.29	6.52-10.05	(31)
III	0.120	0.090-0.155	16.52	13.33-19.61	
IV	0.149	0.113-0.189	31.60	25.36-37.85	
V	0.286	0.216-0.365	43.43	38.39-48.08	
VII	0.295	0.224-0.383	98.59	95.94-100.78	(26)
X	0.723	0.582-0.887	266.52	149.92-451.77	NA
XI	0.856	0.666-1.051	341.68	156.83-602.73	
XII	1.058	0.874-1.248	472.83	224.91-896.91	

Figure 5.6 Phylogenetic analyses with maximum-likelihood ancestral sequences. (**A**) A Bayesian phylogeny of foamy virus (FV) and maximum-likelihood ancestral sequences of FV-like endogenous viruses (FLERVs), rooted according to the tree in **Figure 5.4**, left. Preceding FV names are accession numbers. Arabic numerals on nodes are Bayesian posterior probability node support values. Roman numerals indicate the nodes of which the total per-lineage substitutions to the chimpanzee simian FV (U14327 SFVcpz) were used to construct the time-dependent rate phenomenon (TDRP) model (see **B**). The numbers correspond to those in **Figure 5.4**, left. The scale bar is in units of amino-acid substitutions per site. (**B**) TDRP model and the timescales of FLERVs. The total per-lineage amino acid substitutions (S) from various nodes to SFVcpz are plotted against corresponding evolutionary timescales (T). The S and T estimates are labelled with Roman numerals, referring to nodes in **A**. Solid black dots are median estimates, and the associated 95% highest probability density (95% HPDs) intervals are indicated by error bars. The T estimates of node **I**−**V**, and **VII** were directly inferred from those of their hosts (26, 30, 31). 7,500 power-law TDRP models were fitted to the posterior distributions of the S and T estimates of the nodes **I**-**V**, and **VII** (grey lines). The median model parameter values, adjusted R^2 scores, and corresponding 95% HPDs (in the parentheses) are shown in the bottom right. The model was extrapolated to infer the branching date of lobe-finned fish FLERV lineage (node **X**), the separation date of salamander FLERV lineage (node **XI**), and the age of the entire FV/FLERV clade (node **XII**). (**C**) S estimates and T estimates used in the TDRP model reconstruction (above the line) and the inference of the timescales of FLERVs (below the line). The node numbers refer to those on the viral tree in **A**.

HPD = 225−897) Myr old (**Figure 5.6**). Overall, these similarities in the date estimates suggest that we have successfully avoided the problem of mixed rates of ERV evolution in terminal branches. We note that the date estimate uncertainties are significantly larger than those of our initial date estimates. This is to be expected however, as there were fewer data points available for the TDRP model calibration after the exclusion of endogenous FVs for which ancestral sequences could not be reconstructed due to the lack of data (nice data points in the initial analyses, **Figure 5.5**, compared to six data points in this analysis, **Figure 5.6**). In addition to this, it is likely that the ancestral sequence reconstruction may have also introduced some extra uncertainty into the genetic divergence estimation, contributing further to the increase in the date estimate uncertainties.

5.4 Discussion

Here, we report 36 lineages of novel FLERVs residing in the genomes of salamanders, a frog, ray-finned fish, lobe-finned fish and shark (**Table 5.1**, see also **Table S5.1** in **Electronic Appendix**). These FLERVs are phylogenetically basal to mammalian *true* FVs (**Figure 5.4**), and their functional exogenous viral counterparts are not currently available for direct examination. It is therefore still unclear whether or not we should classify them as FVs. Nonetheless, given the similarity observed between the protein sequences of mammalian FVs and those of salamander/lobe-finned fish FLERVs as well as their phylogenetic positions, we consider that classifying salamander and lobe-finned fish FLERVs as FVs may be appropriate. In contrast, ray-finned fish FLERVs are drastically different from mammalian FVs, both in terms of genome length and gene number. Indeed, to our knowledge, they are the largest known retroviruses ever documented, approximately twice as large as a typical simple retrovirus (~8−9 kb). Furthermore, none of their proteins, apart from Pol, exhibit strong similarity to those of mammalian FVs. These observations support that ray-finned fish FLERVs should be classified as a separate genus in the subfamily *Spumaretrovirinae*. Although the International Committee on Taxonomy of Viruses requires that a well characterised viral member must be available to establish a new genus, which may require exogenous viruses, we tentatively propose the name for these ray-finned fish FV-like viruses '*Gigantospumavirus'.* Our results

should stimulate efforts to search for novel retroviruses that might be circulating among modern-day fish and amphibians.

Our results extend the stable and long-term co-speciation history of FV-like viruses and their hosts beyond mammals (5, 6, 25), showing that the viral progenitors of ray-finned fish FLERVs also co-diverged broadly with their hosts. The co-evolutionary history of ray-finned fish FLERVs, however, is more dynamic than that of mammalian FVs, as indicated by their acquisition of additional accessory genes and that some ray-finned fish genomes even contain two distinct lineages of FLERVs. In the case of salamander FLERVs, we did not find significant evidence supporting an overall virus-host co-speciation; although, since our analyses contained only six lineages of salamander FLERVs, this could be due to the lack of sequence data and/or the low power of the analyses. Overall, our analyses have extended the co-speciation history of this major retroviral lineage from mammalian FVs to cover ray-finned fish FLERVs. To our knowledge, a co-speciation pattern at a scale this large has never been demonstrated in any virus, making this retroviral group a very valuable viral macroevolutionary model.

Our analyses show that FVs and FLERVs from hosts of the same class tend to cluster together, strongly supporting monophyletic clades of mammalian FVs, salamander FLERVs, lobe-finned fish FLERVs, and ray-finned fish FLERVs. This indicates that viral transmissions between vertebrates of different classes are rare, consistent with the overall co-speciation history of FVs/FLERVs. Nonetheless, two cases of cross-class viral transmissions were observed. We found that XtrFLERV, a frog FLERV, is placed within the clade of shark FLERVs, CmiFLERVs, with strong support and not with other amphibian (salamander) FLERVs. This finding suggests that the origin of XtrFLERV involved cross-class transmission of a FV-like virus from a shark to an amphibian. Nevertheless, our result does not preclude a possibility that this conflicting virus/host tree topology is a result of long-term incomplete lineage sorting. Thus, this particular finding should be interpreted with caution.

A more robust cross-class viral transmission involves lobe-finned fish FLERVs. Han and Worobey (12) reported the discovery of a lobe-finned fish FLERV, namely CoeEFV. They found that CoeEFV is placed basal to mammalian FVs and that there is a linear correlation (R^2 = 0.71) between viral and host divergence. Based on these findings, they proposed an ancient co-speciation of CoeEFV and mammalian FVs dating back ~409 Ma, implying a marine origin of FVs. Nevertheless, since there was only one lobe-finned fish FLERV (i.e. CoeEFV) in their study, robust statistical analyses could not be performed to verify the co-speciation, and limited data could obscure a history of viral cross-species transmissions.

At face value, our RT phylogeny seems to suggest that mammalian FVs and salamander FLERVs form a monophyletic clade, and lobe-finned fish FLERVs are basal to this clade, consistent with a co-speciation scenario (**Figure 5.1**). However, the support of this phylogenetic pattern is low (bootstrap clade support = 49%, and Bayesian posterior probability = 0.88), and thus, it should not be over interpreted. Indeed, our phylogenetic analyses of longer Pol protein sequences reveal that lobe-finned fish FLERVs are in fact within the clade of mammalian FVs and salamander FLERVs with robust support (posterior probability > 0.99), which in fact is within the uncertainty of the RT phylogeny. On the basis of the well-supported Pol tree topology, we estimated that lobe-finned fish FLERVs diverged from the mammalian FVs ~263−267 Ma. Overall, this is in conflict with the host diversification pattern, where lobe-finned fish are more basal than amphibians, diverging from mammals ~409 Ma (32) and ~335 Ma (33), respectively. Unlike in the case of XtrFLERV, the fact that the divergence date of lobe-finned fish FLERVs and mammalian FVs is much younger than the hosts' divergence date excludes the possibility of incomplete lineage sorting, and that the viruses truly did not co-speciate with their hosts. Thus, both the phylogenetic placement and temporal origins of CoeEFV are neither consistent with co-speciation nor do they provide evidence for a marine origin of FVs. Instead, the branching date of lobe-finned fish FLERVs falls within the period when therapsids—a group of mammal-like reptiles—diversified and dominated the land (34). Together, our results suggest that lobe-finned fish FV-like viruses do not have a marine origin,

but instead originated from one or a series of cross-class transmissions that happened in the middle Permian, perhaps ultimately from a prehistoric therapsid to a lobe-finned fish.

On the basis of the Pol phylogeny, we estimated the branching of salamander FLERVs to have happened ~342−348 Ma, highly comparable to that of their amphibian hosts ~335 Ma (33), and the age of FVs/FLERVs to be ~455−473 Myr old, strikingly similar to the timescale of their jawed vertebrate hosts, ~465 Myr old (33). Combined with the observed phylogenetic pattern, these date estimates support the hypothesis of a marine origin of FVs. Our results strongly suggest that this group of viruses originated together with their jawed vertebrate hosts in the Ordovician ocean, and underwent a water-to-land evolutionary transition with their hosts, co-evolving with one another for >450 Myr. To our knowledge, this is the oldest date ever directly inferred for any viruses, greater than the age estimate of the oldest known dsDNA viruses by 140 Myr (310 Myr old (35)).

Our results offer key insights into the early history of retroviral evolution as a whole. While retroviruses can be found throughout marine and terrestrial vertebrates, with fish and marine tetrapod ERVs sometimes shown to be basal (1, 2), robust temporal and phylogenetic evidence of their ancient origins has been lacking. The discovery of ancient orthologous ERVs has shed some light on the origin of retroviruses. However, the oldest known orthologous ERVs are only 100 Myr old, providing evidence of their existence during the early diversification of mammals (4). Virus-host co-speciation is another source of information that can be integrated into analyses to infer the age of viruses, but again, the oldest retroviral age estimate that was derived based on virus-host co-speciation pattern is only 100 Myr old (5, 6). Thus, given the ability of retroviruses to jump across distantly related host species (1, 2), no robust evidence of the existence of retroviruses prior to 100 Ma has previously been available and they therefore could be considerably younger than their vertebrate hosts.

By integrating multiple sources of data, evaluating the cospeciation history of FVs, FLERVs, and their hosts across multiple vertebrate classes and incorporating the TDRP model, we were

able to show that this lineage of viruses is ~455−473 Myr old. Since the most recent common ancestor of all retroviruses must be older than that of FVs/FLERVs, our finding provides the first concrete evidence that retroviruses as a whole must be older than jawed vertebrate hosts, extending the oldest age estimate of retroviruses by ~350 Myr years. Furthermore, we also found that fish FLERVs are positioned towards the root of the tree, pointing towards a marine origin of retroviruses. Together, our analyses provide evidence, both phylogenetically and temporally, that retroviruses emerged together with their vertebrate hosts in the ocean, ~460−550 Ma (7–9), in the early Palaeozoic Era, if not earlier.

5.5 Materials and methods

5.5.1 Screening for FV-like sequences

Five publically accessible nucleotide databases were screened for FV-like endogenous retroviruses (FLERVs) by using tBLASTn and the CoeEFV reverse transcriptase (RT) protein as a probe, including the database of GenBank non-redundant nucleotide sequences (nr), expressed sequence tags (est), high throughput genomic sequences (HTGS), whole genome shotgun sequences (wgs), and transcriptome shotgun assembly sequences (TSA). The scope of the screening was limited to vertebrates, excluding mammals (taxid: 40674).

The screening was performed in a stepwise manner. In the first iteration, the number of target contigs was set to 50. Frame-shift mutations were checked, and the resulting multiple BLAST hits were combined as one. When there were more than three RT sequences on one contig, only the best three (ranked under the default tBLASTn settings) were kept for further analyses. Together with RT sequences of known alpha-, beta-, gamma-, delta- epsilon-, lenti-, and spuma-retroviruses, the phylogenetic relationships of these RT sequences were estimated under the maximum-likelihood (ML) framework by using RAxML (36). The RTREV + CAT amino acid substitution model with 25 per site rate categories was employed to construct the phylogeny. Bootstrap clade support values were calculated using 1,000 pseudoreplicates. We used this RT tree to broadly classify viral sequences into various established retrovirus genera.

On the basis of the best estimated ML phylogeny, FV-like RT sequences were identified. The criterion was that they form a monophyletic clade with the RT sequences of mammalian FVs and fish ERVs that have been recognised as FV-like, including CoeEFV, DrFV-1, and platyfishEFV.

To retrieve more RT sequences, the screening process was then repeated with increasing target contig numbers (100, 250, 500, and 1,000) until no additional species were identified as containing FV-like RT sequences. In total, we identified 193 RT sequences as FV-like (**Table S5.1** in **Electronic Appendix**). Potential recombination among these sequences was checked by using a quartet-based recombination detection program VisRD3 (37). 20,000 quartets were randomly sampled to compute the recombination statistic with the extended statistical geometry weighting option. The recombinant-free null distribution of the recombination statistic was built based on 1,000 sets of randomly-shuffled sequences. The window size and the step size was 10% (13 aa) and 5% (6 aa) of the alignment length, respectively. The results showed no significant evidence for recombination. Furthermore, for the final tree, a Bayesian maximum clade credibility phylogeny was also reconstructed to confirm the result, by using MrBayes 3.2.1 (38). The RTREV+I+Γ(4) amino acid substitution model was used, and no molecular clock was imposed. Two chains of MCMC were run for 10,000,000 steps, and parameters were logged every 1,000 steps with the initial 25% discarded as burn in. A metropolis coupling algorithm (3 hot and 1 cold chains) was applied to improve the sampling. Potential scale reduction factors of all parameters were ~1.000, indicating that they were all well sampled from their posterior distributions and had converged. The alignment (130 aa, 251 sequences) is available in **Electronic Appendix**, **Data set S5.1**. The RT tree is shown in **Figure 5.1** and **Figure S5.1** in **Electronic Appendix** (clade support values are shown only for major clades, and nodes with Bayesian posterior support of <0.5 are not annotated).

On the basis of the identification of long terminal repeats (LTRs) and target site duplications, several potentially full-length FLERVs were found in the eastern newt (*Notophthalmus viridescens*; NviFLERV) and midas cichlid (*Amphilophus citrinellus*; AciFLERV). To search for

more FLERVs, we used the RT sequences of NviFLERV, AciFLERV, and SloEFV to screen the five databases again as described above. The numbers of the target contigs were set to 100 for the nr, est, HTGS, and TSA database screening, and 1,000 for the wgs database screening. These settings were used as they were sufficient to identify all species containing FV-like RT sequences in the first round of screening, in which the RT sequence of CoeEFV was used as a probe. BLAST hits from additional species were added to the phylogenetic analyses to determine whether or not they are FV-like. By doing so, we discovered several more potentially full-length FLERVs in annual killifish (*Austrofundulus limnaeus*; AliFLERV).

5.5.2 Characterisation of novel FV-like elements

We first characterised NviFLERVs, AciFLERVs, and AliFLERVs, which are full-length elements. Since we were able to obtain 9 and 23 elements of AciFLERVs and AliFLERVs, we inferred their ML ancestral sequences by using MEGA 6 (39) with the GTR+Γ(4) nucleotide substitution model, and used them for genome annotation. The ancestral sequences of the LTRs were inferred based on both 5'− and 3'−LTRs. The alignments used for the ancestral sequence inferences are available in the **Electronic Appendix** (AciFLERVs: **Data set S5.2**, AliFLERVs: **Data set S5.3**). Potential protein coding regions were identified based on the distribution of stop and start codons, determined by Open Reading Frame Finder (http://www.ncbi.nlm.nih.gov/gorf/orfig.cgi), and were annotated via protein sequence similarity by using BLASTp. Similarity of primer binding sites to tRNAs was inferred by using the tRNA database (trna.bioinf.uni-leipzig.de). To estimate the age of NviFLERV, AciFLERV, and AliFLERV, we used the LTR-dating method. MEGA 6 (39) was used to calculate the number of nucleotide substitutions between available paired LTRs under the Tamura-Nei DNA evolutionary model, which takes the differences between transition and transversion substitution rates as well as the inequality of nucleotide frequencies into account. We then divided the distance between paired LTRs by two to obtain a total per-lineage LTR substitution estimate, which in turn was divided by the rate of the LTR evolution to derive the time it took to accumulate the observed number of substitutions, i.e. their age. We assumed that the LTR rate of evolution is equal to the neutral rate of the host genome. Finally, we used the annotated genomes of NviFLERV,

AciFLERV, and AliFLERV, as well as those of CoeEFV and DrFV-1 to characterise other FLERVs in other vertebrates (**Table S5.1** in **Electronic Appendix**) via protein sequence similarity by using tBLASTn.

5.5.3 Phylogenetic analyses

To investigate phylogenetic relationships between FVs and FLERVs in more detail, we estimated their phylogeny based on a manually-curated Pol protein alignment (580 aa; **Data set S5.4** in **Electronic Appendix**). Gag and Env proteins were not used because, in the case of amphibian FLERVs, they could not be concatenated with the Pol proteins as they were found on different contigs in all cases, except only for those of NviEELRV-1. In the case of fish FLERVs, they were not used because they could not be aligned to those of mammalian FVs due to high sequence divergence. As a result, only Pol proteins were used to reconstruct the phylogeny. The alignment contains only one FLERV per monophyletic clade of FLERVs of the same host species, determined by the initial RT phylogenetic analyses. Class III retroviruses were used to root the tree. Potential recombination within the alignment (excluding the outgroup) was checked by using VisRD3 (37) as described above. The window size and the step size was 10% (58 aa) and 5% (29 aa) of the alignment length, respectively. The results showed no significant evidence for recombination. The phylogeny was then estimated under the Bayesian framework by using MrBayes 3.2.1 (38) with the best amino acid substitution model, LG+I+Γ(4)+F, determined under the AICc criterion by PartitionFinder 1.1.1 (40). No molecular clock was imposed (that is, a non-clock, and unconstrained tree). Two chains of MCMC were run for 10,000,000 steps, and parameters were logged every 1,000 steps with the initial 25% discarded as burn in. A metropolis coupling algorithm (3 hot and 1 cold chains) was applied to improve the MCMC sampling. Potential scale reduction factors of all parameters were ~1.000, indicating that they were all well sampled from their posterior distributions and had converged.

Furthermore, since the genome of the smooth newt (*Lissotriton vulgaris*) contains only a Gag coding region (see **Results: Characterisation of fragmented FLERVs**), it was not included

in this analyses. To estimate its phylogenetic position, a separate Gag phylogeny of salamander FLERVs was estimated by using the same method. The best amino acid substitution model was determined to be WAG+Γ(4)+F under the AICc criterion by PartitionFinder 1.1.1 (40). Potential scale reduction factors of all parameters were ~1.000. The alignment of Gag (305 aa) is also available in Electronic Appendix (**Data set S5.5**). Again, potential recombination within the alignment was checked by VisRD3 (37). The recombination statistic was computed with the extended statistical geometry weighting option based on all possible 330 quartets, and the null distribution was computed by using 1,000 sets of randomly-shuffled sequences. The window size and the step size was 30 aa and 15 aa, respectively. The results showed no significant evidence for recombination.

5.5.4 Co-speciation analyses

Viral-host co-speciation history was evaluated by comparing the topologies of the viral and host trees by using Jane V4.01 (41). The analyses were divided into three sub analyses: those of mammalian FVs, ray-finned fish FLERVs, and salamander FLERVs. Since there were only two lobe-finned fish and one shark species that contain FLERVs, analyses for these two lineages were not performed. Nodes with less than 0.80 Bayesian posterior probability support were collapsed with their most adjacent basal node to form a polytomy. The host trees used in these analyses are published host trees (26–29). The number of co-speciation events was calculated by using a genetic algorithm, with the number of generations, and population set to 100. To maximise the number of co-speciation events inferred, the (vertex-based) costs were set as follows: co-speciation = −1, duplication = 0, duplication and host switch = 0, loss = 0, and failure to diverge = 0. To assess the probability of observing the inferred number of co-speciation events by chance, the random tip mapping method implemented in Jane v4.01 (41) was used (generation = 100, population = 100, sample size = 500).

5.5.5 Estimating evolutionary timescale of FLERVs

For each of the estimated posterior Bayesian Pol phylogenies, a power-law model describing the relationship between mammalian FV total per-lineage amino acid substitutions (S) and evolutionary timescales (T) was constructed by tracing the chimpanzee simian FV lineage (U14327 SFVcpz) backward in time (**Figure 5.4,** nodes **I−IX**). The T estimates were directly inferred from those of their hosts (26, 30, 31) under the well-established FV-host co-speciation assumption (5, 6) (**Table 5.2,** nodes **I−IX**). The model fitting was performed by using the *lm* function implemented in R 3.2.1 (42). The values of S and T were log-transformed (base 10) prior to the linear model fitting. The model was then extrapolated to estimate the timescale of other nodes from their S estimates. This includes the separation date of the lobe-finned fish FLERV lineage (**Figure 5.4** and **Table 5.2**, node **X**), the separation date of the salamander FLERV lineage (**Figure 5.4** and **Table 5.2**, node **XI**), and the age of the entire clade of FVs/FLERVs (**Figure 5.4** and **Table 5.2**, node **XII**). This process was applied to all of the 7,500 posterior Bayesian phylogenies to obtain their full posterior distributions (**Figure 5.5** and **Table 5.2**).

5.5.6 Phylogenetic analyses by using ML FLERV ancestral sequences

To assess the effects of the mixed rates of virus–host evolution of FLERVs on our date estimation, we also performed phylogenetic analyses by using ML FLERV ancestral sequences. For a particular FLERV lineage, its *pol* nucleotide ancestral sequence was inferred by using MEGA 6 (39), with the GTR+Γ(4) nucleotide substitution model. The immediate outgroup of the clade was used to root the tree. FLERV lineages that consist of one sequence sample were excluded from analyses as ML ancestral reconstruction requires more than one sequence, unless samples from the same host genus (but different species) exist. In such cases, they would be grouped with the other FLERV lineage and their ancestor would be inferred together. In total 23 ML FLERV ancestral sequences were reconstructed. Their Pol protein (629 aa) phylogeny, together with those of extant mammalian FVs, was estimated by using MrBayes 3.2.1 (38) as described above, with the best-fit RTREV+I+Γ(4)+F substitution

model, determined under the AICc criterion by PartitionFinder 1.1.1 (40). Similar to the above, potential recombination within the alignment was checked with VisRD3 (37). 20,000 quartets were randomly sampled to compute the recombination statistic with the extended statistical geometry weighting option. The null distribution of the recombination statistic was computed based on 1,000 sets of randomly-shuffled sequences. The window size and the step size was 63 aa and 31 aa, respectively. The results showed no significant evidence for recombination. The alignment is available in **Electronic Appendix** (**Data set S5.6**). The tree was rooted according to the tree in **Figure 5.4**, left. The position of the root was confirmed by using several Class III retroviruses as outgroups. The TDRP analyses were subsequently performed as described above to estimate the branching date of the lobe-finned fish FLERV lineage, and that of the salamander FLERV lineage, as well as the age of the entire clade of FVs/FLERVs.

5.6 References

1. **Hayward A**, **Cornwallis CK**, **Jern P**. 2015. Pan-vertebrate comparative genomics unmasks retrovirus macroevolution. Proc Natl Acad Sci U S A **112**:464–469.
2. **Herniou E**, **Martin J**, **Miller K**, **Cook J**, **Wilkinson M**, **Tristem M**. 1998. Retroviral diversity and distribution in vertebrates. J Virol **72**:5955–5966.
3. **Aiewsakun P**, **Katzourakis A**. 2015. Endogenous viruses: connecting recent and ancient viral evolution. Virology **479**–**480**:26–37.
4. **Lee A**, **Nolan A**, **Watson J**, **Tristem M**. 2013. Identification of an ancient endogenous retrovirus, predating the divergence of the placental mammals. Philos Trans R Soc Lond B Biol Sci **368**:20120503.
5. **Katzourakis A**, **Gifford RJ**, **Tristem M**, **Gilbert MTP**, **Pybus OG**. 2009. Macroevolution of complex retroviruses. Science **325**:1512.
6. **Katzourakis A**, **Aiewsakun P**, **Jia H**, **Wolfe ND**, **LeBreton M**, **Yoder AD**, **Switzer WM**. 2014. Discovery of prosimian and afrotherian foamy viruses and potential cross species transmissions amidst stable and ancient mammalian co-evolution. Retrovirology **11**:61.
7. **Kumar S**, **Hedges SB**. 1998. A molecular timescale for vertebrate evolution. Nature **392**:917–920.
8. **Hedges SB**, **Blair JE**, **Venturi ML**, **Shoe JL**. 2004. A molecular timescale of eukaryote evolution and the rise of complex multicellular life. BMC Evol Biol **4**:2.
9. **dos Reis M**, **Thawornwattana Y**, **Angelis K**, **Telford MJ**, **Donoghue PCJ**, **Yang Z**. 2015. Uncertainty in the timing of origin of animals and the limits of precision in molecular timescales. Curr Biol **25**:2939–2950.
10. **Han G-Z**, **Worobey M**. 2012. An endogenous foamy virus in the aye-aye (*Daubentonia madagascariensis*). J Virol **86**:7696–7698.
11. **Han G-Z**, **Worobey M**. 2014. Endogenous viral sequences from the Cape golden mole (*Chrysochloris asiatica*) reveal the presence of foamy viruses in all major placental mammal clades. PLoS One **9**:e97931.
12. **Han G-Z**, **Worobey M**. 2012. An endogenous foamy-like viral element in the coelacanth genome. PLoS Pathog **8**:e1002790.

13. **Schartl M**, **Walter RB**, **Shen Y**, **Garcia T**, **Catchen J**, **Amores A**, **Braasch I**, **Chalopin D**, **Volff J-N**, **Lesch K-P**, **Bisazza A**, **Minx P**, **Hillier L**, **Wilson RK**, **Fuerstenberg S**, **Boore J**, **Searle S**, **Postlethwait JH**, **Warren WC**. 2013. The genome of the platyfish, *Xiphophorus maculatus*, provides insights into evolutionary adaptation and several complex traits. Nat Genet **45**:567–572.

14. **Llorens C**, **Muñoz-Pomer A**, **Bernad L**, **Botella H**, **Moya A**. 2009. Network dynamics of eukaryotic LTR retroelements beyond phylogenetic trees. Biol Direct **4**:41.

15. **Aiewsakun P**, **Katzourakis A**. 2016. Time-dependent rate phenomenon in viruses. J Virol **90**:7184–7195.

16. **Aiewsakun P**, **Katzourakis A**. 2015. Time dependency of foamy virus evolutionary rate estimates. BMC Evol Biol **15**:119.

17. **Duchêne S**, **Holmes EC**, **Ho SYW**. 2014. Analyses of evolutionary dynamics in viruses are hindered by a time-dependent bias in rate estimates. Proc Biol Sci **281**:20140732.

18. **Duchêne S**, **Lanfear R**, **Ho SYW**. 2014. The impact of calibration and clock-model choice on molecular estimates of divergence times. Mol Phylogenet Evol **78**:277–289.

19. **Muñoz-López M**, **García-Pérez JL**. 2010. DNA transposons: nature and applications in genomics. Curr Genomics **11**:115–128.

20. **Katzourakis A**, **Tristem M**, **Pybus OG**, **Gifford RJ**. 2007. Discovery and analysis of the first endogenous lentivirus. Proc Natl Acad Sci U S A **104**:6261–6265.

21. **Crawford AJ**. 2003. Relative rates of nucleotide substitution in frogs. J Mol Evol **57**:636–641.

22. **Bell CM**, **Connell BJ**, **Capovilla A**, **Venter WDF**, **Stevens WS**, **Papathanasopoulos M a**. 2007. Molecular characterization of the HIV type 1 subtype C accessory genes *vif*, *vpr*, and *vpu*. AIDS Res Hum Retroviruses **23**:322–330.

23. **Jaillon O**, **Aury J-M**, **Brunet F**, **Petit J-L**, **Stange-Thomann N**, **Mauceli E**, **Bouneau L**, **Fischer C**, **Ozouf-Costaz C**, **Bernot A**, **Nicaud S**, **Jaffe D**, **Fisher S**, **Lutfalla G**, **Dossat C**, **Segurens B**, **Dasilva C**, **Salanoubat M**, **Levy M**, **Boudet N**, **Castellano S**, **Anthouard V**, **Jubin C**, **Castelli V**, **Katinka M**, **Vacherie B**, **Biémont C**, **Skalli Z**, **Cattolico L**, **Poulain J**, **De Berardinis V**, **Cruaud C**, **Duprat S**, **Brottier P**,

Coutanceau J-P, **Gouzy J**, **Parra G**, **Lardier G**, **Chapple C**, **McKernan KJ**, **McEwan P**, **Bosak S**, **Kellis M**, **Volff J-N**, **Guigó R**, **Zody MC**, **Mesirov J**, **Lindblad-Toh K**, **Birren B**, **Nusbaum C**, **Kahn D**, **Robinson-Rechavi M**, **Laudet V**, **Schachter V**, **Quétier F**, **Saurin W**, **Scarpelli C**, **Wincker P**, **Lander ES**, **Weissenbach J**, **Roest Crollius H**. 2004. Genome duplication in the teleost fish *Tetraodon nigroviridis* reveals the early vertebrate proto-karyotype. Nature **431**:946–957.

24. **Kijima TE**, **Innan H**. 2010. On the estimation of the insertion time of LTR retrotransposable elements. Mol Biol Evol **27**:896–904.

25. **Switzer WM**, **Salemi M**, **Shanmugam V**, **Gao F**, **Cong M-E**, **Kuiken C**, **Bhullar V**, **Beer BE**, **Vallet D**, **Gautier-Hion A**, **Tooze Z**, **Villinger F**, **Holmes EC**, **Heneine W**. 2005. Ancient co-speciation of simian foamy viruses and primates. Nature **434**:376–380.

26. **Bininda-Emonds ORP**, **Cardillo M**, **Jones KE**, **MacPhee RDE**, **Beck RMD**, **Grenyer R**, **Price SA**, **Vos RA**, **Gittleman JL**, **Purvis A**. 2007. The delayed rise of present-day mammals. Nature **446**:507–512.

27. **Hertwig ST**. 2008. Phylogeny of the Cyprinodontiformes (Teleostei, Atherinomorpha): the contribution of cranial soft tissue characters. Zool Scr **37**:141–174.

28. **Betancur-r R**, **Broughton RE**, **Wiley EO**, **Carpenter K**, **López JA**, **Li C**, **Holcroft NI**, **Arcila D**, **Sanciangco M**, **Ii JCC**, **Zhang F**, **Campbell MA**, **Ballesteros JA**, **Roa-varon A**, **Willis S**, **Borden WC**, **Hough DJ**, **Lu G**. 2013. The tree of life and a new classification of bony fishes. PLOS Curr Tree Life **5**:1–45.

29. **Zhang P**, **Papenfuss TJ**, **Wake MH**, **Qu L**, **Wake DB**. 2008. Phylogeny and biogeography of the family Salamandridae (Amphibia: Caudata) inferred from complete mitochondrial genomes. Mol Phylogenet Evol **49**:586–597.

30. **Stone AC**, **Battistuzzi FU**, **Kubatko LS**, **Perry GH**, **Trudeau E**, **Lin H**, **Kumar S**. 2010. More reliable estimates of divergence times in *Pan* using complete mtDNA sequences and accounting for population structure. Philos Trans R Soc Lond B Biol Sci **365**:3277–3288.

31. **Perelman P**, **Johnson WE**, **Roos C**, **Seuánez HN**, **Horvath JE**, **Moreira MAM**, **Kessing B**, **Pontius J**, **Roelke M**, **Rumpler Y**, **Schneider MPC**, **Silva A**, **O'Brien SJ**,

Pecon-Slattery J. 2011. A molecular phylogeny of living primates. PLoS Genet **7**:e1001342.

32. **Johanson Z**, **Long J a**, **Talent J a**, **Janvier P**, **Warren JW**. 2006. Oldest coelacanth, from the early Devonian of Australia. Biol Lett **2**:443–446.
33. **Broughton RE**, **Betancur-R. R**, **Li C**, **Arratia G**, **Ortí G**. 2013. Multi-locus phylogenetic analysis reveals the pattern and tempo of bony fish evolution. PLoS Curr 1–38.
34. **Benson RBJ**. 2012. Interrelationships of basal synapsids: cranial and postcranial morphological partitions suggest different topologies. J Syst Palaeontol 1–24.
35. **Thézé J**, **Bezier A**, **Periquet G**, **Drezen J-M**, **Herniou E a.** 2011. Paleozoic origin of insect large dsDNA viruses. Proc Natl Acad Sci **108**:15931–15935.
36. **Stamatakis A**, **Hoover P**, **Rougemont J**. 2008. A rapid bootstrap algorithm for the RAxML Web servers. Syst Biol **57**:758–771.
37. **Lemey P**, **Lott M**, **Martin DP**, **Moulton V**. 2009. Identifying recombinants in human and primate immunodeficiency virus sequence alignments using quartet scanning. BMC Bioinformatics **10**:126.
38. **Ronquist F**, **Huelsenbeck JP**. 2003. MrBayes 3: Bayesian phylogenetic inference under mixed models. Bioinformatics **19**:1572–1574.
39. **Tamura K**, **Stecher G**, **Peterson D**, **Filipski A**, **Kumar S**. 2013. MEGA6: molecular evolutionary genetics analysis version 6.0. Mol Biol Evol **30**:2725–2729.
40. **Lanfear R**, **Calcott B**, **Kainer D**, **Mayer C**, **Stamatakis A**. 2014. Selecting optimal partitioning schemes for phylogenomic datasets. BMC Evol Biol **14**:82.
41. **Conow C**, **Fielder D**, **Ovadia Y**, **Libeskind-Hadas R**. 2010. Jane: a new tool for the cophylogeny reconstruction problem. Algorithms Mol Biol **5**:16.
42. **R Core Team**. 2014. R: a Language and environment for statistical computing. R Found Stat Comput **1**:409.

Chapter six: Conclusions and final thoughts

6.1 Conclusions and final thoughts

As highlighted throughout the entire thesis, the discovery of endogenous viral elements (EVEs), also known as *viral genomic fossils*, provides us with unique opportunities, enabling us to directly examine how extinct viruses might have evolved in their distant past. EVEs often fill in the gaps in the analyses of extant viruses that can only inform us about the past of their ancestors and not those of extinct viruses that do not leave any modern-day decedents. The focus of this thesis is to investigate how foamy viruses (FVs) and their (ancient) relatives evolved and interacted with their hosts over geological timescales, how the rates of viral evolution appear to vary temporally, and finally, by using FVs as an evolutionary model, this thesis provides concrete evidence for the ancient origin of retroviruses.

The research conducted in this thesis has provided numerous insights into the long-term evolutionary history of FVs. Together with the work of others, it is now clear that FVs are, or at least were, capable of infecting the entire clade of eutherians (1, 2). Chapter one, in particular, reveals that FVs have an extended co-speciation history with mammals, dating back ~100 million years (Myr) ago (Ma) to the basal radiation of eutherians. The work conducted in Chapter five, through analyses of FV-like endogenous retroviruses (FLERVs) obtained from a wide range of vertebrates, including amphibians, lobe-finned fish, ray-finned fish, and shark, provides evidence supporting that this major retroviral lineage indeed originated together with their jawed vertebrate hosts in the Ordovician ocean, and they have been co-evolving and co-diverging in concert with one another for >450 Myr. Regarding the host range of these viruses, there are only two vertebrate groups left of which the evidence for FV/FV-like infection is still absent, namely reptiles and birds. The discovery of reptile and bird FVs and/or FV-like viruses, if they indeed exist, would fill in the gaps in our analyses, completing the picture of how this major retroviral group co-evolve with their vertebrate hosts. Despite the lack of these particular data, the work in Chapter five offers the first direct temporal evidence in support of an ancient marine origin of retroviruses, excluding the possibility that they are young but are wide spread among animals due to frequent cross-species transmissions. Indeed, retroviruses as a whole are likely as old as their vertebrate hosts, if not older, originating about half a billion years ago.

It is important to note that the extended co-speciation pattern of FVs/FLERVs is by no means absolute. For example, as demonstrated in Chapter two, although New World monkey FVs appear to co-diverge with their hosts at the level of host species and sub-population, there is no evidence in support of their co-speciation pattern at the level of host genera. Similarly, although a broad co-speciation history between the ray-finned fish and their FV-like viruses could be inferred in Chapter five, a substantial number of viral cross-species transmissions were also evident. In fact, not only can they be transmitted between closely related animals, FVs and the progenitors of FLERVs appear to be capable of jumping between distantly related host species as well. This is evidenced by the ancient FV cross order transmissions of the prosimian aye-aye FV, and the Chinese bat FV, highlighted in Chapter one, which involved movement over a great geographical distance, travelling from Africa to Eurasia via the Indian-landmass continental drift. Viral transmissions between vertebrates of different classes crossing the terrestrial-marine barrier were also observed in Chapter five, including one ultimately from a shark to a frog and another one from a mammal-like reptile therapsid to a lobe-finned fish.

These transmissions of FVs and of FV-like viruses might appear to be anomalies against the clear background of their extended co-speciation pattern; however, I would like to point out that they are far from unexpected. An *in vitro* study has demonstrated that an ape FV, namely prototype FV, is capable of infecting many cell types obtained from a wide range of animals, including Old World monkeys, rodents, carnivores, dolphins, pigs, ruminants, birds and reptiles (3). It also has been shown that FVs of different species likely utilise the same, perhaps promiscuous, receptor(s) for cell-attachment and/or entry (4–6). Under the light of these findings, it is surprising that FVs and FV-like viruses can maintain such an incredibly long co-speciation history with their hosts, and it is unclear why they did not cross transmit more frequently than this in nature. Indeed, FV transmissions among closely related host species are not rare, as evidenced by numerous reports of zoonotic infections of human with Old World monkey and ape FVs (7–16). FV cross-species transmissions among Old World monkeys and

apes are also frequent (9, 17, 18). Despite these observations, a strong co-speciation pattern for Old World monkey and ape FVs could still be found (18), indicating that successful onwards FV cross-species transmissions are rare. This is further supported by a lack of FV secondary transmissions from infected persons to their parents, spouses, and/or children despite frequent physical and/or sexual contacts (7, 12–16). Further analyses are required to shed more light on these discrepancies between what we observe in nature and the results from the lab, in particular, what prevents FV secondary transmissions from happening after the host switches. This question warrants further investigation, having implication for our understanding of virus emergence and transmission.

These observations of FV cross-species transmissions should not discourage the use of FVs as a model of viruses that co-speciate with their hosts however. Chapter three, in particular, shows that, on a broad scale, the pattern of FV co-speciation is itself internally consistent, supporting the validity of the FV co-speciation model. Nonetheless, since the co-speciation pattern is not absolute, it is important that future analyses of FV and FLERV evolution infer their co-speciation history with care, especially when the data is limited, otherwise wrong conclusions might be obtained. As highlighted in Chapter five, the best example of this is perhaps the falsely claimed virus-host co-speciation event involving the progenitor of coelacanth endogenous FV (19). Indeed, the study in this thesis shows that the history of its cross-species transmission was hidden by the lack of data, in particular, those of amphibian FLERVs. The work here emphasises and reiterates the importance of testing the virus-host co-speciation hypothesis by checking that the virus and host divergence dates are consistent. This work also highlights the pitfalls of the use of co-phylogeny analysis and the positive correlation between virus and host genetic distances in the inference the co-speciation pattern.

Another major result in this thesis is the discovery that viral evolutionary rate estimates appear to vary systematically over time. At the beginning of my research, it was clear that viral evolutionary studies that span short timescales like epidemiological studies tend to yield evolutionary rate estimates that are much greater than those obtained from virus-host co-

divergence analyses and/or paleovirological studies, which span geological timescales (18, 20–24). Furthermore, at the time, it had been repeatedly reported that evolutionary rate estimates of cellular organisms appear to co-vary with the timescale over which they are measured, strongly negatively correlated with one another (25). However, determining the precise pattern of the rate decay process in cellular organisms was difficult, hindered by the lack of data. In particular, such analyses require multiple evolutionary rate estimates calculated by using multiple independent calibration points, which were, and still are, difficult to obtain. Inspired by these findings, I decided to investigate the rate dynamics of FVs.

Although not absolute, the extended FV-mammal co-speciation pattern, coupled with the wealth of FV molecular data, allowed us to estimate the rate of FV evolution over various timescales, spanning 7 orders of magnitude. This in turn enabled us to describe and examine the rate dynamics of FVs in unprecedented detail. Chapter three shows that mammalian FVs appear to evolve increasingly slowly the further back in time we look, and their rate-decay pattern can be best described by a simple power-law decay function. Throughout this thesis, this phenomenon is referred to as the '*time-dependent rate phenomenon'* (TDRP). This finding has led to the establishment of the TDRP as a general evolutionary feature of all viruses in Chapter four, which enabled the timescales of FLERVs to be estimated. Furthermore, once the TDRP is accounted for, evolutionary rates of different viral groups can be clearly distinguished from one another, a distinction that has become more and more blurred (26–29). For example, when the timescales of the rate estimation are controlled for, the analyses show that double-stranded DNA viruses evolve slower than single-stranded DNA and reverse-transcribing viruses by ~1 order of magnitude, which in turn evolve slower than single-stranded RNA viruses by ~1 order of magnitude. The TDRP represents the bridge that connects recent and ancient viral evolution, allowing viral evolutionary rate estimates to be appropriately transferred over different timescales in viral evolutionary inference, and also reconciling paleovirological date estimates with analyses of extant viruses. Chapter four also shows that the rates of evolution of all viruses seem to decay at the same speed, decreasing by approximately half for every 3-fold increase in the measurement timescale. I found this

particular result very surprising, given the great diversity of virus replication mechanisms, genomic materials, and natural histories. This finding supports that the underlying causes of the rate-decay process are likely general natural phenomena that all viruses encounter.

While I find it is remarkable that the TDRP can be described very well by a power-law decay function, I do not wish to leave an impression that this simple mathematical model can fully describe the TDRP. Indeed, the power-law model was found to be the best TDRP model, but only among the four investigated models (with the other three being (i) a vertically-translated exponential rate decay model, (ii) a simple exponential rate decay model and (iii) a vertically-translated power-law rate decay model). The TDRP is likely much more complex than a simple power-law rate-decay phenomenon, governed by several biological processes that are yet to be determined. Some of the possible underlying causes of the TDRP are highlighted in this thesis, including the bottlenecking nature of viral transmission, and the highly-subdivided population structure of viruses, as well as the long-term evolutionary constraints on viral genomes, required for proper biological functions and host interaction. It is unclear how these biological processes interact to give rise the TDRP, or indeed whether they can give rise to the TDRP at all, and if so, under which conditions. Much further work is still required to elucidate the finer detail of the TDRP. Analyses of simulated molecular sequences, of which every aspect of their evolution can be precisely controlled, including temporal fluctuation of population size, rate of evolution, and the proportion of neutral, conserved, and adaptive sites, as well as the strength of selection pressure, might offer us a better understanding of the emergence of the TDRP, and in turn, allow us to better construct a more precise description of this phenomenon.

While our understanding of the TDRP is still far from complete, the research presented in Chapter three nevertheless demonstrates that the power-law TDRP model outperforms all currently existing evolutionary rate models, estimating FV evolutionary timescales that are consistent with their co-speciation history. Indeed, by using the TDRP model, analyses in Chapter four could for the first time recover evolutionary timescales of modern lentiviruses that

are consistent with paleovirological evidence and our knowledge of host migration, reconciling analyses of extant viruses and paleovirological studies. Furthermore, as shown in Chapter five, by accounting for the TDRP, we were able to provide both temporal and phylogenetic evidence in support of an ancient oceanic origin of retroviruses, ruling out the possibility that they are young but are widespread among vertebrates because of frequent horizontal transmissions. These results, hopefully, will serve to encourage greater adoption of the TDRP model in viral divergence time estimation in the future.

Although the method of incorporating the TDRP into viral evolutionary analyses described here was relatively crude, it nonetheless provides an important step in connecting recent and ancient viral evolution. Indeed, the date estimates obtained from the analyses are of high uncertainty, likely because they were inferred from the number of substitutions separately from the inference of evolutionary rate, and phylogenetic topologies. In the future, it should be possible to develop a TDRP model under a full Bayesian and/or maximum likelihood framework. Such frameworks would allow prior distributions of the evolutionary rates, and timescales to be fully controlled, and enable all parameters to be jointly estimated. Interaction between and constraints provided by the speciation model, evolutionary rate, and timescale priors might reduce the uncertainty associated with the date estimation.

One method to develop a TDRP model under a full Bayesian and/or maximum likelihood framework is to modify currently existing programs, such as BEAST (30, 31). Several relaxed clocks are implemented in BEAST, including the *relaxed random local clock* (32). This clock enables the rate to vary among branches by multiplying the rate of the parental branch with the *branch-specific rate multiplier*, φ, to give the rate of its child branches (32). However, in the relaxed random local clock, the values of φ among branches are not time-correlated but are independent from one another (32), and thus cannot accommodate the TDRP. This framework can be modified so that it can take the TDRP into account, by forcing φ, and hence the rate of evolution, to gradually increase from the root to the tip with its dynamics defined and parameterised by the TDRP model. Another method, which is perhaps more elegant and

fundamental, is to modify the process of how substitutions accumulate through time directly, slowly increasing in rate from the past to the present, following the TDRP pattern. Since the TDRP is a pervasive evolutionary feature of all viruses, and in fact of cellular organisms also, this tool should be useful for evolutionary biologists in general.

Another methodological improvement, which I think will greatly improve the connection between analyses of extant and extinct viruses, is to develop an evolutionary inference method that can accommodate two different rates of evolution. With the advances in genomic sequencing technology, it is clear that the collection of EVEs will be growing at an ever increasing rate, playing key roles in elucidating the deep history of viruses. However, the rate of evolution of viruses changes abruptly upon becoming endogenous, dropping by several orders of magnitude to match those of their hosts. As a result, viral phylogenies that contain EVEs will have several branches that contain two distinct rates of evolution; one part of the branch will correspond to the fast viral evolutionary rate, and another part corresponds to the slow rate of host evolution. To analyse extent viruses together with EVEs under a single framework necessitates that this phenomenon is taken into account, otherwise the outcome can be severely biased.

As presented in Chapter five, two methods were used to estimate evolutionary timescales of FLERVs while avoiding this problem of mixed virus and host evolutionary rates. The first method was to trace the lineage of an extant virus down the tree in order to obtain substitution numbers for the age estimation. By doing so, all the branches that contain mixed rates of virus and host evolution were avoided. The second approach was to reconstruct maximum-likelihood ancestral sequences of FLERVs and use them for evolutionary analyses, minimising the host evolution in the viral sequence data. This particular method has also been employed by other researchers in the field (33). These methods are however relatively crude and have several problems associated with them. For example, regarding the first method, by tracing just only one viral lineage down the tree, it means that the information from other viral lineages were not used in the age estimation although they are very useful especially in term of

estimating the date uncertainty. In the second approach, the uncertainty of ancestral sequences were not accounted for, assuming that the maximum-likelihood ancestral sequences were the actual ancestral sequences.

A better approach to tackle this problem of mixed virus and host evolutionary rates is to introduce extra parameters into evolutionary analyses to model which branches have two rates, and where within the branches that the change of rate occurs. This '*two-rate*' clock will allow all available evolutionary information that is stored in the sequence data to be fully utilised, and also effectively avoid the problem of ancestral sequence reconstruction, and directly estimate the genetic distances while taking the two different rates of virus and host evolution into account. Not only will this clock be useful for paleovirological analyses, but it will also be useful for the study of pathogens that have a latent phase and a virulent phase, such as *Mycobacterium tuberculosis* and human immunodeficiency virus, between which the rates of evolution can vary significantly.

ଔ

I hope this thesis has succeeded in demonstrating that EVEs are instrumental in unravelling the deep history of viruses, including identifying previously unknown hosts, ancient viral transmission routes, and virus-host interaction in deep time. Paleovirological analyses must be performed with care however, due to the complex nature of the TDRP and the mixed rate of EVE evolution. These two phenomena add complexity to analyses of viral evolution, but also at the same time represent a bridge that connects extant and extinct viruses together. Undoubtedly, our knowledge of long-term viral evolution will continue to be revised and improved as more and more EVEs become available and phylogenetic methods undergo further improvement.

6.2 References

1. **Katzourakis A**, **Aiewsakun P**, **Jia H**, **Wolfe ND**, **LeBreton M**, **Yoder AD**, **Switzer WM**. 2014. Discovery of prosimian and afrotherian foamy viruses and potential cross species transmissions amidst stable and ancient mammalian co-evolution. Retrovirology **11**:61.
2. **Han G-Z**, **Worobey M**. 2014. Endogenous viral sequences from the Cape golden mole (*Chrysochloris asiatica*) reveal the presence of foamy viruses in all major placental mammal clades. PLoS One **9**:e97931.
3. **Hill CL**, **Bieniasz PD**, **McClure MO**. 1999. Properties of human foamy virus relevant to its development as a vector for gene therapy. J. Gen. Virol. **80**:2003–2009.
4. **Berg A**, **Pietschmann T**, **Rethwilm A**, **Lindemann D**. 2003. Determinants of foamy virus envelope glycoprotein mediated resistance to superinfection. Virology **314**:243–252.
5. **Plochmann K**, **Horn A**, **Gschmack E**, **Armbruster N**, **Krieg J**, **Wiktorowicz T**, **Weber C**, **Stirnnagel K**, **Lindemann D**, **Rethwilm A**, **Scheller C**. 2012. Heparan sulfate is an attachment factor for foamy virus entry. J. Virol. **86**:10028–10035.
6. **Herchenröder O**, **Moosmayer D**, **Bock M**, **Pietschmann T**, **Rethwilm A**, **Bieniasz PD**, **McClure MO**, **Weis R**, **Schneider J**. 1999. Specific binding of recombinant foamy virus envelope protein to host cells correlates with susceptibility to infection. Virology **255**:228–236.
7. **Switzer WM**, **Bhullar V**, **Shanmugam V**, **Cong M-E**, **Parekh B**, **Lerche NW**, **Yee JL**, **Ely JJ**, **Boneva R**, **Chapman LE**, **Folks TM**, **Heneine W**. 2004. Frequent simian foamy virus infection in persons occupationally exposed to nonhuman primates. J. Virol. **78**:2780–2789.
8. **Meiering CD**, **Linial ML**. 2001. Historical perspective of foamy virus epidemiology and infection. Clin. Microbiol. Rev. **14**:165–176.
9. **Liu W**, **Worobey M**, **Li Y**, **Keele BF**, **Bibollet-Ruche F**, **Guo Y**, **Goepfert PA**, **Santiago ML**, **Ndjango J-BN**, **Neel C**, **Clifford SL**, **Sanz C**, **Kamenya S**, **Wilson ML**, **Pusey AE**, **Gross-Camp N**, **Boesch C**, **Smith V**, **Zamma K**, **Huffman MA**, **Mitani JC**, **Watts DP**, **Peeters M**, **Shaw GM**, **Switzer WM**, **Sharp PM**, **Hahn BH**. 2008. Molecular ecology

and natural history of simian foamy virus infection in wild-living chimpanzees. PLoS Pathog. **4**:e1000097.

10. **Callahan ME**, **Switzer WM**, **Matthews AL**, **Roberts BD**, **Heneine W**, **Folks TM**, **Sandstrom PA**. 1999. Persistent zoonotic infection of a human with simian foamy virus in the absence of an intact orf-2 accessory gene. J. Virol. **73**:9619–9624.
11. **CDC**. 1997. Nonhuman primate spumavirus infections among persons with occupational exposure--United States, 1996. MMWR. Morb. Mortal. Wkly. Rep. **46**:129–31.
12. **Schweizer M**, **Falcone V**, **Gänge J**, **Turek R**, **Neumann-Haefelin D**. 1997. Simian foamy virus isolated from an accidentally infected human individual. J. Virol. **71**:4821–4824.
13. **Heneine W**, **Switzer WM**, **Sandstrom P**, **Brown J**, **Vedapuri S**, **Schable CA**, **Khan AS**, **Lerche NW**, **Schweizer M**, **Neumann-Haefelin D**, **Chapman LE**, **Folks TM**. 1998. Identification of a human population infected with simian foamy viruses. Nat. Med. **4**:403–407.
14. **Calattini S**, **Betsem EB**, **Froment A**, **Mauclere P**, **Tortevoye P**, **Schmitt C**, **Njouom R**, **Saib A**, **Gessain A**. 2007. Simian foamy virus transmission from apes to humans, rural Cameroon. Emerg Infect Dis **13**.
15. **Switzer WM**, **Tang S**, **Ahuka-Mundeke S**, **Shankar A**, **Hanson DL**, **Zheng H**, **Ayouba A**, **Wolfe ND**, **LeBreton M**, **Djoko CF**, **Tamoufe U**, **Esteban A**, **Heneine W**, **Peeters M**, **Wright LL**, **Muyembe-Tamfum JJ**, **Wemakoy EO**, **Mulembakani P**, **Hoff N a**, **Rimoin AW**. 2012. Novel simian foamy virus infections from multiple monkey species in women from the Democratic Republic of Congo. Retrovirology **9**:100.
16. **Betsem E**, **Rua R**, **Tortevoye P**, **Froment A**, **Gessain A**. 2011. Frequent and recent human acquisition of simian foamy viruses through apes' bites in central Africa. PLoS Pathog **7**:e1002306.
17. **Leendertz FH**, **Zirkel F**, **Couacy-Hymann E**, **Ellerbrok H**, **Morozov VA**, **Pauli G**, **Hedemann C**, **Formenty P**, **Jensen SA**, **Boesch C**, **Junglen S**. 2008. Interspecies transmission of simian foamy virus in a natural predator--prey system. J Virol **82**:7741–

7744.

18. **Switzer WM**, **Salemi M**, **Shanmugam V**, **Gao F**, **Cong M-E**, **Kuiken C**, **Bhullar V**, **Beer BE**, **Vallet D**, **Gautier-Hion A**, **Tooze Z**, **Villinger F**, **Holmes EC**, **Heneine W**. 2005. Ancient co-speciation of simian foamy viruses and primates. Nature **434**:376–380.

19. **Han G-Z**, **Worobey M**. 2012. An endogenous foamy-like viral element in the coelacanth genome. PLoS Pathog. **8**:e1002790.

20. **Firth C**, **Kitchen A**, **Shapiro B**, **Suchard MA**, **Holmes EC**, **Rambaut A**. 2010. Using time-structured data to estimate evolutionary rates of double-stranded DNA viruses. Mol. Biol. Evol. **27**:2038–2051.

21. **Davis PL**, **Holmes EC**, **Larrous F**, **Van der Poel WHM**, **Tjørnehøj K**, **Alonso WJ**, **Bourhy H**. 2005. Phylogeography, population dynamics, and molecular evolution of European bat lyssaviruses. J. Virol. **79**:10487–10497.

22. **Shackelton LA**, **Parrish CR**, **Truyen U**, **Holmes EC**. 2005. High rate of viral evolution associated with the emergence of carnivore parvovirus. Proc. Natl. Acad. Sci. U. S. A. **102**:379–384.

23. **Lemey P**, **Pybus OG**, **Van Dooren S**, **Vandamme A-M**. 2005. A Bayesian statistical analysis of human T-cell lymphotropic virus evolutionary rates. Infect. Genet. Evol. **5**:291–298.

24. **Sironen T**, **Vaheri A**, **Plyusnin A**. 2001. Molecular evolution of Puumala hantavirus. J. Virol. **75**:11803–11810.

25. **Ho SYW**, **Lanfear R**, **Bromham L**, **Phillips MJ**, **Soubrier J**, **Rodrigo AG**, **Cooper A**. 2011. Time-dependent rates of molecular evolution. Mol. Ecol. **20**:3087–3101.

26. **Sanjuán R**. 2012. From molecular genetics to phylodynamics: evolutionary relevance of mutation rates across viruses. PLoS Pathog. **8**:e1002685.

27. **Hanada K**, **Suzuki Y**, **Gojobori T**. 2004. A large variation in the rates of synonymous substitution for RNA viruses and its relationship to a diversity of viral infection and transmission modes. Mol. Biol. Evol. **21**:1074–1080.

28. **Duffy S**, **Shackelton LA**, **Holmes EC**. 2008. Rates of evolutionary change in viruses: patterns and determinants. Nat. Rev. Genet. **9**:267–276.

29. **Biek R**, **Pybus OG**, **Lloyd-Smith JO**, **Didelot X**. 2015. Measurably evolving pathogens in the genomic era. Trends Ecol. Evol. **30**:306–313.

30. **Drummond AJ**, **Suchard MA**, **Xie D**, **Rambaut A**. 2012. Bayesian phylogenetics with BEAUti and the BEAST 1.7. Mol. Biol. Evol. **29**:1969–73.

31. **Bouckaert R**, **Heled J**, **Kühnert D**, **Vaughan T**, **Wu CH**, **Xie D**, **Suchard M a.**, **Rambaut A**, **Drummond AJ**. 2014. BEAST 2: a software platform for bayesian evolutionary analysis. PLoS Comput. Biol. **10**:1–6.

32. **Drummond AJ**, **Suchard MA**. 2010. Bayesian random local clocks, or one rate to rule them all. BMC Biol. **8**:114.

33. **Lefeuvre P**, **Harkins GW**, **Lett J-M**, **Briddon RW**, **Chase MW**, **Moury B**, **Martin DP**. 2011. Evolutionary time-scale of the begomoviruses: evidence from integrated sequences in the Nicotiana genome. PLoS One **6**:e19193.

Appendix I: Discovery of prosimian and afrotherian foamy viruses and potential cross species transmissions amidst stable and ancient mammalian co-evolution

RETROVIROLOGY

RESEARCH Open Access

Discovery of prosimian and afrotherian foamy viruses and potential cross species transmissions amidst stable and ancient mammalian co-evolution

Aris Katzourakis[1*†], Pakorn Aiewsakun[1†], Hongwei Jia[2], Nathan D Wolfe[3,4,5], Matthew LeBreton[6], Anne D Yoder[7] and William M Switzer[2*]

Abstract

Background: Foamy viruses (FVs) are a unique subfamily of retroviruses that are widely distributed in mammals. Owing to the availability of sequences from diverse mammals coupled with their pattern of codivergence with their hosts, FVs have one of the best-understood viral evolutionary histories ever documented, estimated to have an ancient origin. Nonetheless, our knowledge of some parts of FV evolution, notably that of prosimian and afrotherian FVs, is far from complete due to the lack of sequence data.

Results: Here, we report the complete genome of the first extant prosimian FV (PSFV) isolated from a lorisiforme galago (PSFVgal), and a novel partial endogenous viral element with high sequence similarity to FVs, present in the afrotherian Cape golden mole genome (ChrEFV). We also further characterize a previously discovered endogenous PSFV present in the aye-aye genome (PSFVaye). Using phylogenetic methods and available FV sequence data, we show a deep divergence and stable co-evolution of FVs in eutherian mammals over 100 million years. Nonetheless, we found that the evolutionary histories of bat, aye-aye, and New World monkey FVs conflict with the evolutionary histories of their hosts. By combining sequence analysis and biogeographical knowledge, we propose explanations for these mismatches in FV-host evolutionary history.

Conclusion: Our discovery of ChrEFV has expanded the FV host range to cover the whole eutherian clade, and our evolutionary analyses suggest a stable mammalian FV-host co-speciation pattern which extends as deep as the exafroplacentalian basal diversification. Nonetheless, two possible cases of host switching were observed. One was among New World monkey FVs, and the other involves PSFVaye and a bat FV which may involve cross-species transmission at the level of mammalian orders. Our results highlight the value of integrating multiple sources of information to elucidate the evolutionary history of viruses, including continental and geographical histories, ancestral host locations, in addition to the natural history of host and virus.

Keywords: Paleovirology, Endogenous retrovirus, Foamy virus, Co-evolution, PSFVgal, PSFVaye, ChrEFV

* Correspondence: aris.katzourakis@zoo.ox.ac.uk; bis3@cdc.gov
†Equal contributors
[1]Department of Zoology, University of Oxford, Oxford, South Parks Road, Oxford OX1 3PS, UK
[2]Laboratory Branch, Division of HIV/AIDS, National Center for HIV, Hepatitis, STD, and TB Prevention, Centers for Disease Control and Prevention, Atlanta, GA 30333, USA
Full list of author information is available at the end of the article

BioMed Central

© 2014 Katzourakis et al.; licensee BioMed Central Ltd. This is an Open Access article distributed under the terms of the Creative Commons Attribution License (http://creativecommons.org/licenses/by/4.0), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work is properly credited. The Creative Commons Public Domain Dedication waiver (http://creativecommons.org/publicdomain/zero/1.0/) applies to the data made available in this article, unless otherwise stated.

Background

Foamy viruses (FVs) are complex retroviruses in the *Spumaretrovirinae* subfamily [1]. All known contemporary FVs cause persistent but non-virulent, asymptomatic infections exclusively among boreoeutherian mammals (Additional file 1: Table S1) [1]. The first FV was discovered in a macaque in 1954 [2] and was isolated in 1955 [3]. Shortly thereafter, numerous FVs were isolated from other boreoeutherian mammals, including non-human primates (NHPs) [4-13], cats [14], cattle [15], and horses [16], for most of which complete genomes are available [16-25]. Recently, a novel FV was discovered by metagenomics in a bat (*Rhinolophus affinis)* from China (RhiFV) and was partially sequenced [26]; however, whether RhiFV is bat-specific will require confirmation with population and epidemiological studies. Although virtually every NHP sampled has a species-specific FV, a human-specific FV has not been identified. Moreover, all FVs isolated from humans have been demonstrated to originate from zoonotic infections with simian FVs (SFVs) [27-36].

Similar to other retroviruses, FVs occasionally leave 'fossils' in host genomes, which are the relics of past infections. These genomic fossils are called endogenous retroviruses (ERVs) and result from germ line infections that are followed by vertical inheritance and thus are present in every cell of an infected organism. In contrast, exogenous retroviruses and hence exogenous FVs are transmitted horizontally from organism-to-organism via infectious body fluids and may be confined to certain cell types defined by their host tropism. Only two endogenous FVs have been discovered in mammalian genomes to date. One is present in the two-toed sloth (*Choloepus hoffmanni*) genome (SloEFV) which was found to have a genome organization characteristic of all contemporary FVs [37]. The second endogenous FV was found in the aye-aye (*Daubentonia madagascariensis)* genome (PSFVaye) which to date has only been partially characterized [38]. Given a broad range of FV mammalian hosts and the wealth of available whole mammalian genomes, the finding of only two endogenous FVs so far suggests that FV endogenization is a rare event. While a number of defective ERVs found in fish genomes have been reported to exhibit some similarity to known exogenous FVs [39-41], more evidence is required to definitively determine if these ERVs are truly endogenous FVs, or alternatively, are distinct lineages that branched early on in FV or retroviral evolution.

Owing to the availability of sequences from diverse mammals (Additional file 1: Table S1), coupled with the stable pattern of codivergence with their hosts, the evolutionary history of mammalian FVs can be studied in great depth. For example, phylogenetic analysis of extant SFVs revealed strong evidence of SFV-host co-speciation over more than 30 million years (Myr) [42]. The discovery and phylogenetic analysis of SloEFV [37] supported and extended the FV-host co-speciation hypothesis further to the basal radiation of eutherians which occurred more than 100 Myr ago (Ma) [43] and simultaneously expanded FV mammalian host range to cover xenarthrans, which are one of the four superorders of the Eutheria clade (Laurasiatheria, Euarchontoglires and Afrotheria are the other superorders). This is in contrast to lentiviruses, for which sequence availability is comparable to that of FVs, but their evolutionary history is not as well understood and much more difficult to investigate, as they do not always co-speciate with their hosts [44]. Nevertheless, at the present day, evidence of FV infection in, and sequence data from prosimians and afrotherian mammals is still lacking, and this is required to better understand the FV radiation and evolutionary history in eutherians. Here, we report the discovery and describe the first extant exogenous PSFV isolated from a lorisiforme galago (*Otolemur crassicaudatus panganiensis*) (PSFVgal) and a novel ERV, that is potentially an endogenous FV, present in the genome of an afrotherian Cape golden mole (*Chrysochloris asiatica*) (ChrEFV). We also further describe PSFVaye that was previously only partially characterized [38]. Together with the currently available FV sequences, we use detailed molecular analyses to re-examine the co-evolutionary history of mammalian hosts and their FVs.

Results

Characterization of the complete genome of the extant PSFVgal from a galago

The complete genome of PSFVgal (GenBank accession number KM233624) was obtained by long-PCR amplification of overlapping 5′ and 3′ genomic halves using infected HeLa cells and sequence analysis. Genomic inspection shows that PSFVgal exhibits all the typical structural features of mammalian FV genomes, including the presence of *gag*, *pol*, *env*, *bel1*, and *bel2* genes flanked by LTRs (Figure 1), supporting that it is a FV. There are no in-frame stop codons identified in the PSFVgal coding regions. PCR and serological analyses also showed that PSFVgal-like sequences are not present in every tested Galago (7 sera and 10 DNA specimens) or Otolemur species (14 sera and 6 DNA specimens) (Additional file 1: Table S2) using a combination of PSFVgal-specific PCR (0/13), generic PSFVgal-PSFVaye PCR (6/16, 37.5%) and PSFVgal Western blot (WB) testing (16/21, 76%). Combined, these findings strongly support the notion that PSFVgal is an exogenous FV. The length of the complete genome is 12,118 nucleotides (nt). A detailed annotation of the PSFVgal genome is provided in Additional file 2: Figure S1.

Sequence similarity analysis revealed that the LTR length is 1,267 nt. An asparagine-1,2 tRNA-utilizing primer binding site (PBS) was identified downstream of the

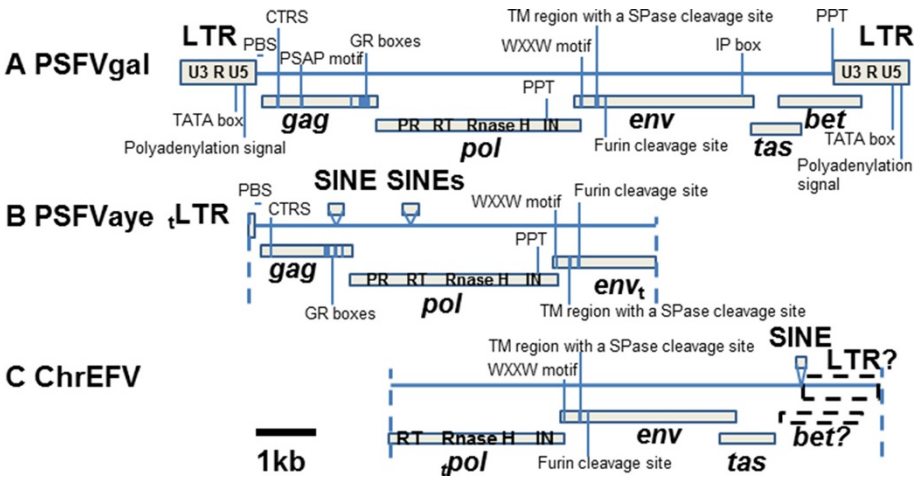


Figure 1 Complete and partial genomic organizations of PSFVgal, PSFVaye and ChrEFV. PSFVgal **(A)**, PSFVaye **(B)**, and ChrEFV **(C)** all exhibit at least some characteristic FV genomic features as indicated (see main text). PSFVaye and ChrEFV are interrupted by a few short interspersed nuclear elements (SINEs). Dashed lines indicate where genomes are truncated. Dotted boxes represent hypothetical domains which may be present but could not be identified. 't'-subscription indicates that the domain is truncated (5'-truncated: preceding the domain name; 3'-truncated: following the domain name). The scale bar (black solid line) represents a nucleotide length of 1 kb. LTR, long-terminal repeat; PBS, primer binding site; CTRS, cytoplasmic retention signal; GR boxes, glycine-arginine rich boxes; PR, protease; RT, reverse transcriptase; IN, integrase; PPT, polypurine tract; WXXW, conserved WXXW site; TM, transmembrane; SPase, signal-peptide peptidase; IP, internal promoter.

5′-LTR before *gag* (TGGCGTCCCTGGGTGGGC, nt 1,270-1,287), but which is $tRNA^{Lys}$ (TGGCGCCCAA CGTGGGGG/C) in all other mammalian FVs. We confirmed the $tRNA^{Asp}$ PBS sequences by population-based sequencing of multiple PCR products amplified on different dates using primers surrounding the PBS with two different primer pairs and two different concentrations (0.1 and 1.0 ug) of day 52 tissue PSFVgal-infected HeLa cell DNA. By comparison with the SFVmac LTR [17], the cap site defining the 5′ boundary of the R region was determined to be 20 nt downstream from the TATA box (TATATAA, nt 928–934) at nt 955 within the 5′-LTR, and the polyadenylation site within the 3′-LTR, which defines the 3′ boundary of the R region, was located at the CA dinucleotide (nt 11,955) 20 nt downstream from the polyadenylation signal (AATAAA, nt 11,930-11,935). Thus, the PSFVgal LTR U3, R, and U5 regions are 954, 150, and 163 nt long, respectively. Two polypurine tracts (PPT: AAGGAGAGGG) for the dual initiation of plus-stand DNA synthesis [45] were located at the 5′ boundary of the 3′-LTR (nt 10,842-10,853) and at the center of the genome toward the 3′ terminus of *pol* (nt 6,068-6,077) as anticipated.

The *gag* gene was predicted to be 1,932 nt long (nt 1,363-3,294) to generate a 644 amino acid (aa) protein. The cytoplasmic retention signal (CTRS), important for particle assembly and viral budding [46], was located at aa 73–90 (GNWGPGDRFARIEVLLRD). The position of a late-assembly domain proline rich PSAP motif, which is essential for primate FV particle assembly and budding [47,48], was located at aa positions 217–220. The PSAP motif is highly conserved in all SFVs but its position within Gag is variable [48]. The three FV-specific glycine-arginine rich (GR I-III) motifs or boxes within the C-terminal region of Gag [49] were also determined (aa 479–493, 541–563, 567–597, respectively).

The Pol protein is 1,156 aa long (nt 3,248-6,718) and protease (PR), reverse transcriptase (RT), RNase H, and integrase (IN) were predicted to be located at aa positions 10–168, 154–355, 581–737, and 742–1,140, respectively. The catalytic centers of PR (DTG) and RT (YVDD) were located at aa 20–22 and 304–307, respectively. RNase H active sites were found to be at D^{589} and D^{659} and IN catalytic motif (DD35E) within the IN core domain (aa 862–972) was at D^{926}-D^{938}-E^{962}.

The Env protein is 999 aa long (nt 6,603-9,602). We found a highly conserved WXXW motif (WLRW, aa 10–13) in the N-terminus which is essential for Gag-Env protein interaction during the budding process and is found in all known extant FVs [50,51]. The internal promoter (TATAAAA) within the *env* gene, which serves for the initiation of *bel1* transcription, is present at nt 9,299-9,305. We also identified a hydrophobic transmembrane region, which contains a putative signal-peptide peptidase cleavage site (TMGWCIGLFCLLLILLFS↓LVIVIL), at aa 81–104. The cleavage occurs within the endoplasmic reticulum to remove the signal peptide [52]. A potential furin-cleavage site was also found (TRPNYTAARSRR↓SVE) at aa 131–145. It is this site at which the FV Env protein is cleaved, either by furin or furin-like proteases, to produce an Env leader protein – a component that interacts with the N-terminus of Gag via its WXXW motif and is absolutely required for FV particle budding [52].

The *bel1* and *bel2* genes were predicted to be present at nt 9,569-10,402 and nt 10,026-11,414, respectively. The putative Bel1 protein (277 aa) exhibits only a weak similarity (<15% identity) to extant FV Bel1 proteins, which could only be shown by comparison with Hidden Markov Model (HMM) of FV Bel1 proteins (HMM profile GyDB (http://www.gydb.org/): Bel1 spumaretroviridae; E = 3.7E-06). In contrast, the putative Bel2 protein (511 aa) exhibits a relatively high similarity with other FV Bel2 proteins with the best BLASTp score to the BFV Bel2 (E = 4E-08; maximum identity = 26%).

Discovery and characterization of PSFVaye and ChrEFV

To search for endogenous FVs, we screened all publically available GenBank whole genome shotgun (WGS) sequences with various FV proteins (Additional file 1: Table S3) using tBLASTn. As previously reported [38], several matches were returned from the WGS assembled sequence of the aye-aye (*Daubentonia madagascariensis)* spanning four non-overlapping contigs (Table 1). Sequence analysis shows that these four contigs represent different parts of a single ERV 'PSFVaye', and this was also confirmed by genome walking. Between the first and the second, and the second and the third contigs are short interspersed nuclear elements (SINEs), and between the third and the fourth contigs are 5 ambiguous nucleotides. We found that PSFVaye contains numerous frameshift and in-frame stop codon mutations and is interrupted by several transposable elements throughout the genome, confirming that PSFVaye is endogenous and not replication competent. The notion that PSFVaye is endogenous was further supported by PCR testing of 18 different aye-ayes using PSFVaye *gag* primers in a single round of PCR (Additional file 1: Table S2); all of the aye-ayes including three infants were strongly PCR-positive which is typical of ERVs. Serums available from 17 of these aye-ayes were all negative for antibodies to the related PSFVgal (Additional file 1: Table S2), indicating they do not code for complete proteins, consistent with other evidence that PSFVaye is endogenous. This *in silico* screening process also returned one contig from the WGS assembled sequence from a Cape golden mole (*Chrysochloris asiatica*) (Table 1). We designated this element 'ChrEFV'. Like PSFVaye, ChrEFV also contains numerous frameshift and in-frame stop codon mutations and is interrupted by several transposable elements, confirming that it is not replication competent as is typical of ancient ERVs. Genomic inspection revealed the putative genes of both PSFVaye and ChrEFV to be in the same order as those of typical exogenous FVs (Figure 1; see Additional file 2: Figure S1 for detailed sequence annotations). Furthermore, it is likely that there is only one copy of each virus present in either host genome since both the aye-aye and Cape golden mole genomes were extensively sequenced (38X [53] and 66X [54] coverage, respectively) but paralogs of each were not found.

Within PSFVaye, via sequence homology, we identified i) a PBS (TGACACCCAATGTGGGGC**;** nt 114–131) with a broad tRNA usage ($tRNA^{Lys,\ Asn,\ Thr,\ Pro}$) predicted, ii) a complete putative *gag* gene (nt 198–1,971) which is interrupted by a SINE (AluJo [55], nt 1,449-1,748), iii) a complete putative *pol* gene (nt 1,941-5,727; PR: nt 1,968-2,459; RT: nt 2,415-3,306; RNase H: nt 3,985-4,467; IN: nt 4,480-5,676) which is also interrupted by SINEs (AluJ Mim [56] & ALU [57], nt 2,871-3,165), and

Table 1 Identification of foamy virus (FV)-like sequences in the aye-aye and Cape golden mole genome

Accession no.	FV genomic region	Position*	BLASTp best hit	Query cover	Max identity	E-value
Aye-aye (*Daubentonia madagascariensis*)						
AGTM011839603	(5′ truncated) 5′-LTR	3208..3318	N/A	N/A	N/A	N/A
	Primer binding site	3321..3338	N/A	N/A	N/A	N/A
	gag	3405..4654	CAA70074 Gag FFV	93%	37%	6E-72
AGTM012070104		131..361				
	pol	323..1247	AAF64414 Pol EFV	95%	48%	0.0
AGTM011519752		90..1996 (end)				
AGTM010361255		1..635				
	(3′ truncated) *env*	520..2763	AFK85016 Env RhiFV	80%	42%	7E-115
Cape golden mole (*Chrysochloris asiatica*)						
AMDV01151999	(5′ truncated) *pol*	<615.. > 3534	AFX98084 Pol SFV	95%	47%	0.0
	env	<3464.. > 6371	AFR79245 Env BFV	96%	30%	5E-121
	*bel1***	6338..6978	-	-	-	-

'N/A' = 'Not applicable'.

*The nucleotide positions are with respect to the contig length and do not take insertion and deletion mutations into account.

**The similarity between ChrEFV Bel1 and other FV Bel1 proteins was determined by comparing the ChrEFV putative Bel1 protein against the Hidden Markov Model (HMM) of FV Bel1 proteins available in the GyDB (http://www.gydb.org/) (HMM profile: Bel1 spumaretroviridae; E: 0.24).

iv) a 3′-truncated putative *env* gene (nt 5,612-7,867). We also found ~113 nt on the 5′ end of the PBS to be uniquely-mapping within the aye-aye genome. This might represent some portion of the 5′ end of the 5′ PSFVaye LTR (nt 1–111). We confirmed the truncation of the putative 5′-LTR and *env* gene and the absence of a 3′-LTR in the PSFVaye genome by using a genome walking of genomic DNA from two different aye-ayes. We found that PSFVaye exhibits several characteristic FV features. For example, its Gag protein contains a degenerate, but still recognizable, putative CTRS (GPR? VGD*WQRICLAFQY, aa 37–54) and three GR I-III boxes (box I: nt 1,248-1,358; box II: nt 1,422-1,781 interrupted by a SINE (nt 1,449-1,748); box III nt 1,851-1,871); the putative *pol* gene harbors a PPT (AAGGACAAAG, nt 5,092-5,101); and the N-terminus of the putative Env contains a highly conserved WXXW motif (WLAW, aa 10–13). A hydrophobic transmembrane region containing a signal-peptide peptidase cleavage site (ILIWIMLFLILFSAILVS↓TLIAVF) was also located in the N-terminus of the putative Env at aa 63–86. We could not identify a furin-cleavage site in the Env of PSFVaye, most likely due to its defective nature. However, by comparing the Env sequence of PSFVaye to those of other exogenous non-defective FVs, we speculate that it should be located between aa 113 and 127 (YFSQAHV*KSRAIHF). Moreover, its putative Gag and Env proteins only exhibit similarities to FV proteins (Table 1) and share no detectible similarities with other (non-FV) retroviral proteins. All these findings strongly suggest that the PSFVaye sequence is an endogenous FV, and not merely a distantly related FV-like ERV.

ChrEFV contains i) a 5′-truncated putative Pol coding gene (nt <1- < 2,937) which only contains full RT (nt 1–597), RNase H (nt 1,270-1,737), and IN coding domains (nt 1,750-2,931), ii) a full putative Env coding gene (nt <2,867- > 5,785) and iii) a full putative Bel1 coding gene (nt 5,752-6,393). Following the putative *bel1* gene is a long non-repetitive sequence of ~1,700 nt containing a SINE (AFROSINE1B [58]). The *bel2* gene and the 3′-LTR of ChrEFV may be located within this region but cannot be identified, most likely due to divergence from the probes we used. Nonetheless, the N-terminus of the ChrEFV Env protein harbors the highly conserved WXXW motif (WMRW, aa 10–13) and only shows significant similarity to FV Env proteins (Table 1). A hydrophobic transmembrane region harboring a putative signal-peptide peptidase cleavage site (VMTSYVS? LILLGIIITA↓SFI TIC, aa 74–97) was also found in the putative Env at the anticipated location, and a potential furin-cleavage site (GQIISNSSSNRRKI) could be located at aa 125–138 within the putative Env as well. Furthermore, the putative Bel1 protein also exhibits some similarity to FV Bel1 proteins which could be demonstrated by using the HMM searching technique (HMM profile GyDB (http://www.gydb.org/): Bel1 spumaretroviridae; E = 0.24; note that this weak support is rather typical and expected given that this is a part of an endogenous FV and that FV Bel1 protein is not conserved among FVs [59]). Although a number of FV-specific genomic features as outlined in the PSFVgal characterization section could not be located, possibly due to the lack of sequence data and/or divergence, this genomic information supports a FV progenitor of ChrEFV.

Aside from the aye-aye genome, currently available partial and complete genomes of other lemurs (*Lemur catta* and *Microcebus murinus*, respectively) do not seem to contain PSFVaye-like orthologs, confirmed by PCR testing of eight lemuriform species and one tarsiiforme species using generic PSFVaye and PSFVgal primers (Additional file 1: Table S2). These results support the hypothesis that endogenization of PSFVaye occurred after the basal diversification of the lemurs ~55-66 Ma [60-62], in the ancestral aye-aye lineage after it diverged from the ancestral lineage of all other living lemurs. Similarly, ChrEFV orthologs could not be found in elephant, hyrax, elephant shrew, and tenrec genomes, implying a maximum age of ~60-85 Myr based on the proposed Afrosoricida basal split date [43,63]. To further estimate the age of these elements, we analyzed the frequencies of in-frame stop codons in their putative Pol proteins with a Monte Carlo simulation approach, using *pol*s of PSFVgal, SFVcpz, SFVagm, SFVsqu, FFV, and SloEFV (Figure 2) as model sequences. Consistently across all simulations, the mean age estimates of PSFVaye and ChrEFV are ~35.2-39.6 and ~65-78 Myr, respectively, congruent with (i.e. less than) their upper-bound age limit.

FV phylogenetic analyses

A retrovirus phylogeny (Additional file 3: Figure S2) was estimated using an alignment of RT protein sequences to establish the phylogenetic positions of PSFVgal, PSFVaye, and ChrEFV among retroviruses. All three new FV sequences are placed with sequences of extant FVs. Together with the genomic characterization above, the phylogeny strongly supports the classification of PSFVgal as a true exogenous FV and that both PSFVaye and ChrEFV are real endogenous FVs. To the best of our knowledge, no exogenous lemuriform and afrotherian FVs have been identified to date. Thus, our finding of PSFVaye, and ChrEFV raises the possibility of the presence of undetected FVs circulating among other lemurs and afrotherians.

Interestingly, while we can confirm that PSFVaye and ChrEFV are endogenous FVs, a number of ERVs found in fish genomes, including the genomes of coelacanth (CoeEFV) [39], zebra fish (DrFV-1) [40], platyfish (platyfishEFV) [41], and cod (CodEFV) [41], have been reported to show some detectible similarity to exogenous

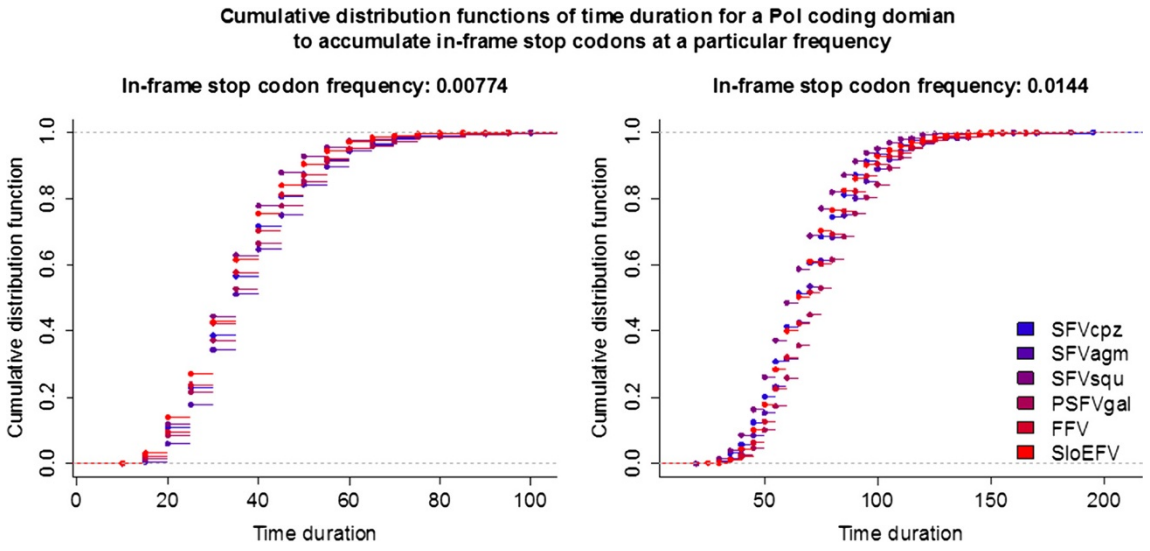


Figure 2 PSFVaye and ChrEFV integration dates estimated from in-frame stop codon frequencies. Cumulative distribution functions (CDFs) of time duration for Pols of SFVcpz, SFVagm, SFVsqu, PSFVgal, FFV, and SloEFV to accumulate in-frame stop codons at the frequencies of 0.00774 (right, stop codon frequency of PSFVaye Pol) and 0.0144 (left, stop codon frequency of ChrEFV Pol). Based on these CDFs, the mean age of PSFVaye is estimated to be ~35.2-39.6 millions of years (Myr) old (95% confidence interval: 15–75 Myr old, mode: 30–35 Myr old, median: 35 Myr old). The mean age estimate of ChrEFV is 65.6-78.2 Myr old (95% confidence interval: 35–125 Myr old, mode: 60–65 Myr old, median: 65–75 Myr old).

FVs, but more evidence is required to definitively determine that these ERVs truly have a FV origin. Since these fish ERVs are phylogenetically basal to all known extant mammalian FVs, it is difficult to say with certainty that the progenitors of these ERVs are "true" FVs, and not distinct lineages that branched earlier in retroviral evolution but which are "FV-like" enough to be detected by sequence similarity using the true FVs as probes. Moreover, in terms of genomic organization, these fish ERVs do not show all the features characteristic of mammalian FV genomes. For example, the two identified accessory genes of CoeEFV show no significant similarity to those of extant mammalian FVs [39], only two out of five determined DrFV-1 proteins, Gag and Pol, are FV-like [40], FV-like accessory genes in platyfishEFV have not been located [41], and only FV-like RT could be found for CodEFV [41]. Although it is known that the accessory genes are the least conserved FV genes [59], and that this partial characterization of their genomes could be due to an ancient divergence of fish and mammalian FVs or simply the lack of sequence data, it raises a possibility that the progenitors of these ERVs may not be FVs but merely class III FV-like viruses. Resolution of this debate will require identification and analysis of FV genomes of other vertebrates like amphibians, reptiles, and birds, if they indeed exist [39]. Phylogenetic analysis of extant fish FVs and ERVs would also help resolve this debate.

To investigate FV phylogenetic relationships in more detail and evaluate the FV-host co-evolutionary history, we performed two analyses: (i) FV-host phylogenetic reconciliation analysis, and (ii) FV-host divergence correlation analysis. A Bayesian phylogeny of FVs was constructed based on an alignment of concatenated Pol-Env protein sequences and compared to the previously published host phylogeny [43] (Figure 3A). Gag protein sequences were not included in the alignment since (i) the Gag sequence of RhiFV is not available and (ii) the alignable region of Gag for the rest of the FVs is relatively short (~180 aa). Potential recombination among the sequences within the alignment was assessed using a quartet-based recombination detection program VisRD3 [64], and the results showed no significant evidence for recombination, both at nucleotide and protein levels (nucleotide: $p = 0.621$; protein: $p = 0.495$). Furthermore, we also found that neither of the separate Pol nor Env alignments reject the topology of the best phylogeny inferred from the concatenated Pol-Env alignment (Figure 3A), congruent with the results from the recombination test (approximately unbiased test: $p\text{-}AU_{Pol} = 0.987$; $p\text{-}AU_{Env} = 0.700$; Additional file 4: Figure S3). The two phylogenies show remarkably similar topologies, reflecting the well-established evolutionary history of stable FV co-speciation with their hosts [37,42]. In total, 14/17 (82.4%) potential FV-host co-speciation events were inferred among the 17 FV and host sequences using the co-phylogeny reconstruction software Jane v4.0 [65]. The deepest co-speciation event occurred early on in eutherian diversification (Figure 3A) corresponding to the Exafroplacentalia-Afrotheria split. The reconstruction that maximizes the

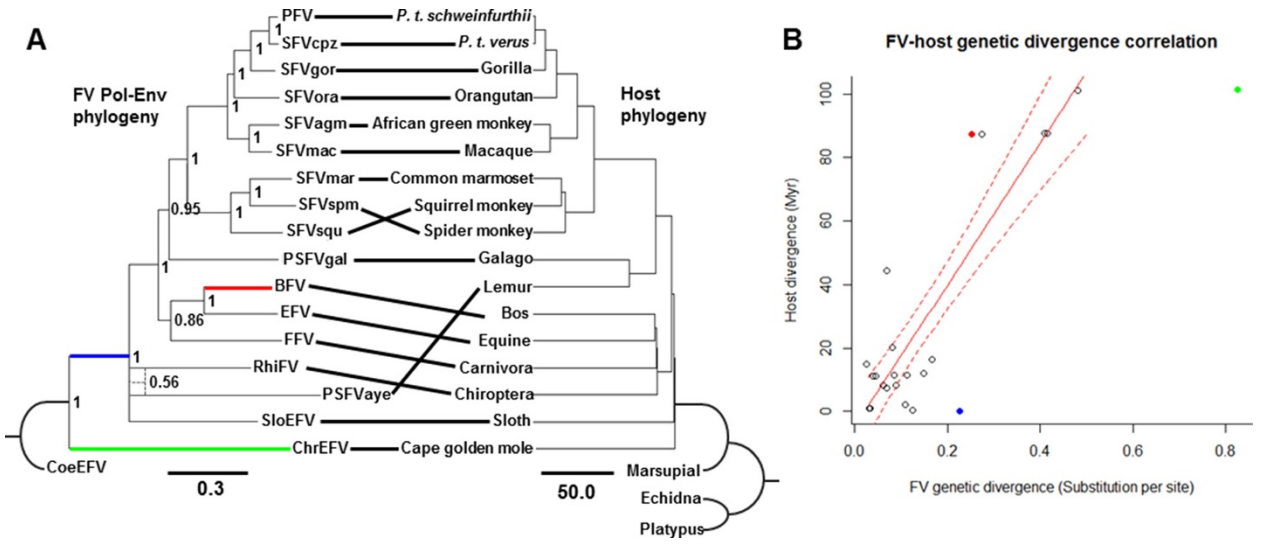


Figure 3 Co-evolution of foamy viruses (FVs) and their mammalian hosts. (A) FV Bayesian phylogeny (left) is compared to that of their hosts (right, previously published in [43]). Curved branches indicate outgroups. Numbers on branch nodes are Bayesian posterior probabilities. Solid lines between the two trees indicate FV-host associations. FV and host phylogeny scale bars are in units of amino-acid substitutions per site and million years, respectively. Weak support for the sister-taxon relationship of RhiFV and PSFVaye (dotted branches, posterior probability = 0.56) was inferred. Among 17 FV and host sequences, 3 apparent mismatches have been inferred using co-phylogenetic reconstruction. Branches corresponding to FV-host co-evolution were identified and used in FV-host divergence correlation analysis **(B)**. Branch lengths of FV tree and host divergence time were determined to have a linear correlation with a statistically strong support for its coefficient (linear regression: N = 23, $R^2 = 0.823$, p = 3.48E-8), represented by the solid red line. Dotted lines represent the 95% confidence interval of the estimated linear relationship. Outliers determined by Cook's distances (solid dots), including the bovine foamy virus (BFV) branch (red), the ChrEFV branch (green), and the ancestral branch leading to the exafroplacentalian FVs (blue), were not included in the linear regression. The colors of the outliers correspond to the colors of the branches in **A**.

number of co-speciation events suggests that the most recent common ancestor (MRCA) of PSFVaye and RhiFV was present in the MRCA of their hosts and requires a duplication of the virus lineage at the base of the exafroplacentalian mammal clade (Additional file 5: Figure S4). Based on what we know about FV biology, it is extremely unlikely that a FV will diversify into two lineages within a single natural host, i.e. in the absence of host diversification. We therefore postulate that this viral lineage duplication represents an ancient host speciation event, where one of the resulting species is not sampled in our dataset. We therefore adopt a conservative approach and do not count the PSFVaye and RhiFV co-speciation event, constraining the number of co-speciation events in our reconstruction to 13 (76.5%). This estimate relies on the assumption of the existence of this un-sampled host lineage. Nevertheless, 13 co-speciation events are still greater than expected by chance (random tip mapping: sample size = 1000, $p < 0.001$), indicating that the FV-host co-speciation history is very stable. Moreover, there is a strong linear correlation between FV and host divergence which extends from the present to the exafroplacentalian/exafroplacentalian FV radiation (linear regression: N = 23, $R^2 = 0.823$, p = 3.48E-8; Figure 3B), confirming previous estimates [37]. This result suggests that across this protracted time period the accumulation of FV sequence divergence occurred in proportion to host divergence. Combined, these findings strongly support stable FV co-divergence with their mammalian hosts for more than 100 Myr throughout the Cretaceous and Cenozoic eras as has been previously proposed [37].

Despite ChrEFV having an inferred co-speciation event (posterior probability = 1), the ChrEFV branch lengths, including both the terminal branch and the internal branch leading to other FVs, are longer than would be expected based on the FV-host divergence linear relationship (Figure 3B; identified as outliers, green and blue dots, respectively, with Cook's distance investigation). Given that ChrEFV is basal to all known extant FVs, this finding challenges the notion that ChrEFV is an ERV that has a FV origin. Nevertheless, ChrEFV is a long terminal branch in the inferred FV phylogeny and has accumulated numerous neutral genetic changes since its ancient endogenisation. Coupled with long-branch attraction, this 'pseudogene effect' [66] could cause artificial inflation of the branch length with a concomitant deep placement of ChrEFV within the FV phylogeny. Thus, given the congruent ChrEFV/host phylogenies and the FV-like features of ChrEFV outlined above, it is still likely that ChrEFV is a genuine endogenous FV and not merely a class III FV-like ERV. Resolution of ChrEFV evolutionary history requires further analysis with other extant afrotherian, marsupial or monotreme FVs

when and if they are identified. The results from these analyses could also elucidate the earlier events of mammalian FV diversification. To the best of our knowledge, this is the first (endogenous) afrotherian FV to be discovered, extending the FV host range to cover the whole eutherian clade.

PSFVgal was found to be a sister taxon of SFVs with robust support (posterior probability = 0.95), consistent with a FV-host co-speciation history. Additionally, we also identified a number of extant FVs from other lorises, including the silvery greater galago (*Otolemur monteiri monteiri*), southern lesser bush baby (*Galago senegalensis moholi*), and a potto (*Perodicticus potto*) (Additional file 1: Table S2), and partially sequenced their genomes (IN core domain: ~259 nt, 85 aa). When these new sequences were included in our phylogenetic analysis, they were all found to be distinct FVs forming a clade with PSFVgal with relatively high statistical support (posterior probability = 0.85) within which the branching orders generally mirror those of their lorisiforme hosts (Additional file 6: Figure S5). These results suggest that co-speciation between exogenous lorisiforme FVs and their hosts is very stable, extending to the species level. Together, our results indicate an ancient distribution and evolution of lorisiforme FVs in continental Africa dating to the divergence between strepsirrhine primates (lemurs and lorises) and the anthropoid primates about 70–95 Ma [43,62,63]. The absence of PSFV in some lorises (Additional file 1: Table S2) may be due to the testing of limited numbers of individuals and/or testing of captive animals that were captured as infants and thus were likely PSFV-negative at that time.

In contrast, PSFVaye was not a sister-taxon of PSFVgal as would be expected, showing a conflict in the co-evolutionary history of PSFVaye (Figure 3A). Instead, the sequence is robustly placed outside the boreoeutherian FV clade (posterior probability = 1), but still remains together with exafroplacentalian FVs (posterior probability = 1). The same evolutionary history was inferred for RhiFV (Figure 3A); instead of being grouped together with other fereungulata FVs (bovine, equine, and feline FVs) as would be expected if it were to co-speciate with its host, it is placed robustly outside the boreoeutherian FV clade (posterior probability = 1), but still remains together with exafroplacentalian FVs (posterior probability = 1). We compared our best estimated phylogeny against hypothetical alternative phylogenies where PSFVaye and/or RhiFV co-speciate with their hosts using approximately unbiased tests [67], and found that the co-speciation hypotheses of PSFVaye and RhiFV are both rejected (Additional file 4: Figure S3-A). Interestingly, our analysis also inferred a sister group relationship between PSFVaye and RhiFV although statistical support for this relationship is extremely weak (posterior probabilities = 0.56, Figure 3A).

Discussion

Consistent with previous findings [37,42], our analyses suggest an extremely stable FV-host co-speciation history across a timescale spanning millions of years. Interestingly, this is in contrast to what is suggested by various *in vitro* experiments. For example, an *in vitro* study has shown prototype FV (PFV) to be capable of infecting not only primate, but also rodent, laurasiatheria, avian and reptile cells, with an extremely broad range of susceptible cell types [68]. Furthermore, *in vitro* investigations of FV cell-attachment and entry also suggested that FVs of different species likely utilize the same, perhaps promiscuous, receptor molecule(s) for cell-attachment and/or entry [69-71]. Although these findings might in part help to explain FV cross-species transmissions (FV speciation events by means of jumping from one host species to another, without the host speciating at the time of the jump) observed in nature (see below), one would expect to see more host-switches happening given these findings. Analyses of several antiretroviral restriction factors have shown that they are specific to particular FVs. For example, while the tripartite motif protein 5αs (TRIM5αs) of most New World monkeys (NWMs) have been shown to be able to restrict some NWM FVs, PFV and a SFV from macaque, TRIM5αs from apes cannot, but instead can restrict feline FV [19,72]. Inhibition of apolipoprotein B mRNA-editing, enzyme-catalytic, polypeptide-like 3C (APOBEC3C) by FVs has also been shown to be species-specific [73]. Whether this species-specific FV-host antagonistic interaction is one of the factors underlining the stable co-speciation history or the result of it is still unclear. To date, the factors that determine this extremely stable pattern of FV-host co-speciation in nature are still very poorly understood.

Although we found the pattern of FV-host co-speciation to be very strong and stable, it is not absolute. Against a clear background of FV-host co-divergence are a small number of mismatches in FV-host evolutionary history. One involves NWM FVs which form a clade that is sister to catarrhine primate FVs, reflecting the branching order of their hosts. However, the branching orders within the NWM FV clade clearly do not parallel that of their platyrrhine hosts [60,74,75] (Figure 4), with SFVmar from a common marmoset being more closely related to SFVspm from a spider monkey than to SFVsqu from a squirrel monkey, as would be expected under a co-evolutionary scenario. We compared our inferred phylogeny against an alternative phylogeny in which NWM FVs co-speciate with their hosts using approximately unbiased tests [67], and found that the co-speciation picture is rejected (Additional file 4: Figure S3-B). These results are consistent with results from a previous study that used an alignment of Gag protein sequences [19] and implied an

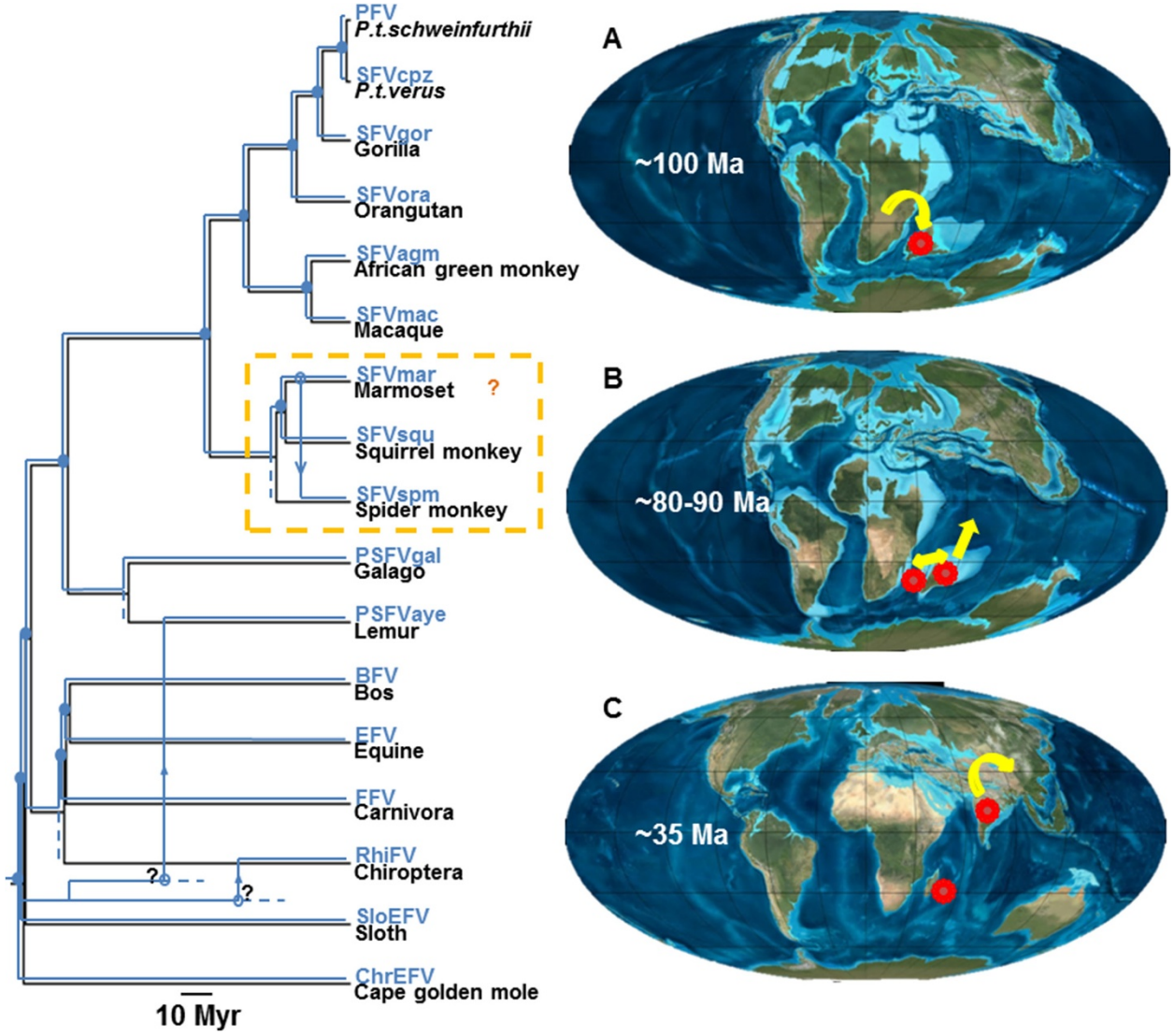


Figure 4 Hypothesis explaining the mismatches in foamy virus (FV)-host evolutionary history. FV phylogeny (blue) was superimposed upon host phylogeny (black) and is scaled to host divergence times. The scale bar is in units of millions of years. Solid circles represent FV-host co-speciation events. Open-circles represent possible host switching events of which the inferred switching directions are indicated by arrows. A hypothetical scenario of host switching (as indicated by an orange '**?**') within the New World monkey FV clade is shown in an orange box. An alternative scenario is shown in Additional file 5: Figure S4. For PSFVaye and RhiFV, we speculate that, in the very early history of mammalian evolution, eutherians came to the Madagascar-India landmass with their FVs (red circles) upon which they established stable populations **(A)**. The landmass was then split into the Madagascar and India landmasses resulting in the FV population splitting into two separate groups ~80-90 Ma **(B)**; one FV variant remained on Madagascar and the other was transported to Laurasia via the Indian landmass continental drift **(C)**. The former gave rise to the PSFVaye progenitor while the latter gave rise to RhiFV. Directions of infected FV host population movements are indicated by yellow arrows. However, it is unknown which ancestral species introduced FV to the lemur and bat lineages and when the transmissions occurred (as indicated by a black '**?**'). Additional sequence data from other mammals, especially bats and lemurs, are required to further resolve these aspects of FV evolutionary history.

ancestral NWM FV host switch (Figure 4 and Additional file 5: Figure S4, orange boxes). Marmosets, squirrel monkeys, and spider monkeys are closely related, occupy large overlapping geographic ranges [19], and are commonly used in biomedical research. It is thus plausible to imagine a scenario under which heterologous NWM FV infections might have occurred in the past, either in the wild or during captivity. This is unsurprising given that this type of FV cross-species transmission between closely related primate species has already been documented [30]. In contrast to our results, a phylogenetic analysis of a wider range of NWM FVs using short *pol* nucleotide sequences (276 nt) suggested that these three NWM FVs co-speciate with their hosts and that host switches

occurred elsewhere [76]. Analysis of additional NWM FV complete genomes will be necessary to distinguish these possibilities.

A more extreme case of host switching across taxonomic orders may explain the phylogenetic placement of PSFVaye and RhiFV, where both are placed robustly outside the boreoeutherian FV clade (posterior probability = 1). Nevertheless, one possibility for the inferred placement of PSFVaye may be that it is an artefact due to neutral genetic changes that have accumulated since its endogenization [38]. To examine whether or not neutral evolution alone could be responsible for the observation, Bayesian phylogenies of concatenated Pol-Env protein alignments containing artificially and uniformly mutated SFVagm, SFVsqu, and PSFVgal sequences were inferred. These sequences were mutated *in silico* with an overestimate of the mammalian neutral substitution rate of 10E-9 substitutions per site per year (s/n/y) over a mean estimate of PSFVaye endogenization date of 40 Myr ago (Figure 2). The substitution process was assumed to be neutral, homogenous, and independent across sites and base types. Re-aligning the sequences after the simulation using MUSCLE implemented in MEGA6 [77] does not change the relative position of mutated sequences in other SFV sequences. Our results suggest that, even with this biased substitution rate setting of 10E-9 s/n/y, the neutral genetic change accumulation alone is insufficient to explain the inferred phylogenetic position of PSFVaye. Our *in silico* analysis only increased the branch length and decreased the branch support without changing the tree topology (Additional file 7: Figure S6). It is thus unlikely that the placement of PSFVaye is an artefact. Similarly, the placement of RhiFV is also likely to be genuine since it is an extant exogenous FV, supported by the absence of in-frame stop codons or frameshift mutations, and that it was found to be present in viral particles and not host genomic material [26]. In addition, the RhiFV lineage remains in the same phylogenetic position even after removal of PSFVaye from the phylogenetic analysis.

We then asked whether there is a reasonable FV host switching model that can explain both (i) the deep phylogenetic placement of PSFVaye and RhiFV which is basal to all known exafroplacentalian FVs, and (ii) the dispersed geographical distribution of their hosts (Madagascar and China, respectively). We propose the following scenario to explain the phylogenetic relationships of PSFVaye and RhiFV within the context of the evolutionary and geographical history of FV hosts (Figure 4). Approximately 180–160 Ma, Gondwana was split into two separate landmasses: the Madagascar-India-Antarctica (MIA) landmass and the Africa-South America (ASA) landmass [78,79]. The MIA landmass then split into the Madagascar-India and the Antarctica landmasses about 135 Ma [78,79]. Thereafter, but still in the very early period of eutherian evolutionary history, eutherians arrived onto the Madagascar-India landmass upon which their FVs established stable populations (Figure 4A). Since then, these early eutherian FVs evolved independently from, and hence have distant evolutionary relationships with, FVs circulating on the African landmass, which later would give rise to all extant SFVs, lorisiforme FVs and fereungulata FVs. The India landmass was then split from the Madagascar landmass about 80–90 Ma [78,79] (Figure 4B), separating the FV populations into two groups: one circulating in Madagascar which gave rise to the progenitor of PSFVaye and the second transferred to Laurasia via the India landmass continental drift which independently gave rise to RhiFV (Figure 4C). However, the model does not suggest which ancestral species introduced FV to the lemur and bat lineages and when these transmissions occurred. Nonetheless, both PSFVaye and RhiFV are placed robustly with exafroplacentalian FVs but outside the boreoeutherian FV clade (posterior probability = 1 and 1, respectively). This suggests that the original hosts were exafroplacentalians, and not boreoeutherians, and that the host switches occurred between species across deep clades long diverged from each other. Sequence data of FVs from additional mammals, especially bats and lemurs, are required to examine further the evolutionary histories of PSFVaye and RhiFV.

Our hypothesis helps to explain the deep phylogenetic placement and dispersed geographical distribution of PSFVaye and RhiFV. In addition, this hypothesis also gives an explicit prediction that PSFVaye and RhiFV should share a MRCA prior to 80–90 Ma based on the estimated Madagascar-India landmass split date [78,79]. Indeed, although not well-supported (posterior probability = 0.56), our best phylogenetic estimate suggests that PSFVaye and RhiFV have a sister-group relationship (Figure 3A). Assuming that our inferred PSFVaye-RhiFV relationship is correct, we estimated the tMRCA of PSFVaye and RhiFV to be 93 (85–101) Myr based on the linear relationship of FV-host diversity (Figure 3B). Although slightly high, our estimate range largely overlaps with the date range derived from this hypothesis. The accumulated neutral changes in the PSFVaye sequences may explain both the weak branch support and the slightly overestimated divergence date.

Conclusions

Here, we report the characterization of the complete PSFVgal genome obtained from a galago, and describe in more detail the partial genome of PSFVaye from the aye-aye. The genomic organization of PSFVgal and PSFaye is characteristic of FVs, and they are phylogenetically placed within the FVs. The defective nature of the PSFVaye genome, coupled with its robust amplification from genomic DNA and an absence of antibodies to the genetically related PSFVgal in infected animals, indicates that it is endogenous. In contrast, PSFVgal has a complete

viral transcriptome, elicits a FV-specific immune response, and is not present in all individuals, consistent with other exogenous FVs. We also report the discovery and describe a novel ERV present in the Cape golden mole genome, ChrEFV. Genomic inspection and analysis suggests that ChrEFV is also an endogenous FV, and its phylogenetic placement is consistent with a history of co-speciation between FVs and their mammalian hosts. However, ChrEFV has a high level of accumulated sequence divergence compared to other FVs, likely an artefact of relaxation of evolutionary constraints since becoming endogenous.

We found a general, stable pattern of mammalian FV-host co-divergence which extends as deep as the exafroplacentalian basal diversification, spanning more than 100 Myr. To date, it is still poorly understood why FVs stably co-speciate with their natural hosts. Furthermore, we also identified two possible cases of host switching in the evolutionary history of FVs. The first was observed in NWMs, as previously shown [19], which may have happened during captivity or in the wild and which has previously been reported to rarely occur in Old World monkeys and apes [30]. However, others have shown congruent NWM and SFV phylogenies using a larger distribution of species-specific sequences [76]. The second involves PSFVaye and RhiFV which may involve cross-species transmission at the level of mammalian orders. We propose a scenario based upon geographical knowledge of continental drift and hypothetical migration of ancient eutherians to explain this observation. Our results highlight the value of integrating multiple sources of information to elucidate the evolutionary history of mammals and their viruses, including continental and geographical histories, ancestral host locations, in addition to the natural history of host and virus.

Methods

Nonhuman primate (NHP) samples and nucleic acid preparation

NHP specimens (serum, frozen or fresh whole blood, tissues) were purchased from the Duke Lemur Center, were collected on an opportunistic basis from seven U.S. zoos following approved animal use protocols, or dried blood spots (DBS) via NHP hunters in Cameroon from freshly hunted NHP bushmeat in a study approved by Institutional Animal Care and Use Committees and the Cameroon government (Additional file 1: Table S2). Hunters were educated about the risks associated with direct contact with NHPs and about appropriate prevention measures. Prior to processing, all specimens were stored at −20°C or −80°C, except DBS which were stored in nytran ziplock bags with dessicant at room temperature. DNA lysates or nucleic acids were prepared from blood specimens and DBS as previously described [28]. Nucleic acids were extracted from archived tissue specimens from the Duke Lemur Center (liver, spleen, kidney) using a Qiagen QIAmp DNA mini kit. DNA concentrations were determined using a Nanodrop spectrophotometer and DNA integrity was confirmed with ß-actin PCR.

Isolation of PSFVgal

HeLa cells were infected with PSFVgal (SFV-5; ATCC# VR-644), originally isolated from the throat swab of an *Otolemur crassicaudatus panganiensis*, and grown in complete DMEM medium supplemented with 10% FBS, 2 mM L-glutamine, 100 ug/ml streptomycin and 100 U/ml penicillin (Invitrogen). PSFVgal is the only viral isolate analyzed in our study. Cell cultures were incubated at 37°C and 5% CO2 and were split when 90% confluent. Cellular DNA was prepared using a Qiagen kit and quantified with a Nanodrop spectrophotometer when cytopathic effect was observed in >50% of the cells.

Western blot (WB) assay

PSFVgal-infected and uninfected cells were pelleted by centrifugation at 1,500 rpm for 10 min and washed 2X with phosphate-buffered saline. Antigen for WB testing was prepared by treating cell pellets with WB sample buffer (Invitrogen) at 100°C for 10 min. Protein concentrations were determined using the BioRad DC Protein Assay kit and 150 ug of infected and uninfected HeLa cell lysates were applied separately to 4-12% polyacrylamide gels (Invitrogen) for WB analysis. Serum specimens were tested using a 1:50 dilution as previously described [80].

PCR-amplification of 5′ and 3′ PSFVgal genome halves, plasmid cloning, and sequence analysis

To obtain the full-length genomic sequence of PSFVgal we first PCR-amplified small regions in the LTR and polymerase (*pol*) gene by using nested PCR with degenerate SFV primers and conditions provided elsewhere [80,81]. PSFVgal-specific primers were then designed from the LTR and *pol* fragments to amplify two overlapping genomic halves using PCR and a Roche Expand 20 kb PCR System Kit following the manufacturer's instructions. The Expand kit contains thermostable Taq DNA polymerase and a thermostable DNA polymerase with proofreading activity to minimize PCR-induced sequence artefacts. Briefly, the PCR reaction mixture was prepared in 50 ul reaction volumes containing 500 uM of each dNTP, 400 nM of each forward and reverse primer, and 5 units of Expand 20 kb enzyme mix using 0.5 ug of PSFVgal tissue culture DNA. The LTR-*pol* fragment was amplified using the primers PSFVgal-LF1 5′ GGC TTG GAT AAT TAA TTG TTA GAT GCT CTG 3′ and PSFVgalpolF2R 5′ GTT CCA AAC GTA TGC CCC TCT CCT T 3′. The PSFVgal *pol*-LTR fragment primers are PSFVgalpolF1 5′ GTC AGC ATT CAC CTC TTC CAC CTT G 3′ and PSFVGAL-LR1 5′ GAC TTA TTT ATT ACT GCA AGA CCC GAG AGG G 3′. Following denaturation at 92°C for 2 min, amplification consisted of 10 cycles at 92°C for 10

secs, 60°C for 30 secs, and 68°C for 6 mins, followed by 20 cycles at 92°C for 10 secs, 60°C for 30 secs, and 68°C for 6 mins with an additional 10 sec per cycle.

PCR products were visualized with 0.8% agarose gel electrophoresis and amplicons of the expected size were collected using a QIAGEN gel purification kit. Purified amplicons were cloned into the pCR-XL-TOPO vector (Invitrogen) for genome sequencing. Two plasmids were obtained; pCR-XL-TOPO- PSFVgal-LP containing the LTR-*pol* insert (5,148-bp) and pCR-XL-TOPO- PSFVgal-PL containing the *pol*-LTR insert (6,201-bp). The PSFVgal plasmid DNAs were purified using the Purelink Hipure Plasmid Filter Purification Kit (Invitrogen) and sequenced in both directions using a 3130XL Genetic Analyzer (Applied Biosystems). Complete genomes were assembled using the software program Vector NTI v11.1. 5′ and 3′ LTRs were determined based on overlapping sequences and comparison with available FV genome sequences.

Homology of PBS sequences to tRNAs was inferred using the transfer RNA database (http://trnadb.bioinf.uni-leipzig.de/). Confirmation of the tRNA PBS sequence was done by PCR amplification with primers flanking that region and using 0.1 and 1.0 ug of PSFVgal-infected HeLa tissue culture DNA. Primary and nested primers were used to independently amplify two products of different lengths and amplification was done separate from that to obtain the 5′ halve of the genome containing the PBS. Primary PCR primer sequences are F982 (5′-GAC CAG TGT GAG ATT GGT GTC TC-3′) and R1547 (5′-CCA GTT GCC TCC TGT GAT TCT AAC-3′) and the internal primers are F1133 (5′-CTC CTG GTT GAG GAC AAG GGA AC-3′) and R1475 (5′-CCA ATT TGT GCT CGT GGC ACT GG-3′) and standard PCR conditions were used. Numbers in the primer names reference their locations in the PSFVgal genome. PCR products from both primary and nested and 0.1 and 1.0 ug DNA inputs were sequenced.

Discovery and characterization of ChrEFV and PSFVaye

Publically available GenBank whole genome shotgun (WGS) sequences were screened using tBLASTn and various FV protein sequences (Additional file 1: Table S3). As previously reported by Han and Worobey [38], several matches were returned from the WGS assembled sequence of the aye-aye, *Daubentonia madagascariensis,* spanning four non-overlapping contigs. The adjacency of these four contigs and sequences flanking the *gag* and *env* regions, especially the LTR and host integration sequences, were determined using a Seegene genome walking kit (Seegene, MD, USA). Three rounds of PCR were performed using universal primers provided in the kit (DW2-ACP, DW2-ACPN, UNIP2) with PSFVaye-specific primers (1st PCR primers: upstream PSFVAYE-1174R/DW2-ACP; downstream, PSFVAYE-6122 F/DW2-ACP; 2nd PCR primers: upstream, PSAGF2REV/DW2-ACPN; downstream, PSAER2FOR/DW2-ACPN; 3rd PCR primers: upstream, PSAGF1REV/UniP2; downstream, PSAER1FOR/UniP2). Positive bands were cloned and Sanger sequenced using an ABI7700 instrument. Furthermore, this *in silico* screening process also returned one contig from the WGS assembled sequence of a Cape golden mole (*Chrysochloris asiatica*), designated 'ChrEFV' here. Both were annotated via sequence homology. The presence of simple repetitive elements was determined by CENSOR (http://www.girinst.org/censor/).

PCR testing of Prosimian genomic DNA specimens for FV

1 ug of prosimian DNA was PCR tested for FV *pol* sequences using a combination of nested primer sets or a single round of PCR to detect PSFVaye *gag* sequences depending on the species (Additional file 1: Table S2). The first PCR assay is PSFVgal-specific and is based on the PSFVgal genome and uses the first and second round PCR primers SIF1GAL 5′ CTT GCT GTG CAG AGC AGT CAC AAG GT 3′ and SIR1GAL 5′ GTT TTA TTT CAC TGT TTT TCC GTT CCA C 3′ and SIF3GAL 5′ CCA AGT CTG GAT GCA GAG CTT ATC CA 3′ and SIR3-GAL 5′ ACT TTG GGG GTG ATA CGG AGT ACT 3′ to generate 712-bp and 635-bp fragments, respectively. The second assay was designed using an alignment of complete Old World and New World primate SFV genomes and uses the primary primers SIF5N 5′ TAC ATG GTT ATA CCC CAC KAA GGC TCC TCC 3′ and SIR5N 5′ AAT AAW GGA TAC CAC TTT GTA GGT CTT CC 3′ and nested primers SIP4N 5′ TGC ATT CCG ATC AAG GAT CAG CAT T 3′ and SIR1NN 5′ GTT TTA TYT CCY TGT TTT TCC TYT CCA CCA T 3′ to generate to generate 282-bp and 141-bp *pol* sequences, respectively. The third assay was designed using an alignment of PSFVgal and PSFVaye genomes and uses the primary primers 3′ FVPF05 5′ KKM TAY TGG TGR CCT AAT ATG 3′ and FVPFR5 5′ GGT CTW CCA ACY ART AGT TTA G 3′ and nested primers FVPF01 5′ CCT TTT GAT AAA ATY TAT ATG G 3′ and FVPR01 5′ CAS CTT TCC ACT ACT TTG G 3′ to generate 483-bp and 292-bp products, respectively. To detect PSFVaye *gag* sequences the primers PSAGF1 5′ AAG ACC CTT GCT GCC TAA TGT TGG 3′ and PSAGR1 5′ TAT TTG TAA CCA GGG CTT GAC CAG 3′ were used to amplify a 475-bp sequence. 5 ul of primary PCR product was used as template in the nested PCR reaction. Selected PCR products were visualized by agarose gel electrophoresis analysis, extracted using a QIAquick Gel Extraction kit, and sequenced in both directions using an ABI7700 instrument.

Estimating the integration dates of PSFVaye and ChrEFV with Monte Carlo simulation

To estimate the age of PSFVaye and ChrEFV, we analyzed the frequencies of in-frame stop-codons in their Pol

protein sequences. Excluding a hypothetical last stop codon, PSFVaye (1,163 codons) and ChrEFV (975 codons) Pols contain 9 and 14 in-frame stop codon mutations, which are equivalent to stop codon frequencies of 0.00774 and 0.0144, respectively. Using a Monte Carlo simulation approach and *pol*s of PSFVgal, SFVcpz, SFVagm, SFVsqu, FFV, and SloEFV as model sequences, we estimated two cumulative distribution functions (CDFs) of time duration for in-frame stop codons to be accumulated at these two frequencies for each model sequence. Each *pol* model sequence was mutated *in silico* over various hypothetical time durations from 5 Myr to 200 Myr with an increment of 5 Myr. The mutation (substitution) process was assumed to be neutral, homogenous, and independent across all sites and base types with the rate of 2.2E-9 s/n/y, an average substitution rate of mammalian genomes [82]. 1,000 simulations were performed for each period of time and the frequency of in-frame stop codons was then calculated for each simulation. Two separate CDFs of time duration were constructed for each sequence model using stop-codon frequencies of 0.00774 and 0.0144. In total, we constructed six CDFs for each of the two stop-codon frequencies. Means, medians, modes and 95% confidence intervals of the time duration for in-frame stop codons to accumulate at the two frequencies were calculated.

Phylogenetic analysis

To investigate the phylogenetic relationships of PSFVgal, PSFVaye, and ChrEFV with other retroviruses, a consensus unrooted phylogenetic tree was constructed using a manually curated alignment of 240 RT protein sequences (162 aa) from alpha-, beta-, gamma-, delta-, epsilon-, lenti-, and spuma-like retroviruses using RAxML 7.2.8-HPC2 on XSEDE [83]. The alignment is available from the authors upon request. The LG + G + F model was determined to fit the data best by using ProtTest 2.4 [84]. Bootstrap support values for the branching order were calculated using 5,000 pseudoreplicates.

To further investigate the phylogenetic relationships of PSFVgal, PSFVaye, and ChrEFV with other FVs, a Bayesian FV phylogeny was estimated using MrBayes 3.2.1 [85] and a manually-curated alignment of concatenated Pol-Env sequences (893 and 603 amino acids, respectively). We used CoeEFV [39] to root our mammalian FV tree since it is the most immediate outgroup of mammalian FVs known to date although it is still debatable whether this ERV is a real endogenous FV or not. Gag sequences were not included in the alignment since (i) the Gag of RhiFV is not available and (ii) the alignable region of Gag for the rest of the FVs is relatively short (~180 aa). Potential recombination among the sequences within the alignment were checked using VisRD3 [64] both at nucleotide and protein levels. For both analyses, the null distribution was built based on 1,000 sets of randomly-shuffled sequences and the extended statistical geometry (Hamming) weighting option was selected. In the nucleotide analysis, the window size and step size was 300 and 60 nt, respectively; while in the protein analysis, the window size and step size was 100 and 20 aa, respectively. The results showed no significant evidence for recombination. The alignment is available from the authors upon request. The protein phylogeny reconstruction was performed using best available protein-specific model partitions of rtREV + Γ(4) + F (Pol) and WAG + I + Γ(4) + F (Env) as determined by ProtTest 2.4 [84] under the AIC criterion. The MCMC was run for 5,000,000 steps with the initial 25% discarded as burn-in. Trees and parameters were logged every 100 steps. Parameter value convergence and sampling independency were manually inspected and had effective sample sizes >1,000.

We evaluated the FV-host co-speciation hypothesis using two analyses: (i) FV-host phylogenetic reconciliation analysis and (ii) FV-host divergence correlation analysis. Phylogenetic reconciliation analysis of FV and host trees was performed using Jane v4.0 [65] (Genetic algorithm: number of generations = 100, population size = 100). The vertex-based cost mode was used, and the costs were set as follows: co-speciation = −1, duplication = 0, duplication & host switch = 0, loss = 0, and failure to diverge = 0. This setting was adopted in order to maximize the number of co-speciation events. The probability of observing the inferred number of potential co-speciation events by chance was calculated by random tip mapping in Jane v4.0 [65] (Genetic algorithm: number of generation = 100, population size = 100, sample size = 1,000). This method of analyzing phylogenetic incongruence cannot formally differentially weight transmission at different taxonomic levels however, and the derivation of p-values relies upon the interchangeability of branches. Thus, it does not discriminate between distant and close interspecies transmissions. Once the potential FV-host co-speciation events were located, we then identified FV-host corresponding branches that represent FV-host co-evolutionary histories for the FV-host divergence correlation analysis. In this test, we investigated whether or not the lengths of these corresponding branches were linearly correlated, i.e. whether or not the accumulation of FV sequence divergence occurred in proportion to host divergence, using linear regression implemented in MATLAB [86]. Outliers were determined by using Cook's distance inspection. Outliers were removed one at a time using a threshold value of 3x mean of the observed Cook's distances and the model was re-fitted until no outliers were found. The model was then used to estimate how long the parental branch of RhiFV and PSFVaye branches (0.0564 amino-acid substitutions per site) is in units of time. Given that the radiation of exafroplacentalians was 101.1 Ma [43], we then re-calculated when PSFVaye and RhiFV shared a MRCA to be 93.342 (85.095-101.589) Ma.

Our inferred phylogeny suggests two possible cases of FV host-switches: one involves PSFVaye and RhiFV and the other involves NWM FVs. Approximate unbiased (AU) tests using multi-scale bootstrapping [67] were performed to investigate whether or not our Pol-Env/Pol/Env alignments reject the co-speciation model for PSFVaye and RhiFV, and NWM FVs (Additional file 4: Figure S3). Four completing alternative phylogenetic placements of PSFVaye and RhiFV (Additional file 4: Figure S3-A), and two alternative placements of NWM FVs (Additional file 4: Figure S3-B) are compared to each other. The site-wise log likelihoods used in the tests were computed using PAML 4.7a [87]. The best available amino acid substitution model, as determined by ProtTest 2.4 [84] under the AIC criterion, was used (Pol: rtREV + Γ(4) + F; Env: WAG + Γ(4) + F). A molecular clock was not imposed, and ambiguous sites were included. AU tests were performed in Consel [88] with default settings.

Additional Bayesian FV phylogenies were constructed using the same settings as described above. One included a number of short fragments of other lorisiforme FV Pol sequences (IN core domain sequences, 85 aa) to further investigate phylogenetic relationships of lorisiforme FVs with other FVs. The additional Bayesian phylogenies included several *in silico* mutated sequences to examine the effects of neutral genetic changes on the inferred phylogenetic relationships. The alignment was not re-aligned after the sequence simulation (see main text). Both alignments are available from the authors upon request.

Additional files

Additional file 1: Table S1. Mammalian foamy viruses. **Table S2.** Distribution of simian foamy virus in prosimians. **Table S3.** GenBank accession numbers of protein sequences used as probes to search for integrated mammalian foamy viruses (FVs) as well as for phylogenetic analyses.

Additional file 2: Figure S1. Detailed descriptions of PSFVgal, PSFVaye, and ChrEFV complete or partial genomes. Protein open reading frame (ORF) and putative reading frame (RF) locations were determined by sequence similarity to other FVs and by searching against the GyDB (http://www.gydb.org/). Short interspersed nuclear elements (SINEs) were determined by CENSOR (http://www.girinst.org/censor/). TATA boxes, polyadenylation signals, primer binding sites (PBS), and polypurine tracts (PPT) are indicated by bold type and solid lines. Long terminal repeats (LTRs), Gag, Pol, Env, Bel1, and Bel2 proteins are highlighted in grey, green, purple, blue, orange, and red, respectively. Several conserved domains and catalytic centers are highlighted in darker colors. Putative insertion and deletion mutations are indicated by red dashes. The beginning and the end of ORFs and RFs are indicated by arrows and square-arrows, respectively.

Additional file 3: Figure S2. A consensus unrooted retrovirus reverse transcriptase phylogeny. The phylogeny was built using an alignment of reverse transcriptase proteins of several retrovirus genera: alpharetrovirus (α, light green), betaretrovirus (β, dark green), gammaretrovirus (γ, blue), deltaretrovirus (δ, orange), epsilonretrovirus (ε, red), lentiretrovirus (Lenti, yellow), and spuma-like retrovirus (Spuma-like, purple). The final alignment length consisted of 240 taxa and 162 amino acids in length. The tree was constructed using RAxML 7.2.8-HPC2 on XSEDE [83]. The best substitution model was determined by ProtTest 2.4 [84] to be the LG + G + F model. The scale bar is in units of amino acid substitution per site. Numbers on nodes are bootstrap support values estimated using 5,000 pseudoreplicates. PSFVgal, PSFVaye, and ChrEFV are indicated in bold type and with asterisks. JSRV, jaagsiekte sheep retrovirus; MMTV, mouse mammary tumor virus; HERV, human endogenous retrovirus (ERV); LPDV, lymphoproliferative disease virus; EIAV, equine infectious anemia virus; BIV, bovine immunodeficiency virus; HIV, human immunodeficiency virus; SIV, simian immunodeficiency virus; BLV, bovine leukemia virus; HTLV, human T-lymphotropic virus; PFV, prototype FV; CoeEFV, coelacanth EFV; ZFERV, zebrafish ERV; SpeV, sphenodon ERV; MeEV, Meles endogenous virus (EV); VuEV, Vulpes EV; MLV, murine leukemia virus; MDEV, Mus dunni EV; MuRRS, murine retrovirus-related sequence; Xen, Xenopus laevis ERV; WDSV, walleye dermal sarcoma virus.

Additional file 4: Figure S3. Foamy virus (FV)-host evolutionary conflicts. Our analyses suggested two major FV-host evolutionary conflicts: one is of PSFVaye and RhiFV (**A**), and the other is of New World monkey (NWM) FVs (**B**). **A.1** shows our best estimated phylogeny in which both PSFVaye and RhiFV are placed robustly outside the Boreoeutherian FV clade but still remain together with exafroplacentalian FVs (see main text). **A.2-4** show alternative phylogenies in which either PSFVaye or RhiFV (**A.2 & A.3**) or both (**A.4**) co-speciate with their hosts. We used approximately unbiased (AU) [67] tests to compare these alternative phylogenies against one another given Pol-Env/Pol/Env alignments, based on comparisons of site-wise log-likelihood scores (LLH) computed using PAML 4.7a [87]. The AU tests were performed in Consel [88]. The same analyses were performed to compare our best estimates of the phylogenetic placement of NWM FVs which show FV-host evolutionary conflicts (**B.1**) against topologies that do not (**B.2**). Results are summarized in the table; only total LLH scores and AU probabilities (p-AUs) are shown.

Additional file 5: Figure S4. Inferred foamy virus (FV)-host co-speciation events using co-phylogeny reconstruction software Jane v4.0 [65]. The FV Pol-Env protein phylogeny (blue and green) was superimposed upon a previously published host phylogeny (black, published in [43]) and is scaled to host divergence times. The scale bar is in units of millions of years. Concatenated Pol-Env protein alignments (893 and 603 amino acids, respectively) were used to infer the FV trees. Solid circles represent FV-host co-speciation events, open circles represent host switching events, and an asterisk represents a lineage duplication event in the absence of host diversification which gave rise to two FV lineages sharing host species. The thick lines in the FV phylogeny correspond to the evolutionary period spanning about 30 Myr, during which the two lineages shared their host species. A hypothetical scenario of host switching of New World monkey FVs (as indicated by an orange '**?**') is shown in an orange box. An alternative scenario is shown in Figure 4. (JPEG 43 kb)

Additional file 6: Figure S5. Co-speciation between lorisiforme foamy viruses (FVs) and their hosts. FV phylogeny (left) was estimated using an alignment of short integrase protein sequences of lorisiformes (indicated with asterisks, 85 amino acids (aa)) and concatenated polymerase-envelope protein sequences of other eutherian FVs (893 aa and 603 aa, respectively). The phylogeny was inferred using Bayesian methods in the program MrBayes 3.2.1 [85]) and rooted with CoeEFV. Numbers at branch nodes are posterior probabilities and the scale bar is in units of amino-acid substitutions per site. The topology of the lorisiforme FV phylogenetic relationships is compared to that of their prosimian hosts (right, previously published in [43]). Solid lines between the two trees indicate FV-host associations.

Additional file 7: Figure S6. Investigation of the effects of neutral genetic changes on foamy virus (FV) phylogenetic relationships. We mutated the polymerase and envelope nucleotide sequences of SFVagm, SFVsqu, and PSFVgal (indicated with asterisks) *in silico* with a substitution rate of 10E-9 substitutions per site per year over a hypothetical period of 40 Myr. The mutated gene sequences were then translated into protein sequences and the FV phylogeny was re-built as described using Bayesian methods using MrBayes 3.2.1 [85]. The substitution process was assumed to be neutral, homogenous, and independent across all sites and base types. Numbers at branch nodes are posterior probabilities and the scale bar is in units of amino acid substitutions per site.

Competing interests
The authors declare that they have no competing interests.

Authors' contributions
AK and WMS conceived the project, AK, PA and WMS performed the sequence analyses, PA and AK performed the evolutionary analyses, HJ performed the laboratory work and assembled the PSFVgal genome, and NDW, ML, ADY provided samples. AK, PA and WMS wrote the paper. All authors read and commented on the manuscript. All authors read and approved the final manuscript.

Acknowledgements
We thank the seven U.S. zoos, including the San Diego Zoo, Zoo New England, the Gladys Porter Zoo, the Baltimore Zoo, the Detroit Zoo, and two other zoos that wish to remain anonymous, and the Duke Lemur Center for providing us with specimens for the study. We also thank David Sintasath and Michelle Sturgeon for help preparing nucleic acids from the potto DBS. Funding for this study was provided in part by the Faucett Family and Rawji Foundations and Global Viral Forecasting (ANRS12182/12255 and NIH RO1 AI50529), graciously supported by the U.S. Department of Defense Armed Forces Health Surveillance Center, Division of Global Emerging Infections, Surveillance Operations (AFHSC GEIS), the Defense Threat Reduction Agency Cooperative Biological Engagement Program (DTRA-CBEP), Department of Defense HIV/AIDS Prevention Program (DHAPP), Google.org, the Skoll Foundation, and the U.S. Agency for International Development (USAID) Emerging Pandemic Threats Program, PREDICT project, under the terms of Cooperative Agreement Number GHN-A-OO-09-00010-00. Use of trade names is for identification only and does not imply endorsement by the U.S. Department of Health and Human Services, the Public Health Service, or the Centers for Disease Control and Prevention. The findings and conclusions in this report are those of the authors and do not necessarily represent the views of the Centers for Disease Control and Prevention. P.A. is funded by the Royal Thai Government. AK is funded by the Royal Society. We thank R. Blakey for tectonic maps. This is DLC publication number 1273.

Author details
[1]Department of Zoology, University of Oxford, Oxford, South Parks Road, Oxford OX1 3PS, UK. [2]Laboratory Branch, Division of HIV/AIDS, National Center for HIV, Hepatitis, STD, and TB Prevention, Centers for Disease Control and Prevention, Atlanta, GA 30333, USA. [3]Metabiota, One Sutter, Suite 600, San Francisco, CA 94104, USA. [4]Program in Human Biology, Stanford University, Stanford, CA 94305, USA. [5]Global Viral, One Sutter, Suite 600, San Francisco, CA 94104, USA. [6]Mosaic (Environment, Health, Data, Technology), Yaoundé, Cameroon. [7]Department of Biology, Duke University, Durham, NC 27708, USA.

Received: 19 February 2014 Accepted: 2 July 2014
Published: 4 August 2014

References

1. Linial ML: **Foamy viruses are unconventional retroviruses.** *J Virol* 1999, **73:**1747–1755.
2. Enders JF, Peebles TC: **Propagation in tissue cultures of cytopathogenic agents from patients with measles.** *Proc Soc Exp Biol Med* 1954, **86:**277–286.
3. Rustigian R, Johnston P, Reihart H: **Infection of monkey kidney tissue cultures with virus-like agents.** *Proc Soc Exp Biol Med* 1955, **88:**8–16.
4. Johnston PB: **A second immunologic type of simian foamy virus: monkey throat infections and unmasking by both types.** *J Infect Dis* 1961, **109:**1–9.
5. Stiles GE, Bittle JL, Cabasso VJ: **Comparison of simian foamy virus strains including a new serological type.** *Nature* 1964, **201:**1350–1351.
6. Johnston PB: **Taxonomic features of seven serotypes of simian and ape foamy viruses.** *Infect Immun* 1971, **3:**793–799.
7. Rogers NG, Basnight M, Gibbs CJ, Gajdusek DC: **Latent viruses in chimpanzees with experimental kuru.** *Nature* 1967, **216:**446–449.
8. Hooks JJ, Gibbs CJ, Chou S, Howk R, Lewis M, Gajdusek DC: **Isolation of a new simian foamy virus from a spider monkey brain culture.** *Infect Immun* 1973, **8:**804–813.
9. Hooks JJ, Gibbs CJ: **The foamy viruses.** *Bacteriol Rev* 1975, **39:**169–185.
10. Heberling RL, Kalter SS: **Isolation of foamy viruses from baboon (Papio cynocephalus) tissues.** *Am J Epidemiol* 1975, **102:**25–29.
11. McClure MO, Bieniasz PD, Schulz TF, Chrystie IL, Simpson G, Aguzzi A, Hoad JG, Cunningham A, Kirkwood J, Weiss RA: **Isolation of a new foamy retrovirus from orangutans.** *J Virol* 1994, **68:**7124–7130.
12. Bieniasz PD, Rethwilm A, Pitman R, Daniel MD, Chrystie I, McClure MO: **A comparative study of higher primate foamy viruses, including a new virus from a gorilla.** *Virology* 1995, **207:**217–228.
13. Marczynska B, Jones CJ, Wolfe LG: **Syncytium-forming virus of common marmosets (Callithrix jacchus jacchus).** *Infect Immun* 1981, **31:**1261–1269.
14. Riggs JL, Oshirls, Taylor DO, Lennette EH: **Syncytium-forming agent isolated from domestic cats.** *Nature* 1969, **222:**1190–1191.
15. Malmquist WA, Van der Maaten MJ, Boothe AD: **Isolation, immunodiffusion, immunofluorescence, and electron microscopy of a syncytial virus of lymphosarcomatous and apparently normal cattle.** *Cancer Res* 1969, **29:**188–200.
16. Tobaly-Tapiero J, Bittoun P, Neves M, Guillemin MC, Lecellier CH, Puvion-Dutilleul F, Gicquel B, Zientara S, Giron ML, de Thé H, Saïb A: **Isolation and characterization of an equine foamy virus.** *J Virol* 2000, **74:**4064–4073.
17. Kupiec JJ, Kay A, Hayat M, Ravier R, Périès J, Galibert F: **Sequence analysis of the simian foamy virus type 1 genome.** *Gene* 1991, **101:**185–194.
18. Renne R, Friedl E, Schweizer M, Fleps U, Turek R, Neumann-Haefelin D: **Genomic organization and expression of simian foamy virus type 3 (SFV-3).** *Virology* 1992, **186:**597–608.
19. Pacheco B, Finzi A, McGee-Estrada K, Sodroski J: **Species-specific inhibition of foamy viruses from South American monkeys by New World Monkey TRIM5α proteins.** *J Virol* 2010, **84:**4095–4099.
20. Herchenröder O, Renne R, Loncar D, Cobb EK, Murthy KK, Schneider J, Mergia A, Luciw PA: **Isolation, cloning, and sequencing of simian foamy viruses from chimpanzees (SFVcpz): high homology to human foamy virus (HFV).** *Virology* 1994, **201:**187–199.
21. Thümer L, Rethwilm A, Holmes EC, Bodem J: **The complete nucleotide sequence of a New World simian foamy virus.** *Virology* 2007, **369:**191–197.
22. Verschoor EJ, Langenhuijzen S, van den Engel S, Niphuis H, Warren KS, Heeney JL: **Structural and evolutionary analysis of an orangutan foamy virus.** *J Virol* 2003, **77:**8584–8587.
23. Schulze A, Lemey P, Schubert J, McClure MO, Rethwilm A, Bodem J: **Complete nucleotide sequence and evolutionary analysis of a gorilla foamy virus.** *J Gen Virol* 2011, **92:**582–586.
24. Winkler I, Bodem J, Haas L, Zemba M, Delius H, Flower R, Flügel RM, Löchelt M: **Characterization of the genome of feline foamy virus and its proteins shows distinct features different from those of primate spumaviruses.** *J Virol* 1997, **71:**6727–6741.
25. Renshaw RW, Casey JW: **Transcriptional mapping of the 3' end of the bovine syncytial virus genome.** *J Virol* 1994, **68:**1021–1028.
26. Wu Z, Ren X, Yang L, Hu Y, Yang J, He G, Zhang J, Dong J, Sun L, Du J, Liu L, Xue Y, Wang J, Yang F, Zhang S, Jin Q: **Virome analysis for identification of novel mammalian viruses in bat species from Chinese provinces.** *J Virol* 2012, **86:**10999–11012.
27. Achong BG, Mansell PW, Epstein MA, Clifford P: **An unusual virus in cultures from a human nasopharyngeal carcinoma.** *J Natl Cancer Inst* 1971, **46:**299–307.
28. Switzer WM, Bhullar V, Shanmugam V, Cong M-E, Parekh B, Lerche NW, Yee JL, Ely JJ, Boneva R, Chapman LE, Folks TM, Heneine W: **Frequent simian foamy virus infection in persons occupationally exposed to nonhuman primates.** *J Virol* 2004, **78:**2780–2789.
29. Meiering CD, Linial ML: **Historical perspective of foamy virus epidemiology and infection.** *Clin Microbiol Rev* 2001, **14:**165–176.
30. Liu W, Worobey M, Li Y, Keele BF, Bibollet-Ruche F, Guo Y, Goepfert PA, Santiago ML, Ndjango J-BN, Neel C, Clifford SL, Sanz C, Kamenya S, Wilson ML, Pusey AE, Gross-Camp N, Boesch C, Smith V, Zamma K, Huffman MA, Mitani JC, Watts DP, Peeters M, Shaw GM, Switzer WM, Sharp PM, Hahn BH: **Molecular ecology and natural history of simian foamy virus infection in wild-living chimpanzees.** *PLoS Pathog* 2008, **4:**e1000097.
31. Nemo GJ, Brown PW, Gibbs CJ, Gajdusek DC: **Antigenic relationship of human foamy virus to the simian foamy viruses.** *Infect Immun* 1978, **20:**69–72.
32. Brown P, Nemo G, Gajdusek DC: **Human foamy virus: further characterization, seroepidemiology, and relationship to chimpanzee foamy viruses.** *J Infect Dis* 1978, **137:**421–427.
33. Callahan ME, Switzer WM, Matthews AL, Roberts BD, Heneine W, Folks TM, Sandstrom PA: **Persistent zoonotic infection of a human with simian foamy virus in the absence of an intact orf-2 accessory gene.** *J Virol* 1999, **73:**9619–9624.

34. CDC: **Nonhuman primate spumavirus infections among persons with occupational exposure–United States, 1996.** *MMWR Morb Mortal Wkly Rep* 1997, **46:**129–131.
35. Schweizer M, Falcone V, Gänge J, Turek R, Neumann-Haefelin D: **Simian foamy virus isolated from an accidentally infected human individual.** *J Virol* 1997, **71:**4821–4824.
36. Heneine W, Switzer WM, Sandstrom P, Brown J, Vedapuri S, Schable CA, Khan AS, Lerche NW, Schweizer M, Neumann-Haefelin D, Chapman LE, Folks TM: **Identification of a human population infected with simian foamy viruses.** *Nat Med* 1998, **4:**403–407.
37. Katzourakis A, Gifford RJ, Tristem M, Gilbert MTP, Pybus OG: **Macroevolution of complex retroviruses.** *Science* 2009, **325:**1512.
38. Han G-Z, Worobey M: **An endogenous foamy virus in the aye-aye (Daubentonia madagascariensis).** *J Virol* 2012, **86:**7696–7698.
39. Han G-Z, Worobey M: **An endogenous foamy-like viral element in the coelacanth genome.** *PLoS Pathog* 2012, **8:**e1002790.
40. Llorens C, Muñoz-Pomer A, Bernad L, Botella H, Moya A: **Network dynamics of eukaryotic LTR retroelements beyond phylogenetic trees.** *Biol Direct* 2009, **4:**41.
41. Schartl M, Walter RB, Shen Y, Garcia T, Catchen J, Amores A, Braasch I, Chalopin D, Volff J-N, Lesch K-P, Bisazza A, Minx P, Hillier L, Wilson RK, Fuerstenberg S, Boore J, Searle S, Postlethwait JH, Warren WC: **The genome of the platyfish, Xiphophorus maculatus, provides insights into evolutionary adaptation and several complex traits.** *Nat Genet* 2013, **45:**567–572.
42. Switzer WM, Salemi M, Shanmugam V, Gao F, Cong M-E, Kuiken C, Bhullar V, Beer BE, Vallet D, Gautier-Hion A, Tooze Z, Villinger F, Holmes EC, Heneine W: **Ancient co-speciation of simian foamy viruses and primates.** *Nature* 2005, **434:**376–380.
43. Bininda-Emonds ORP, Cardillo M, Jones KE, MacPhee RDE, Beck RMD, Grenyer R, Price SA, Vos RA, Gittleman JL, Purvis A: **The delayed rise of present-day mammals.** *Nature* 2007, **446:**507–512.
44. Charleston MA, Robertson DL: **Preferential host switching by primate lentiviruses can account for phylogenetic similarity with the primate phylogeny.** *Syst Biol* 2002, **51:**528–535.
45. Kupiec JJ, Sonigo P: **Reverse transcriptase jumps and gaps.** *J Gen Virol* 1996, **77**(Pt 9)**:**1987–1991.
46. Eastman SW, Linial ML: **Identification of a conserved residue of foamy virus Gag required for intracellular capsid assembly.** *J Virol* 2001, **75:**6857–6864.
47. Patton GS, Morris SA, Chung W, Bieniasz PD, McClure MO: **Identification of domains in gag important for prototypic foamy virus egress.** *J Virol* 2005, **79:**6392–6399.
48. Stange A, Mannigel I, Peters K, Heinkelein M, Stanke N, Cartellieri M, Göttlinger H, Rethwilm A, Zentgraf H, Lindemann D: **Characterization of prototype foamy virus gag late assembly domain motifs and their role in particle egress and infectivity.** *J Virol* 2005, **79:**5466–5476.
49. Schliephake AW, Rethwilm A: **Nuclear localization of foamy virus Gag precursor protein.** *J Virol* 1994, **68:**4946–4954.
50. Lindemann D, Pietschmann T, Picard-Maureau M, Berg A, Heinkelein M, Thurow J, Knaus P, Zentgraf H, Rethwilm A: **A particle-associated glycoprotein signal peptide essential for virus maturation and infectivity.** *J Virol* 2001, **75:**5762–5771.
51. Wilk T, Geiselhart V, Frech M, Fuller SD, Flügel RM, Löchelt M: **Specific interaction of a novel foamy virus Env leader protein with the N-terminal Gag domain.** *J Virol* 2001, **75:**7995–8007.
52. Geiselhart V, Bastone P, Kempf T, Schnölzer M, Löchelt M: **Furin-mediated cleavage of the feline foamy virus Env leader protein.** *J Virol* 2004, **78:**13573–13581.
53. Perry GH, Reeves D, Melsted P, Ratan A, Miller W, Michelini K, Louis EE, Pritchard JK, Mason CE, Gilad Y: **A genome sequence resource for the aye-aye (Daubentonia madagascariensis), a nocturnal lemur from Madagascar.** *Genome Biol Evol* 2012, **4:**126–135.
54. **The draft genome of Chrysochloris asiatica.** http://www.ncbi.nlm.nih.gov/nuccore/AMDV00000000.1.
55. Jurka J: **Origin and evolution of Alu repetitive elements**. In *Mol Biol Intell Unit impact short interspersed Elem host genome.* Edited by Maraia RJ. Austin: R.G. Landes Company; 1995:25–41.
56. Jurka J: **SINE elements from the mouse lemur.** *Repbase Reports* 2010, **10:**2171.
57. Jurka J, Smith T: **A fundamental division in the Alu family of repeated sequences.** *Proc Natl Acad Sci U S A* 1988, **85:**4775–4778.
58. Jurka J: **SINE elements from African elephants.** *Repbase Reports* 2010, **10:**491.
59. Bodem J: **Regulation of foamy viral transcription and RNA export.** *Adv Virus Res* 2011, **81:**1–31.
60. Perelman P, Johnson WE, Roos C, Seuánez HN, Horvath JE, Moreira MAM, Kessing B, Pontius J, Roelke M, Rumpler Y, Schneider MPC, Silva A, O'Brien SJ, Pecon-Slattery J: **A molecular phylogeny of living primates.** *PLoS Genet* 2011, **7:**e1001342.
61. Horvath JE, Weisrock DW, Embry SL, Fiorentino I, Balhoff JP, Kappeler P, Wray GA, Willard HF, Yoder AD: **Development and application of a phylogenomic toolkit: resolving the evolutionary history of Madagascar's lemurs.** *Genome Res* 2008, **18:**489–499.
62. Yoder AD, Yang Z: **Divergence dates for Malagasy lemurs estimated from multiple gene loci: geological and evolutionary context.** *Mol Ecol* 2004, **13:**757–773.
63. Springer MS, Murphy WJ, Eizirik E, O'Brien SJ: **Placental mammal diversification and the Cretaceous-Tertiary boundary.** *Proc Natl Acad Sci U S A* 2003, **100:**1056–1061.
64. Lemey P, Lott M, Martin DP, Moulton V: **Identifying recombinants in human and primate immunodeficiency virus sequence alignments using quartet scanning.** *BMC Bioinformatics* 2009, **10:**126.
65. Conow C, Fielder D, Ovadia Y, Libeskind-Hadas R: **Jane: a new tool for the cophylogeny reconstruction problem.** *Algorithms Mol Biol* 2010, **5:**16.
66. McAllister BF, Werren JH: **Phylogenetic analysis of a retrotransposon with implications for strong evolutionary constraints on reverse transcriptase.** *Mol Biol Evol* 1997, **14:**69–80.
67. Shimodaira H: **An approximately unbiased test of phylogenetic tree selection.** *Syst Biol* 2002, **51:**492–508.
68. Hill CL, Bieniasz PD, McClure MO: **Properties of human foamy virus relevant to its development as a vector for gene therapy.** *J Gen Virol* 1999, **80**(Pt 8)**:**2003–2009.
69. Berg A, Pietschmann T, Rethwilm A, Lindemann D: **Determinants of foamy virus envelope glycoprotein mediated resistance to superinfection.** *Virology* 2003, **314:**243–252.
70. Plochmann K, Horn A, Gschmack E, Armbruster N, Krieg J, Wiktorowicz T, Weber C, Stirnnagel K, Lindemann D, Rethwilm A, Scheller C: **Heparan sulfate is an attachment factor for foamy virus entry.** *J Virol* 2012, **86:**10028–10035.
71. Herchenröder O, Moosmayer D, Bock M, Pietschmann T, Rethwilm A, Bieniasz PD, McClure MO, Weis R, Schneider J: **Specific binding of recombinant foamy virus envelope protein to host cells correlates with susceptibility to infection.** *Virology* 1999, **255:**228–236.
72. Yap MW, Lindemann D, Stanke N, Reh J, Westphal D, Hanenberg H, Ohkura S, Stoye JP: **Restriction of foamy viruses by primate Trim5alpha.** *J Virol* 2008, **82:**5429–5439.
73. Perkovic M, Schmidt S, Marino D, Russell RA, Stauch B, Hofmann H, Kopietz F, Kloke B-P, Zielonka J, Ströver H, Hermle J, Lindemann D, Pathak VK, Schneider G, Löchelt M, Cichutek K, Münk C: **Species-specific inhibition of APOBEC3C by the prototype foamy virus protein bet.** *J Biol Chem* 2009, **284:**5819–5826.
74. Opazo JC, Wildman DE, Prychitko T, Johnson RM, Goodman M: **Phylogenetic relationships and divergence times among New World monkeys (Platyrrhini, Primates).** *Mol Phylogenet Evol* 2006, **40:**274–280.
75. Schneider H: **The current status of the New World monkey phylogeny.** *An Acad Bras Cienc* 2000, **72:**165–172.
76. Muniz CP, Troncoso LL, Moreira M, Soares E, Pissinatti A, Bonvicino CR, Seuánez HN, Sharma B, Jia H, Shankar A, Switzer WM, Santos AF, Soares M: **Identification and characterization of highly divergent simian foamy viruses in a wide range of new world primates from Brazil.** *PLoS One* 2013, **8:**e67568.
77. Tamura K, Stecher G, Peterson D, Filipski A, Kumar S: **MEGA6: molecular evolutionary genetics analysis version 6.0.** *Mol Biol Evol* 2013, **30:**2725–2729.
78. Hawkesworth C, Kelley S, Turner S, Le Roex A, Storey B: **Mantle processes during Gondwana break-up and dispersal.** *J Afr Earth Sci* 1999, **28:**239–261.
79. de Wit MJ: **MADAGASCAR: heads it's a continent, tails it's an island.** *Annu Rev Earth Planet Sci* 2003, **31:**213–248.
80. Hussain AI, Shanmugam V, Bhullar VB, Beer BE, Vallet D, Gautier-Hion A, Wolfe ND, Karesh WB, Kilbourn AM, Tooze Z, Heneine W, Switzer WM: **Screening for simian foamy virus infection by using a combined antigen Western blot assay: evidence for a wide distribution among Old World primates and identification of four new divergent viruses.** *Virology* 2003, **309:**248–257.
81. Wolfe ND, Switzer WM, Carr JK, Bhullar VB, Shanmugam V, Tamoufe U, Prosser AT, Torimiro JN, Wright A, Mpoudi-Ngole E, McCutchan FE, Birx DL,

Folks TM, Burke DS, Heneine W: **Naturally acquired simian retrovirus infections in central African hunters.** *Lancet* 2004, **363:**932–937.
82. Kumar S, Subramanian S: **Mutation rates in mammalian genomes.** *Proc Natl Acad Sci U S A* 2002, **99:**803–808.
83. Stamatakis A, Hoover P, Rougemont J: **A rapid bootstrap algorithm for the RAxML Web servers.** *Syst Biol* 2008, **57:**758–771.
84. Abascal F, Zardoya R, Posada D: **ProtTest: selection of best-fit models of protein evolution.** *Bioinformatics* 2005, **21:**2104–2105.
85. Ronquist F, Huelsenbeck JP: **MrBayes 3: Bayesian phylogenetic inference under mixed models.** *Bioinformatics* 2003, **19:**1572–1574.
86. MATLAB: *MATLAB and Statistics Toolbox Release R2012a.* Massachusetts: The MathWorks Inc; 2010.
87. Yang Z: **PAML 4: phylogenetic analysis by maximum likelihood.** *Mol Biol Evol* 2007, **24:**1586–1591.
88. Shimodaira H, Hasegawa M: **CONSEL: for assessing the confidence of phylogenetic tree selection.** *Bioinformatics* 2001, **17:**1246–1247.

doi:10.1186/1742-4690-11-61
Cite this article as: Katzourakis *et al.*: Discovery of prosimian and afrotherian foamy viruses and potential cross species transmissions amidst stable and ancient mammalian co-evolution. *Retrovirology* 2014 **11**:61.

Submit your next manuscript to BioMed Central and take full advantage of:

- **Convenient online submission**
- **Thorough peer review**
- **No space constraints or color figure charges**
- **Immediate publication on acceptance**
- **Inclusion in PubMed, CAS, Scopus and Google Scholar**
- **Research which is freely available for redistribution**

Submit your manuscript at
www.biomedcentral.com/submit

BioMed Central

Appendix II: Wide distribution and ancient evolutionary history of simian foamy viruses in New World primates

Ghersi *et al. Retrovirology (2015) 12:89*
DOI 10.1186/s12977-015-0214-0

RETROVIROLOGY

RESEARCH Open Access

Wide distribution and ancient evolutionary history of simian foamy viruses in New World primates

Bruno M. Ghersi[1*], Hongwei Jia[2], Pakorn Aiewsakun[3], Aris Katzourakis[3], Patricia Mendoza[4], Daniel G. Bausch[1,5], Matthew R. Kasper[1], Joel M. Montgomery[1,6] and William M. Switzer[2*]

Abstract

Background: Although simian foamy viruses (SFV) are the only exogenous retroviruses to infect New World monkeys (NWMs), little is known about their evolutionary history and epidemiology. Previous reports show distinct SFVs among NWMs but were limited to small numbers of captive or wild monkeys from five (*Cebus*, *Saimiri*, *Ateles*, *Alouatta*, and *Callithrix*) of the 15 NWM genera. Other studies also used only PCR testing or serological assays with limited validation and may have missed infection in some species. We developed and validated new serological and PCR assays to determine the prevalence of SFV in blood specimens from a large number of captive NWMs in the US (n = 274) and in captive and wild-caught NWMs (n = 236) in Peruvian zoos, rescue centers, and illegal trade markets. Phylogenetic and co-speciation reconciliation analyses of new SFV polymerase (*pol*) and host mitochondrial cytochrome *B* sequences, were performed to infer SFV and host co-evolutionary histories.

Results: 124/274 (45.2 %) of NWMs captive in the US and 59/157 (37.5 %) of captive and wild-caught NWMs in Peru were SFV WB-positive representing 11 different genera (*Alouatta*, *Aotus*, *Ateles*, *Cacajao*, *Callithrix*, *Cebus*, *Lagothrix*, *Leontopithecus*, *Pithecia*, *Saguinus* and *Saimiri*). Seroprevalences were lower at rescue centers (10/53, 18.9 %) compared to zoos (46/97, 47.4 %) and illegal trade markets (3/7, 8/19, 42.9 %) in Peru. Analyses showed that the trees of NWM hosts and SFVs have remarkably similar topologies at the level of species and sub-populations suggestive of co-speciation. Phylogenetic reconciliation confirmed 12 co-speciation events ($p < 0.002$) which was further supported by obtaining highly similar divergence dates for SFV and host genera and correlated SFV-host branch times. However, four ancient cross-genus transmission events were also inferred for Pitheciinae to Atelidae, *Cacajao* to ancestral *Callithrix* or *Cebus* monkeys, between *Callithrix* and *Cebus* monkeys, and *Lagothrix* to *Alouatta*.

Conclusions: We demonstrate a broad distribution and stable co-speciation history of SFV in NWMs at the species level. Additional studies are necessary to further explore the epidemiology and natural history of SFV infection of NWMs and to determine the zoonotic potential for persons exposed to infected monkeys in captivity and in the wild.

Keywords: Retrovirus, Simian foamy virus, Co-evolution, Co-speciation, Nonhuman primates, South America, Peru, Neotropical

*Correspondence: bghersi@tulane.edu; bis3@cdc.gov
[1] U.S. Naval Medical Research Unit No. 6, Lima, Peru
[2] Laboratory Branch, Division of HIV/AIDS Prevention, National Center for HIV/AIDS, Viral Hepatitis, STD, and TB Prevention, Centers for Disease Control and Prevention, 1600 Clifton Rd., MS G-45, Atlanta, GA 30329, USA
Full list of author information is available at the end of the article

BioMed Central

© 2015 Ghersi et al. This article is distributed under the terms of the Creative Commons Attribution 4.0 International License (http://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution, and reproduction in any medium, provided you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons license, and indicate if changes were made. The Creative Commons Public Domain Dedication waiver (http://creativecommons.org/publicdomain/zero/1.0/) applies to the data made available in this article, unless otherwise stated.

Background

Foamy virus (FV), or spumavirus, comprises the only genus of the *Spumavirinae* subfamily of retroviruses [1, 2]. FVs have been reported in several mammalian species, including nonhuman primates (NHPs), cats, cows, horses, and sheep [3–5]. Simian foamy viruses (SFVs) were first described in 1954 as contaminants in primary monkey kidney cultures [6] and since then have been identified in many Old World and New World primate species using a variety of laboratory methods [5, 7]. SFV is the only exogenous retrovirus known to infect New World monkeys. FV is considered non-pathogenic in natural and experimental hosts but systematic, longitudinal studies have not been conducted to verify the apparent non-pathogenicity. Humans can be zoonotically infected with a variety of SFVs originating from Old World monkeys and apes (OWMA) through occupational and natural exposures but demonstrate an apparently asymptomatic though persistent infection [5, 8, 9]. SFV proviral DNA has been shown to be present at low copy numbers in peripheral blood mononuclear cells (PBMCs) and tissues from healthy and immune suppressed animals and infected humans [1, 10–12]. Isolation and/or detection of SFV from the oral mucosa of infected humans and NHPs has also been demonstrated [12–15]. The presence of virus in the oral mucosa and the seroconversion of NHPs at adulthood, a period more prone for biting, supports the hypothesis that transmission occurs via saliva through biting or licking [16, 17]. Moreover, most humans infected with SFV reported NHP bite or scratch exposures with higher prevalences seen in persons with severe bite wounds [10, 18–21].

Phylogenetic analysis has shown species-specific distribution of SFV in OWMA (Catarrhini), indicating a long co-evolution with their natural hosts [22]. However, little is known about the evolutionary history and distribution of SFV in New World primates (Platyrrhini) with the majority of studies done using only animals bred and housed in the US and evidence of infection was only demonstrated using serology [1, 5, 23–25]. Recently, complete SFV genomes have been reported for each of three captive New World monkeys (NWM), including a squirrel monkey (*Saimiri* species), a spider monkey (*Ateles* species), and a common marmoset (*Callithrix jacchus*) [26, 27]. All three NWM SFVs were distinct from those in OWMA and shared less than 50 % amino acid identity in the structural and enzymatic proteins, suggesting that serological assays used to detect SFV from OWMA may be less sensitive for detecting SFV from NWM. While all three NWM SFVs clustered together phylogenetically, a co-evolutionary history could not be verified since only a single sequence from each species is available. In addition, all three were captive animals and cross-species infections from other monkey species could not be excluded.

More recently, with collaborators in Brazil we identified SFV in 18 species of neotropical monkeys from Brazil using PCR-amplification of short (192-bp), highly conserved polymerase (*pol*) sequences, including capuchin (*Cebus* species), owl (*Aotus* sp.), marmoset (*Callithrix* sp.), tamarin (*Saguinus* sp.), squirrel (*Saimiri* sp.), titi (*Callicebus* sp.), saki (*Chiropotes* sp.), and howler (*Alouatta* sp.) monkeys [28]. However, there was not enough phylogenetic information in the highly conserved *pol* sequences in this study to fully resolve the evolutionary histories of all the NWM SFVs from Brazil. Another limitation of the study was the lack of serological testing, which may underestimate the reported prevalence. The authors demonstrated co-evolution of SFV from five NWM species using longer *pol* sequences (520-bp) obtained from *Alouatta* and *Cebus* monkeys in Brazil and *pol* sequences available from complete SFV genomes from spider, squirrel and marmoset monkeys at GenBank. One recent study also demonstrated SFV infection in a small number of three different NWM species captive in the US, including howler, capuchin, and squirrel monkeys [29]. Although these results are informative, the natural history and geographical and species distribution of SFV outside of Brazil and in captive animals elsewhere is thus incomplete.

At least 90 Platyrrhine species live in Central and South America belonging to three families (*Pitheciidae, Atelidae, and Cebidae*), eight subfamilies [*Callitrichinae* (n = 42), *Cebinae* (n = 14), *Aotinae* (n = 11), *Pitheciinae* (n = 43), *Saimirinae* (n = 10), *Alouattinae* (n = 19), *Callicebinae* (n = 29), and *Atelinae* (n = 24)] [30, 31], and nineteen genera: *Callithrix*, *Mico*, *Callibella*, *Cebuella*, *Leontopithecus*, *Saguinus*, *Callimico*, *Cebus*, *Saimiri*, *Aotus*, *Callicebus*, *Pithecia*, *Chiropotes*, *Cacajao*, *Alouatta*, *Ateles*, *Brachyteles*, *Lagothrix*, and *Oreonax* [30, 31]. Peru is considered a mega diverse country; with more than 500 species of mammals, 39 of which are primates [32]. Thus, a rich retroviral diversity in Neotropical primates would be expected in Peru like that observed in Brazil, and in OWMAs in Africa and Asia [11, 28, 33]. To better understand the prevalence, geographic distribution, genetic diversity, and evolutionary history of SFV in neotropical primates we tested convenience serum and dried blood spots from primates kept at zoos, rescue centers and illegal trade markets in Peru and in NWMs kept in US zoological gardens and research institutions. Evidence of SFV infection was determined using a combination of serologic and PCR assays followed by sequence analysis to infer phylogenetic and co-evolutionary relationships.

Methods

Study populations and sample preparation

Primates housed at four zoos, four rescue centers and one illegal trade market in five areas of Peru were sampled as part of another study to examine microbial infection in these animals (Fig. 1). Three of the zoos are located in Lima and one is located in the rainforest region of Pucallpa; the primate rescue centers are located near Puerto Maldonado

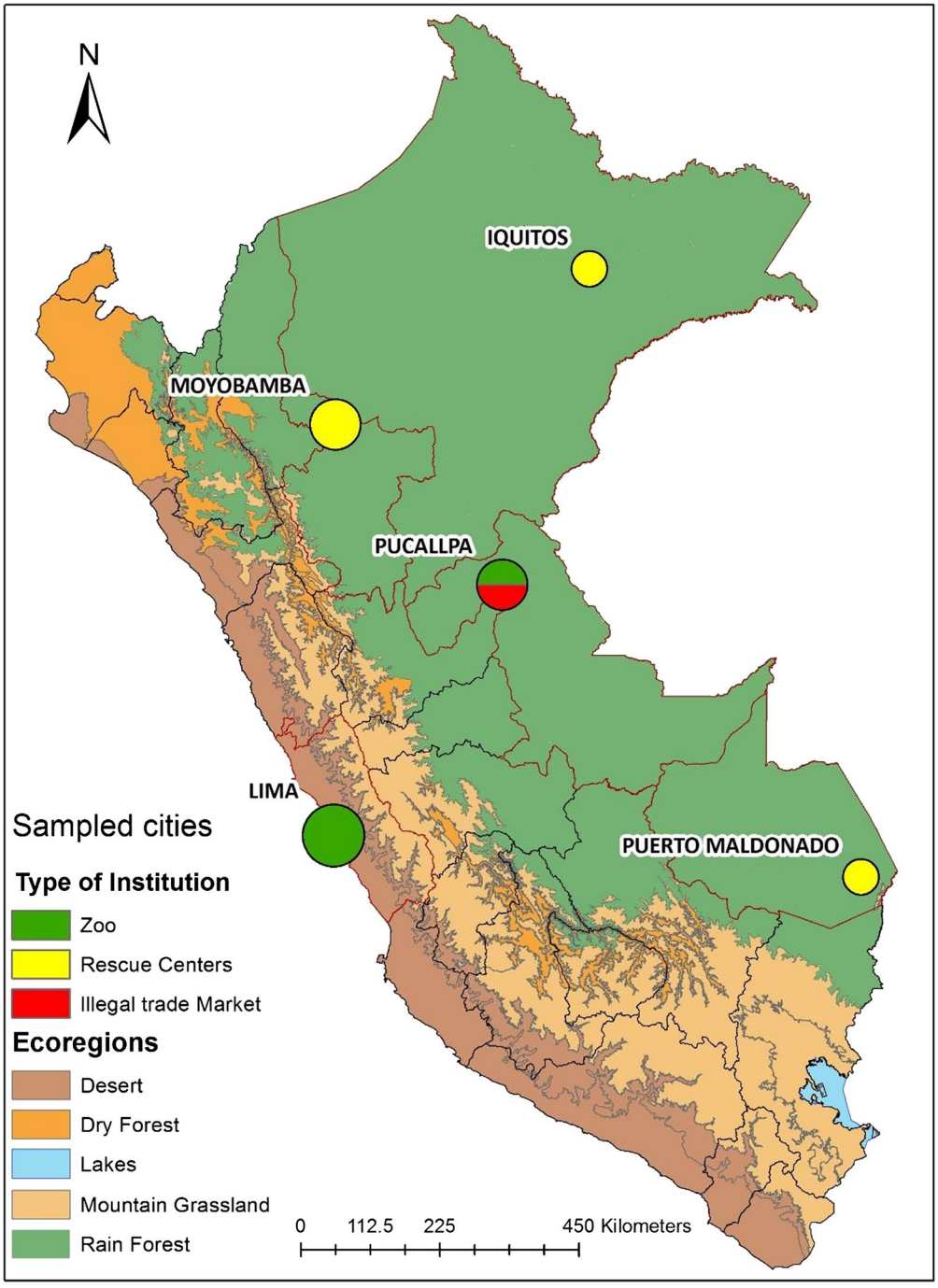


Fig. 1 Primate location at zoos, rescue centers and one illegal trade market in Peru. Three zoos participating in the study are located in Lima and one is located in the rainforest region of Pucallpa; one primate rescue center is located near Puerto Maldonado in the southern rainforest, one in Iquitos and two near Moyobamba in the northern lower and upper rainforests. The illegal trade market was located in the rain forest region of Pucallpa

(n = 1), in Iquitos (n = 1) and near Moyobamba (n = 2) in the southern and northern lower and upper rainforests, respectively. The illegal trade market was also located in Pucallpa. All animal work in Peru was approved by the Institutional Animal Care and Use Committee (IACUC) of the US Naval Medical Research Unit No. 6 (NAMRU 6) and approved by the Peruvian Ministry of Agriculture. Specimens were also collected from NWMs at seven US zoos and five research institutions following IACUC approval at each respective institution. Convenience blood samples were collected opportunistically during regular health exams from 211 Neotropical primates in Peru and from 274 captive NWM in the US (Tables 1, 2, 3). Whole blood was collected using standard procedures from the femoral vein with a vacutainer tube without additives. Approximately 125 µl of blood droplets were placed on a dry blood spot (DBS) card or on Whatman FTA filter paper. The remaining blood sample was allowed to clot and the serum was aliquoted and stored at −80 °C. EDTA-treated blood specimens obtained from captive NWMs in the US were processed for plasma and peripheral blood mononuclear cells (PBMCs) as previously described [34]. Plasma samples archived at −80 °C from persons infected with human T-lymphotropic virus (HTLV, n = 65), human immunodeficiency virus (HIV, n = 118), and HIV/HTLV-negative US blood donors (n = 237) that previously all tested negative for OWMA SFV were available for assay development.

Tissue culture isolation and propagation of SFV

PBMCs from a captive spider monkey (*Ateles* species, asp) were stimulated with 10 % IL2 for 48 h and co-cultured with canine thymocytes (Cf2Th) cells using methods reported in detail elsewhere until cytopathic effect (CPE) was observed [34]. Tissue culture supernatant was filtered using a 0.45 uM filter and passaged on fresh Cf2Th cells to propagate SFVasp. SFV from a common marmoset (*Callithrix jacchus*, *SFVcja*) was obtained from the American Type Culture Collection (ATCC VR-919) and propagated in Cf2Th cells [18]. Infection of Cf2Th cells was confirmed by DNA PCR using primers and conditions described below. Infected and uninfected Cf2Th cells were harvested and crude protein lysates were prepared, quantified and used in serological assays as before [34].

Serological assay development, validation, and specimen testing

Given the high genetic diversity between NWM and OWMA SFVs we developed a new Western blot (WB) assay to detect antibodies to NWM SFV with procedures

Table 1 High sensitivity and specificity of new western blot (WB) and PCR assays for the detection of simian foamy virus (SFV) in New World monkeys

Family	Scientific name	Common name	n	WB Pos/PCR Pos	WB Neg/PCR Neg	WB Pos/PCR Neg	WB Neg/PCR Pos
Atelidae	*Alouatta palliata*	Mantled howler	1	1	–[a]	–	–
	Alouatta seniculus sara	Red howler	3	3	–	–	–
	Ateles species	Spider monkey	1	1	–	–	–
	Ateles belzebuth hybridus	Colombian brown spider	1	1	–	–	–
	Ateles fusciceps robustus	Brown-headed spider	5	3	2	–	–
	Ateles geoffroyi	Black-handed spider	17	15	2	–	–
Cebidae	*Aotus trivirgatus*	Northern gray-necked owl	10	0	10	–	–
	Callithrix jacchus	Common marmoset	10	1	9	–	–
	Cebus albifrons	White-fronted capuchin	3	3	–	–	–
	Cebus apella	Brown capuchin	28	24	2	2	–
	Saguinus oedipus	Cotton-top tamarin	10	0	10	–	–
	Saimiri species	Squirrel monkey	1	0	1	–	–
	Saimiri boliviensis boliviensis	Bolivian squirrel	6	3	3	–	–
	Saimiri boliviensis peruviensis	Peruvian squirrel	7	3	4	–	–
	Saimiri sciureus	Common squirrel monkey	1	–	–	–	1
Pitheciidae	*Cacajao rubicundus*	Red uakari	2	2	–	–	–
	Pithecia pithecia	White-faced saki	1	1	–	–	–
Total			107	61 (57 %)	43 (40.2 %)	2 (1.9 %)	1 (0.9 %)

PCR testing using diagnostic primers to detect 141-bp polymerase sequences in genomic DNA specimens

[a] Dashes indicate an absence of specimen(s) with results in this category

Table 2 High sensitivity and specificity of a new EIA for the detection of simian foamy virus (SFV) antibodies in New World monkeys in the US

Family	Scientific name	Common name	n	WB		EIA			
				Pos (%)	Neg (%)	TP (%)	FN (%)	TN (%)	FP
Atelidae	*Ateles belzebuth hybridus*	Columbian brown spider	2	2 (100)	–[a]	2	–	–	–
	Ateles fusciceps robustus	Brown-headed spider	12	9 (75)	3 (25)	8	1	3	–
	Ateles geoffroyi frontatus	Black-handed spider	19	13 (68.4)	6 (31.6)	13	–	6	–
	Ateles geoffroyi vellerosus	Mexican spider	14	9 (64.3)	5 (35.7)	9	–	5	–
	Ateles paniscus chamek	Peruvian black spider	9	7 (77.8)	2 (22.2)	7	–	2	–
	Ateles species	Spider monkey	20	14 (70)	6 (30)	14	–	6	–
	Alouatta caraya	Black howler	5	5 (100)	–	5	–	–	–
	Alouatta palliata	Mantled howler	2	1 (50)	1 (50)	1	–	1	–
	Alouatta seniculus sara	Red howler	2	2 (100)	–	2	–	–	–
	Alouatta seniculus straminea	Golden howler	2	1 (50)	1 (50)	1	–	1	–
Cebidae	*Aotus trivirgatus*	Northern gray-necked owl	11	1 (9.1)	10 (90.9)	1	–	10	–
	Callithrix geoffroyi	Geoffrey's marmoset	4	2 (50)	2 (50)	2	–	2	–
	Callimico goeldii	Goeldii's marmoset	31	–	31 (100)	–	–	31	–
	Cebus albifrons	White-fronted capuchin	3	3 (100)	–	2	1	–	–
	Cebus apella	Tufted capuchin	40	37 (92.5)	3 (7.5)	37	–	3	–
	Leontopithecus rosalia	Golden lion tamarin	2	2 (100)	–	1	1	–	–
	Saguinus bicolor	Pied tamarin	1	–	1 (100)	–	–	1	–
	Saguinus imperator	Emperor tamarin	1	–	1 (100)	–	–	1	–
	Saguinus labiatus	Red-bellied tamarin	1	–	1 (100)	–	–	1	–
	Saguinus midas	Golden-handed tamarin	2	–	2 (100)	–	–	2	–
	Saguinus mystax	Moustached tamarin	30	–	30 (100)	–	–	30	–
	Saguinus oedipus	Cotton-top tamarin	17	1 (5.9)	16 (94.1)	1	–	16	–
	Saimiri boliviensis boliviensis	Bolivian squirrel monkey	7	4 (57)	3 (43)	1	3	3	–
	Saimiri boliviensis peruviensis	Peruvian squirrel monkey	7	–	7 (100)	–	–	7	–
	Saimiri sciureus	Common squirrel monkey	11	1 (9.1)	10 (90.9)	–	1	10	–
Pitheciidae	*Cacajao rubicundus*	Red uakari	2	2 (100)	–	2	–	–	–
	Callicebus moloch	Dusky titi monkey	5	–	5 (100)	–	–	5	–
	Pithecia pithecia	White-faced saki	12	8 (66.7)	4 (33.3)	7	1	4	–
Total			274	124 (45.2)	150 (54.8)	116 (93.6)	8 (6.5 %)	150 (100 %)	–

WB western blot testing, *TP* true positive, *FN* false negative, *TN* true negative, *FP* false positive

[a] Dashes indicate an absence of specimen(s) with results in this category

used successfully to detect a broad diversity of OWMA SFV [21]. Thus, for the WB assay we used antigens from two genetically diverse NWM SFVs, SFVasp and SFVcja, that share about 65 % genetic identity [26] to allow broad serologic detection of SFV in all three Platyrrhine families. Protein concentrations of the lysates were determined using the BioRad DC Protein Assay (Hercules, CA, USA). Plasma or serum samples were diluted 1:50 and reacted separately to 150 µg of infected and uninfected cell lysates overnight at 4 °C after protein separation through 4–12 % polyacrylamide gels and transfer to Nytran membranes, as previously described [34]. Seroreactivity was detected using peroxidase-conjugated protein A/G (Pierce, Rockford, IL, USA) and chemiluminescence (Amersham, Uppsala, Sweden) [34]. Seroreactivity to both Gag p68 and p72 precursor proteins with an absence of similar reactivity to antigen from uninfected Cf2Th cells was interpreted as seropositive. Specimens without reactivity to either Gag protein were considered seronegative.

The SFV WB assay was validated with serum and plasma samples from 15 different species of NWMs (n = 107, Table 1). The infection status of these primates was determined by PCR analysis using newly developed generic *pol* primers as described below. Specificity of the WB assay was also determined using sera from 118 persons infected with HIV-1/2, 65 persons with HTLV-1/2 infection and 237 HIV/HTLV-negative sera from

Table 3 Distribution of simian foamy virus in captive and wild caught neotropical monkeys from Peru

Origin	Family	Scientific name	Common name	n	Serology		PCR	
					EIA (%)	WB (%)	141-bp pol (%)	495-bp pol (%)
Illegal trade market	*Atelidae*	*Ateles paniscus chamek*	Peruvian spider monkey	1	0/1	ND	0/1	0/1
		Lagothrix lagotricha	Common wooly monkey	2	0/2	ND	0/1	0/1
	Cebidae	*Aotus species*	Owl monkey	1	1/1 (100)	1/1 (100)	0/1	0/1
		Cebus albifrons	White-fronted capuchin	2	1/2 (50)	1/1 (100)	0/2	0/2
		Saguinus fusicollis	Brown mantled tamarin	1	–[a]	–	0/1	0/1
		Saguinus species	Tamarin	1	–	–	0/1	0/1
		Saimiri sciureus	Common squirrel monkey	4	–	–	1/4 (25)	1/4 (25)
	Pitheciidae	*Pithecia monachus*	Geoffrey's Monk saki	1	1/1 (100)	1/1 (100)	0/1	0/1
Subtotals				13	3/7 (42.6)	3/3 (100)	1/12 (8.3)	1/12 (8.3)
Rescue centers	*Atelidae*	*Alouatta seniculus*	Howler monkey	2	–	–	0/2	ND
		Ateles bezelbuth	Long-haired spider monkey	1	0/1	ND[b]	ND	ND
		Ateles paniscus chamek	Peruvian spider monkey	28	2/14 (14.3)	2/2 (100)	1/16 (6.3)	3/3 (100)
		Lagothrix lagotricha	Common wooly monkey	39	8/36 (22.2)	8/8 (100)	1/5 (20)	2/2 (100)
	Cebidae	*Cebus albifrons*	White-fronted capuchin	2	–	–	0/2	ND
		Cebus apella	Tufted capuchin	4	1/2 (50)	0/1	0/3	0/1
		Saimiri boliviensis peruviensis	Peruvian squirrel monkey	2	–	–	1/2 (50)	1/1 (100)
Subtotals				78	11/53 (20.7)	10/11 (90.9)	3/30 (10)	6/7 (85.7)
Zoos	*Atelidae*	*Alouatta seniculus*	Howler monkey	3	3/3 (100)	3/3 (100)	1/3 (33.3)	1/3 (33.3)
		Ateles belzebuth	Long-haired spider monkey	3	0/3	ND	ND	ND
		Ateles paniscus chamek	Peruvian spider monkey	7	5/6 (83.3)	4/4 (100)	1/2 (50)	0/1
		Ateles paniscus	Red-faced spider monkey	1	0/1	ND	ND	ND
		Ateles species	Spider monkey	2	0/2	ND	ND	ND
		Lagothrix cana	Peruvian wooly monkey	2	2/2 (100)	2/2 (100)	0/2	ND
		Lagothrix lagotricha	Common wooly monkey	14	6/14 (42.8)	5/6 (83.3)	3/11 (27.3)	2/11 (18.2)
	Cebidae	*Aotus nancymae*	Peruvian red-necked owl monkey	1	0/1	0/1	ND	ND
		Aotus nigriceps	Peruvian night owl monkey	2	0/2	0/2	ND	ND
		Aotus species	Owl monkey	3	0/3	ND	ND	ND
		Callithrix pygmea	Pygmy marmoset	1	1/1 (100)	0/1	–	–
		Cebus albifrons	White-fronted capuchin	7	3/7 (42.8)	2/4 (50)	0/3	ND
		Cebus apella	Tufted capuchin	40	29/32 (90.6)	29/29 (100)	11/36 (30.6)	4/13 (30.8)
		Saguinus fusicollis	Brown mantled tamarin	15	0/3	ND	0/12	ND
		Saguinus labiatus	Red-bellied tamarin	1	1/1 (100)	1/1 (100)	–	–
		Saimiri boliviensis peruviensis	Peruvian squirrel monkey	4	0/4	ND	ND	ND
		Saimiri sciureus	Common squirrel monkey	6	0/6	0/6	ND	ND
		Saimiri species	Squirrel monkey	1	1/1 (100)	0/1	–	–
	Pitheciidae	*Callicebus discolor*	Red titi monkey	1	0/1	0/1	–	–
		Callicebus oenanthe	Rio Mayo titi monkey	1	–	–	0/1	ND
		Callicebus species	Titi monkey	1	0/1	0/1	–	–
		Pithecia monachus	Geoffrey's Monk saki	3	0/3	ND	ND	ND

Table 3 continued

Origin	Family	Scientific name	Common name	n	Serology		PCR	
					EIA (%)	WB (%)	141-bp pol (%)	495-bp pol (%)
Subtotals				119	51/97 (52.5)	46/62 (74.2)	16/65 (24.6)	7/28 (25)
Totals				210	65/157 (41.4)	59/76 (77.6)	20/107 (18.7)	14/55 (25.5)

All monkeys at the wet markets and rescue centers were wild caught

WB western blot testing

[a] Dashes indicate either serum or FTA specimens were not available for testing

[b] ND, testing not done based on EIA results and/or availability of certain specimen types, i.e. WB testing not done on EIA-negative and/or if FTA cards were not collected for nucleic acid preparation

US blood donors. Cross-reactivity of selected sera from different SFV-infected NWM genera to SFV antigens derived from Old World monkeys and apes (SFVagm, African green monkey) and SFVcpz (chimpanzee) was also performed by WB testing to further evaluate assay specificity.

We also developed a new enzyme immunoassay (EIA) to facilitate rapid screening of a large number of specimens for antibodies to NWM SFV. Serum or plasma samples were diluted 1:100 in assay diluent and tested in duplicate in two different microtiter wells coated with crude cell lysates from Cf2Th cells infected with SFVasp and SFVcja in a single well and uninfected Cf2Th lysates in a separate well to assess assay specificity. Each specimen was tested in duplicate. Replicate sample optical density (OD) values were averaged and adjusted ODs of reactivity to SFV minus that to the uninfected antigens were calculated as described before [35].

SFV and host PCR

Genomic DNA was extracted from FTA/DBS cards using the QIAamp DNA Mini Kit (QIAGEN) as described by the manufacturer. DNA lysates were prepared from PBMCs from monkeys captive in the US and the integrity of the extracted DNA was validated using ß-actin PCR as described in detail elsewhere [18]. All DNA samples were first screened for SFV sequences using a novel semi-nested PCR that utilizes generic *pol* primers. These primers were designed using an alignment of sequences from the three complete SFV genomes available at GenBank from marmoset (SFVmar), squirrel (SFVsqu), and spider monkey (SFVspm) (accession numbers GU356395, GU356394, and EU010385, respectively) [26, 27].

For the first PCR, 0.5 µg of DNA was applied to 50 µl of reaction mixture containing 1× buffer with 1.5 mM $MgCl_2$, 800 uM dNTPs, 2 ng/µl of primary primers (SIF5N 5′ TAC ATG GTT ATA CCC CAC KAA GGC TCC TCC 3′ and SIR5N 5′ AAT AAW GGA TAC CAC TTT GTA GGT CTT CC 3′) and 1.25 U AmpliTaq DNA polymerase (Applied Biosystems) with the following conditions to generate a 282-bp sequence: five cycles at 94 °C for 1 min; 37 °C for 1 min and 72 °C for 1 min, then 35 cycles at 94 °C for 1 min; 50 °C for 1 min and 72 °C for 1 min, with a final extension at 72 °C for 1 min. For the second PCR, 2.5 µl of primary product was added to 50 µl nested PCR mixture containing the same concentration of components as the first PCR except using the nested primers SIP4N (5′ TGC ATT CCG ATC AAG GAT CAG CAT T 3′ and SIR1NN (5′ GTT TTA TYT CCY TGT TTT TCC TYT CCA CCA T 3′) to generate a 141-bp *pol* sequence. Nested PCR conditions were 40 cycles at 94 °C for 1 min; 50 °C for 1 min and 72 °C for 1 min. 20 µl of nested PCR product of each sample was loaded onto a 1.8 % agarose gel for electrophoresis analysis. Samples positive using the generic *pol* primers were subjected to additional PCR testing to obtain longer fragments containing adequate sequence information for resolution by phylogenetic analysis. Primary (SNF3 5′ GAT AAR TTG GCW RYM CAA GGW AGT TAT 3′ and SNR3 5′ GAR GTR AAT GCT GAT CCT TGA TCG GAA T 3′) and semi-nested PCR primers (SNF3 and SNR4 5′ GAA GGA GCY TTH GTG GGG TAT AAC CA 3′) were used to amplify 581 and 495-bp *pol* sequences, respectively, using 35 cycles of standard PCR and an annealing temperature of 50 °C.

Primate host species taxonomic classification was determined by analysis of 975-bp cytochrome *B* (*CytB*) mitochondrial DNA (mtDNA) sequences obtained by one-step PCR using primers L14724 (5′ CGA AGC TTG ATA TGA AAA ACC ATC GTT G 3′) and Mus15398 (5′ GAA TAT CAG CTT TGG GTG TTG RTG 3′) as previously described [28].

Statistical analysis

MedCalc v12.5.0 was used to perform receiver operator curve analysis of EIA data and infer assay cutoff values and for 2 × 2 tables for determining assay sensitivity, specificity, positive predictive value (PPV), negative predictive value (NPV), and accuracy. Online statistical tools available at VassarStats (http://vassarstats.net/)

were used to determine associations using Chi square or Fisher exact probability tests. When known, NWMs were stratified by age with adults being >4 years old, subadults 3–4 years old, juveniles 1 <3 years old, and infants <1 year old.

Sequence analyses

Amplified products were purified, quantified, and sequenced on both strands using the Big Dye v.3.1 sequencing kit (Life Technologies, Carlsbad, USA) and an automated ABI 3130XL Genetic Analyzer and edited with SeqMan v7.0 (DNASTAR, Madison, USA). New SFV and *CytB* sequences were aligned with those available from NWM retrieved from GenBank by using the Clustal W program implemented in MEGA v6 [36]. Quality of the alignments was verified using the program GUIDANCE (http://guidance.tau.ac.il/). Percent nucleotide identities were determined using the program Geneious Pro v6.1.3.

Recombination detection

The absence of genetic recombination in the *pol* alignments was confirmed using Bootscan, Geneconv, MaxChi, Chimera, and RDP within the program RDP v3 using the parameter defaults [37].

Phylogenetic reconciliation analysis

A Bayesian phylogeny of 74 NWM SFVs was estimated from a manually-curated *pol* alignment (412 nt) by using MrBayes 3.2.1 [38] without imposing a molecular clock which may influence and/or bias the tree topology estimation. Six OWMA SFV sequences (PFV; Y07725, SFVcpz; U04327, SFVgor; HM245790, SFVora; AJ544579, SFVmac; NC_010819, and SFVagm; M784895) were included as an outgroup. Nine sequences in the alignment were 3351 nt long (PFV, SFVcpz, SFVgor, SFVora, SFVmac, SFVagm, SFVmar; GU356395, SFVspm; EU010385, and SFVsqu; GU356394) and were included as 'backbones' to increase the power of deep node separation. The GTR+I+Γ(4) nucleotide substitution model was used. Two independent Markov chain Monte Carlo (MCMC) chains were run for 50 million steps with the initial 25 % discarded as burn-in. Trees and parameters were logged every 2500 steps thereafter. A metropolis coupling algorithm was applied to improve the MCMC samplings, using the setting of 3 hot and 1 cold chains. A Bayesian phylogeny of 156 NWM hosts was also estimated by using the same protocol from a manually-curated *CytB* alignment (618 nt). However, unlike the SFV phylogeny, the host phylogeny was rooted according to the tree in Perelman et al. [30]. All alignments are available from the authors upon request. The convergence of estimated parameter values was diagnosed using potential scale reduction factors (PSRFs). PSRFs of all parameters were ~1.000, indicating that they were all well sampled from their posterior distributions and had converged.

We then compared the topologies of the consensus NWM SFV and host phylogenies to infer potential co-speciation events by using the co-phylogeny reconstruction software Jane v4 with the following settings: generation number = 50 and population size = 100 [39]. In total, four reconciliation analyses were performed: (1) at the genus level, (2) species level—conservative tree collapsing, (3) species level—overall, and (4) all sequences (see "Results and discussion" for details). In the last analysis, NWM SFVs for which the corresponding host sequences were not available were excluded from the analysis. The vertex-based cost mode was used with costs set to maximize the number of co-speciation events (co-speciation = −1, duplication = 0, duplication and host switch = 0, loss = 0, and failure to diverge = 0). To assess the probability of observing the inferred co-divergence number by chance, a null distribution was calculated by using the random tip mapping method implemented in Jane v4 with the settings of generation number = 50, population size = 100, and sample size = 500 [39].

Evaluation and refining the co-speciation model

To evaluate the inferred co-speciation events and refine the co-evolution model, we first inferred the dates for some of the co-speciation events that could be mapped conclusively onto the tree, directly from the host dates estimated in [30]. Four dates were inferred in total, including (1) the NWM-OWMA FV separation date, (2) the branching date of the *Cebus xanthosternos* SFV lineage, (3) the branching date of the *Ateles chamek* SFV lineage, and (4) the separation date between the *Alouatta belzebul* and *Alouatta sara* SFV lineages. These dates were then used to estimate the divergence dates of other nodes in the SFV tree, under the Bayesian phylogenetic framework using BEAST 1.8.2 [30, 40]. The BEAST analysis used the SRD06 nucleotide substitution model, the Yule speciation process, a relaxed log-normal molecular clock and the default prior settings. The tree topology was fixed to the one we obtained from the MrBayes analyses. The Bayesian MCMC was run for 50 million steps with the initial 25 % discarded as burn-in. Trees and parameters were logged every 2500 steps thereafter. Parameter value convergence and sampling independency were manually inspected using the program Tracer included in the BEAST package. Effective sample sizes of all parameters were >200, indicating that they were all well sampled and had converged. We then compared the estimated SFV evolutionary timescales to those of their hosts and refined our SFV-host co-evolutionary model. The tree with the maximum product of the posterior

clade probabilities (maximum clade credibility tree) was constructed from the posterior distribution of the sampled trees with the program TreeAnnotator v.1.8.2. Node heights were calculated from the posterior distribution of the trees and viewed in FigTree v.1.3.1.

To further evaluate the co-evolutionary model, an SFV-host divergence correlation analysis was performed by manually identifying a subset of SFV-host co-diverging branches based on the refined model, and examining if the SFV and host divergences are linearly correlated. The linear correlation and its *p* value were calculated using the *lm* function implemented in R (https://cran.r-project.org). The coefficient of determination (R^2) and the *p* values were also derived.

GenBank accession numbers

All new *CytB* and SFV sequences generated during our study have been deposited at GenBank with the accession numbers KR902362-KR902495.

Results and discussion

SFV NWM PCR, WB and EIA validation

To determine the sensitivity of the diagnostic NWM SFV PCR assay both SFVasp and SFVcja tissue culture DNA lysates were diluted 10 fold with lysis buffer from 10^{-1} to 10^{-8} representing 0.1 µg to 0.01 pg cellular DNA and PCR tested. The diagnostic primers were able to detect both SFVasp and SFVcja sequences between 0.1 and 1.0 pg (10^{-6} to 10^{-7} dilutions). DNA lysates prepared from PBMCs collected from a human without exposure to NWMs were consistently negative in all assay runs. We also determined the copy number sensitivity of the diagnostic primers using SFVasp and SFVcja *pol* plasmid clones to be 10 and 1 copies each, respectively, which is similar to that reported recently using an SFVsqu *pol* plasmid [29]. To evaluate the ability of these primers to detect a broad range of SFV in NWMs we screened PBMC lysates from 107 captive NWMs representing nine genera and 17 species (Table 1). SFV sequences were detected in 61 individual animals (57 %) from seven genera in each of the three NWM families *Pitheciidae*, *Atelidae*, and *Cebidae*, including *Cebus apella* (24/28, 85.7 %), *Cebus albifrons* (3/3, 100 %), *Alouatta palliata* (1/1), *Alouatta seniculus sara* (3/3, 100 %), *Callithrix jacchus* (1/10, 10 %), *Ateles* species (1/1), *Ateles belzebuth hybridus* (1/1, 100 %), *Ateles geoffroyi* (15/17, 88.2 %), *Ateles fusciceps robustus* (3/5, 60 %), *Saimiri boliviensis* (3/6, 50 %), *Saimiri boliviensis peruviensis* (3/7, 42.9 %), *Cacajao rubicundus* (2/2), and *Pithecia pithecia* (1/1). These results demonstrate the ability of the diagnostic primers to broadly detect SFV diversity in a wide range of divergent NWM species.

Next, plasma from all 107 NWMs was tested using the new WB assay for comparison with the PCR results to infer the sensitivity and specificity of each assay. Representative WB results for a variety of NWM species are shown in Fig. 2. Very little cross-reactivity of NWM sera to OWMA SFV antigens was observed, supporting the specificity of the NWM SFV WB test (Fig. 2b). Concordant WB and PCR results were obtained for 104 specimens, giving a 97.2 % accuracy of each assay for SFV detection (Table 1). Two specimens from *Cebus apella* were WB-positive but PCR-negative and conversely one sample from a *Saimiri sciureus* was WB-negative but PCR-positive. Repeat WB testing of this animal using a serum sample collected 3 years later gave concordant results (data not shown). Thus, the WB assay was found to have a sensitivity, specificity, PPV, NPV, and accuracy of 96.8, 99.8, 98.4, 99.6, and 99.6 %, respectively. The sensitivity, specificity, PPV, NPV, and accuracy of the

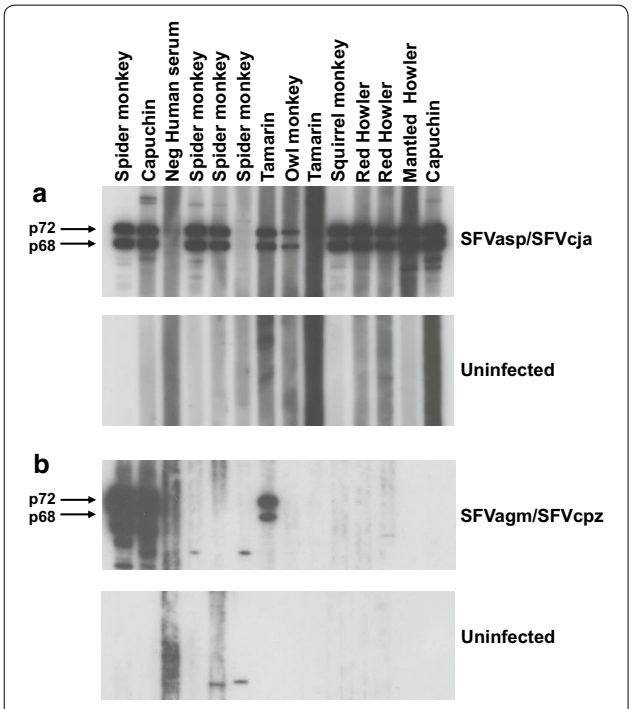


Fig. 2 Detection of plasma/serum antibodies to SFV from spider monkey (SFVasp) and marmoset monkey (SFVcja) using a combined antigen western blot (WB) assay. **a** *Upper panel* shows seroreactivity of representative neotropical primate samples to the combined NWM SFV antigens from spider monkey (*asp Ateles* species) and marmoset (*cja Callithrix jacchus*) tissue cultures; *lower panel* shows reactivity to crude cell lysate antigens from uninfected canine thymocytes (Cf2Th). **b** *Upper panel* shows seroreactivity of representative Neotropical primate samples to combined Old World monkey (*AGM* African green monkey) and ape (*cpz* chimpanzee) SFV antigens except *lanes 1* and *2* are plasma samples from SFVagm- and SFVcpz-infected humans as positive controls; *lower panel* shows reactivity to crude cell lysate antigens from uninfected Cf2Th. Seroreactivity was defined as those specimens with reactivity specific to the diagnostic Gag doublet proteins in the combined viral antigens. *Lane 3* is a pedigreed negative human plasma control

diagnostic PCR test was 98.4, 95.6, 96.8, 97.7, and 97.2 %, respectively.

To efficiently screen large numbers of specimens for evidence of NWM SFV infection we also developed a microtiter-based EIA. For this purpose, we expanded screening of serum and plasma specimens available from an additional 167 NWMs that were also tested by WB analysis for a total of 274 specimens (Table 2). PBMC DNA was not available from the majority of these 167 animals for PCR testing. Of these, 124 (45.2 %; 39.3–51.4 % 95 % CI) were WB-positive and 150 (54.8 %) were WB-negative (Table 2). EIA specificity was also determined using a total of 417 specimens from HIV-1-infected persons (n = 56), HIV-1/2 infections (n = 59), HTLV-1/2 infection (n = 65), and US blood donors (n = 237) that all tested negative in the WB assay (Table 3).

An adjusted OD ≥0.235 was set as a cutoff value for seroreactive samples using receiver operator curves (ROC) generated in the MedCalc software program based on assay validation with the WB-confirmed specimens (Tables 1, 2). Using this cutoff, the EIA assay sensitivity, specificity, PPV, PNV, and accuracy was 93.6, 97.7, 91.0, 98.4, and 96.9 %, respectively. The EIA gave concordant positive and negative results for all NWM specimens except for false negative results for eight WB-positive samples (6.5 %) from four squirrel, one capuchin, one saki, one spider, and one tamarin monkey (Table 2). All eight specimens showed weak seroreactivity in the WB test. False-positive results were obtained with 13 WB-negative human specimens (3.1 %), including seven blood donors, three HTLV-positive, and three HIV-positive samples (Table 4).

SFV prevalence and distribution in captive NWM monkeys in the US and in captive and wild monkeys from Peru

The overall SFV prevalence in the 274 captive seroreactive monkeys from the US was 45.2 % (39.3–41.5 % 95 % CI) and ranged from 0 to 100 %, but included species with small numbers of representatives such as Emperor tamarins (0/1, 0 %) and brown spider monkeys (2/2, 100 %) (Table 2). For those species with at least 10 animals, SFV prevalence was greatest in tufted capuchins (37/40, 92.5 %; 78.5–98 % 95 % CI) followed by brown-headed spider monkeys (9/12, 75 %; 42.8–93.3 % 95 % CI), spider monkeys (14/20, 70 %; 45.7–87.2 % 95 % CI), black-handed spider monkeys (13/19, 68.4 %; 43.5–86.4 % 95 % CI), white-faced sakis (8/12, 66.7 %; 35.4–88.7 % 95 % CI), Mexican spider monkeys (9/14, 64.3 %; 35.6–86 % 95 % CI), common squirrel monkeys (1/11, 9.1 %; 0.5–42.9 % 95 % CI), northern grey-necked owl monkeys (1/11, 9.1 %; 0.5–42.9 % 95 % CI), and cotton-topped tamarins (1/17, 5.9 %; 0.3–30.8 % 95 % CI). SFV was absent in mustached tamarins (0/30), whose ages ranged from 2.6 to 10.5 years old, and also in Goeldii's marmosets (0/31; ages not available).

180 sera and 178 DBS or FTA-prepared blood samples were collected from 18 primate species at zoos (n = 130), rescue centers (n = 78) and illegal trade markets (n = 7) in Peru (Fig. 1; Table 3). All monkeys at the trade markets and rescue centers were wild caught. Overall, 59/157 (37.6 %, 30.1–45.7 % 95 % CI) of these NWMs had antibodies against SFV. For animals with serum or plasma available, a higher SFV seroprevalence was observed at zoos (46/97, 47.4 %; 37.2–57.8 % 95 % CI) and at illegal trade markets (3/7, 42.68.1–64.6 %95 % CI) than at rescue centers (11/53, 18.9 %; 9.9–32.4 % 95 % CI) (Table 3), though this difference was not statistically significant ($p = 0.337$ and 0.228, respectively). However, the higher SFV prevalence at zoos was significant compared to that found at rescue centers ($p = 0.001$). For the three species with more than 10 animals, the highest prevalence was seen in *C. apella* (90.6 %), *L. lagotricha* (42.8 % at zoos and 22.2 % at rescue centers), and *A. chamek* (14.3 %). Our observed wide distribution of SFV in the three NWM families is similar to that reported in Brazil [28]. However, for the first time we identify SFV infection of

Table 4 High specificity of a new EIA for the detection of simian foamy virus (SFV) antibodies in humans

Population	n	WB		EIA			
		Pos (%)	Neg (%)	TP	FN	TN (%)	FP (%)
HIV-1-infected	56	–[b]	56 (100)	–	–	55 (98.2)	1 (1.8)
HIV-1/2-infected	59	–	59 (100)	–	–	57 (96.6)	2 (3.4)
HTLV-1/2-infected	65	–	65 (100)	–	–	62 (95.4)	3 (4.6)
US blood donors[a]	237	–	237 (100)	–	–	230 (97.1)	7 (2.9)
Total	417	0	417 (100)	–	–	404 (96.9)	13 (3.1)

WB western blot testing, *TP* true positive, *FN* false negative, *TN* true negative, *FP* false positive

[a] Blood donors previously tested negative for antibodies to SFV from Old World monkey and apes

[b] Dashes indicate an absence of specimen(s) with results in this category

L. lagotricha and *L. cana*, *Ateles chamek* and *A. belzebuth*, *Pithicea monachus*, *Saguinus labiatus*, and *Saimiri boliviensis* which were either not sampled or were under sampled in the Brazilian study (Table 3 and [28]). The overall WB prevalence of SFV in NWM at the zoos in Peru (47.4 %) is comparable to that seen in monkeys at US zoos (45.2 %). However, our SFV prevalence rates are somewhat higher that those reported in Brazil (14–30 %) but which used only PCR testing and also which had numerous species in that study which were under sampled [28]. It is also not clear what impact the absence of ages for some animals in our study may have had on the observed prevalences.

SFV seroprevalence in captive NWM in the US was nearly identical in male (41/73, 56.2 %) and female (58/98, 59.2 %) animals. Prevalence increased with age in both males and females and ranged from zero percent in two infants, 30–50 % in juveniles, 50–58 % in sub-adults, and 55–64 % in adults. Age and sex were not available for all monkeys, including the negative Goeldii's marmosets. The SFV prevalence in males (19.4 %, 14/72) compared to females (26.1 %, 23/88) in Peru was not statistically significant ($p = 0.209$) with gender available for most animals at zoos (male = 47, female = 36) and rescue centers (male = 25, female = 52) only. These results are similar to the equal distribution reported in the captive NWMs in the US (Table 2), in captive adult NWMs in Brazil, and other studies of SFV-infected OWMA [1, 17, 28, 41]. However, the SFV prevalence in both males and females in US zoos was significantly higher than those combined in zoos and rescue centers in Peru ($p < 0.0001$).

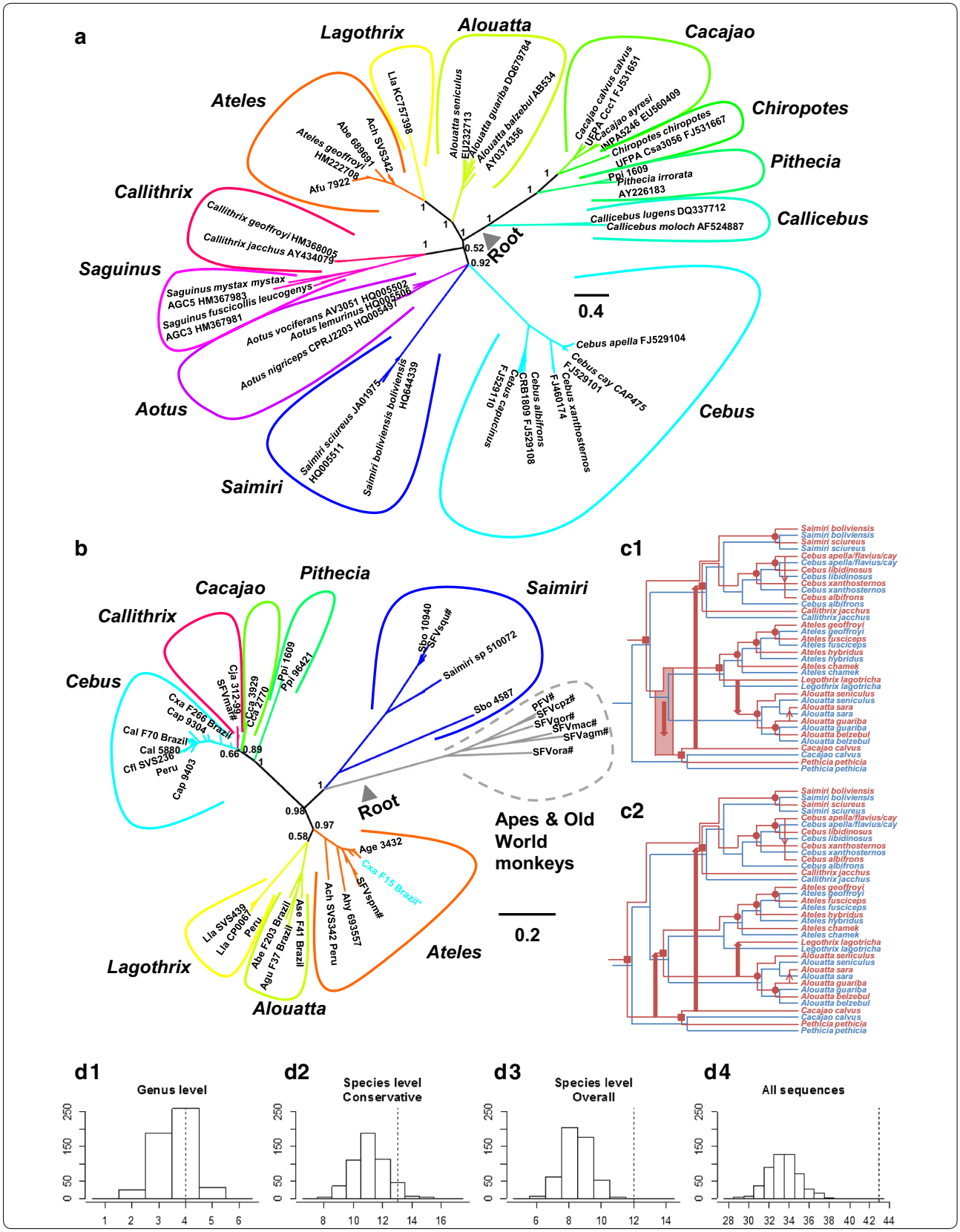
Seroprevalence was higher in adult (17/53, 32.1 %) and sub-adult (1/1, 100 %) animals at rescue centers and zoos in Peru than in juvenile (4/24, 16.7 %) monkeys, but the totals may be too low for a statistically informative comparison. Sex and ages were not recorded for the monkeys captured at the illegal trade market which may limit estimating the overall prevalence in this setting in Peru. Nonetheless, the finding of higher SFV prevalences in adult animals than in juveniles is consistent with that reported for OWMA [1, 17, 41], indicating an increased risk of SFV transmission associated with aggressive behaviors, such as biting and scratching, that occur as monkeys approach sexual maturity.

SFV provirus DNA was detected in 20/107 (18.7 %) FTA samples from Peruvian NWMs by PCR using the highly degenerate *pol* primers (Table 3). The majority of PCR-positive specimens were from animals housed at zoos (16/65, 24.6 %), followed by rescue centers (3/30, 10.0 %). SFV PCR detection was distributed across a wide range of species, including *Cebus apella* (n = 11), *Lagothrix lagotricha* (n = 4), *Ateles chamek* (n = 2), *Alouatta seniculus* (n = 1), *Saimiri boliviensis* (n = 1), and *Saimiri sciureus* (n = 1). One *L. lagotricha* and both PCR-positive *Saimiri* species were from rescue centers. These results are similar to the 24.1 % PCR prevalence reported in mostly captive NWMs from Brazil [28], but are about half that (57.3 %) observed in captive US NWMs in our current study (Table 1).

Co-evolutionary history of NWM SFVs and their hosts

To investigate the co-evolutionary history of NWM SFVs and their hosts, we first estimated their phylogenies, and subsequently compared the two topologies to reconstruct possible co-phylogenetic histories. The host tree contained confirmed SFV-positive primate species and randomly selected SFV-negative monkeys for which DNA specimens were available. The tree was reconstructed by using phylogenetic analysis of 618-bp *CytB* sequences from 158 taxa, of which 32 and 44 are from Peru and US zoo monkeys, respectively, and was rooted according to the phylogeny in [30]. Thirteen *CytB* sequences were from Brazilian NWMs reported in a recent paper investigating SFV diversity [28]. We found that *CytB* sequences of the same host genus cluster together forming monophyletic clades, and that the topology of the estimated tree is generally similar to the one obtained by Perelman et al. [30] (Fig. 3a; Additional file 1: Figure S1, right) with one exception; while we estimated *Aotus* to be a sister clade to the *Saimiri-Cebus* clade, Perelman et al. found *Aotus* to be a sister clade of the *Callithrix-Saguinus* clade. This, however, would not affect our SFV-host co-phylogenetic analysis since our SFV tree does not contain any *Aotus* FVs. Others also have shown that the placement of *Aotus* in NWM phylogeny is not completely resolved [31].

Phylogenetic analysis of the short *pol* SFV sequences (85-bp without primer sequences) was not performed since there is not enough phylogenetic information present in the alignment to accurately resolve the branching orders. Similar results have been reported using slightly longer *pol* sequences (138-bp) that overlaps our diagnostic PCR region [28]. In this study, we successfully amplified longer *pol* sequences (495-bp) from 14 of the 20 (70 %) Peruvian monkeys, including five *L. lagotricha*, five *C. apella*, two *Saimiri* species, one *Ateles chamek*, and one *Alouatta seniculus*. We also obtained 42 additional 495-bp *pol* sequences using blood specimens from NWMs captive in the US, which were used for the PCR and serologic validation assays, and from the SFVcja and SFVspm tissue culture DNAs. SFV *pol* sequences (~383-bp) from three captive NWM species (howler, capuchin, and squirrel) reported by Stenbak et al. [29] only overlap our *pol* sequences by about 197 nucleotides and were from species already in our dataset; thus they were not included in our analyses. The phylogeny of these 58 new *pol* sequences with those available from GenBank,


a
Lagothrix
Alouatta
Cacajao
Ateles
Chiropotes
Pithecia
Callithrix
Callicebus
Saguinus
Aotus
Cebus
Saimiri
Root
0.4
b
Pithecia
Cacajao
Saimiri
Callithrix
Cebus
Root
Apes & Old World monkeys
0.2
Lagothrix
Alouatta
Ateles
c1
c2
d1
Genus level
d2
Species level Conservative
d3
Species level Overall
d4
All sequences

(See figure on previous page.)

Fig. 3 Co-speciation history of New World monkeys (NWM) and simian foamy viruses (SFVs). **a** A consensus Bayesian phylogeny of NWM hosts, estimated from an alignment of cytochrome *B* nucleotide sequences (156 sequences, 618 nt) by using MrBayes 3.2.1 [38]. The tree was rooted according to the phylogeny in [30]. **b** A consensus Bayesian phylogeny of NWM FVs, estimated from an alignment of polymerase nucleotide sequences (74 sequences, 412 nt), and rooted using six ape and Old World monkey SFVs. 'Backbone' sequences are indicated with '*hash*'. 'Cxa F15 Brazil' (written in *blue* and indicated with an *asterisk* '*') was isolated from a *Cebus xanthosternos* monkey in Brazil, but was found to be placed well within the clade of *Ateles* SFVs. The roots of the trees are indicated by *grey triangles*. *Numbers* on nodes are posterior probability node supports. The *scale bars* are in the units of substitutions per site. The comprehensive SFV and host trees are shown in Additional file 1: Figure S1. **c** Competing models for the co-evolutionary histories of NWM SFVs (*red*) and their hosts (*blue*) at the level of species inferred by Jane v4 [39]. The directions of transmissions are indicated by *red arrows*. *Small arrows* indicate cross-species transmissions, and *large arrows* indicate cross-genus transmissions. The *red transparent bars* show the uncertainty of the cross-species transmission timing. Four co-speciation events at the genus level are indicated by *solid red squares*, and those at the species level are indicated by *solid red circles*. Note that, the trees are not scaled to time. **d** The distributions of the number of co-speciation events expected to occur by chance, estimated by using random tip mapping method implemented in Jane v4 [39] (sample size = 500); **d1** genus level; **d2** species level, conservative tree collapsing method; **d3** species level, overall; and **d4** all sequences. The *dotted line* indicates the actual observed number of co-speciation events inferred by Jane v4 [39]. See Additional file 2: Table S1 for a complete list of species codes used in the study; PFV is primate foamy virus which is the new name given to HFV (human foamy virus)

including 13 from Brazilian NWMs [28], is shown in Fig. 3b and Additional file 1: Figure S1 (left).

We noticed that there are some inconsistencies between the SFV tree estimated herein and the one obtained in our previous study that was based on an analysis of short *pol* sequences (17 sequences, 276 nt) [28]. In our previous study, we found 'Cxa F266' isolated from a *C. xanthosternos* clustered with a *Callithrix* SFV, suggesting that this virus may represent a recent cross-genus transmission. However, in this work we found Cxa F266 to be more closely related to *Cebus* SFVs than to *Callithrix* SFVs which is consistent with a phylogenetic analysis of long terminal repeat-*gag* sequences (18 sequences, 265 nt) that was also conducted in our previous study [28]. Combined, it is therefore more likely that Cxa F266 is in fact a *Cebus* SFV, and hence does not represent a cross-genus transmission. Another inconsistency found is that while our present study showed that SFV from a spider monkey (SFVspm) is more closely related to SFV from a common marmoset (SFVmar) than to SFV from a squirrel monkey (SFVsqu) (which is also consistent with previous Gag and Pol protein analyses [26, 42]), our previous work [28] showed that SFVspm is more closely related to SFVsqu than SFVmar. These two disparities are likely because the *pol* sequences used in our previous study were too short and/or the number of SFV sequences used in the analysis was too low.

Overall, phylogenetic analyses show that SFVs isolated from the same host genus tend to cluster together forming monophyletic clades. SFV 'Cxa F15 Brazil' is the only exception to this; as shown in a previous study [28], this virus was isolated from a *Cebus xanthosternos* monkey but instead clustered within the *Ateles* SFV clade. This likely represents a recent cross-genus SFV transmission which has been reported previously in captive and wild OWMAs [22, 43, 44]. By comparing the topologies of the viral and host trees, another three cross-genus transmissions were inferred: (1) from an ancestral Atelidae monkey to an ancestral Pitheciinae monkey (Fig. 3c1) or vice versa (Fig. 3c2), (2) from an ancestral *Cacajao* monkey to the lineage giving rise to *Cebus* and *Callithrix* monkeys, and (3) from an ancestral *Lagothrix* monkey to an ancestral *Alouatta* monkey (Fig. 3c1) or vice versa (Fig. 3c2). Four potential co-speciation events at the level of viral genera were also inferred: (1) the divergence of the *Pithecia* SFV lineage from *Cacajao* SFV lineage, (2) the separation of the *Ateles* SFV lineage from *Lagothrix* (Fig. 3c1) or *Alouatta* (Fig. 3c2) SFV lineage, (3) the split of an ancestral *Cebus* SFV from an ancestral *Callithrix* SFV, and (4) the divergence of the *Saimiri* SFV lineage from *Atelidae* (Fig. 3c1) or *Pitheciidae* (Fig. 3c2) SFV lineage. Nevertheless, it is noteworthy that the clade of *Lagothrix* and *Alouatta* is weakly supported (posterior probability = 0.58, Fig. 3b). An alternative scenario could imply that *Atelidae* SFVs and their hosts co-diverged with one another, decreasing the number of cross-genus transmissions by one and thereby increasing the number of co-divergence events by one. These four (or five) co-divergence events, however, are not greater than expected to occur by chance [random tip mapping: sample size = 500, p (4 co-speciation events) = 0.516, p (5 co-speciation events) = 0.062, Fig. 3d1].

SFV transmissions among closely related NWMs are also relatively common. *Cebus* FVs are a clear example of this. Phylogenetic analyses show that *C. apella*, *C. flavius*, and *C. cay* FVs, as well as their hosts, cluster together without forming clear phylogenetic structures (Additional file 1: Figure S1), indicative of frequent cross-species transmissions among these closely related monkeys with overlapping habitats. A few transmissions within the clades of *Ateles* monkeys and *Saimiri* monkeys were also observed (Additional file 1: Figure S1). To examine whether or not SFVs stably co-speciate with their NWM hosts at the species level, we collapsed the clades

comprising host and viral sequences of the same species into one representative sequence and compared the trees. In total, 13 co-speciation events were inferred. Again, this is not greater than expected to occur by chance (random tip mapping: sample size = 500, p = 0.108, Fig. 3d2). Nonetheless, it is important to note that this method is highly conservative, retaining all transmissions in the trees. Since our samples were collected from a market, zoos, and rescue centers where cross-species transmissions can readily occur due to extreme close proximity, it is possible that this method might be too conservative and the noise from spurious recent cross-species transmissions might overwhelm the signal of the FV-host co-speciation history.

To examine this possibility, we pruned the FV tree further such that it contains only one FV per one host species, of which the phylogenetic placement represents the majority of its kind. We then compared this pruned FV tree to the host species tree. 12 potential co-speciation events were inferred in total (Fig. 3c), which are greater than expected to occur by chance (random tip mapping: sample size = 500, p < 0.002, Fig. 3d3). Combined with the results above, this finding suggests that, overall, NWM SFVs stably and broadly co-diverge with their hosts at the species level over the long timescale; however, cross-species and genus transmissions are not rare. Lastly, we compared the entire SFV tree with the host tree (Additional file 1: Figure S1). In total, 43 potential co-speciation events were inferred, which is greater than expected to occur by chance (Random tip mapping sample size = 500, p < 0.002, Fig. 3d4). These findings suggest that NWM FVs stably co-speciate with their hosts at the level of subpopulation also, and are consistent with the results from a previous study that was based on OWMA FVs [18].

Evaluating and refining the co-speciation model

As discussed above, many of the inferred co-divergence events cannot be placed conclusively onto the trees. To further evaluate the co-speciation events and to determine which alternative scenarios are more likely, we time-calibrated the SFV tree using the dates of some of the co-speciation events that could be mapped conclusively onto the trees, directly inferred from the host timescales estimated in [29]. Interestingly, although our reconciliation analyses suggested that the *Pithecia-Cacajao* SFV lineage separation and the *Cebus-Callithrix* SFV lineage separation are FV-host co-divergence events, the inferred dates suggested otherwise. While the branching dates of the *Pithecia* and *Cebus* SFV lineages were inferred to be ~13.29, and ~19.95 Ma, respectively, the split of the *Pithecia* SFV lineage was topologically determined to be before that of the *Cebus* SFV lineage with strong support (posterior probability = 0.89, Fig. 3b; Additional file 1: Figure S1). This implies that at least one of these two SFV splits was incorrectly determined as a co-speciation event, and therefore could not be used to calibrate the tree. In total, 4 dates were used for time-calibration (Table 5): (1) the separation date of the NWM-OMWA FVs [~43.47 Ma (95 % HPD = 38.55–48.36), [30]], (2) the branching date of the *Cebus xanthosternos* SFV lineage [~1.95 Ma (95 % HPD = 0.91–3.31),

Table 5 Time to most recent common ancestor (tMRCA) mean estimates for Haplorrhini and simian foamy virus (SFV) polymerase (*pol*) sequences

Branch node	tMRCA SFV *pol*	tMRCA simian phylogeny[a]	Fossil estimate[a]
Simiiformes	39.85 (34.90–44.89)	43.47 (38.55–48.36)[b]	43 ± 4.5
Platyrrhini	22.31 (14.84–31.04)	24.82 (20.55–29.25)	23.5 ± 3.0
Atelidae (*Ateles/Alouatta* split)	10.77 (7.45–15.04)	16.13 (10.52–21.35)	NA[c]
Atelidae (*Ateles/Lagothrix* split)	10.77 (7.45–15.04)	11.25 (7.25-15.46)	NA
Ateles	6.28 (4.56–8.05)	5.07 (2.87–7.50)[b]	NA
Alouatta (excluding *Alouatta palliatta*)	5.54 (3.99–7.27)	4.94 (2.93–7.26)[b]	NA
Cebus (*C. apella/C. xanthosternos* split)	3.45 (2.61–4.35)	1.95 (0.91–3.31)[b]	NA
Cebus/Callithrix	5.20 (3.08–6.06)	19.95 (15.66–24.03)	NA
Pitheciinae	8.11 (4.88–12.33)	13.69 (9.24–18.34)	NA
Saimirinae[d]	16.89 (10.18–24.93)	2.24 (1.05–3.73)	NA

Using an alignment of 412 nt for 74 SFV taxa. Millions of years (MY) ago. Medians inferred using Bayesian methods and a relaxed molecular clock; ranges in parentheses are 95 % highest posterior density intervals

[a] Dating and fossil estimates from Perelman et al. [30]

[b] Dates used to calibrate the SFV tree

[c] NA, not available

[d] Analysis excluded SFV_Sbo4587 which had an unusually long branch length

[30]], (3) the branching date of the *Ateles chamek* SFV lineage [~5.07 Ma (95 % HPD = 2.87–7.50), [30]], and (4) the separation date between the *Alouatta belzebul* and *Alouatta sara* SFV lineages [~4.95 Ma (95 % HPD = 2.93–7.26), [30]]. The time-calibrated SFV tree is shown in Fig. 4a.

Overall, our analysis estimated the times to most recent common ancestors (tMRCAs) of SFVs to be comparable to those of their hosts (Table 5), and the nucleotide substitution rate was calculated to be ~2.14×10^{-8} (95 % HPD = 1.72×10^{-8}–2.62×10^{-8}) substitutions per site per year (s/n/y), which is also similar to previously estimated rates of SFV evolution (~1.7×10^{-8} s/n/y [22] and ~ 7.79×10^{-9} s/n/y [28]). The estimated date for the separation of the *Pithecia* and *Cacajao* SFVs [~8.11 Ma (95 % HPD = 4.88–12.33)] was found to be relatively

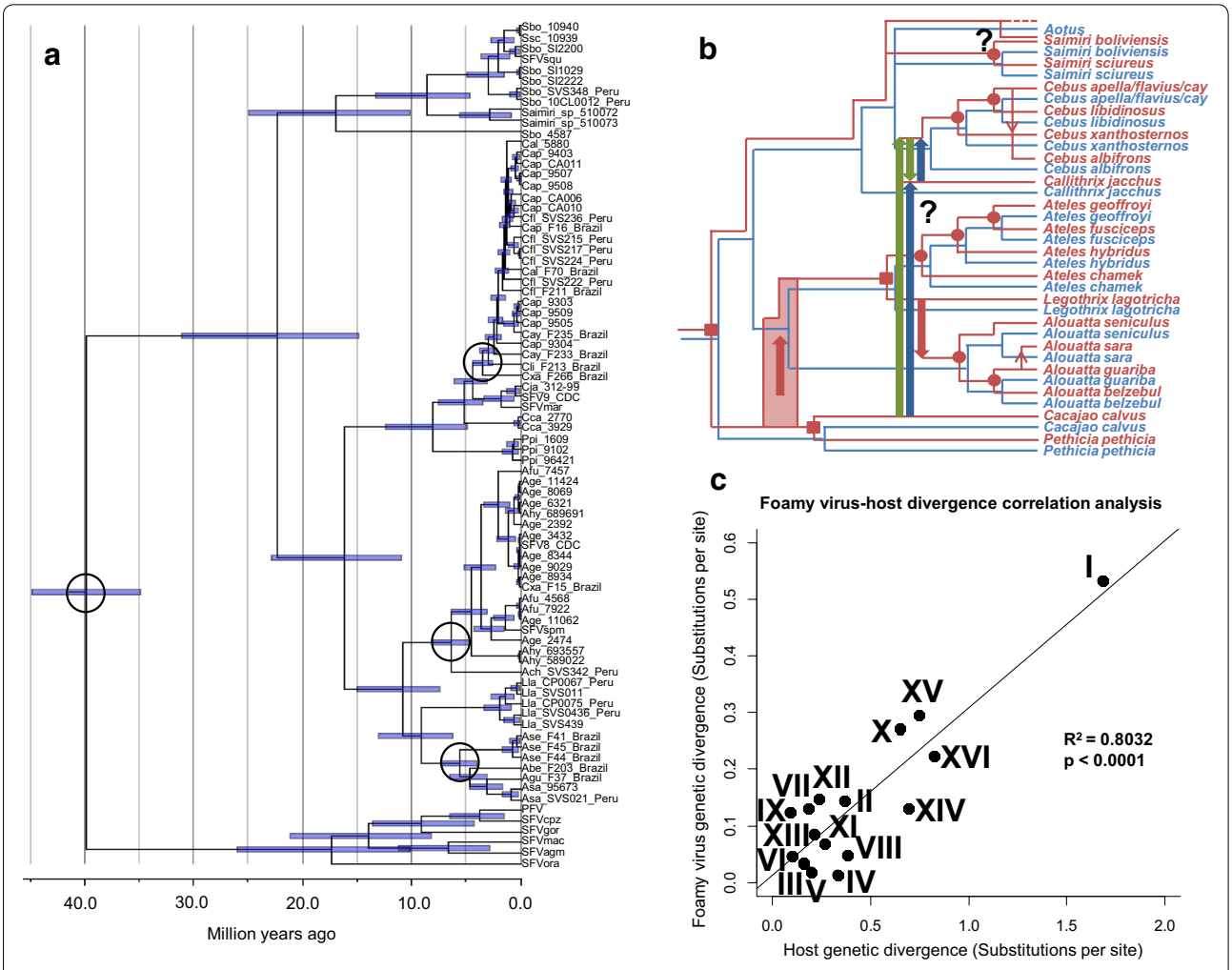


Fig. 4 Evaluating and refining simian foamy virus (SFV)-host co-evolutionary model. **a** Calibrated maximum clade credibility Bayesian SFV phylogeny, estimated from an alignment of polymerase nucleotide sequences (74 sequences, 412 nt), using BEAST 1.8.0 [40]. The topology was fixed to the one we obtained from the MrBayes analyses. The calibrating nodes are encircled. The *bars* represent the uncertainty of the estimated node heights. The timescale is in millions of years. See Additional file 2: Table S1 for a complete list of species codes used in the study; PFV is primate foamy virus which is the new name given to HFV (human foamy virus). **b** Refined co-evolutionary history of SFVs (*red*) and their hosts (*blue*). *Red dotted branch* represents a possible *Aotus* ghost FV lineage that has not been sampled in our study. The directions of cross transmissions are indicated by *arrows*. *Small arrows* indicate cross-species transmission events, and *large arrows* indicate cross-genus transmission events. The *red transparent bar* shows the uncertainty of the cross-species transmission timing. Two alternative possible SFV cross-species transmission scenarios involving *Cacajao*, *Cebus*, and *Callithrix* monkeys are shown in *green* and *blue*. '*Question marks*' indicate ambiguous cross-species transmission pathways. Co-speciation events at the genus level are indicated by *solid red squares*, and those at the species level are indicated by *solid red circles*. The trees are not scaled to time. **c** SFV-host divergence correlation analysis. *Black dots* represent co-diverging branches identified under the refined SFV-host co-evolutionary model (see panel **b**). A well supported linear correlation was found (linear regression: N = 16, $R^2 = 0.8032$, $p < 0.0001$), represented by a *solid black line*. The *dots* are labelled with *roman numerals* (I–XVI), referring to branches in Additional file 1: Figure S1

more comparable to that of their hosts [~13.69 Ma (95 % HPD = 9.24–18.34), [30]] than that of the *Callithrix* and *Cebus* SFVs [~4.36 Ma (95 % HPD = 3.08–6.06)] which was estimated to happen much later than the host split [~19.95 Ma (95 % HPD = 15.66–24.03), [30]]. These findings thus suggest that the *Pithecia-Cacajao* SFV separation is likely a co-speciation event as initially inferred, and the split between the *Cebus* and *Callithrix* SFVs is not, but rather represents a cross-genus transmission between the two host monkeys (Fig. 4b, short blue or green large arrows). These results also imply that the initially inferred cross-genus transmission from an ancestral *Cacajao* monkey to the *Cebus* and *Callithrix* MRCA (Fig. 3c) was erroneous; it is either a cross-genus transmission from an ancestral *Cacajao* to an ancestral *Callithrix* monkey (Fig. 4b, long blue large arrow), or to an ancestral *Cebus* monkey (Fig. 4b, long green large arrow). To distinguish between these two competing alternative scenarios would require additional SFV sequence data from other NWMs, such as *Mico*, *Cebuella*, and *Callimico* monkeys.

Our phylogenetic reconciliation analyses also suggested that the branching of the *Ateles* SFV lineage is an FV-host co-speciation event, but it could be mapped either to the *Ateles-Lagothrix* monkey split event (Fig. 3c1) or to the *Ateles-Alouatta* monkey split event (Fig. 3c2). Our additional analyses estimated the evolutionary timescale of *Atelidae* SFVs to be ~10.77 Myr old (95 % HPD = 7.45–15.04). This is comparable to the separation date of the *Ateles* and *Lagothrix* monkeys [~11.25 Ma (95 % HPD = 7.25–15.46), [30]], suggesting that the former scenario is more likely.

Lastly, our analyses estimated the evolutionary timescale of *Saimiri* SFV radiation [~16.89 Ma (95 % HPD = 10.18–24.93) to be ~7.5× that of their hosts (~2.24 Ma (95 % HPD = 1.05–3.73), [30]]. This finding suggests that the *Saimiri* SFV clade likely contains one, or more, NWM SFVs lineages that arose from a cross-species transmission event, but for which the ancestral SFV lineages were not sampled. Assuming a stable FV-host co-speciation history, we hypothesized that the ancestral FV 'ghost lineage' may have been an *Aotus* SFV (Fig. 4b), since the split between *Aotus* and *Saimiri* monkeys were inferred to occur ~19.95 Ma (95 % HPD = 15.66–24.03) [29], comparable to our estimated *Saimiri* SFV evolutionary timescale. To further examine this hypothesis, *Aotus* SFV sequences are required. However, while several *Aotus* species have been identified with antibodies to NWM SFV antigens, SFV sequences have not been reported from these seropositive animals for phylogenetic analysis to test this hypothesis (Table 4; [28]). Furthermore, we also found that Saimiri SFVs form a separate, ancestral lineage to all other NWM SFVs (Figs. 3, 4a), and that the two diverged from one another ~22.31 Ma (95 % HPD = 14.84–31.04), comparable to the date of the basal NWM diversification [~24.82 Ma (95 % HPD = 20.55–29.25), [30]]. This date in turn supports the hypothesis that the split between *Saimiri* SFVs and the rest is likely a co-speciation event, corresponding to the basal radiation of NWMs. The refined SFV-host co-evolutionary history is shown in Fig. 4b.

An SFV-host divergence correlation analysis (Fig. 4c) was also performed to further evaluate our refined model of SFV-host co-evolution. Unlike the phylogenetic reconciliation analyses, this analysis takes both the topologies of the virus and host trees as well as their branch lengths into account. In this analysis, we identified SFV-host co-diverging branches based on the refined model, and examined if the genetic divergences of the two are linearly correlated. We found that, given the model, the SFV genetic divergence is proportional to that of their hosts (linear regression: $R^2 = 0.8032$; $p < 0.0001$, Fig. 4c). This indicates that the divergence of SFVs and their hosts are internally consistent, and simultaneously supporting our model as well as the stable NWM SFV-host co-speciation hypothesis. We note that the number of co-speciation events is reduced by one, and the number of cross-species transmission events is increased by one under the refined model. Nonetheless, these changes do not alter the conclusion about the stable co-speciation history, as the probability for the observed 11 co-speciation events by chance is still significant ($p = 0.008$). Our results extend those of others demonstrating an ancient coevolution of OWMA SFV to include co-speciation of NWM SFV. However, unlike OWMA SFV that has only rare and relatively recent cross-species transmissions, NWM SFVs may have had at least four ancient cross-genus transmission events in their evolutionary history. Sequence analysis of additional NWM SFVs, especially those from *Aotus* species, are required to further evaluate this scenario.

Zoonotic potential of NWM SFV

The high prevalence and distribution of SFV in many NWM species reported herein and in other studies [28, 29] highlights the potential zoonotic infection risks for persons handling neotropical monkeys in captivity as pets or at zoological and research institutions and via hunting and butchering in the wild. An estimated 28,000 NHPs are also reported to be collected in Peru each year [45], most for use in biomedical research [46], but some of the larger NWMs including capuchin, spider, and woolly monkeys are hunted for bushmeat (http://www.careforthewild.com). These activities increase opportunities for zoonotic transmission of SFV [9]. Nonetheless, only a single study has reported evidence of human infection with NWM SFV in 11.6 % seroreactive but PCR

negative persons [29]. These findings suggest exposure without infection or nonspecific seroreactivity to the antigen used in their assay. Alternatively, limited validation of the PCR primers used in the testing may have affected the assay sensitivity for detecting divergent NWM SFV in the seroreactive persons. Host restriction may also have contributed to the lack of productive infection in those seroreactive primate workers with Trim5α and APOBEC3 having been shown to have a broad antiretroviral activity [47] may contribute to preventing productive infection of humans with some NWM SFVs. However, human Trim5α expressed in Cf2Th cells could inhibit only SFVsqu but not SFVspm or SFVmar suggesting a limited viral suppression potential of NWM SFV by human Trim5α [26]. Little is known about the ability of human APOBEC3 to inhibit NWM SFVs, with most work limited to OWMA SFV, but which showed that the SFV Bet protein from a variety of OWMAs can counteract human APOBEC3 activity, suggesting that the Bet of NWM SFVs may have similar neutralizing activity [48–50]. Additional studies of larger numbers of workers and also of persons who hunt NWMs using the new serological and PCR assays described here are required to further evaluate the risk of human infection with neotropical monkey SFVs.

Conclusions

We demonstrate here, and expand on previous studies in Brazil, a broad range of neotropical primate species harboring SFV in captive animals from the US and Peru and in wild-caught NWMS from Peru. We identified at least seven novel and divergent SFVs found in *Ateles chamek* (Peruvian spider monkey) and two species of *A. geoffroyi* (black-handed and Mexican spider monkeys), *Cacajao rubicundus* (uakari monkey), *Pithecia pithecia* (saki monkey), *Saimiri sciureus* (squirrel monkey) and *S. boliviensis peruviensis* (Peruvian squirrel monkey), and *Lagothrix lagotricha* (brown wooly monkey). Co-speciation between SFV and their hosts in all three NWM families was inferred but included ancient cross-genus transmissions of SFV from Pitheciinae to Atelidae, *Cacajao* to *Cebus*/*Callithrix*, between *Cebus* and *Callithrix*, and *Lagothrix* to *Alouatta*. Our results provide an important future reference for time-calibrating information essential for NWM SFV evolutionary timescale inference, as well as evolutionary insight of NWM SFVs. Further studies are needed to expand our understanding of the evolutionary history of broader NWM SFVs and to clarify the possible "ghost" lineage identified by our analyses. Furthermore, the new serological and molecular tools reported here will facilitate an assessment of the risk of zoonotic SFV infection in persons naturally or occupationally exposed to NWMs.

Additional files

Additional file 1: Figure S1. Co-speciation history of New World monkeys (NWM) and simian foamy viruses (SFVs). A consensus NWM tree (left) estimated from a cytochrome-B nucleotide alignment (156 sequences, 618 nt) is compared to a consensus NWM SFV tree (right) estimated from a polymerase nucleotide alignment (80 sequences, 412 nt). Both trees were constructed under the Bayesian phylogenetic framework by using MrBayes 3.2.1 [38], and molecular clocks were not imposed. The host tree was rooted according to the tree in [30]. The FV tree was rooted with ape and Old World monkey SFVs. The 'backbone' polymerase sequences in the SFV tree are labelled with '**#**'. The scale bars are in the units of substitutions per site, and numbers on nodes are posterior probabilities. Grey lines indicate SFV-host associations. SFVs without associating hosts are labelled in grey, and were excluded from the phylogenetic reconciliation analysis. Thick branches are co-diverging branches used in the SFV-host divergence correlation analysis, labelled with roman numerals (**I**-**XVI**), referring to dots in Fig. 4C. See Supplementary Table S1 for a complete list of species codes used in the study.

Additional file 2: Table S1. Three letter species codes used for New World monkeys samples in the study and used to label taxa in the phylogenetic analyses (Figs. 3, 4, and S1).

Authors' contributions

BMG, JMM, DGB, and WMS conceived the project; PA, AK, HJ, and WMS performed the sequence analyses; HJ performed the laboratory work and BMG, PM, DGB, MRK, JMM, and WMS provided samples. BMG, PA, AK and WMS wrote the paper. All authors read and commented on the manuscript. All authors read and approved the final manuscript.

Author details

[1] U.S. Naval Medical Research Unit No. 6, Lima, Peru. [2] Laboratory Branch, Division of HIV/AIDS Prevention, National Center for HIV/AIDS, Viral Hepatitis, STD, and TB Prevention, Centers for Disease Control and Prevention, 1600 Clifton Rd., MS G-45, Atlanta, GA 30329, USA. [3] Department of Zoology, University of Oxford, Oxford OX1 3PS, UK. [4] Wildlife Conservation Society, Lima, Peru. [5] Tulane School of Public Health and Tropical Hygiene, New Orleans, LA, USA. [6] Centers for Disease Control and Prevention, Atlanta, GA 30333, USA.

Acknowledgements

We are thankful for the authorization for collection, exportation and processing of samples from the Peruvian Ministry of Agriculture (RD N 0363-2010-AG-DGFFS-DGEFFS and RD N 411-2010-AG-DGFFS-DGEFFS; export permit N 001312, Access to genetic material contract 0016-2014-MINAGRI-DGFFS/DGEFFS). We also thank the staff at all the Peru and US zoos, rescue centers and research institutions that provided the archived and opportunistic blood specimens. This work would not be possible without the collaborations of Marieke Rosenbaum, Nancy Cavero, Tatiana Quevedo, Catalina Hermoza, Milagros Ramos, Helene Collongues and Raul Bello. Use of trade names is for identification only and does not imply endorsement by the US Department of Health and Human Services, the Public Health Service, or the Centers for Disease Control and Prevention. This work was supported by the US Department of Defense, Global Emerging Infections Surveillance and Response System, the US Centers for Disease Control and Prevention and the United States Agency for International Development (USAID) Emerging Pandemic Threats PREDICT. P. A. is funded by the Royal Thai Government. A. K. is funded by the Royal Society. The findings and conclusions in the report are those of the authors and do not necessarily reflect the official policy or position of the Department of the Navy, Department of Defense, the Centers for Disease Control and Prevention, or the US Government.

Competing interests

The authors declare they have no competing interests.

Received: 28 April 2015 Accepted: 4 October 2015
Published online: 29 October 2015

References

1. Murray SM, Linial ML. Foamy virus infection in primates. J Med Primatol. 2006;35:225–35.
2. Linial ML, Fan H, Hahn B, Löwer R, Neil J, Quackenbush S, Rethwilm A, Sonigo P, Stoye J, Tristem M. Retroviridae. In: Fauquet CM, Mayo MA, Maniloff J, Desselberger U, Ball LA, editors. Virus taxonomy, 7th report of the International Committee on taxonomy of viruses. London: Elsevier/ Academic Press; 2004. p. 421–40.
3. Flanagan M. Isolation of a spumavirus from a sheep. Aust Vet J. 1992;69:112–3.
4. Achong BG, Mansell PW, Epstein MA. A new human virus in cultures from a nasopharyngeal carcinoma. J Pathol. 1971;103:P18.
5. Meiering CD, Linial ML. Historical perspective of foamy virus epidemiology and infection. Clin Microbiol Rev. 2001;14:165–76.
6. Enders JF, Peebles TC. Propagation in tissue cultures of cytopathogenic agents from patients with measles. Proc Soc Exp Biol Med. 1954;86:277–86.
7. Linial ML. Foamy viruses are unconventional retroviruses. J Virol. 1999;73:1747–55.
8. Khan AS. Simian foamy virus infection in humans: prevalence and management. Expert Rev Anti Infect Ther. 2009;7:569–80.
9. Switzer WM, Heneine W. Foamy virus infection of humans. In: Liu D, editor. Molecular detection of human viral pathogens, vol 1. Boca Raton: CRC Press, Taylor and Francis Group; 2011. p. 131–46.
10. Betsem E, Rua R, Tortevoye P, Froment A, Gessain A. Frequent and recent human acquisition of simian foamy viruses through apes' bites in central Africa. PLoS Pathog. 2011;7:e1002306.
11. Switzer W, Ahuka-Mundeke S, Tang S, Shankar A, Wolfe N, Heneine W, Peeters M, Ayouba A, Mulembakani P, Rimoin A. Simian foamy virus (SFV) infection from multiple monkey species in women from the Democratic Republic of Congo. Retrovirology. 2011;8:A233.
12. Boneva RS, Switzer WM, Spira TJ, Bhullar VB, Shanmugam V, Cong ME, Lam L, Heneine W, Folks TM, Chapman LE. Clinical and virological characterization of persistent human infection with simian foamy viruses. AIDS Res Hum Retroviruses. 2007;23:1330–7.
13. Schweizer M, Falcone V, Gange J, Turek R, Neumann-Haefelin D. Simian foamy virus isolated from an accidentally infected human individual. J Virol. 1997;71:4821–4.
14. Falcone V, Leupold J, Clotten J, Urbanyi E, Herchenroder O, Spatz W, Volk B, Bohm N, Toniolo A, Neumann-Haefelin D, Schweizer M. Sites of simian foamy virus persistence in naturally infected African green monkeys: latent provirus is ubiquitous, whereas viral replication is restricted to the oral mucosa. Virology. 1999;257:7–14.
15. Rua R, Betsem E, Gessain A. Viral latency in blood and saliva of simian foamy virus-infected humans. PLoS One. 2013;8:e77072.
16. Blewett EL, Black DH, Lerche NW, White G, Eberle R. Simian foamy virus infections in a baboon breeding colony. Virology. 2000;278:183–93.
17. Calattini S, Wanert F, Thierry B, Schmitt C, Bassot S, Saib A, Herrenschmidt N, Gessain A. Modes of transmission and genetic diversity of foamy viruses in a Macaca tonkeana colony. Retrovirology. 2006;3:23.
18. Switzer WM, Bhullar V, Shanmugam V, Cong ME, Parekh B, Lerche NW, Yee JL, Ely JJ, Boneva R, Chapman LE, et al. Frequent simian foamy virus infection in persons occupationally exposed to nonhuman primates. J Virol. 2004;78:2780–9.
19. Heneine W, Switzer WM, Sandstrom P, Brown J, Vedapuri S, Schable CA, Khan AS, Lerche NW, Schweizer M, Neumann-Haefelin D, et al. Identification of a human population infected with simian foamy viruses. Nat Med. 1998;4:403–7.
20. Calattini S, Betsem EB, Froment A, Mauclere P, Tortevoye P, Schmitt C, Njouom R, Saib A, Gessain A. Simian foamy virus transmission from apes to humans, rural Cameroon. Emerg Infect Dis. 2007;13:1314–20.
21. Jones-Engel L, Engel GA, Schillaci MA, Rompis A, Putra A, Suaryana KG, Fuentes A, Beer B, Hicks S, White R, et al. Primate-to-human retroviral transmission in Asia. Emerg Infect Dis. 2005;11:1028–35.
22. Switzer WM, Salemi M, Shanmugam V, Gao F, Cong ME, Kuiken C, Bhullar V, Beer BE, Vallet D, Gautier-Hion A, et al. Ancient co-speciation of simian foamy viruses and primates. Nature. 2005;434:376–80.
23. Hooks JJ, Gibbs CJ Jr, Chou S, Howk R, Lewis M, Gajdusek DC. Isolation of a new simian foamy virus from a spider monkey brain culture. Infect Immun. 1973;8:804–13.
24. Marczynska B, Jones CJ, Wolfe LG. Syncytium-forming virus of common marmosets (*Callithrix jacchus jacchus*). Infect Immun. 1981;31:1261–9.
25. Johnston PB. Taxonomic features of seven serotypes of simian and ape foamy viruses. Infect Immun. 1971;3:793–9.
26. Pacheco B, Finzi A, McGee-Estrada K, Sodroski J. Species-specific inhibition of foamy viruses from South American monkeys by New World Monkey TRIM5{alpha} proteins. J Virol. 2010;84:4095–9.
27. Thumer L, Rethwilm A, Holmes EC, Bodem J. The complete nucleotide sequence of a New World simian foamy virus. Virology. 2007;369:191–7.
28. Muniz CP, Troncoso LL, Moreira MA, Soares EA, Pissinatti A, Bonvicino CR, Seuanez HN, Sharma B, Jia H, Shankar A, et al. Identification and characterization of highly divergent simian foamy viruses in a wide range of new world primates from Brazil. PLoS One. 2013;8:e67568.
29. Stenbak CR, Craig KL, Ivanov SB, Wang X, Soliven KC, Jackson DL, Gutierrez GA, Engel G, Jones-Engel L, Linial ML. New World simian foamy virus infections in vivo and in vitro. J Virol. 2014;88:982–91.
30. Perelman P, Johnson WE, Roos C, Seuanez HN, Horvath JE, Moreira MA, Kessing B, Pontius J, Roelke M, Rumpler Y, et al. A molecular phylogeny of living primates. PLoS Genet. 2011;7:e1001342.
31. Schneider H, Sampaio I. The systematics and evolution of New World primates—a review. Mol Phylogenet Evol. 2015;82 Pt B:348–57. doi:10.1016/j.ympev.2013.10.017.
32. Pacheco V, Cadenillas R, Salas E, Tello C, Zeballos H. Diversidad y endemismo de los mamiferos del Peru. Revista Peruana de Biologia. 2009;16:5–32.
33. Wolfe ND, Heneine W, Carr JK, Garcia AD, Shanmugam V, Tamoufe U, Torimiro JN, Prosser AT, Lebreton M, Mpoudi-Ngole E, et al. Emergence of unique primate T-lymphotropic viruses among central African bushmeat hunters. Proc Natl Acad Sci USA. 2005;102:7994–9.
34. Hussain AI, Shanmugam V, Bhullar VB, Beer BE, Vallet D, Gautier-Hion A, Wolfe ND, Karesh WB, Kilbourn AM, Tooze Z, et al. Screening for simian foamy virus infection by using a combined antigen Western blot assay: evidence for a wide distribution among Old World primates and identification of four new divergent viruses. Virology. 2003;309:248–57.
35. Wolfe ND, Switzer WM, Carr JK, Bhullar VB, Shanmugam V, Tamoufe U, Prosser AT, Torimiro JN, Wright A, Mpoudi-Ngole E, et al. Naturally acquired simian retrovirus infections in central African hunters. Lancet. 2004;363:932–7.
36. Tamura K, Stecher G, Peterson D, Filipski A, Kumar S. MEGA6: molecular evolutionary genetics analysis version 6.0. Mol Biol Evol. 2013;30:2725–9.
37. Martin DP, Lemey P, Lott M, Moulton V, Posada D, Lefeuvre P. RDP3: a flexible and fast computer program for analyzing recombination. Bioinformatics. 2010;26:2462–3.
38. Ronquist F, Teslenko M, van der Mark P, Ayres DL, Darling A, Hohna S, Larget B, Liu L, Suchard MA, Huelsenbeck JP. MrBayes 3.2: efficient Bayesian phylogenetic inference and model choice across a large model space. Syst Biol. 2012;61:539–42.
39. Conow C, Fielder D, Ovadia Y, Libeskind-Hadas R. Jane: a new tool for the cophylogeny reconstruction problem. Algorithms Mol Biol. 2010;5:16.
40. Drummond AJ, Rambaut A. BEAST: Bayesian evolutionary analysis by sampling trees. BMC Evol Biol. 2007;7:214.
41. Hood S, Mitchell JL, Sethi M, Almond NM, Cutler KL, Rose NJ. Horizontal acquisition and a broad biodistribution typify simian foamy virus infection in a cohort of *Macaca fascicularis*. Virol J. 2013;10:326.
42. Kehl T, Tan J, Materniak M. Non-simian foamy viruses: molecular virology, tropism and prevalence and zoonotic/interspecies transmission. Viruses. 2013;5:2169–209.
43. Leendertz FH, Zirkel F, Couacy-Hymann E, Ellerbrok H, Morozov VA, Pauli G, Hedemann C, Formenty P, Jensen SA, Boesch C, Junglen S. Interspecies transmission of simian foamy virus in a natural predator–prey system. J Virol. 2008;82:7741–4.
44. Liu W, Worobey M, Li Y, Keele BF, Bibollet-Ruche F, Guo Y, Goepfert PA, Santiago ML, Ndjango JB, Neel C, et al. Molecular ecology and natural history of simian foamy virus infection in wild-living chimpanzees. PLoS Pathog. 2008;4:e1000097.
45. Bodmer RE, Lozano EP. Rural development and sustainable wildlife use in Peru Desarrollo Rural y Utilización de Vida Silvestre en Perú. Conserv Biol. 2001;15:1163–70.
46. Maldonado AM, Nijman V, Bearder SK. Trade in night monkeys Aotus spp. in the Brazil–Colombia–Peru tri-border area: international wildlife trade regulations are ineffectively enforced. Endanger Species Res. 2009;9:143–9.

47. Wolf D, Goff SP. Host restriction factors blocking retroviral replication. Annu Rev Genet. 2008;42:143–63.
48. Perkovic M, Schmidt S, Marino D, Russell RA, Stauch B, Hofmann H, Kopietz F, Kloke BP, Zielonka J, Strover H, et al. Species-specific inhibition of APOBEC3C by the prototype foamy virus protein bet. J Biol Chem. 2009;284:5819–26.
49. Lochelt M, Romen F, Bastone P, Muckenfuss H, Kirchner N, Kim YB, Truyen U, Rosler U, Battenberg M, Saib A, et al. The antiretroviral activity of APOBEC3 is inhibited by the foamy virus accessory Bet protein. Proc Natl Acad Sci USA. 2005;102:7982–7.
50. Russell RA, Wiegand HL, Moore MD, Schafer A, McClure MO, Cullen BR. Foamy virus Bet proteins function as novel inhibitors of the APOBEC3 family of innate antiretroviral defense factors. J Virol. 2005;79:8724–31.

Submit your next manuscript to BioMed Central and take full advantage of:

- **Convenient online submission**
- **Thorough peer review**
- **No space constraints or color figure charges**
- **Immediate publication on acceptance**
- **Inclusion in PubMed, CAS, Scopus and Google Scholar**
- **Research which is freely available for redistribution**

Submit your manuscript at
www.biomedcentral.com/submit

BioMed Central