Paleovirology and the evolution of virus-host gene exchange



Amr Aswad
St Hilda's College
University of Oxford

A thesis submitted for the degree of Doctor of Philosophy

Acknowledgments

To my parents, without whom I would have nothing,
To Mazen, with whom I learnt to questioned everything,
To Yiannis, without whom I would have been alone,
To Christina, without whom I wouldn't have taken the leap,
To Aris, without whom I wouldn't have grown,
To Franziska, without whom I would be writing this,
Thank you.

Abstract

Paleovirology is the study of viruses at large evolutionary timescales through the investigation of endogenous viral elements (EVEs) that are found in host genomes. EVEs are the result of a heritable, genomic integration of a viral genome in a host germline cell. The presence of retroviral sequences in animal genomes has been recognized since the 1970's, but discoveries made in the last five years have revealed that viruses of all genome types can endogenise. Moreover, although most EVEs are non-functional relics, there are a number of examples of functional EVEs that are co-opted to benefit the host. There remain some methodological challenges in the identification and analysis of EVEs, and we have yet to fully understand the evolutionary processes involved in the rare instances of co-option. In this thesis, I develop novel techniques to identify and characterize large herpesviruses, which lead to the discovery of the first endogenous herpesvirus in the genome of the Philippine tarsier. The same methods were redeployed to identify a new family of dsDNA virus in the genome data of their hosts. Among the 15 different new species of virus, at least four viruses in four different host fish exhibit characteristic features of EVEs such as non-sense mutations and the accumulation of transposable elements. Moreover, more than half of the predicted genes in a full-length salmon virus have no detectable similarity to known genes. The second investigative theme of this thesis explored the idea that EVEs can sometimes be the result of adaptive gene transfer, and that this is the conceptual equivalent of viral genes that are host-derived. Considering a subset of rare, functional EVEs that act as antiviral genes, I developed an evolutionary framework to understand such gene transfers as a biological strategy within the evolutionary arms race between viruses and hosts. In viruses, I conducted a systematic genome-wide evolutionary analysis of host-derived genes, which revealed a consistent shift towards higher purifying selection when switching from host to virus genome. Finally, this thesis also considered the evolution of gene transfer between viruses, by investigating the intriguing history of a retroviral superantigen gene with similarity to genes in unrelated herpesviruses. Through the evolutionary reconstruction of a large group of mammalian EVEs, I was able to prove the independent horizontal gene transfer of superantigens from retroviruses to herpesviruses from two different donor lineages. This shows that viral gene capture can occur in an evolutionary manner, and demonstrates the utility of paleovirological methods in understanding the evolution of pathogenicity. Altogether, this thesis highlights the results of synthesizing methods from paleovirology with metagenomic techniques, demonstrating that genomic databases are themselves rich with novel viral biodiversity that hold key insights for virology, genomics and evolution.

Declaration of Author Contributions

Amr Aswad (AA)
Aris Katzourakis (AK)
Thanos Mourikis (TM)

Chapter one, two, four and five: AA and AK conceived the idea. AA wrote the paper, conducted the analysis and performed the experiments. AK supervised the project.
Chapter one: AA and AK conceived the idea. AA wrote the paper. AK supervised the project.
Appendix one: AA and AK conceived the idea. AA and AK wrote the paper. AK supervised the project.
Appendix two: AA, TM and AK conceived the idea. AA, TM and AK wrote the paper. AK supervised the project.

Citations for Chapters in Print

Introduction: Aswad, Amr, Katzourakis A (2016) Paleovirology: The Study of Endogenous Viral Elements. In: Scott C. Weaver, Mark Denison, Marilyn Roossinck MV, editor. *Virus Evolution: Current Research and Future Directions.* Norfolk: Caister Academic Press.
Chapter one: Aswad A, Katzourakis A (2014) The first endogenous herpesvirus, identified in the tarsier genome, and novel sequences from primate rhadinoviruses and lymphocryptoviruses. *PLoS Genetics* 10: e1004332.
Chapter three: Aswad A, Katzourakis A (2012) Paleovirology and virally derived immunity. *Trends in Ecology and Evolution* 27: 627–636.
Chapter five: Aswad A, Katzourakis A (2015) Convergent capture of retroviral superantigens by mammalian herpesviruses. *Nature Communications* 6: 8299.
Appendix one: Katzourakis A, Aswad A (2014) The origins of giant viruses, virophages and their relatives in host genomes. *BMC Biology* 12: 51.
Appendix two: Mourikis T, Aswad A, Katzourakis A (In press) Paleovirology: The Study of Endogenous Viral Elements In: Elsevier Encyclopedia of Evolution. Academic press.

INTRODUCTION 9

BACKGROUND 9
EVE BIOLOGY 10
IDENTIFYING & AUTHENTICATING ENDOGENOUS VIRUSES 11
STEP 1: VIRUS-HOST SEQUENCE COMPARISON 11
STEP 2: DISCRIMINATING BETWEEN TRUE EVES AND FALSE POSITIVES AND DETERMINING THE FIXATION IN THE POPULATION 12
STEP 3: PHYLOGENETIC RECONSTRUCTION AND EVOLUTIONARY ANALYSIS 14
EVES FROM VIRUSES OF ALL KNOWN GENOME TYPES 15
GROUP I: DOUBLE-STRANDED DNA VIRUSES 15
GROUP II: SINGLE-STRANDED DNA VIRUSES 17
GROUP III: DOUBLE-STRANDED RNA VIRUSES 17
GROUP IV: POSITIVE SINGLE-STRANDED RNA VIRUSES 18
GROUP V: NEGATIVE SINGLE-STRANDED RNA VIRUSES 18
GROUP VI: RETROTRANSCRIBING RNA VIRUSES 19
GROUP VII: RETRO-TRANSCRIBING DOUBLE-STRANDED DNA VIRUSES 20
THESIS OVERVIEW 22
REFERENCES 23

CHAPTER ONE: THE FIRST ENDOGENOUS HERPESVIRUS, IDENTIFIED IN THE TARSIER GENOME, AND NOVEL SEQUENCES FROM PRIMATE RHADINOVIRUSES AND LYMPHOCRYPTOVIRUSES 27

ABSTRACT 27
AUTHOR SUMMARY 27
INTRODUCTION 28
RESULTS 30
A HERPESVIRUS ENDOGENOUS VIRAL ELEMENT IN TARSIUS SYRICHTA 32
EXOGENOUS VIRAL GENOMES IN PRIMATE GENOME DATA 38
PPANHVL GENOME CHARACTERISATION 42
DMADHVL GENOME CHARACTERIZATION 45
DISCUSSION 50
MATERIALS AND METHODS 53
SEQUENCE COLLATION 54
DATABASE MINING 54
CONTIG VISUALISATION 55
PCR AMPLIFICATION 55
ALIGNMENT AND PHYLOGENETIC RECONSTRUCTION 57
REFERENCES 57

CHAPTER TWO: UNUSUAL LARGE DSDNA VIRUSES DISCOVERED AMONG GENOME DATA OF 15 DIFFERENT FISH 61

ABSTRACT 61
AUTHOR SUMMARY 61
INTRODUCTION 62
RESULTS 64
DATABASE MINING AND PHYLOGENETIC RECONSTRUCTION 64
COMPARATIVE ANALYSIS AND SEQUENCE CHARACTERIZATION 66
DISCUSSION 69
MATERIALS & METHODS 73
GENOME DATABASE MINING 73
ALIGNMENT & PHYLOGENETICS 74
CONTIG ANNOTATION, SYNTENY MAPPING & VISUALIZATION 74

ACKNOWLEDGMENTS 75
REFERENCES 75

CHAPTER THREE: PALEOVIROLOGY AND VIRALLY DERIVED IMMUNITY 77

ABSTRACT 77
ENDOGENOUS VIRAL ELEMENTS ARE RELICS THAT INFORM NATURAL HISTORY 78
THE PERSISTENT AND INFLUENTIAL EVOLUTIONARY CONSEQUENCES OF EVES 78
MOLECULAR STRATEGIES OF EDIS CAN BE CATEGORISED 79
RESISTING VIRAL INFECTION BY BLOCKING CELL ENTRY 80
ENDOGENOUS GAG GENES DISRUPT VIRAL REPLICATION AFTER CELL ENTRY 80
ANTICIPATORY STRATEGIES OFFERED BY EDIS 82
EDI EVOLUTION: INTERPRETING SELECTION PATTERNS BY DIRECT AND INDIRECT PALEOVIROLOGY 85
THOUGHTS ON THE FATE OF EDIS: ARE THEY PREDESTINED FOR FAILURE? 86
EDIS ARE AN EXTREME EXAMPLE OF VIRUS-HOST EVOLUTIONARY ENTANGLEMENT 87
THE BIG PICTURE: EDI-LIKE EVOLUTIONARY MECHANISMS ARE UBIQUITOUS IN NATURE 89
FINDING EDIS: INFORMED BIOINFORMATICS STRATEGIES AND PERSPECTIVES FROM ECOLOGY 90
FINDING EDIS WILL ELUCIDATE THE NATURE AND VALIDITY OF THE PHENOMENA 91
REFERENCES 92

CHAPTER FOUR: HOST GENES EVOLVE UNDER STRONGER PURIFYING SELECTION IN VIRAL GENOMES 96

ABSTRACT 96
INTRODUCTION 97
RESULTS 99
MAJOR HISTOCOMPATIBILITY COMPLEX, INTERLEUKIN 6 AND INTERLEUKIN 17 99
DIHYDROFOLATE REDUCTASE, CD200 IMMUNOGLOBULIN & CHEMOKINE C-C MOTIF LIGAND 3 (CCL3) 102
CD59 GLYCOPROTEIN & CORE 2 B-1,6-N-ACETYLGLUCOSAMINYLTRANSFERASE-MUCIN (C2GNT-M) 103
PHOSPHORIBOSYLFORMYLGLYCINAMIDINE SYNTHASE & COMPLEMENT CONTROL PROTEIN 104
OTHER GENES 108
POST-CAPTURE EVOLUTIONARY CHANGES 112
Analysis of Selection 112
SEQUENCE COMPOSITION & STRUCTURAL MODELING 116
DISCUSSION 117
METHODS 122
SEQUENCE COLLATION, ORTHOLOGY DETECTION & ALIGNMENT 122
TAXON CHOICE & PHYLOGENETIC RECONSTRUCTION 123
SELECTION ANALYSIS 124
SEQUENCE COMPOSITION ANALYSIS, STRUCTURAL MODELING & CO-PHYLOGENY ANALYSIS 125
REFERENCES 126

CHAPTER FIVE: CONVERGENT CAPTURE OF RETROVIRAL SUPERANTIGENS BY MAMMALIAN HERPESVIRUSES 129

ABSTRACT 129
INTRODUCTION 130
RESULTS 132
CHARACTERIZATION OF THE HERPESVIRUS SAG LOCI 132
SURVEY AND ALIGNMENT CONSTRUCTION OF SAG-CONTAINING ERVS 132
RECONSTRUCTION OF THE EVOLUTIONARY HISTORY OF SAG 133

RECONSTRUCTING THE SAG EVOLUTIONARY TIMELINE 136
SELECTION ANALYSIS OF RETROVIRAL AND HERPESVIRAL SAGS 139
DISCUSSION **141**
METHODS **146**
SEQUENCE COLLECTION & PHYLOGENETIC RECONSTRUCTION. 146
LTR DATING & BEAST ANALYSIS 147
SELECTION ANALYSIS 147
REFERENCES **148**

DISCUSSION: STATE OF THE ART AND FUTURE DIRECTIONS **150**

CHAPTER ONE AND TWO: THE DISCOVERY OF HERPESVIRUS EVES REVEAL THE INTERDISCIPLINARY POTENTIAL OF PALEOVIROLOGY **150**
CHAPTER THREE, FOUR AND FIVE: EVES AS A SPECIAL CASE OF HORIZONTAL GENE TRANSFER BETWEEN DIFFERENT VIRUSES AND HOSTS **152**
THE CAUSES AND CONSEQUENCES OF EVES **154**
CURRENT AND FORESEEABLE CHALLENGES IN PALEOVIROLOGY **156**
CONCLUSION **158**
REFERENCES **159**

APPENDIX ONE: THE ORIGINS OF GIANT VIRUSES, VIROPHAGES AND THEIR RELATIVES IN HOST GENOMES **162**

ABSTRACT **162**
THE DISCOVERY OF GIANT VIRUSES **163**
EVOLUTIONARY ORIGINS OF VIRAL GIANTRY **163**
GIANT VIRUSES ARE SUSCEPTIBLE TO VIRAL INFECTION BY RELATIVES OF EUKARYOTIC DNA TRANSPOSONS **165**
A QUESTION OF GENE FLOW AND ITS EVOLUTIONARY CONSEQUENCES **166**
THE ENTANGLEMENT OF GIANT VIRUSES, VIROPHAGES AND THEIR HOSTS **167**
REFERENCES **167**

APPENDIX TWO: THE ENDOGENISATION OF RETROVIRUSES AND THEIR LASTING INFLUENCE ON HOST EVOLUTION **169**

ABSTRACT **169**
INTRODUCTION **170**
CLASSIFICATION AND IDENTIFICATION OF ERVS **172**
FROM EXOGENOUS TO ENDOGENOUS: HOW AN ERV IS FORMED **172**
GERMLINE INTEGRATION AND FIXATION 172
INTRAGENOMIC PROLIFERATION 173
THE INFLUENCE OF ERVS ON HOSTS **173**
THE EFFECT OF ERVS ON GENOME ARCHITECTURE 173
THE EFFECT OF ERVS ON GENE EXPRESSION 174
ERVS, RECOMBINATION AND THE EMERGENCE OF NEW VIRAL TYPES 175
OCCASIONAL BENEFICIAL EFFECTS OF ERVS ON HOSTS 176
ERVS AS TOOLS IN THE STUDY OF THE HOST-VIRUS RELATIONSHIP **177**
ERVS IN THE STUDY OF VIRUSES AND VIRUS-HOST INTERACTION AT EVOLUTIONARY TIMESCALES **178**
REFERENCES **180**

Introduction

Background

Usage of the term paleovirology to denote the study of ancient viruses has risen in the last decade, reflecting the establishment of a distinct specialty of evolutionary virology. The precise definition continues to expand since it has emerged from infancy, but predominantly involves the analysis of viral remnants in host genomes. Under the general term endogenous viral elements (EVEs), these 'viral fossils' are the consequence of a germline infection that results in the heritable genome integration of a virus (or part thereof) [1]. Alternatively, ancient viruses can be studied by examining the lasting effect they have had on the evolution of antiviral genes, a method described as indirect paleovirology in contrast to the direct approach of investigating EVEs [2].

The first EVEs were endogenous retroviruses identified in chickens and mice in the late 60's and early 70's, even before retroviruses were formally named [3]. Endogenous retroviruses (ERVs) are by far the most abundant group of EVEs due to an obligatory genome integration step in their replication cycle. Retroviruses posses the reverse transcriptase enzyme that is responsible for converting their RNA genomes into cDNA, which is inserted into the host genome by means of an integrase. This is in contrast to other EVEs, for which a route to endogenisation is not obvious; many viruses do not ordinarily integrate into the host genome or even enter the nucleus during replication. Moreover, other RNA viruses do not have enzymes that can generate cDNA or integrate the viral genome, and most DNA viruses lack an integration mechanism. Nonetheless, it was recently revealed that there are endogenous representatives of viruses from all seven main genome types and replication strategies [1]. This finding was a significant milestone in the emergence of paleovirology, and was followed by a series of comprehensive reviews in quick succession that synthesized research from different fields under a new paleovirological umbrella. This included the decades of established ERV research, advances in genomics and virology, discoveries of non-retroviral EVEs and relevant studies of the evolution of virus-host interactions [2,4–6].

This chapter will offer a synthesis of paleovirology research, with a focus on the most recent non-retroviral EVE work. We will describe a set of techniques that encompasses the latest methods and approaches in paleovirology, which are intimately linked to the nature of the data and the interpretation of results. Viruses are classified according to the guidelines of the International Committee on Taxonomy of Viruses (ICTV), which has a

hierarchal structure similar to that of zoological taxonomy. Additionally, viruses can be categorized according to their method of mRNA synthesis under the Baltimore Classification system, which recognizes seven different viral groups [7].

EVE biology

The integration of ancient viruses that became EVEs cannot be observed and there are no clear sources of evidence for how it could have occurred [6]. We can, however, speculate on the plausible mechanisms of integration by drawing form from our knowledge of virology, cell biology and genomics. Understanding how the germline integration of a virus becomes an EVE requires us to consider consequences to the progeny within the context of host population genetics. Despite the remarkably diverse replication strategies employed by viruses, they share a common fundamental life cycle: a viral particle or virion infects a host cell in order to parasitize the molecular machinery needed for reproduction. Ordinarily, a virus is specific to a host species (or closely related species), and is usually adapted to target a particular cell type; a phenomenon known as host/cellular tropism. Interestingly, EVEs that have been described are not necessarily related to viruses with germline-specific tropism. Moreover, it is not clear where a germline invasion would be most likely to occur; possible targets include either of the two gametes or their precursor cells, the zygote or an early-stage embryo. Once inside the cell the number of steps preceding integration for viruses will vary depending on their genome type. While dsDNA viruses are already in a state capable of integrating into a host genome, single-stranded DNA viruses first require the synthesis of a complementary strand. For RNA viruses, the only means of producing complementary dsDNA genomes is through the catalytic action of reverse transcriptase.

Once it becomes endogenous, an EVE will conceivably reach fixation in the population rapidly if it offers a beneficial function to the host. Several such cases of co-opted EVEs have been documented in the literature, and they vary widely in terms of the how they benefit the host, as well as the mechanism in which this is achieved. In the case of ERVs, the genes of most loci are inactive relics that cannot encode proteins and many are in fact removed through recombination of their long terminal repeats (LTRs) that leaves behind ‘solo-LTRs’. However, both full-length ERVs and solo LTRs can affect transcriptional regulation in the genome by introducing promoters and alternative splice sites. While this can be detrimental to the host, it has also shown to be beneficial in some cases. For example, an LTR is responsible for the beneficial elevation of transcription of the mouse cytidine deaminase APOBEC3 [8]. Another mechanism by which EVEs can benefit the

host is through coopted genes, such as the retrovirus-derived *Syncytin* gene in mammals. Multiple *Syncytins* have been identified in mammals, and they are responsible for the cell-to-cell fusion required in placental formation [9]. EVE-derived immunity genes (EDIs) have also been described in a number of mammals, where an EVE has been co-opted to function against the exogenous viral counterpart that they originate from [5].

Identifying & Authenticating Endogenous Viruses

Identifying EVEs involves a combination of sequence similarity searches and evolutionary analyses. A region of a host genome that is similar to a known viral sequence is likely to be an EVE, except in the case of viral genes that originate from host-gene capture. In order to identify novel EVEs, sequence data gathered from extant exogenous viruses are compared against a host genome using a sequence similarity-searching algorithm such as BLAST [10].

Step 1: Virus-host sequence comparison

Most EVEs are inactivated relics that are no longer subject to the sequence-preserving effect of purifying selection. This means that an EVE sequence is likely to only share low levels of similarity to a query sequence from an extant virus, particularly at the nucleotide level. However, due to the redundancy of the genetic code, the protein sequence takes longer to diverge, and so amino acid comparison is a more sensitive means of identifying EVEs (Figure 1a). The strategy for choosing a set of query sequences depends on the EVE being sought, the candidate host genome(s), and the amount of available processing power. The simplest and fastest method would involve using a single viral protein sequence to query a six-frame translation of a single host genome using tBLASTn (Figure 1a). A more comprehensive search of the same genome for EVEs related to the chosen virus type would use protein sequences from multiple virus species. The best strategy here would be to maximize diversity by choosing representatives from across the (extant) viral tree, or better yet, using all available proteins.

Rather than searching a single genome, multiple hosts can be targeted in a single search, as can multiple sets of viral queries. With unlimited computational resources, one can conceivably search every available host genome using every known viral protein. However, there will be an upper limit to what can be achieved by BLAST and similar programs, since sequence divergence due to genetic drift will eventually render the EVE unrecognizable [11]. This sets a limit on the age of EVEs that can be found, with the exception of those that are co-opted by the host for at least some time after

endogenisation [5]. In such cases, the functional sequence in an EVE is maintained by purifying selection and therefore 'preserved' in the genomic fossil record for a longer time. The alternative to using viral proteins to directly compare against host genomes is to use consensus sequences or sequence profiles that capture a lot of the viral diversity in a single query (Figure 1a). For example, programs such as HMMER [12] use a hidden markov model approach to compare individual sequences to sequence-profiles that capture both the sequence itself and also account for variability at each site. This contains information about how a set of sequences can evolve, such as the position of common insertions or deletions or the probability of an amino acid given the identity of the adjacent site. This allows us to detect faint patterns of similarity to viral queries and search for EVEs beyond the hypothetical age/divergence limit of BLAST. The efficacy of these tools has been demonstrated in viral genome annotation, where distant homologs of apparently unique genes were identified [11].

Step 2: Discriminating between true EVEs and false positives and determining the fixation in the population

A virus is considered an EVE once inherited vertically by the host from parent to a viable offspring. By this definition however, many more EVEs will have existed briefly, having never reached fixation in the population to be maintained and observed today. A similar set of analyses can be used to investigate the authenticity of an EVE as well as whether it is fixed in the host population. The most straightforward way to achieve both of these objectives is to identify homologous EVE loci. An orthologous EVE in another species or individual indicates that the loci are derived from a common ancestor, which indicates that they were vertically inherited. For EVEs that have not been co-opted, this also allows us to estimate the date of the integration by measuring the genetic distance between loci and dividing half of this by the host neutral rate of evolution. The probability that it is fixed is higher for older EVEs, if the individual being sampled is not part of a geographically isolated sub-population. This technique offers a conservative minimum bound for the age of an EVE, which can be compared to the expected time it would need to reach fixation according to the Hardy-Weinberg principle. Alternatively, the same dating technique can be applied to date the duplication event of paralogous EVEs in the same individual, which is a conservative minimum age of the EVE. This approach was used to date the integration of a lentivirus in the genome of the European rabbit at over 10 million years [13].

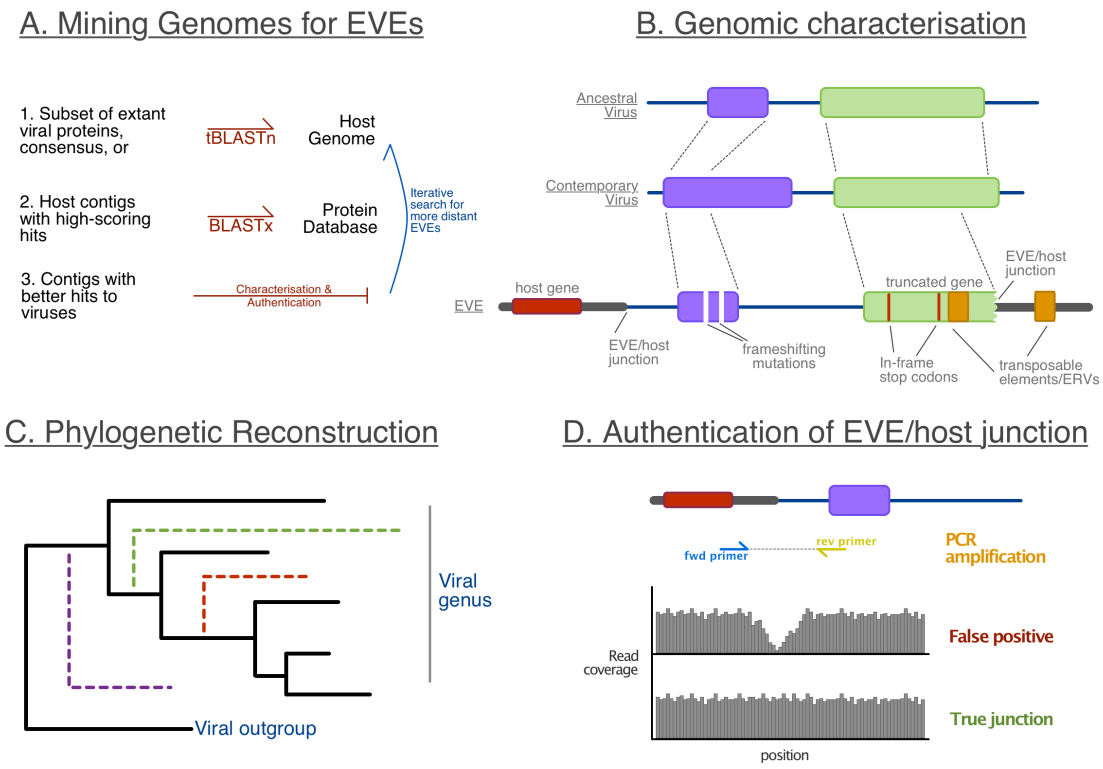


Figure 1 Identifying, characterising and analysing endogenous viral elements
Finding and authenticating EVEs involves a multifaceted bioinformatic strategy that is informed by the nature of the data. ***Panel A*:** A strategic subset of extant viral proteins is used to query a genomic database of one or more hosts using the tBLASTn algorithm that searches a 6-frame translation of the target genome. From the resulting list of hits, a reciprocal BLASTx of the nucleotide sequences obtained from the genome(s) in question is performed against a general protein database to identify and roughly annotate each contig. Contigs from this search that contain hits to viruses, which out-score those to anything else, are shortlisted as putative EVEs for further analysis. ***Panel B*:** The extant viral query used to search host genomes for an EVE may not be identical to the ancestral virus that became the EVE. For instance in this hypothetical example, one of the genes was significantly shorter in the ancestral virus compared to the contemporary relative. For any contig suspected to contain an EVE, there are a number of genomic features we would expect to observe. Neutral genetic drift will often result in a number of in-frame stop codons and frame-shifting mutations. Recombination or errors at the point of integration could have caused the virus to become truncated, and the prolonged existence in a host genome will often result in the insertion of transposable elements such as SINES or LINEs into the EVE. In the most ideal scenario, the EVE/host junction will be clearly identifiable, with unambiguous host and viral genes on either side. ***Panel C*:** Phylogenetic reconstruction of the EVE is an essential part of an evolutionary analysis into the natural history of the virus in question. Various possible placements of an EVE in a tree of viral relatives are shown as dashed lines. The resulting topology that would be most easily interpretable is shown in red, where an EVE resolves deep within an existing viral group. In some cases however, long branch attraction can lead to the potentially incorrect placement of an EVE (shown here in green). The purple line shows another common result in paleovirology, where an EVE resolves outside known viral groups; this could either mean that it belongs to a previously unrecognised group, or the results could be used as justification for the expansion of an existing group to include the new diversity introduced by the EVE. ***Panel D:*** The authentication of the EVE/host junction can be achieved by confirmatory PCR and sequencing of a fragment amplified using primers designed on either side of the junction. Alternatively, or in addition, we can consult the raw sequence data of the locus, which will reveal whether or not the putative EVE is genuinely inserted into the genome or if the junction is just an assembly artifact.

These methods depend on inferring either authenticity or fixation from the age of a putative EVE. If it is not possible to directly obtain an estimate of the integration date, we can also look for indirect evidence for the age of an EVE. A characteristic feature of most EVEs is the degradation of their sequence, either in the form of gene loss or nonsense mutations. This is because the population frequency of an EVE that is either neutral or

only slightly detrimental to the host will be governed by genetic drift [5]. This can be used as evidence to distinguish an EVE from an exogenous virus, particularly if it is very old and has had time to accumulate neutral mutations. However, evidence of degradation alone cannot exclusively prove that a viral sequence is an EVE, because the low fidelity of most viral polymerases means that a proportion of an exogenous viral population will exhibit similar patterns. Furthermore, host restriction factors are also known to hypermutate exogenous viruses [14], and so the coding capacity of a putative EVE can only act to supplement other sources of evidence. For example, an important step in the authentication process is the identification and characterization of the EVE/host junction, where one can discriminate the integration site by recognizing the difference between a host region and an unambiguously viral sequence (Figure 1b). This can be verified by the targeted amplification and sequencing of the junction, and additional assurance can be sought by interrogating the genome assembly for evidence of high coverage at the junction (Figure 1d). This is because viruses often have repetitive sequences that can be falsely connected to a repetitive region of the host genome by automatic contig assembly software. For example, an endogenous herpesvirus in the tarsier genome is integrated into the repetitive telomeric region, and a combination of high assembly coverage and targeted sequencing was used to confirm the authenticity of the integration site [15].

Step 3: Phylogenetic reconstruction and evolutionary analysis

In addition to characterizing the genomic locus, the authenticity of an EVE can be achieved as part of an evolutionary analysis that involves phylogenetic reconstruction. Although an EVE with remote similarity can still be characterized at the genomic level, phylogenetic analysis is a crucial step in the confident authentication of an EVE, as well as the key to understanding its natural history. Broadly speaking, EVEs that are undetectable by BLAST are also unlikely to yield robust phylogenetic results, since it will be difficult to align them with confidence to other viral sequences. In a phylogeny composed of viral sequences related to the query used to find the EVE as well as a number of more distantly related viruses, a true EVE should exhibit more sequence similarity to some viruses than to others (Figure 1c). Although not always the case (particularly for young insertions), the accumulation of mutations in an EVE as a result of neutral genetic drift can cause an EVE to have a longer terminal phylogenetic tree branches compared to viral sequences. This can sometimes return misleading results, because such long branches could resolve incorrectly in the tree (using some methods long branches tend to converge), making it difficult to precisely pinpoint the relationship of the EVE to extant viruses. In some cases, an EVE will resolve distantly to known viral

groups, rather than within them, which can either mean that the phylogenetic range of the group needs to be expanded, or can be interpreted as evidence for a novel group of either extinct or unsampled viruses (Figure 1c).

EVEs from viruses of all known genome types

A number of EVEs are known that are related to each of the seven main groups of viruses as classified by the Baltimore system. Examples of EVEs from each of the groups have revealed aspects of viral evolutionary history, as has their host distributions and abundances. For example, the fact that ERVs are the most abundant type of EVE reveals that there might be a causal relationship between a viral replication strategy and its propensity to endogenise. For each of the seven groups, we will highlight examples of EVEs as a means of discussing the range of insights that have been gained in paleovirology, as well as methodological challenges that are often unique to a particular group.

Group I: Double-stranded DNA viruses

The largest viruses we know of possess dsDNA genomes, probably due to the biochemical stability of dsDNA compared to other nucleic acids. Thirty-three families of dsDNA viruses are currently recognized by the ICTV [16], but only a small number of EVEs have been recognized. Their size is probably a barrier to endogenisation in that an insertion is more likely to be detrimental/fatal to the host the lager it is, because of the simultaneous introduction of a number of genes, regulatory elements and potential recombination sites. The scarcity of EVEs related to dsDNA viruses is also compounded by the bioinformatic challenges associated with searching for them. For example, while the number of host genomes in public databases continues to increase, the majority of them are in draft assembly states, meaning that large EVEs will be split over multiple short contigs. Moreover, dsDNA viruses are known to frequently co-opt host genes, which further complicates their identification of EVEs.

The *Polydnaviridae* (PDV) are a family of dsDNA virus EVEs that are associated with wasps in the *Ichneumonoidea* or *braconidae*, and are accordingly classified into the genus *Bracovirus* or *Ichnovirus,* respectively [17,18]. The endogenisation of PDVs into wasp genomes has occurred multiple times in the evolutionary history of these wasps, leading to functional, vertically inherited EVEs that produce and package viral genes. These wasps parasitize lepidopteran larvae by injecting their eggs along with infectious PDV particles that ‘protect’ them from the caterpillar immune system [18]. These PDVs were

probably once viruses that became domesticated in the wasp genome, but they do not replicate independently. The PDV 'macrolocus' is distributed across the wasp genome, and is differentially packaged into particles that do not individually constitute a complete polydnavirus genome. Interestingly, the PDV has undergone multiple duplication events that have resulted in the emergence of gene families, and the insertion of wasp genes into the PDV locus has been shown to have occurred over 86 million years ago [18]. These characteristics highlight the ancient host-specific adaptations undergone by PDVs that characterize them as components of the genome, driven by selective pressure on the host rather than the virus.

A well-studied group of dsDNA viruses, in the order *Herpesvirales*, has a wide host range with particular host groups infected by each of the three constituent families. The *Alloherpesviridae* infect amphibians and fish, while mammals, birds and reptiles are susceptible to members of the *Herpesviridae*. The third family, *Malacoherpesviridae,* is populated by an oyster Herpesvirus [19], but partially described malacoherpesvirus-like sequences were identified amidst the genome data of the lancelet (*Branchiostoma lanceolatum)*. This could be evidence of a herpesvirus EVE in amphioxous, or else represents an exogenous virus that contaminated the sequencing sample [20]. Similarly, a herpesvirus related to roseoloviruses was identified in the genome of the Tarsier, a basal primate found in the Philippines, Malaysia, Indonesia and Brunei [15]. In this study, the host-EVE junction was confirmed using experimental data and the sequences exhibited the nonsense mutations that are typical of anciently integrated EVEs. Although less ancient, the closely related *Human Herpesvirus 6* has been shown to integrate into the germline and is inherited in ~1% of people [21].

A fascinating group of dsDNA viruses are the *Megavirales* (pending ICTV ratification) or nucleo-cytoplasmic large DNA viruses (NCLDVs) [22]. Among the megaviruses are the largest known viruses, which can exceed the size of the smallest cells, with genomes up to 2.5 MB in the case of the pandoravirus [23]. Shortly following their discovery, it was found that some giant viruses are themselves susceptible to infection by viruses (known as virophages) that specifically parasitize the replication machinery of the host that was recruited by the giant virus known as a 'viral factory'[24]. Unlike satellite viruses that require co-infection of a cell with a helper virus, virophages negatively affect the production of giant virus particles by ~80% [25]. EVEs known as mavericks, described in the genome of the giant virus' host amoeba, likely originate from the recurring integration of virophages [26,27]. Interestingly, infection of a giant virus-infected amoeba with a virophage results in the partial recovery of the amoeba [25]. This could explain the

repeated endogenisation of virophages in amoeba genomes, which could be co-opting them to defend against infection giant viruses [26,27] (Appendix 1).

Group II: Single-stranded DNA viruses

EVEs related to group II viruses are among the most abundant and widely represented non-retroviral EVEs in animal genomes, particularly those related to the family *Parvoviridae*. Well over a hundred different *Parvovirinae* and *Densovirinae* (the two subfamilies) have been identified in placental and non-placental mammals, fish, birds, a tunicate, crustaceans, insects and a flatworm [1,28–30]. An orthologous pair of parvovirus EVEs in mice and rats suggests that the integration event occurred at least 30 million years ago, offering a minimum age for the virus group for the first time [1]. While nearly all of the vertebrate EVEs exhibit signs of degradation, an intact open reading frame of the NS1 gene was identified in a *Dependovirus*-like EVE in elephants [1], as well as a number of different invertebrate *Parvoviridae*-EVEs [28].

EVEs have also been described from a number of other families of ssDNA viruses. The exogenous viruses in the family *Circoviridae* have a host range limited to vertebrates. In addition to finding EVEs in species previously not known to be circovirus hosts, a diverse range of non-vertebrate genomes has been shown to harbor EVEs with circovirus similarity, including mites, nematodes, gastropods, hydrozoans and giardia [31,32]. Additionally, EVEs related to the plant-infecting *Geminiviridae* have been identified in several *Nicotania* species including many tobacco cultivars (*Nicotiana tabacum*), the black cottonwood (*Populus trichocarpa*) and green algae [32–34].

Group III: Double-stranded RNA viruses

Group three viruses possess linear dsRNA genomes with between 1-4 segments or 10-12 in the case of the *Reoviridae*. The first dsRNA EVE to be identified, in the genome of the disease vector mosquito *Aedes aegypti,* is almost identical to segment 5 of the exogenous Liaoning virus isolated from another *Aedes* species [1]. The presence of a host-flanking sequence and a large deletion excludes the possibility that the element is a contaminant, but the extremely high similarity to a currently circulating virus suggests that the integration occurred very recently. Indeed, it maybe that not all *A. aegyptii* genomes contain the EVE, raising the possibility that this is an unfixed genome invasion.

Since the discovery of endogenous Liaoning virus, a number of studies revealed the widespread presence of dsRNA-related EVEs, particularly those related to the fungi- and plant-infecting *Partitiviridae* and *Totiviridae* [35]. In addition to elements identified in four

plants and four fungi, EVEs were identified in genomes that are not natural partitivirus/totivirus hosts. A partitivirus EVE was found in the genome of a gastropod, and a number of totuvirus-like EVEs we found in a nematode and protozoan [36].

Group IV: Positive single-stranded RNA viruses

EVEs related to RNA viruses with positive-sense, single-stranded genomes have been identified in both animals and plants [1,37–41]. Much of the diversity of group IV viruses is distributed among those that infect plants, but eight different families of mainly human and non-human vertebrate viruses are known. In the case of plants, EVEs with similarity to the plant-specific potyviruses were described in the genome of the common grapevine (*Vitis vinifera*) [39]. In the Western honeybee genome (*Apis melifera*), an EVE with similarity to the *Dicistoviridae* was identified and named the Israeli acute paralysis virus (IAPV) [42]. Interestingly, the IAPV is one of the extremely rare demonstrable examples of a viral endogenisation event that has yet to reach fixation in the population, having been identified in only 30% of bee populations [42]. In addition to the ~1% of humans known to carry a heritable integration of human herpesvirus 6 [21,43], we also know of an ERV present in some koala populations [44], which invaded the genome less than 50,000 years ago and continues to proliferate [45]. Similarly, retroviruses that are still capable of intragenomic proliferation can give rise to new EVE paralogs that are polymorphic in the host population, such as the human ERV family HERV-K (HML2) [46].

The *Flaviviridae* includes a variety of human disease-causing viruses that are transmitted by arthropod vectors, and in the case of two different mosquito species (*A. aegypti* and *A. albopictus*) EVEs were identified that are phylogenetically similar to the *Aedes* flaviviruses cell fusing agent and Kamiti River virus [1,40,41]. It was also found that in the medaka genome, there is a putative flavivirus EVE with similarity to *Tamana Bat Virus*, although the authors could not rule out the possibility of contamination [37].

Group V: Negative single-stranded RNA viruses

Among the group V viruses is the order *Mononegavirales*, which is composed of five families that mainly infect humans and other vertebrates, all but one of which are represented in the viral fossil record. *Bornaviridae*-like EVEs are especially abundant and are found in numerous and diverse hosts, including mammals, marsupials, reptiles, insects, fish and spiders [1,37,47,48]. Many of these EVEs are only represented by the bornavirus nucleoprotein gene, and are referred to as endogenous borna-like N (EBLN). While there have been multiple independent insertions of bornavirus EVEs, orthologous

loci in different species allow us to obtain a minimum bound of their integration date from estimates of their hosts' speciation dates. For example, a set of orthologous EBLNs in afrotherians can be dated to at least 65 million years old based on the divergence date of their hosts [48]. The identification of orthologous EBLNs in strepsirrhine primates indicated that a particular EBLN integration occurred over 54 million years ago and was selectively maintained for some time in all the extant host lineages, with coding capacity maintained in baboon and all great apes except orangutans [1,48]. Although a specific function has yet to be ascribed to any of these potentially functional ORFs, a recent study in the thirteen-lined ground squirrel demonstrated that EBLNs can act as a restriction factor of Bornavirus replication [49]. This pioneering work has paved a route for translational research into *Borna Disease Virus*, which is a known zoonotic agent that can infect humans.

EVEs related to *Filoviridae* have also been identified in a range of mammals [1,50,51], which is surprising given that their exogenous counterparts are limited to primates and bats in central Africa [6]. The phylogeographic pattern of exogenous and endogenous filoviruses has been considered as evidence of large gaps in the sampling of exogenous filoviruses, because known exogenous viruses are more closely related to marsupials than to endogenous filoviruses of placental mammals [1,51]. *Rhabdovirus*-like EVEs were initially identified in the genomes of the black-legged tick (*I. scapularis*) and mosquitoes (*A. aegypti*) [1], and there is some evidence for rhabdovirus EVEs in other arthropods [52]. In addition to the *Mononegavirales,* there are a number of EVEs related to ssRNA negative strand viruses that are not assigned to a viral order. A short sequence in the black-legged tick genome resembling viruses from the genus *Quarjavirus* are the only orthomyxovirus-like EVEs described to date [1]. This EVE is particularly interesting since the influenza viruses are also contained within the *Orthomyxoviridae*, and insights into their long-term evolution will be informative to intervention strategies.

Group VI: Retrotranscribing RNA viruses

Retroviruses represent the vast majority of the genomic fossil record (Appendix 2). For example, while ~3% of the Human genome is composed of protein-coding genes, it is estimated that ERVs represent up to 8% [53]. Part of this abundance can be explained by their replication cycle that involves a genome integration step. Additionally however, retroviruses can continue to proliferate in the genome after the initial endogenisation event. The consequent evolutionary dynamics of ERVs in the genome are therefore distinct from those resulting from the endogenisation of other viral groups.

The repertoire of ERVs in mammalian genomes is structured into 'families' that originate from a single insertion. An ERV can proliferate by re-infecting the germline cell, but also intra-genomically if it is incapable of producing virions because of the loss/damage of some of their genes [54]. In the latter case, this means that we can track the natural history of a retroviral lineage that has only ever existed within the confines of a genome. Interestingly, the number of ERVs in the genomes of different mammals varies widely and the mechanisms behind this non-uniformity are not fully understood. An increased capacity to proliferate intragenomically in some species could explain the differences observed in some cases. It has been shown that the loss of the envelope gene (*env*) in the IAP family of murine ERVs has resulted in an increased rate of proliferation despite an inability to re-infect [55].

The fact that the loss of *env* has occurred multiple times, independently, suggests that the process might be adaptive [55]. Evolving towards intragenomic replication might be favorable to an ERV due to the harmful effects of *env* expression to the host cell, as well as the consequent harmful effect of viral infection to the host in general [55]. Similarly, another observation in the abundance and distribution of ERVs could be explained as an adaptive response to the harmful effects of ERVs. In mammals, the number of ERVs in a species correlates negatively with larger body size [56]. This could be explained as an adaptive response by larger hosts that are under a higher selective pressure to control the mutagenic effects of ERVs. This could be due to the increases in risk associated with having a higher number of cells that also undergo a greater number of cell divisions since larger mammals live longer [56].

Group VII: Retro-transcribing double-stranded DNA viruses

The two families of group VII viruses are the *Caulimoviridae* that infect plants and the vertebrate-infecting *Hepadnaviridae*, both of which possess dsDNA genomes that replicate via an RNA intermediate. This step in the replication cycle has also earned them the term pararetroviruses in reference to their reverse transcription, but unlike retroviruses they do not require a genome integration step.

A hepadnavirus EVE was identified in the genome of the black-legged tick [1], despite the fact that hepadnaviruses are not known to infect insects. This EVE could belong to an unknown vertebrate lineage of hepadnavirus that crossed species facilitated by the parasitic nature of ticks. Another set of hepdavirus EVEs also expanded the previously known host range of Hepadnaviruses to include reptiles. Two independent insertions of

hepadnvirus-like EVEs were identified in the genome of speckled rattlesnakes (*Crotalus mitchellii* species complex), one of which is also shared by the king cobra (*Ophiophagus hannah*).

An EVE related to avian hepadnaviruses was identified the genome of the zebrafinch by comparison to the duck hepatitis B virus (DHBV). The EVE is spread over multiple chromosomes but altogether represents almost 80% of the hypothesized ancestral virus [1,57]. The endogenous zebrafinch HBV (eZHBV) is distributed across four different chromosomes, which means that the ancestral insertion could have been fragmented by rearrangements. Moreover, one of the fragments described was shown to have duplicated post-insertion, allowing the estimation of a minimum bound for the date of the integration. By dividing the molecular divergence of the duplicate fragments by an estimate of the host neutral evolutionary rate, we can estimate a minimum date of the duplication between 3.8 and 7.5 million years ago [57].

The date of the hepadnavirus genome invasion was estimated to be between 19 and 40 million years ago by finding an ortholog of eZHBV in the distantly related dark-eyed junco (*Junco hyemalis*), thereby inferring the date from estimates of the birds' speciation date [57]. In addition to the Dark-eyed junco, fragments of endogenous avian hepdnaviruses were identified in the genomes of the Black-throated finch (*Poephila cincta*), Scaly-breasted monia (*Lonchura punctulata*) and Gouldian finch (*Chloebia gouldiae*), representing other orthologs of two distinct eZHBV integrations [57]. While these EVEs were all fragmented, two full-length avian hepadnavirus EVE were identified in the budgerigar (*Melopsittacus undulatus*), one of which also underwent a partial duplication that increased its length by ~30% [58]. Further details into the natural history of these EVEs emerged as the number of identified eZHBV accumulated. The combined strategic sampling for new loci and an analysis of their integration dates revealed that the oldest eZHBV must have endogenised before the divergence of neoavian birds, suggesting that the hepadnaviruses are at least ~74-94 million years old [59].

Four of the seven genera of *Caulimoviridae* have been identified in plant genomes [60]. Multiple EVEs related to *Banana streak virus* (*Badnavirus*) were identified in the genome of banana (*Musa sp.*) [61]. In rice (*Oryza sp.*) a *Tungrovirus* EVE was identified and a *Petuvirus*-like EVE was found in the genome of petunia (*Petunia hybrida*) and the trifoliate orange (*Poncirus trifoliata*), which is closely related to citruses [62,63]. *Cavemovirus*-like

EVEs have also been identified in several species of tobacco (*Nicotiana sp.*), potato (*Solanum sp.*) and tomato (*Lycopersicon sp.*) [64–66].

Thesis Overview

The examples throughout this introduction represent highlights of recent work in paleovirology, with an emphasis on the viral diversity that has been uncovered in animal and plant genomes. Work on this doctoral thesis began very shortly after it was established that apart from retroviruses, virus of all known genome types and replication strategies had been identified in animal genomes [1]. At the time, only the small EVEs had been found, due to the difficulty of mining genome data for large viruses, which have complex genomes that are well over 100Kbps in length. I decided to undertake the challenge of finding out if a large dsDNA virus has the ability to stably integrate in a vertebrate genome, which was an obvious next step in direct paleovirology research. In chapter one, I present the results of the discovery and characterization of the first herpesvirus EVE, found in the genome of the Philippine tarsier [15]. As well as characterization of this EVE and confirmation of the host junction, the chapter details the methodology I developed to tackle the problem of finding large EVEs in draft assembly data. The chapter also includes details of novel herpesvirus sequence data in the genome of two other primates, which illustrates the utility of my approach for viral discovery in general. Indeed this was discussed in a subsequent review and proposed as a formal technique in herpesviral discovery [70]. Chapter two is a test of this idea, where I used principles from paleovirology to identify a new family of large dsDNA viruses that infect fish.

Alongside the discovery that all viral groups have EVE representatives, the study of EVEs as a stand alone disciple began to take shape with the emergence of different approaches and perspectives, as well as avenues of enquiry that developed from ground breaking discoveries. For instance, the distinction between ‘direct’ and ‘indirect’ paleovirology was established to differentiate between the inference of past viral evolution versus the *direct* observation of ‘paleoviruses’ integrated in host genomes [2,67]. Indirect paleovirology allow us to infer the evolution of viruses by investigating the impact they have had on their host. For instance, we can infer the existence and minimum age of unknown ancient retroviruses that drove adaptive changes in host anti-retroviral genes that exhibit signatures of positive selection [2]. In addition to indirect paleovirology, there was renewed interest in the study of potentially beneficial EVEs, partly as result of the discovery of intact bornavirus genes in humans and other species [47] that appeared to

have been evolutionarily preserved [1]. This was made clear in a number of reviews on paleovirology that also discussed the implications of this to the arms-race between viruses and their host [68,69]. Chapter three of this thesis is a contribution to this aspect of paleovirology, where I consider the various examples of EVEs that have been co-opted to serve an antiviral function in the host. By synthesizing concepts and methodological strategies from both direct and indirect paleovirology, I develop a evolutionary framework for studying EVE-derived immune genes (EDIs) and discuss the future directions and implications of what may be revealed [5]. Conversely, chapter four and five are both studies that investigate the impact of gene capture by viruses. In chapter three, I investigate the remarkable evolutionary history of a retroviral gene that has been captured by herpesviruses twice independently. In a sense, this can also be thought of as an EVE, except that the genome is that of another virus rather than the host. Similarly, chapter four is also an investigation of horizontally acquired genes in herpesviruses, but this time we consider the large repertoire of host-derived genes that are repeatedly captured. In this work, we discuss the possible mechanisms involved in the evolution of these genes, and uncover a surprising underlying pattern of selection.

References

1. Katzourakis A, Gifford RRJ (2010) Endogenous viral elements in animal genomes. PLoS Genet 6: e1001191.
2. Patel MR, Emerman M, Malik HS (2011) Paleovirology—ghosts and gifts of viruses past. Curr Opin Virol 1: 304–309. doi:10.1016/j.coviro.2011.06.007.
3. Weiss RA (2006) The discovery of endogenous retroviruses. Retrovirology 3: 67. doi:10.1186/1742-4690-3-67.
4. Feschotte C, Gilbert C (2012) Endogenous viruses: insights into viral evolution and impact on host biology. Nat Rev Genet 13: 283–296. doi:10.1038/nrg3199.
5. Aswad A, Katzourakis A (2012) Paleovirology and virally derived immunity. Trends Ecol Evol 27: 627–636. doi:10.1016/j.tree.2012.07.007.
6. Holmes EC (2011) The Evolution of Endogenous Viral Elements. Cell Host Microbe 10: 368–377. doi:10.1016/j.chom.2011.09.002.
7. Baltimore D (1971) Expression of animal virus genomes. Bacteriol Rev 35: 235–241.
8. Sanville B, Dolan MA, Wollenberg K, Yan Y, Martin C, et al. (2010) Adaptive evolution of Mus Apobec3 includes retroviral insertion and positive selection at two clusters of residues flanking the substrate groove. PLoS Pathog 6: e1000974. doi:10.1371/journal.ppat.1000974.
9. Lavialle C, Cornelis G, Dupressoir A, Esnault C, Heidmann O, et al. (2013) Paleovirology of “syncytins”, retroviral env genes exapted for a role in placentation. Philos Trans R Soc Lond B Biol Sci 368: 20120507. doi:10.1098/rstb.2012.0507.
10. Camacho C, Coulouris G, Avagyan V, Ma N, Papadopoulos J, et al. (2009) BLAST+: architecture and applications. BMC Bioinformatics 10: 421. doi:10.1186/1471-2105-10-421.
11. Kuchibhatla DB, Sherman WA, Chung BYW, Cook S, Schneider G, et al. (2013) Powerful Sequence Similarity Search Methods and In-Depth Manual Analyses Can Identify Remote Homologs in Many Apparently “Orphan” Viral Proteins. J Virol 88: 10–20. doi:10.1128/JVI.02595-13.
12. Finn RD, Clements J, Eddy SR (2011) HMMER web server: interactive sequence similarity searching. Nucleic Acids Res 39: W29–W37. doi:10.1093/nar/gkr367.
13. Katzourakis A, Tristem M, Pybus OG, Gifford RJ (2007) Discovery and analysis of the first endogenous lentivirus. PNAS 104: 6261–6265. doi:10.1073/pnas.0700471104.
14. Zhang H, Yang B, Pomerantz RJ, Zhang C, Arunachalam SC, et al. (2003) The cytidine deaminase CEM15 induces hypermutation in newly synthesized HIV-1 DNA. Nature 424: 94–98. doi:10.1038/nature01707.
15. 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. doi:10.1371/journal.pgen.1004332.
16. King AMQ, Adams MJ, Lefkowitz EJ, editors (2012) Virus Taxonomy: Ninth Report of the International Committee on Taxonomy of Viruses. San Diego: Elsevier. 1327 p.
17. Espagne E, Dupuy C, Huguet E, Cattolico L, Provost B, et al. (2004) Genome sequence of a polydnavirus: insights into symbiotic virus evolution. Science 306: 286–289. doi:10.1126/science.1103066.
18. Herniou EA, Huguet E, Thézé J, Bézier A, Periquet G, et al. (2013) When parasitic wasps hijacked viruses: genomic and functional evolution of polydnaviruses. Philos Trans R Soc Lond B Biol Sci 368: 20130051. doi:10.1098/rstb.2013.0051.
19. Davison AJ, Eberle R, Ehlers B, Hayward GS, McGeoch DJ, et al. (2009) The order Herpesvirales. Arch Virol 154: 171–177. doi:10.1007/s00705-008-0278-4.
20. Savin KW, Cocks BG, Wong F, Sawbridge T, Cogan N, et al. (2010) A neurotropic herpesvirus infecting the gastropod, abalone, shares ancestry with oyster herpesvirus and a herpesvirus associated with the amphioxus genome. Virol J 7: 308. doi:10.1186/1743-422X-7-308.
21. Arbuckle JH, Medveczky MM, Luka J, Hadley SH, Luegmayr A, et al. (2010) The latent human herpesvirus-6A genome specifically integrates in telomeres of human chromosomes in vivo and in vitro. PNAS 107: 5563–5568. doi:10.1073/pnas.0913586107.
22. Colson P, De Lamballerie X, Yutin N, Asgari S, Bigot Y, et al. (2013) “Megavirales”, a proposed new order for eukaryotic nucleocytoplasmic large DNA viruses. Arch Virol 158: 2517–2521. doi:10.1007/s00705-013-1768-6.
23. Philippe N, Legendre M, Doutre G, Couté Y (2013) Pandoraviruses: amoeba viruses with genomes up to 2.5 Mb reaching that of parasitic eukaryotes. Science (80-).
24. La Scola B, Desnues C, Pagnier I, Robert C, Barrassi L, et al. (2008) The virophage as a unique parasite of the giant mimivirus. Nature 455: 100–104. doi:10.1038/nature07218.
25. Campos R, Boratto P (2014) Samba virus: a novel mimivirus from a giant rain forest, the Brazilian Amazon. Virol J.
26. Katzourakis A, Aswad A (2014) The origins of giant viruses, virophages and their relatives in host genomes. BMC Biol 12: 51. doi:10.1186/s12915-014-0051-y.
27. Fischer M, Suttle C (2011) A virophage at the origin of large DNA transposons. Science (80-) 332: 231–234.
28. Liu H, Fu Y, Xie J, Cheng J, Ghabrial SA, et al. (2011) Widespread endogenization of densoviruses and parvoviruses in animal and human genomes. J Virol 85: 9863–9876. doi:10.1128/JVI.00828-11.
29. 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. doi:10.1093/gbe/evu163.
30. 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. doi:10.1128/JVI.01789-10.
31. 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. doi:10.1128/JVI.01789-10.
32. Liu H, Fu Y, Li B, Yu X, Xie J, et al. (2011) Widespread horizontal gene transfer from circular single-stranded DNA viruses to eukaryotic genomes. BMC Evol Biol 11: 276. doi:10.1186/1471-2148-11-276.
33. Ashby MK, Warry A, Bejarano ER, Khashoggi A, Burrell M, et al. (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.
34. Bejarano ER, Khashoggi A, Witty M, Lichtenstein C (1996) Integration of multiple repeats of geminiviral DNA into the nuclear genome of tobacco during evolution. PNAS 93: 759–764.
35. Chu H, Jo Y, Cho WK (2014) Evolution of endogenous non-retroviral genes integrated into plant genomes. Curr Plant Biol. doi:10.1016/j.cpb.2014.07.002.
36. Liu H, Fu Y, Jiang D, Li G, Xie J, et al. (2010) Widespread horizontal gene transfer from double-stranded RNA viruses to eukaryotic nuclear genomes. J Virol 84: 11876–11887. doi:10.1128/JVI.00955-10.
37. 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. doi:10.1371/journal.ppat.1001030.
38. Chiba S, Kondo H, Tani A, Saisho D, Sakamoto W, et al. (2011) Widespread endogenization of genome sequences of non-retroviral RNA viruses into plant genomes. PLoS Pathog 7: e1002146. doi:10.1371/journal.ppat.1002146.
39. 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. doi:10.1016/j.virol.2004.11.007.
40. 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. doi:10.1186/1743-422X-6-93.
41. Crochu S, Cook S, Attoui H, Charrel RN, De Chesse R, et al. (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. doi:10.1099/vir.0.79850-0.
42. Maori E, Tanne E, Sela I (2007) Reciprocal sequence exchange between non-retro viruses and hosts leading to the appearance of new host phenotypes. Virology 362: 342–349. doi:10.1016/j.virol.2006.11.038.
43. Arbuckle JH, Pantry SN, Medveczky MM, Prichett J, Loomis KS, et al. (2013) Mapping the telomere integrated genome of human herpesvirus 6A and 6B. Virology 442: 3–11. doi:10.1016/j.virol.2013.03.030.
44. Tarlinton RE, Meers J, Young PR (2006) Retroviral invasion of the koala genome. Nature 442: 79–81. doi:10.1038/nature04841.
45. Ishida Y, Zhao K, Greenwood AD, Roca AL (2014) Proliferation of Endogenous Retroviruses in the Early Stages of a Host Germ Line Invasion. Mol Biol Evol: msu275 – . doi:10.1093/molbev/msu275.
46. Belshaw R, Dawson ALA, Woolven-Allen J, Redding J, Burt A, et al. (2005) Genomewide screening reveals high levels of insertional polymorphism in the human endogenous retrovirus family HERV-K(HML2): implications for present-day activity. J Virol 79: 12507–12514. doi:10.1128/JVI.79.19.12507-12514.2005.
47. Horie M, Honda T, Suzuki Y, Kobayashi Y, Daito T, et al. (2010) Endogenous non-retroviral RNA virus elements in mammalian genomes. Nature 463: 84–87. doi:10.1038/nature08695.
48. 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. doi:10.1098/rstb.2012.0499.
49. Fujino K, Horie M, Honda T, Merriman DK, Tomonaga K (2014) Inhibition of Borna disease virus replication by an endogenous bornavirus-like element in the ground squirrel genome. PNAS 111: 13175–13180. doi:10.1073/pnas.1407046111.
50. Taylor DJ, Dittmar K, Ballinger MJ, Bruenn J (2011) Evolutionary maintenance of filovirus-like genes in bat genomes. BMC Evol Biol 11: 336. doi:10.1186/1471-2148-11-336.
51. Taylor D, Leach R, Bruenn J (2010) Filoviruses are ancient and integrated into mammalian genomes. BMC Evol Biol 10: 193.
52. Fort P, Albertini A, Van-Hua A, Berthomieu A, Roche S, et al. (2011) Fossil Rhabdoviral Sequences Integrated into Arthropod Genomes: Ontogeny, Evolution, and Potential Functionality. Mol Biol Evol 29: 381–390. doi:10.1093/molbev/msr226.
53. 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. doi:10.1128/JVI.74.8.3715-3730.2000.
54. Jern P, Coffin JM (2008) Effects of retroviruses on host genome function. Annu Rev Genet 42: 709–732. doi:10.1146/annurev.genet.42.110807.091501.
55. Magiorkinis G, Gifford RJ, Katzourakis A, De Ranter J, Belshaw R (2012) Env-less endogenous retroviruses are genomic superspreaders. PNAS 109: 7385–7390. doi:10.1073/pnas.1200913109.
56. Katzourakis A, Magiorkinis G, Lim AG, Gupta S, Belshaw R, et al. (2014) Larger mammalian body size leads to lower retroviral activity. PLoS Pathog 10: e1004214. doi:10.1371/journal.ppat.1004214.
57. Gilbert C, Feschotte C (2010) Genomic fossils calibrate the long-term evolution of hepadnaviruses. PLoS Biol 8: 12. doi:10.1371/journal.pbio.1000495.
58. Liu W, Pan S, Yang H, Bai W, Shen Z, et al. (2012) The first full-length endogenous hepadnaviruses: identification and analysis. J Virol 86: 9510–9513. doi:10.1128/JVI.01164-12.
59. 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. doi:10.1038/ncomms2798.
60. Staginnus C, Richert-Pöggeler KR (2006) Endogenous pararetroviruses: two-faced travelers in the plant genome. Trends Plant Sci 11: 485–491. doi:10.1016/j.tplants.2006.08.008.
61. Iskra-Caruana M-L, Duroy P-O, Chabannes M, Muller E (2014) The common evolutionary history of badnaviruses and banana. Infect Genet Evol 21: 83–89. doi:10.1016/j.meegid.2013.10.013.
62. Kunii M, Kanda M, Nagano H, Uyeda I, Kishima Y, et al. (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. doi:10.1186/1471-2164-5-80.
63. Richert-Pöggeler KR, Noreen F, Schwarzacher T, Harper G, Hohn T (2003) Induction of infectious petunia vein clearing (pararetro) virus from endogenous provirus in petunia. EMBO J 22: 4836–4845. doi:10.1093/emboj/cdg443.
64. Hansen CN, Harper G, Heslop-Harrison JS (2005) Characterisation of pararetrovirus-like sequences in the genome of potato (Solanum tuberosum). Cytogenet Genome Res 110: 559–565. doi:10.1159/000084989.
65. Jakowitsch J, Mette MF, van Der Winden J, Matzke MA, Matzke AJ (1999) Integrated pararetroviral sequences define a unique class of dispersed repetitive DNA in plants. PNAS 96: 13241–13246.
66. Lockhart BE, Menke J, Dahal G, Olszewski NE (2000) Characterization and genomic analysis of tobacco vein clearing virus, a plant pararetrovirus that is transmitted vertically and related to sequences integrated in the host genome. J Gen Virol 81: 1579–1585.
67. Emerman M, Malik H (2010) Paleovirology—modern consequences of ancient viruses. PLoS Biol.

68. Feschotte C, Gilbert C (2012) Endogenous viruses: insights into viral evolution and impact on host biology. Nat Rev Genet.
69. Malik HS (2012) Retroviruses push the envelope for mammalian placentation. PNAS 109: 2184–2185. doi:10.1073/pnas.1121365109.
70. Houldcroft CJ, Breuer J (2015) Tales from the crypt and coral reef: the successes and challenges of identifying new herpesviruses using metagenomics. Front Microbiol 6: 188. doi:10.3389/fmicb.2015.00188.

Chapter One: The First Endogenous Herpesvirus, Identified in the Tarsier Genome, and Novel Sequences from Primate Rhadinoviruses and Lymphocryptoviruses

Abstract

Herpesviridae is a diverse family of large and complex pathogens whose genomes are extremely difficult to sequence. This is particularly true for clinical samples, and if the virus, host, or both genomes are being sequenced for the first time. Although herpesviruses are known to occasionally integrate in host genomes, and can also be inherited in a mendelian fashion, they are notably absent from the genomic fossil record comprised of endogenous viral elements (EVEs). Here, we combine paleovirological and metagenomic approaches to both explore the constituent viral diversity of mammalian genomes and search for endogenous herpesviruses. We describe the first endogenous herpesvirus from the genome of the Philippine tarsier, belonging to the *Roseolovirus* genus, and characterize its highly defective genome that is integrated and flanked by unambiguous host DNA. From a draft assembly of the aye-aye genome, we use bioinformatic tools to reveal over 100,000bp of a novel *Rhadinovirus*; the first lemur gammaherpesvirus, closely related to Kaposi's sarcoma-associated virus. We also identify 58 genes of *Pan paniscus Lymphocryptovirus 1*, the bonobo equivalent of human Epstein-Barr virus. For each of the viruses, we postulate gene function via comparative analysis to known viral relatives. Most notably, the evidence from gene content and phylogenetics suggests that the aye-aye sequences represent the most basal of known rhadinoviruses, and indicates that tumorigenic herpesviruses have been infecting primates since their emergence in the late Cretaceous. Overall, these data show that a genomic fossil record of herpesviruses exists despite their extremely large genomes, and expands the known diversity of *Herpesviridae*, which will aid the characterization of pathogenesis. Our analytical approach illustrates the benefit of intersecting evolutionary approaches with metagenomics, genetics and paleovirology.

Author Summary

Herepsviridae is a family of DNA viruses that have characteristically large and complex genomes. This defining feature is also responsible for bioinformatic challenges that complicate herpesvirus genomics, and why an endogenous herpesvirus remains elusive. Given that several species of herpesvirus are clinically relevant to humans, there is a pressing demand for techniques capable of generating and managing large quantities of herpesvirus genome data. This is coupled with a need to explore herpesvirus diversity in order to understand pathogenesis within an evolutionary context. Lessons from the study of ancient viral integrations have also highlighted the need to include information offered by paleoviruses. Using perspectives from paleovirology and metagenomics, we identify three herpesviruses within the genome data of their primate hosts, including the first endogenous herpesvirus. All three viruses are closely related to important human pathogens and two of them are entirely new species. Both comparative molecular biology and evolutionary analysis were applied to examine our results for their clinical relevance. Furthermore, we demonstrate how this analytical approach was also used for the data collection itself, by treating nucleotide databases in their entirety as a single metagenomic resource.

Introduction

Herpesviruses are double-stranded DNA viruses that maintain pervasive life-long infections that are intimately linked to their hosts' biology. They exhibit a long history of host co-evolution [1–3], resulting in viruses that are highly adapted to their natural host. This specialization is responsible for an intricately balanced host/virus relationship, resulting in characteristically asymptomatic infections [4]. Disruption to this balance can occur during periods of immunosuppression, resulting in pathogenesis ranging in severity from mild blisters to cancer [4,5]. Imbalance can also result in disease if a cross-species transmission introduces the virus to a new host [6]. This contrast between a benign state and potentially fatal pathogenesis means that herpesviruses represent a unique system to study the mechanisms of disease development, as well as the influences of co-evolution on viral emergence. A thorough appreciation of herpesvirus biology and evolutionary dynamics, however, requires access to whole genomic data, as well as strategies for novel herpesvirus discovery. The complexity that makes herpesviruses an ideal study group is also responsible for exacerbating the challenge of sequencing them. Although hundreds of species have been recorded, their genome size and complexity (often including host derived sequences), ensures that only a fraction of these have been sequenced, severely limiting their analytical potential.

In recent years, more detailed macroevolutionary studies of viruses have become possible via endogenous viral elements (EVEs) that are ancient in origin [7]. These genomic relics are the result of accidental host germline integration events by viruses that eventually reach population fixation. In essence these EVEs 'fossilize' in the species' genome and are thereafter transmitted vertically as part of the host gene repertoire. By virtue of their replication mechanism, retroviruses represent the vast majority of this viral fossil record, as they necessarily integrate into the host genome [8]. This relative abundance is due in part to the difficulty in searching for EVEs from other groups, because identifying fossils of larger viral genomes with complicated natural histories is far more bioinformatically challenging. Herpesviruses are emblematic of these difficulties due to their low sequence conservation, lack of consistent gene order and evidence of host-gene recombination [3]. Indeed, in the most comprehensive EVE survey conducted to date, herpesviruses were excluded from the search [7].

Despite remaining elusive thus far, there is nonetheless reason to suspect the existence of a herpesvirus EVE (HEVE), as several herpesviruses are known to undergo strategic somatic genome integration to establish their latent state. Known examples include

Epstein-Barr virus (EBV) [9,10], Marek's Disease virus (MDV) [11–13] and human herpesvirus 6 (HHV6) [14,15], although integrated EBV viruses represent a small fraction of infected cells. HHV6 is particularly noteworthy as it has been shown that the virus is capable of germline integration, which has led to speculation that HHV6 is an EVE at an early stage of the fossilization process that has yet to reach an appreciable frequency in the population [7]. Less than 1% of humans (UK and US data) carry a heritable HHV6 (ciHHV6) integrated into telomeric regions of various chromosomes [15,16]. Both HHV6 and MDV genomes posses host satellite telomeric repeats (TMRs) within the repeat regions found at either end of their genomes. Despite being controversial for many years, a large body of work has established the telomere-specific integration of both HHV6 and MDV that occurs via a poorly characterized process of homologous recombination between viral and host TMRs [10,11,15].

Paleovirology as a discipline has only become possible due to innovations in sequencing technology that allow the rapid sequencing of whole genomes. The same advances have also unlocked avenues in metagenomics, which are especially valuable in clinical settings such as mapping the human virome [17]. Paleovirology can be seen as contributing to this effort, if we consider the virome as not only the viruses that are either in or on species, but also the collection of viruses embedded in their genome. Indeed the technical challenges are not dissimilar, as in both cases the objective is to identify and characterise minute proportions of viral nucleic acid from data that is overwhelmed with host sequences. The only difference is that unlike host genome sequencing efforts, metagenomics applied to the human virome actively welcomes 'contaminants', as they are the subjects of study. Similar to metagenomic techniques, though admittedly with a narrower scope, viral identification has been achieved by deep sequencing of clinical samples that subsequently filter out host sequences using a reference [18]. This approach is limited, however, to situations where reference sequences exist for the host genome and often also that of the virus. Although this issue is also a limitation in paleovirological surveys, since viral references are used to probe genome records, the sequences need only be similar enough for initial screening. Furthermore, *a priori* knowledge of the host genome is not necessary for EVE identification, because paleovirology employs evolutionary techniques as part of the process of characterising sequences. As a consequence of their rarity, new EVEs are sought by querying entire databases, rather than individual genomes. This maximises the chance of EVE detection, treating the nucleotide content of databases as though they were a single resource, or metagenome.

Recognising that paleovirology and viral metagenomics share many of the same goals and techniques, we set out to establish the benefits of approaching viromics with this interdisciplinary perspective in mind. This means that the ultimate objective was to explore the viral diversity through bioinformatic means alone, as a proof-of-principle that demonstrates the hidden value of online genomic databases. Because HHV6 is the only known herpesvirus capable of germline chromosomal integration, we hypothesized that a closely related HEVE should be detectable using HHV6 as an initial probe. As a result, we describe novel *Herpesviridae*-like (HVL) sequences from the genome of the aye-aye (*Daubentonia madagascariensis)*, bonobo (*Pan paniscus*) and Philippine Tarsier (*Tarsius syrichta*). We recognise the viral sequences of *P. paniscus* as the previously described *Pan paniscus Lymphocryptovirus 1* (PpanLCV1)*,* for which only 3,190 bp had been identified [19]. Our analyses demonstrate that two of these (*T. syrichta* and *D. madagascariensis*) sequence sets represent partial genomes of entirely new virus species, in mammals that we did not know to be herpesvirus hosts. Most notably our analyses reveal that the *T. syrichta* sequences represent the first endogenous herpesvirus whose ancestral exogenous relative closely resembled HHV6. Each set of sequences we identified is referred to throughout the text as PpanHVL (*Pan paniscus Herpesviridae*-like), DmadHVL (*Daubentonia madagascariensis Herpesviridae*-like) and TsyrHVL (*Tarsius syrichta Herpesviridae*-like). The TsyrHVLs are additionally referred to as the tarsier herpesvirus endogenous viral element (THEVE), when discussing the endogenous virus.

Results

We used consensus sequences of 33 NCBI viral protein clusters (table S1) to initially screen the mammalian genomes in the WGS database, which includes 14 primate genomes from 9 species. Of the 114 hits, we shortlisted those of bonobo, aye-aye and Philippine tarsier for further investigation, with 76 viral hits reported on 22 contigs (results summary in table S2). We identified viral hits to Epsein-Barr virus in human, which is a well-characterised and fully sequenced herpesvirus, and we therefore did not investigate them further. Each of shortlisted genomes (table 1) were then searched in more detail using a larger query set of all proteins from fully sequence viral genomes of genera that showed similarity to the primates in the first BLAST round. From the resulting hit list, we performed a reciprocal BLASTx search of 165 contigs against viral proteins in the NCBI nr database, thereby eliminating 54 false positives (table S3). Contigs were visualised by creating conceptual schematics of each sequence set by mapping the homologous regions for each sequence against an appropriate reference genome. Herpesvirus genomes fall into 6 general layouts with respect to their unique coding regions in addition

to the number and orientation of repetitive blocks [4]. Although it is not a rule, herpesviruses of the same lineage tend to exhibit the same overall genome architecture and closely related herpesviruses (e.g., belonging to the same genus) tend to maintain a similar gene set and organisation (figure S1) [4]. We therefore used viruses closely related to each HVL sequence set as a guide to mapping and visualising the contigs.

		Pan paniscus	*Daubentonia madagascariensis*	*Tarsius syrichta*
WGS Genome records	**Sex/animal name**	**Female 'Ulindi'**	**Male 'Goblin'**	**Female (name unkown)**
	Institute	Leipzig Zoo	Duke Lemur Centre	Duke Lemur Centre
	NCBI BioSample ID	SAMEA1029457/SAMEA1029458	SAMN00690380	SAMN02445010
	Tissue	Blood	Liver (necropsy)	N/A
	Coverage	26×	38×	48×
	WGS project ID	AJFE0	AGTM0	ABRT0
	Sequencing platform	GS FLX (454)	Illumina GA II	Sanger/Illumina mixed
	Total sequence length	2,869,206,676	2,855,365,987	3,453,847,770
	Total assembly gap length	143,237,068	0	48,109,210
	Gaps between scaffolds	0	N/A	0
	Number of scaffolds	10,868	N/A	337,188
	Scaffold N50	10,124,892	N/A	401,181
	Number of contigs	121,235	3,231,305	492,902
	Contig N50	66,775	3,653	38,165
	Total number of chromosomes and plasmids	1	0	0
	Reference/credits (According to WGS record)	[76]/Kay Prüfer *et. al.*	[77]/George H. Perry *et. al.*	Wesley Warren, Richard K. Wilson & the Genome Institute, Washington Univ. School of Medicine

Details of the genome records that were investigated for viral sequences. Information was obtained from the NCBI assembly website: http://www.ncbi.nlm.nih.gov/assembly/. N50 is the contig length at which all contigs of that length or longer represent 50% of the total of the lengths of all contigs.
doi:10.1371/journal.pgen.1004332.t001

Table 1 Genomes containing Herpesvirus-like sequences.

To determine the taxonomic and evolutionary relationship of these HVLs to *Herpesviridae*, a series of phylogenetic trees were reconstructed. Previous phylogenetic studies have reproducibly established the common evolutionary origin for each of the subfamilies in *Herpesviridae* [2,4]. Because of the complicated natural history for herpesvirus genomes, single-gene trees do not necessarily reflect the true evolutionary relationships. The dependence of phylogenetic accuracy on gene availability poses a challenge for novel virus characterization, because new sequences need to be compared against the greatest range of available diversity possible. Doing this, however, necessitates the inclusion of partially sequenced Herpesviruses, many of which are identified by only a single gene, or less. This trade off between phylogenetic accuracy and detail was taken into consideration when investigating the HVLs, and we employed multiple phylogenetic strategies, rather than depending on a single tree, to arrive at the best possible characterisation.

A herpesvirus endogenous viral element in Tarsius syrichta

Phylogenetic evidence indicated that the *Tarsius syrichta* HVL (TsyrHVL) sequences are most closely related to the Roseoloviruses (figure 1A/B), and so they were characterized with the HHV6A genome as a reference (figure 1C). Crucially, the TsyrHVL sequences include the junction between the viral-like region and host chromosome, which was necessary to identify it as a *bona fide* integration. Roughly 40.7Kb upstream of the viral junction in contig ABRT02391417.1 shows high sequence similarity to primate genomes and contains characteristic macroscopic features including LINEs, SINEs and a class 1 endogenous retroviral insertion (Figure 2). The junction itself consist of an array of (ATTGGG) telomeric repeats, as has been shown to be the case for chromosomally integrated HHV6 in the telomere [15]. To ensure that the junction was not the result of erroneous assembly, we independently confirmed the validity of the junction via PCR (figure 2) from a different animal than that of the WGS record.

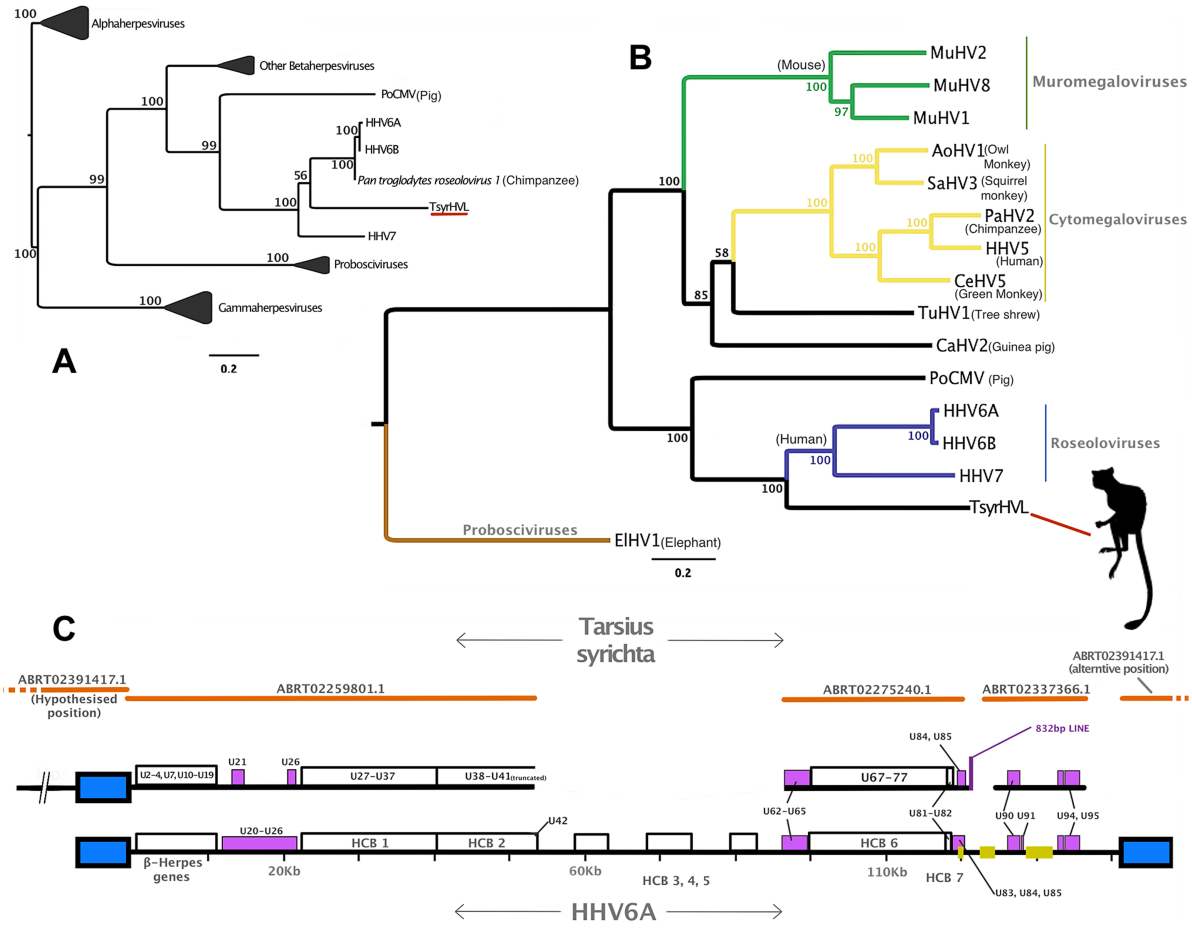


Figure 1 Phylogenetic and genomic analysis of the tarsier endogenous herpesvirus. Panel A: DNA polymerase tree showing the placement of the TsyrHVL sequence within the *Betaherpesvirinae*. Only the lineage leading to the node including TsyrHVL is shown and the rest are collapsed for clarity, and the size of the collapsed clade is arbitrary. Panel B represents the phylogeny reconstructed from a concatenated amino acid alignment of 6 core genes (terminase, large tegument, uracil-DNA glycosylase, kinase, capsid protein and helicase). Unclassified betaherpesviruses are shown as black branches, whereas those belonging to defined genera are indicated in colour. The rooting at *Proboscivirus* was determined according to the phylogeny in panel A. Numbers at each node in both Panel A and B represent bootstrap support. Panel C

shows a schematic of the tarsier sequences mapped to HHV6 as a reference. Orange lines indicate wgs contigs obtained from NCBI and GenBank IDs are annotated. Contig ABRT02391417.1 is represented on both sides, since it consists entirely of the DR region, although it aligns with ABRT02259801.1 with only 5 differences, and both placements are plausible. Blue box indicates the virus' terminal direct repeat (DR) regions. Yellow boxes represent the major internal repeat regions. Because the genomes are so large, it is not feasible to represent the complete coding content. Instead major herpesvirus core blocks (HCB) are indicated (as in reference [5]), and genes that are relevant to discussion points in the main text are also annotated. Abbreviations for Panel A and B are THEVE: Tarsier Herpesvirus Endogenous Viral Element, HHV6A/6B/7/5: *human herpesvirus 6A/6B/7/5*, MuHV1/2/8: *Murid herpesvirus 1/2/8*, AoHV1: *Aotine herpesvirus 1*, SaHV3: *Saimiriine herpesvirus 3*, PaHV2: *Panine herpesvirus 2*, CeHV5: *Cercopithecine herpesvirus 5*, TuHV1, PoCMV: *Porcine cytomegalovirus*.

It was important to amplify a single, large fragment that crossed the viral/host junction and also included both the viral and host flanks. Given the length of the product sought, we devised a semi nested-PCR strategy to overcome the fact that the DNA used was of low quality. Primer pairs for a 2,363bp fragment were initially used on genomic DNA, and although no band was visible by gel electrophoresis, low concentration was detected by spectrophotometry. These PCR products were then amplified using semi-nested primers, and a ‘touchdown’ PCR program, yielding high concentrations of the correct products (*see* Materials and Methods).

Similarly to HHV6, TsyrHVL includes a direct repeat (DR) region that contains 2 TMR motif regions. The TMR at the TsyrHVL/host junction is almost twice as large as its counterpart, which is a more comparable length to those in HHV6 (~300bp). This could have resulted from the mechanism of integration, and scenarios that would create a double-length TMR have been hypothesised for HHV6, albeit not with an identical outcome to the TsyrHVL [10]. Contig ABRT02259801.1 overlaps with ABRT02391417.1 by 337bp. The overlap region consists of 124bp from the main DR region, followed by the TMR array that continues past the overlap. There are 4 gaps and 1 mismatch that could be the result of assembly error, which occurs most frequently at contig ends. Alternatively, it is possible that ABRT02391417.1 contains the other DR from the other side of the viral integration, and only aligns to ABRT02259801.1 because the TMRs in both DR regions are identical. Both scenarios are valid in the absence of additional information and represented in figure 1C for clarity. In order to estimate the date of integration, we took advantage of the fact that at the time of integration, the right-most TMR in each DR would have been composed of a series of perfect hexanucleotide repeats. Thus, any differences from this repeat structure would have occurred in TysrHVL due to neutral evolution after its endogenistation. 294bp from the viral side TMR was used, beginning and ending with an intact TAACCC motif (the other TMR is far too divergent to have been the original perfect repeat). We inferred the maximum likelihood pairwise divergence of the observed TMR from the hypothetical TMR ancestor using the GTR model of substitution, at 0.165

substitutions/site. As we are measuring the divergence from an ancestral sequence rather than the divergence between contemporary repeats, the distance value was converted into time directly rather than being divided by two first. Using the mammalian neutral rate of evolution of $2.2x10^{-9}$ [20], the insertion is estimated to be ~76 million years old. Using the estimate for bushbabies of $3x10^{-9}$ [21], the date is significantly younger at ~56my, and maybe more accurate given that they are both relatively small prosimians with similar life histories. Nonetheless, both dates are plausible being after the tarsier divergence of ~80my [22], consistent with the absence of orthologous insertions (thus far) in other primates. The approximate date, however, should be interpreted with caution, since the repetitive nature of the TMR could mean that the pattern and rate of substitution is significantly non-neutral.

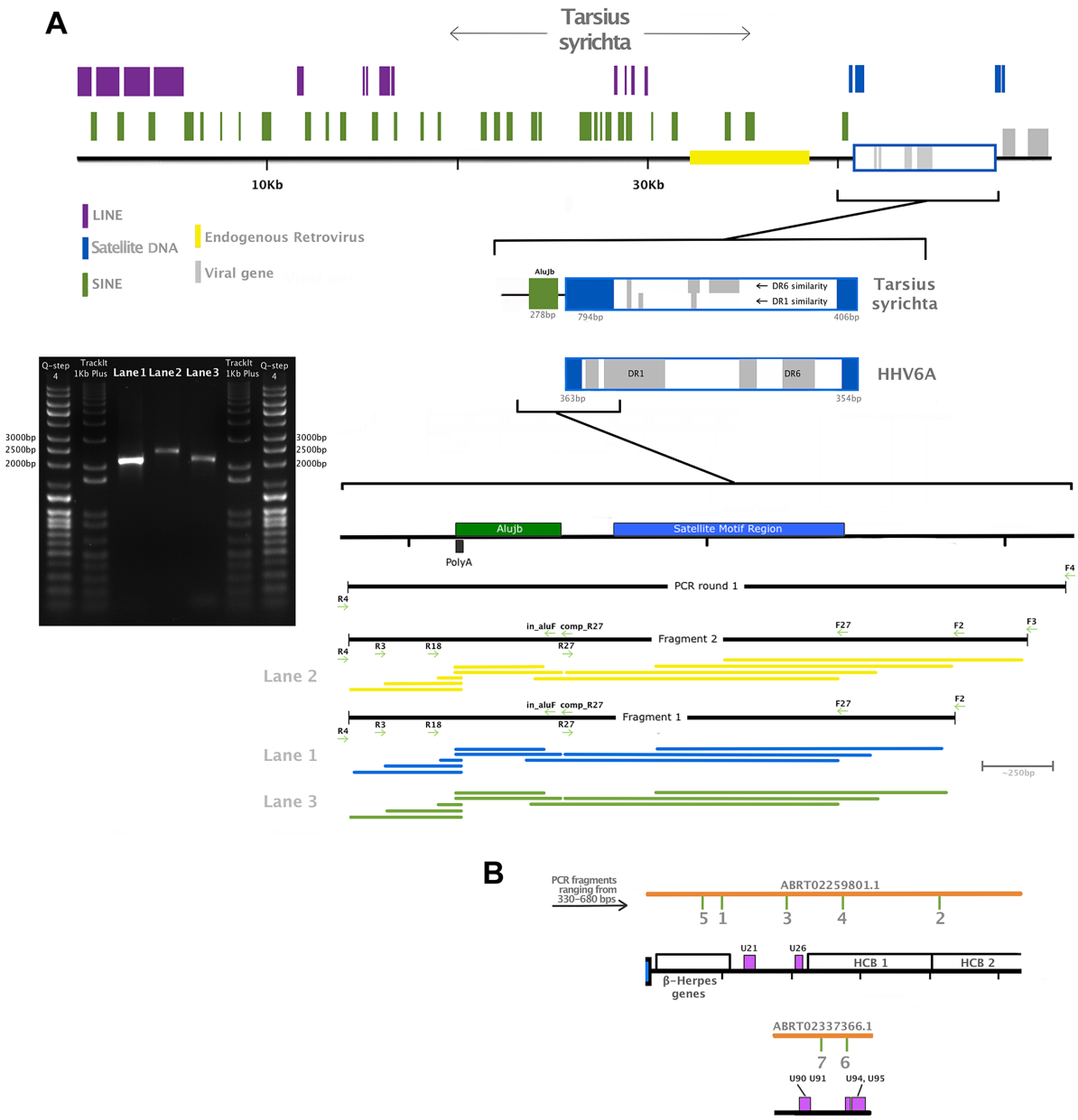


Figure 2 Validation of the endogenous status of TsyrHVLs.
Panel A is a close up of contig ABRT02391417.1, showing RepeatMasker-detected repetitive elements, an endogenous retrovirus and the satellite DNA telomeric repeat motif (TTAGGG)n that is characteristic of chromosomally integrated HHV6. A zoomed-in representation of the junction is also depicted, with regions highlighted according to the colour key, as well as a map of primers used for the amplification and location of sequenced fragments. Two independent PCR reactions were run in the first instance using the primer pair F4/R4 and genomic DNA, which was semi-nested in the second round of amplification. A single fragment using primers F2/R4 was obtained from both first-round amplicons, and a larger fragment from primer F3/R4 was possible from one of them. A 1% agarose gel electrophoreses is shown indicating the approximate size of the fragments: F2/R4 amplified a 2,034 bp fragment shown in Lane 1 and 3, while the F3/F4 fragment was 2,277 bp (Lane 2). Each fragment was sequenced using all visible primers (F4 was only possible for the largest amplicon). The coverage map indicates the sequence obtained from each primer, trimmed for quality. The final contig after quality trimming was 2,159 bp, and included 5 nucleotide differences and 2 indels relative to the tarsier genome record. A proportion of these will be true polymorphisms, while others may have resulted from polymerase error during either sequencing or amplification, (in both our fragment and the published sequence). Panel B shows the location of small fragments amplified to confirm the presence of the unique viral region.

Given that the TMR motif is detectable on the same contig that the left-most viral genes are, we think that the DR in our data is probably the left DR. Unusually however, the

orientation of the ATTGGG motif in the TMRs is the reverse complement to the expected, assuming it should be the same as HHV6. This could either mean that the exogenous viral ancestor to the TsyrHVLs is different to HHV6, or else be the result of the integration mechanism. Because the TsyrHVLs have been in the tarsier genome long enough to accumulate many nonsense mutations, we cannot take advantage of genomic features of Roseoloviruses that may have informed us on the integration event. This includes the fact that the left and right TMR regions within each DR in exogenous Roseoloviruses are not identical, and instead are either perfect or imperfect. In the TsyrHVLs, both are degraded and so cannot be distinguished, and other features such as pac1 and 2 signalling sites are mutated beyond recognition.

A phylogenetic tree was reconstructed using a short region of the conserved Herpesviral gene DNA polymerase in order to capitalize on partially sequenced herpesviruses for taxonomic breadth (Figure 1A). Although this low-resolution phylogeny is not reliable for detailed evolutionary comparison, it established the sub-family placement of TsyrHVL within *Betaherpesvirinae* with strong bootstrap support. A more robust tree of only Betaherpesviruses was obtained using a concatenation of six highly conserved core genes (terminase, large tegument, uracil-DNA glycosylase, kinase, capsid protein and helicase, Figure 1B). The *Roseolovirus* genus currently consists of HHV6, HHV7 and *Pan troglodytes Roseolovirus 1* and the tarsier sequences group closest to this lineage. Whereas it resolves basally to the Roseoloviruses, the genome organization and gene content is almost identical to HHV6, including Roseolovirus-specific genes (e.g. U21 and U85), thereby supporting their designation as a basal member of the group.

Unlike ciHHV6, nearly all genes in TsyrHVL have lost coding capacity through multiple frame-shifts and stop codons. These cannot be sequencing errors as the tarsier genome was sequenced to ~48x coverage, and is rather characteristic of fossilized EVEs as their evolutionary longevity depends on either being beneficial or nearly neutral to the host [23,24]. Consequently, unless the EVE is a currently active captured gene, it will drift at the host neutral rate, accumulating nonsense mutations [7,8] just as appears to have occurred in the TsyrHVLs.

Contig ABRT02259801.1 begins with the satellite motif followed by a short coding gap and similarity to a conserved block of 13 Betaherpesvirus genes (U2-4, U7, U10-U19) that are collinear with HHV6 (figure 1C). U2-4 are related to the US22 gene family that perform functions in replication and pathogenesis [5]. U11 encodes tegument pp150, the major

antigen recognised by HHV6-specific IgG [5]. U12 encodes a G-protein coupled receptor whose function is still unclear, but it is a chemokine receptor that might be involved in promoting cell growth and inhibiting apoptosis, as well as promoting cell migration [25]. Following this block in HHV6, there are 6 genes (U20-26) spanning 6835bp, but only U21 and U26 were detectable by BLAST, in a region that is approximately half the size of the HHV6 equivalent. After this, the TsryHVL contains similarity to Herpesviral core block 1 (HCB1), which is composed of 11 conserved genes (U27-37). Of these, all but the small capsid protein and DNA packaging protein UL33 appear to have lost coding capacity in TsyrHVL. ABRT02259801.1 ends with most of HCB2, which in HHV6 is composed of U38-U42. Clear similarity can be detected to all but one of these genes, as it ends within U41, most likely due to the draft state of the assembly. This is also true for the TsyrHVL equivalent of HCB 3-5, which were not found among the data.

Contig ABRT02275240.1, is mostly composed of HCB6 (genes U67-77) that is preceded by similarity to non-core HHV6 genes U62-65. All TsyrHVL HCB6 genes have lost coding capacity, except U68 and U72 that encode a tegument protein and envelope glycoprotein M. In HHV6, U62 is situated immediately downstream of the first exon of DNA packaging terminase subunit 1 (U60). Although it is absent in the TsyrHVL, the second exon is situated following three disrupted ORFs similar to U63-U65. HCB6 is succeeded by a 2,590bp coding gap in HHV6A that is only 1,612bp in TsyrHVL, followed by U79-82. In HHV6A and HHV6B, the next coding sequence would be U83-U85, but the TsyrHVL is more similar to HHV7, in that the chemokine gene is absent. ABRT02275240.1 ends with similarity to U86, the Immediate-Early IE2 gene that is involved in the virus' latent cycle. In HHV6 IE2 is followed by a coding gap that contains one of three major internal repeat regions, which is the most likely cause of assembly failure. The HHV6A genes 90-100 are the last four genes after U86 and encode IE1, membrane protein UL114, U94/rep, U95 and the Roseolovirus-specific glycoprotein Q. ABRT02337366.1 exhibits similarity to all except for membrane protein. Interestingly, U94/rep is a captured adeno-associated virus gene [5,26], which means that the TsyrHVL U94/rep represents an EVE within an EVE.

In addition to the host/virus junction, we also sequenced 7 fragments from the unique viral region to confirm that not only the junction is present (figure 2B). As well as a nucleotide match to the reference sequence, two of the amplified fragments included the characteristic nonsense mutations we would expect to find in endogenous elements. Fragment 1 is a 686bp partial gene U10, which encodes protein UL31 and contains the same in frame stop-codon as in contig ABRT02259801.1. Fragment 2 is 530bp spanning

the whole length of gene U35 and the first 90 bp of U36. Fragment 3 is a 471bp region the spans the coding gap between genes U17 and U18, and indeed does not appear similar to any proteins. The 325 bp of Fragment 4 includes part of the gene U28, which encodes subunit 1 of ribonucleotide reductase. The last amplicon obtained from ABRT02259801.1 is fragment 5, which is 456bp of gene U7. Two fragments were amplified from ABRT02337366.1. Fragment 6 is a partial sequence of the captured parvovirus *rep* gene, and includes the first 114 amino acids, as well as 1 of the 4 frame shifts found in ABRT02337366.1, and one in-frame stop codon. Fragment 7 is 462bp, spanning part of one of 3 internal repeat region (IR3).

Exogenous viral genomes in primate genome data

The evidence from DmadHVL and PpanHVL analyses indicate that they are most likely exogenous viral contaminants, since none of the sequences exhibited non-sense mutations in genes or the acquisition of transposable elements, which are characteristic features of paleoviruses. A DNA polymerase phylogeny established the subfamily placement of both HVLs as gammaherpesviruses (figure 3A and 4C, respectively). There are two main *Gammaherpesvirinae* lineages recognised as Υ_1 and Υ_2 rooted in between [27]. The Υ_1 lineage corresponds to the *Lymphocryptovirus* genus which to date is known to exclusively infect primates and include the PpanHVLs (figure 4A/B). The Υ_2 herpesviruses have a wider host range and the 100% bootstrap support at the root of the lineage place the DmadHVLs within them (figure 3A). A more detailed tree was reconstructed from a concatenated alignment of six core genes (DNA polymerase, glycoprotein B, Helicase/Primase, ssDNA binding protein, transport protein and major capsid protein), revealing that the DmadHVLs are neither a member of *Percavirus* nor *Macavirus* (known to infect Perissodactyls and Ruminants respectively), but are rather a sister group to the currently recognised *Rhadinovirus* (figure 3B). Despite numerous studies, the topology of the clade has yet to be fully resolved; like the rest of the herpesviruses, it involves broad patterns of co-evolution with their hosts but also frequent deviations that complicate phylogenetic interpretation [2–4,27–29]. In *Rhadinovirus*, BoHV4 and MuHV4 are unusual in that they group away from the ungulate and rodent clades respectively [2]. Furthermore, there are two lineages of primate *Rhadinovirus*, despite all other primate Gammaherpesviruses belonging to the Lymphocryptovirus clade. Considering that the topology for the rest of *Gammaherpesvirinae* is congruent with that of the host, the *Rhadinovirus* lineage likely emerged via a cross-species transmission event [2]. According to this hypothesis, the most parsimonious way to explain the topological placement of both BoHV4 and MuHV4 is to invoke a further two species transfer events,

from primates. Excluding them from the phylogeny leaves a primate Rhadinovirus clade with a topology that matches that of their hosts, and the placement of DmadHVLs basal to the group supports this evolutionary scenario. This suggests that the initial transmission would have occurred early in the evolution of primates, with a minimum root age of ~87 million years (Strepsirrhine divergence) [22]. Alternative explanations would require multiple cross species transmissions, once for each of the Rhadinoviral lineages (at least five for the phylogeny in figure 3B).

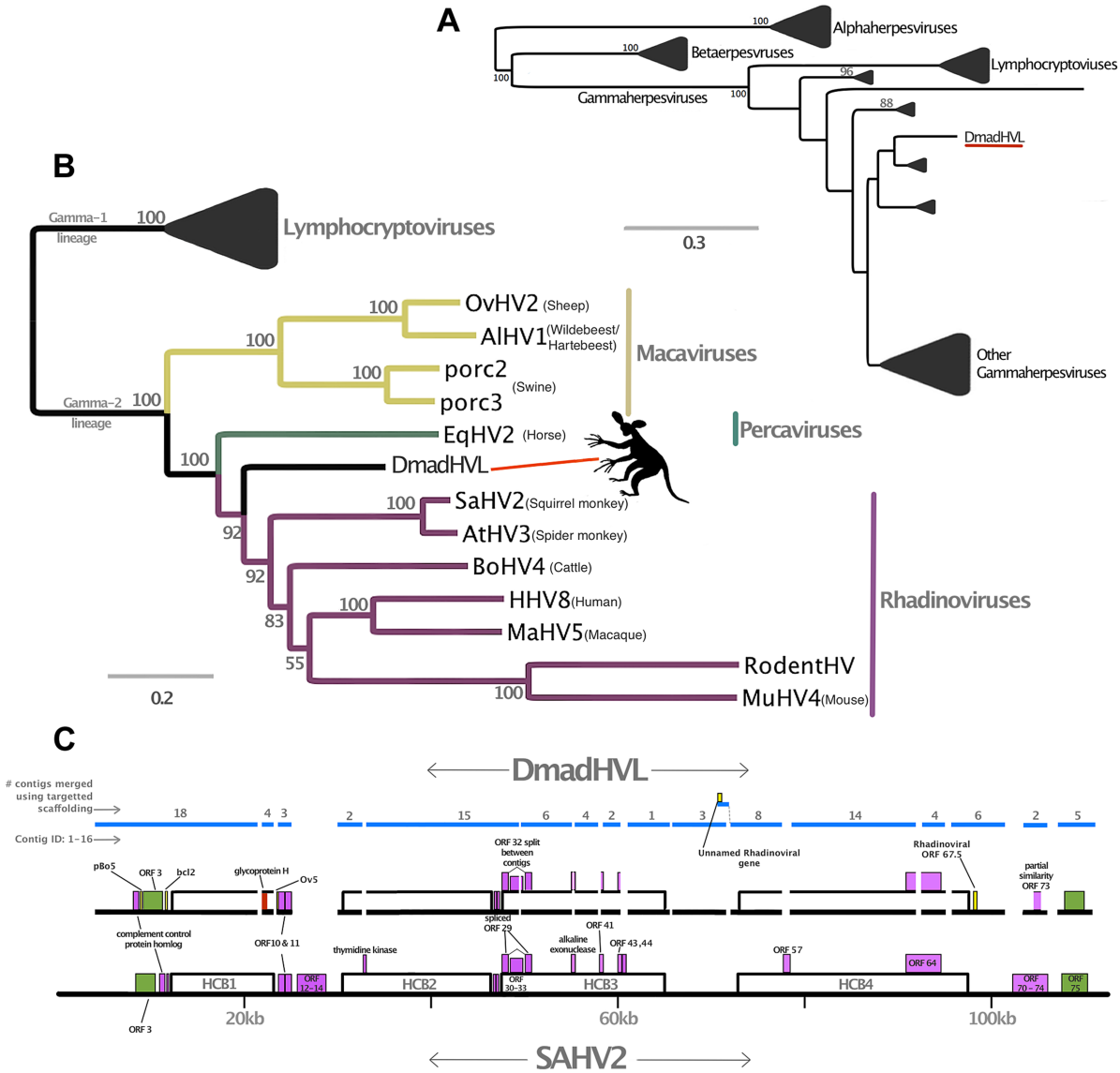


Figure 3 Phylogenetic and genomic analysis of the Daubentonia madagascariensis rhadinovirus.

Panel A is a maximum likelihood amino acid phylogeny of DNA polymerase, indicating the subfamily placement of the DmadHVL sequence as a gammaherpesvirus. Numbers at each node represent bootstrap support and only those above 50% are shown. Lineages other than those leading to DmadHVL are collapsed for clarity. In Panel B, the tree shown is a maximum likelihood phylogeny estimated using a concatenated alignment of 6 core genes (terminase, large tegument, uracil-DNA glycosylase, kinase, capsid and helicase). Coloured clades represent the different genera within gamma-2 herpesviruses, and bootstrap support is shown for each node. Panel C shows DmadHVL sequences mapped to Saimiriine herpesvirus 2 (SaHV2) as a guide, and major repeat blocks as well as noteworthy genomic differences and genes discussed in the main text are highlighted in coloured boxes. The green blocks are the FGAM synthase coding sequences, which are found at the termini. The red box annotated as glycoprotein H is presumed to be an assembly error. Pink boxes are discussed genes present in SaHV2, while the yellow ORFs are those found in different viruses. The blue lines indicate the different sequences that are a composite of multiple wgs contigs, the number of which is indicated above each sequence. The composite DmadHVL sequences discussed in the main text are numbered from 1–16 in a left-right direction. The DmadHVL virus genome appears to have a slightly larger region spanning herpes core block (HCB) 3 and HCB4, and so contig 11 is drawn to represent this. The scale of the schematic is approximate. Abbreviations for Panel B are porc2/3: OvHV2: Ovine herpesvirus 2, AlHV1: Alcelaphine herpesvirus 1, Porcine lymphotropic herpesvirus 2/3, EqHV2: Equine herpesvirus 2, RodHV: Rodent herpesvirus peru, MuHV4: Murid herpesvirus 4, AtHV3: Ateline herpesvirus 3, SaHV2: Saimiriine herpesvirus 2, BoHV4: Bovine herpesvirus 4, HHV8: human herpesvirus 8, MaHV5: Macacine herpesvirus 5.

A comparison of host and virus branch lengths reveals that old world monkey and ape Lymphocryptoviruses are more similar than expected assuming co-evolution (there are no

other instances of herpesvirus sequences from different natural hosts being so similar) [30], and the lineage exhibits strikingly low rates of evolution compared to Υ_2 [2]. The topology of the second lineage has been much more difficult to resolve and deviates from the co-speciation expectation, but has nonetheless been closely scrutinised in previous studies and split into multiple genera [31]. The bonobo sequences clearly group within the genus *Lymphocryptovirus* (figure 4B), the type species of which is the human pathogen human herpesvirus 4 commonly known as Epstein-Barr Virus (EBV). The sequence of glycoprotein B and DNA polymerase have already been identified for a bonobo Lymphocryptovirus, and the *Pan paniscus* HVLs (PpanHVLs) sequences in this study appear to be the same virus considering the phylogeny in figure 4A and when comparing particular genes. For example, the nucleotide sequence of glycoprotein B gene is 97 and 98% identical to the corresponding chimpanzee and bonobo sequence, respectively, compared to 89% against EBV. In total, the contigs characterised here extended the genomic sequence for *Pan paniscus Lymphocryptovirus 1* (panLCV1) by over 78 Kbps. Although nearly 95% of the adult human population is infected with EBV, we can be confident that this is not a human contaminant because of the comparatively higher sequence similarity of the PpanHVLs to panLCV1 than to EBV. This is further supported by the phylogenetic placement of the PpanHVLs in the tree shown in figure 4A, consistent with studies that have identified three distinct clades within the Υ_1 viruses that roughly correspond to major primate divergence events; new world monkeys, old world monkeys and hominoid host [32]. Figure 3B shows MaHV4 and PpanHVLs as a monophyletic group, which is inconsistent with a history of co-evolution. Interestingly, previous work has identified that MaHV4 and HHV4 are more closely related than expected, given the divergence date estimates for macaques and humans [19,30]. In the figure 4A tree, the well-supported topology of the Υ_1 lineage reflects a topology more consistent with co-divergence. The clade containing all known old world monkey infecting viruses also contains two separate lineages of Hominoid hosts. This discordance is further complicated by the fact that Clade B does not reflect their hosts' speciation pattern. The unusual grouping of hominoid viruses has been speculated to be the result of two cross-species transfers within the last 12 million years; the first must have occurred from a monkey host to a hominoid, which either happened independently in clade B_1 and B_2, or else resulted from a secondary transfer between hominoids. It is also thought that the B_2 lineage must have appeared in the last million years due to transmission from macaques to orang-utans and gibbons in Indonesia [2].

PpanHVL genome characterisation

The bonobo sequences are composed of 9 contigs totalling 78178bp, and constitute approximately 45% of the expected total length of the virus, assuming a similar genome size to EBV. They contain ORFs that are homologous to 58 of the 94 Epstein-Barr Virus coding sequences, in precise co-linearity.

Figure 5A describes the correspondence between EBV and contig AJFE01003225.1, spanning 44,035bp with 89% identity. It encodes several genes of the lytic cycle; the 5' end begins with a partial coding sequence similar to EBNA1, the only gene that is expressed during both latent and lytic states. It is homologous to a region in EBV that is flanked by the short internal repeat regions IR3 and IR4, suggesting that AJFE01003225.1 terminates where it does because the PpanHVL virus also contains repetitive sequences at the same loci; genome assemblers struggle to consolidate short reads that span such low complexity regions.

EBNA1 is known to disrupt cellular antiviral/tumour suppressing ability and is expressed in association with several malignancies (reviewed in [5,33,34]). Next, there are ORFs similar to the early lytic cycle BKRF2, BKRF3 and BKRF4, expression of which is controlled epigenetically by BZLF1 (absent in the PpanHVLs), which targets methylated CpG-containing binding sites [35]. The next ORFs are homologues of BBLF4, BBRF1, BBRF2, BBLF2/3 and BBRF3. BBLF2/3 and BBLF4 are known to interact with DNA polymerase (BALF5) and BSLF1 [36], as part of the lytic replication complex [37]. A PpanHVL BALF5 homologue is present (figure 5B), but BSLF1 is presumed to be in the missing regions. The next three ORFs (BBLF1, BGLF4 and BGLF5) are similar to a complex EBV region that is involved in lytic replication that are transcribed into mRNAs with the same 3' polyA signal [38]. The remainder of the contig encodes a number of genes that maintain the precise synteny with EBV so far exhibited, and a homologue of BcRF1 is particularly interesting with respect to its disease association. BcRF1 is a captured host pleiotropic cytokine Il-10 that performs immunoevasive functions. It has been shown to reduce thymocyte enhancement and mast cell proliferation in mice, and can also cause a decrease of components in the antigen-processing pathway (reviewed in [39]).

Contig AJFE01005652.1 (figure 5B) contains homologues of glycoprotein B and DNA polymerase genes, which are also the genes with near identity to the only two previously described PpanLCV1 sequences. There is also similarity to BARF1, a secreted lytic

protein involved in EBV immune evasion, by inhibiting the release of Interferon-α from monocytes [40]. Contig AJFE01001185.1 exhibits similarity to 7 EBV genes, including latent membrane antigen LMP-2A and BNLF2a, which is an important inhibitor of the HLA class I antigen presentation pathway (unique to Lymphocryptoviruses of Old Word primates [41]). EBV contains a block of internal repeats, part of which is detectible on contig AJFE01002662.1, which also encodes the last 254aa of BWRF1.

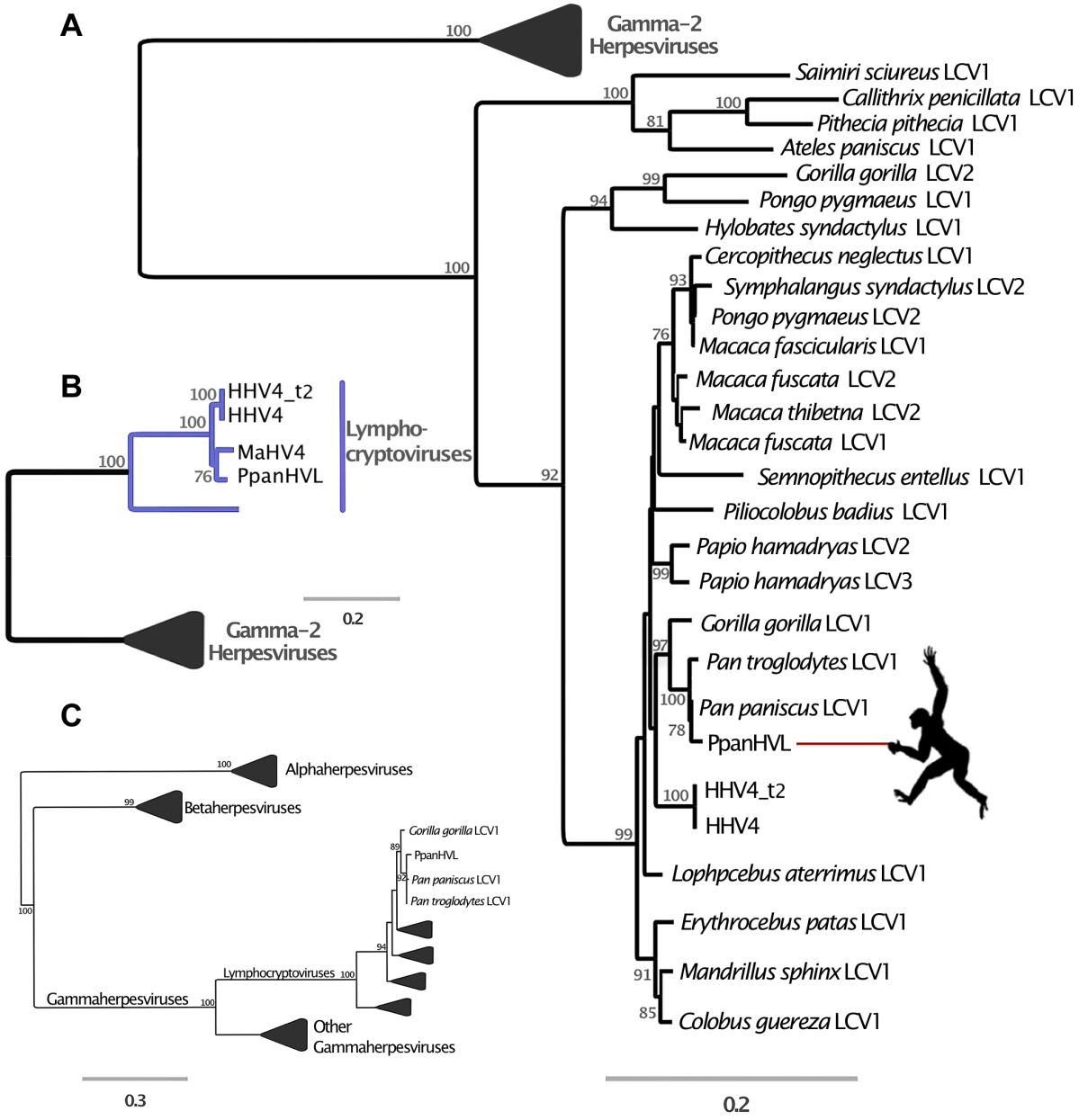


Figure 4 Phylogenetic analyses of Pan paniscus lymphocryptovirus.
In all panels, the trees depicted represent Maximum likelihood phylogenies reconstructed using different gene sets. Node values represent Bootstrap support but only nodes with over 70% are annotated. The panel A tree was reconstructed using a concatenated amino acid alignment of DNA polymerase and glycoprotein B, for which there are many representative taxa in Gammaherpesvirinae. In panel B the tree was reconstructed from a concatenated alignment of 6 core genes (terminase, large tegument, uracil-DNA glycosylase, kinase, capsid and helicase), and in both Panel A and B the gamma-2 lineage is collapsed for clarity. Panel C shows the phylogeny reconstructed using a conserved region of the DNA polymerase gene in order to ascertain the subfamily placement of the PpanHVL sequences. Abbreviations: LCV: Lymphocryptovirus, HHV4: Human herpesvirus 4, HHV4_t2: Human herpesvirus 4 type 2, MaHV4: Macacine herpesvirus 4, CaHV3: Callitrichine herpesvirus 3.

The contigs depicted in figure 5C represent a 17944bp contiguous stretch containing 8 ORFs with exact correspondence to the equivalent EBV region. The first of these, BHRF1 is known to inhibit cell death by interacting with several apoptotic inducers [42]. Thereafter there is similarity to BFLF2 and BFLF1 that are involved in virion maturation [43,44], and

then homologues of BFLF1 and BFRF1A, involved in DNA packaging - their deletion results in genome-less virions, making them an ideal vaccine target since a CD4+ T-cell response is maintained [45]. The ORFs after these are similar to BFRF2, BFRF3 and BPLF1, which deubiquitinates (thereby disrupting) the proliferating cell nuclear antigen [46]. The last two contigs in figure 5C include similarity to BOLF1, a core herpesvirus tegument gene that contains sequence similarity to the human HLA class II allele DPB1*0201. Patients with oligoarticular Juvenile Idiopathic Arthritis with this allele variant experience a cytotoxic response to self HLA-DPB1*0201, which is likely to be the result of T-cell cross recognition of BOLF1 [47]. Of the 449 amino acids present in the PpanHVL contig, over 70% are identical. A bulk of the difference between them lies in a 19aa stretch that includes the region of similarity between EBV and HLA-DPB1*0201.

DmadHVL genome characterization

The DmadHVL contigs were particularly fragmented; the scaffolding data had not been released and the contig N50 was much smaller than the other two genomes (table1). For this reason, we devised a targeted scaffolding strategy to link contigs that appeared close together based on the reference virus. This involved an iterative sequence-search for the edges of each contig, by reformatting the mate-pair reads into separate BLAST databases. The resulting assembly of mate pairs to the edge of their respective contigs allowed extension of the edge by tens of bases at a time. These new edges were then used to search for overlaps with contigs from the original wgs draft assembly that had previously not been identified by the original tBLASTn search. Scaffolding software and contig extension techniques do exist, but they are designed with the aim to improve the genome assembly as a whole. The brute-force approach employed here takes advantage of Illumina sequence reads, bridging the gaps between contigs in a sequence-specific manner. The labour-intensive dependence on ‘eye-ball’ quality assessment, means that it is inappropriate for large-scale contig extension, and would be difficult to automate fully. Indeed, it was often the case that multiple sets of read pairs were equally well aligned to the edge, and the decision process was aided by knowledge of the expected coding content. This means that the technique is particularly poor for contig edges that are non-coding. The sequence for the composite contigs can be found in supporting information S1.

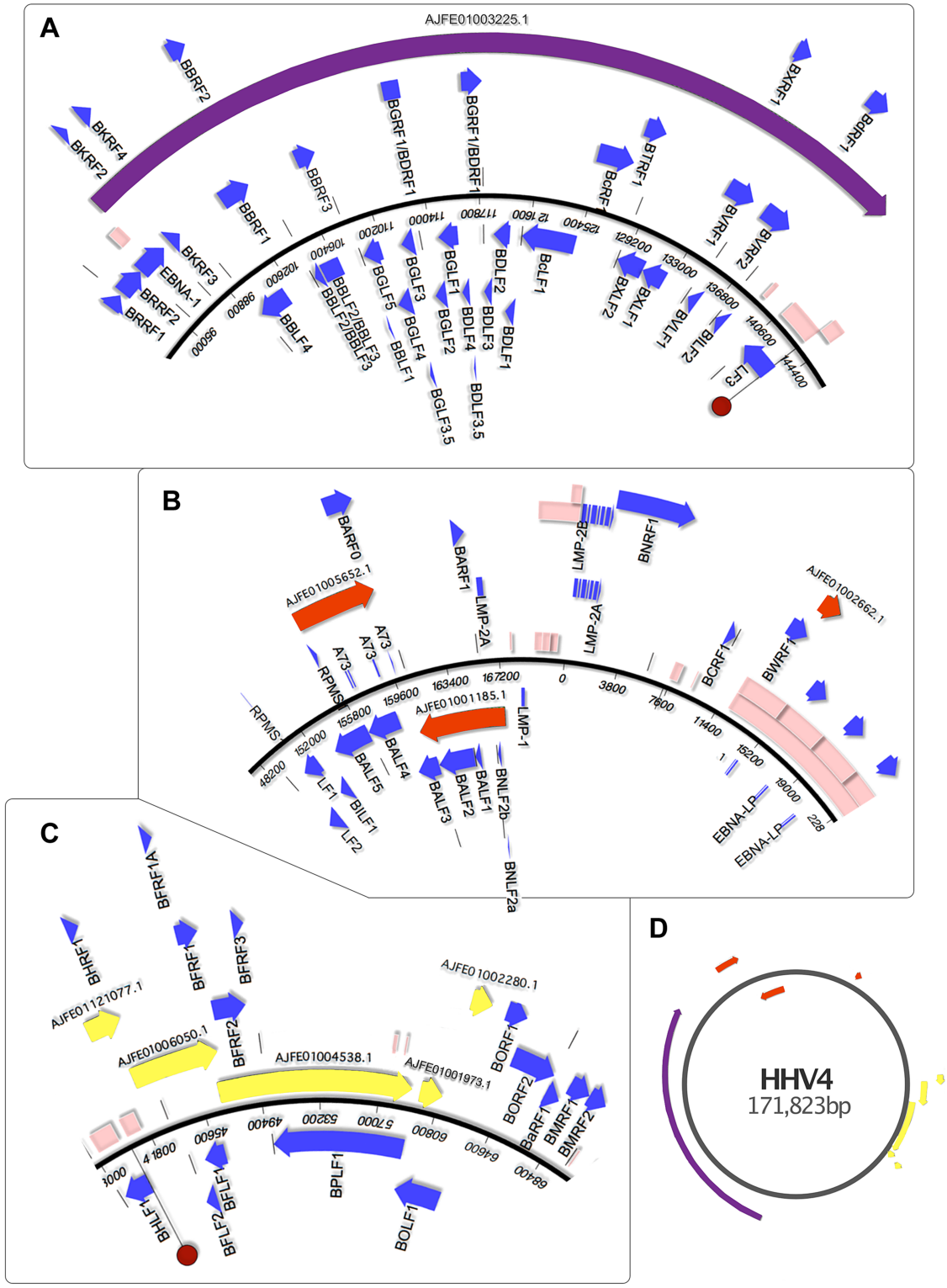


Figure 5 Genomic mapping of Pan Paniscus viral sequences.
The viral sequences found in the Pan paniscus genome appear to represent parts of the P. paniscus LCV1, which had been previously been identified but only partially sequenced. Because of the extreme genetic similarity to human herpesvirus 4 (HHV4) and identical gene set, ORF visualization was possible by aligning the contigs directly to the HHV4 genome. The contigs were separated into panels A–C to for clarity, and a zoomed out layout of the contigs is shown in panel D. In each panel, the blue-boxed arrows indicate the ORFs of HHV4, which are identically positioned in the PpanHVLs. Pink boxes represent repetitive sequence regions. In panel A and C, the dark red circle indicates the HHV4 origins of replication. It is interesting to note the repetitive sequences in regions of HHV4 that correspond to nearly all the edges of the PpanHVL contigs. Assembly algorithms are known to struggle in the reconstruction of low complexity sequences, which strongly suggests that for the P. paniscus LCV1 genome, repetitive sequences are located in the same place as they are in HHV4.

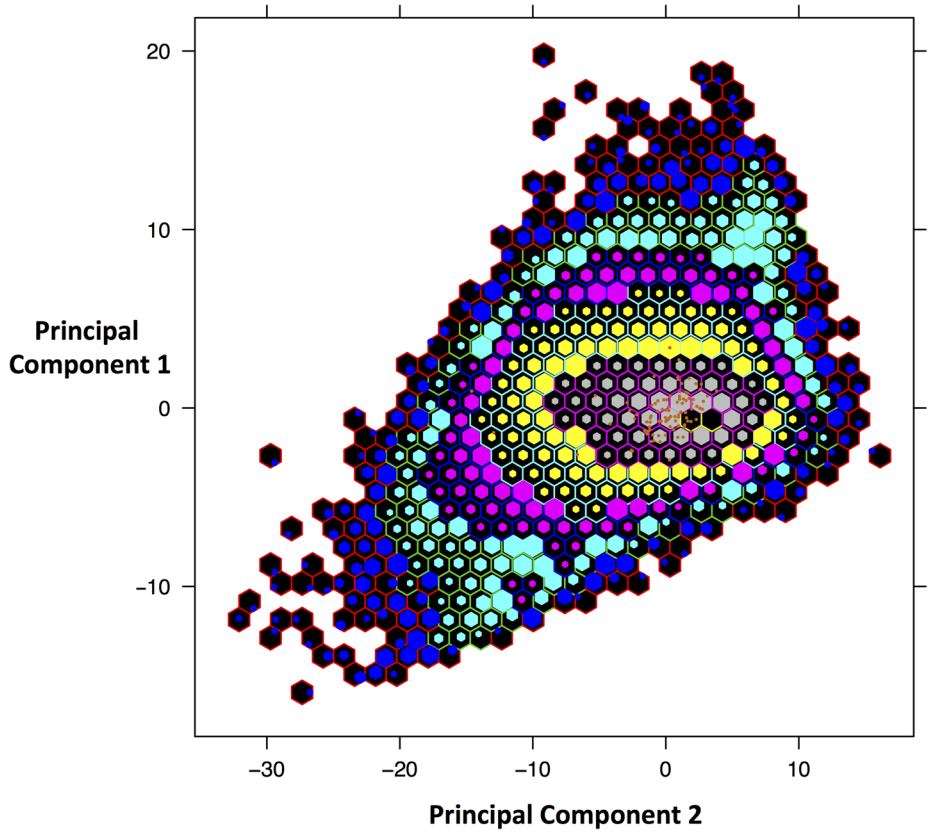


Figure 6 Principal Components Analysis of D. madagascariensis contigs.
Plot of scores from the first 2 principal components that account for 77% of the variance. The variables are: read coverage depth, all four mononucleotide and sixteen dinucleotide frequencies, mean, median, minimum and maximum insert size of reads. Scores were binned in order to more easily view the distribution. Orange points represent the contigs identified as viral, which have been over-plotted to identify where in the distribution the lie. Each hexagon represents a different bin and the size of the internal hexagon represents how 'full' that bin is in terms of the number of contigs. The placement of the internal hexagon describes the mean value of scores in that bin. The number of contigs contained within each bin is represented by the following colour scheme: blue with red border: 1–9, turquoise with green border: 10–99, pink with blue border: 100–999, yellow with turquoise border: 1000–9999, grey with pink border: 10,000–99,999, Black with yellow border: 100,000–999,999.

The phylogenetic analyses and evidence from sequence similarity indicate that the DmadHVLs probably represent part of a previously undescribed Rhadinovirus. As well as not detecting any nonsense mutations, we compared the nucleotide composition and mate-pair characteristics of viral contigs to the rest of the genome, in order to assess any differences that might better indicate they were part of an endogenous or exogenous virus. A principal components analysis of the assembled contigs was performed using 28 variables (figure 6). This revealed that the viral and non-viral contigs do not significantly differ, but it is nonetheless reassuring that the sequences cluster together, which is supportive of their shared identity. The overall genomic architecture of viruses in this group consists of a long unique region flanked by terminal repeat sequences, and the DmadHVLs adhere to this plan. Though it was not possible to determine the termini fully, regions in contig 1 and 16 both exhibit similarity to v-FGAM synthase, which is also found

in other Rhadinovirus genome extremities (figure 3C). In Contig 1, the 5' region contains similarity to the complement control protein, which is involved in active immune evasion in the closely related human herpesvirus 8 (HHV8), the etiological agent of Kaposi's sarcoma [48]. The corresponding gene in SaHV-2 is situated between v-FGAM (ORF 3) and the single-stranded binding protein. Instead, the DmadHVLs possess an ORF at that position with similarity to bcl2, a family of cellular apoptosis regulators. Similar host-derived bcl2-like genes can be found in other Rhadinoviruses, such as ORF16 in HHV8, and are known to either promote or inhibit apoptosis [49]. The DmadHVL version appears more similar to versions in *Porcine Lymphotropic herpesvirus* 1 and 2, *Ovine herpesvirus 2* and *Equine herpesvirus 2,* none of which are Rhadinovirus, albeit belonging to *Gammaherpesvirinae.* Another region of interest on contig 1 is ORF2, which shows similarity to BoHV4 pB05. The first three genes of the herpes core block 1 (HCB1) can be found in the DmadHVLs, but contig 1 ends before the end of glycoprotein B.

In every other Rhadinovirus, the Glycoprotein B and polymerase genes are invariably found in tandem and in the same coding direction. Contig 2 contains polymerase (ORF2) but is preceded by ORF1, which is a truncated homologue of glycoprotein H. It could be that this is due to erroneous assembly, or even a sequencing error, since we would expect glycoprotein H further downstream according to the genome plan of other Rhadinoviruses. Indeed, we find that the first ORF in contig 5 shows similarity to glycoprotein H. Contig 3 shows similarity to the first two ORFs after HCB1 that are conserved in several other herpesviruses and likely derived from host-captured dUTPase, thereafter becoming internally duplicated in different lineages independently [50]. The 5' end of contig 3 contains a short stretch of sequence similarity to Ov5, a G-protein coupled receptor found in the *Ovine herpesvirus 2*, *Alcelaphine herpesvirus 1*, and EBV BILF1, which share distant similarity to host interleukin 8 receptor [51]. The viral protein interferes with RNA-dependent protein kinase metabolism, which ordinarily acts to arrest cellular protein synthesis and induce apoptosis as an anti-viral defence. It was shown that the equivalent gene in SaHV2 was captured from the host [52]. Although the region is very short, it is unlikely to be a false positive, since it would be expected at that position given the colinearity with SaHV2.

In contig 4, ORF2 and 4 are similar to conserved core herpesvirus tegument protein and the protease/minor capsid protein, respectively. ORF 1 and 2 are also conserved in several herpesviruses, though their function has not been characterised. The third ORF of HCB2 falls in the data gap between contig 4 and 5. On contig 5, other than glycoprotein H,

the subsequent 8 ORFs lie in a contiguous stretch of 13,440bp. They are similar to the remaining HCB2 ORFs 23-26 as well as ORF27 and 28 and the first HCB3 ORFs 29-32. ORF32 is the DNA cleavage and packaging gene and is split between the two contigs 5 and 6, before the second exon of ORF 29. Alkaline exonuclease, a conserved herpesvirus gene superfamily straddles the gap between contig 6 and 7. The bulk of the gene is in contig 6 with 73% identity to SaHV2. Whereas its function in nearly all characterised genomes is involved in DNA replication and packaging, an additional function in HHV8 grants the ability to interfere at the mRNA level to shut off host gene expression [53]. Although none of the mutations implicated in HHV8's gain-of-function are shared by ORF1 contig 6, the host shut-off ability is attributed to structural differences, implying that it might not be limited to these specific positions/residues [54]. It would be interesting to know whether any such activity is present in the DmadHVL alkaline exonuclease.

Contig 8, 9 and 10 have the same HCB3 coding content and co-linearity with SaHV2, although the HCB3 ORF44 is entirely missing in the gap between 8 and 9. Interestingly, the first ORF on contig 9 exhibits higher similarity to the HHV8 ORF45, which blocks the activity of cellular interferon regulatory factor 7, an important part of the cell antiviral defence strategy [55]. Contig 10 contains genes with similarity to the inter-HCB region, including the RTA transactivator that is involved in the switch from a latent to lytic infection state [56]. Whereas Contig11 does contain the first 6 genes of HCB4 as expected, there is a larger preceding region than in SaHV2. In the first ~1.2kb there is similarity to a gene of unknown function found in *porcine lymphotropic herpesvirus 1* (genbank: AAM22152.1). The gap preceding contig 12 is presumed to contain HCB4 ORF57. Contig 12, 13 and 14 are 23,152bp combined, and consist of separate genes; the DmadHVL ORF64 homolog is split over contig 12 and 13. Contig 14 begins with the last HCB4 ORF65-67, and then contains ORF67.5, a gene found in some Rhadinoviruses but not SaHV2, the most similar of which is the Bovine herpesvirus 4 version. The contig ends with homologs of ORF68 and 69. Contig 15 contains similarity to ORF73, the latent nuclear antigen, a highly immunogenic product found in several other virus (e.g. EBNA1 in EBV), but is most similar to the HHV8 version. LANA is essential for episomal genome maintenance [57], and is involved in a multitude of cellular interactions [58] and mechanisms of pathogenesis; It can upregulate cellular proliferation as well as prolong the cell cycle and inhibit apoptosis [57,59–61].

Discussion

Each of the HVLs sets were characterised with publically accessible data and open source software, demonstrating the utility of combining metagenomics and paleovirology to the study of viral diversity. Given the difficulty of sequencing herpesvirus genomes, our results highlight the value of mining genomic sequence databases in this way. They exemplify the advantage of paleovirological approaches to genome analysis that view nucleotide repositories as a single resource, containing more biodiversity than the sum of its contributing organisms. In the era of high throughput sequencing methods, all genomic data has the potential to be treated as metagenomic and it is important to highlight that the viruses described here were obtained not through technical innovations, but rather, by reframing the strategic approach and borrowing the paleovirological perspective.

The TsyrHVLs represent a Roseolovirus-like endogenous viral element, and originates from a previously undescribed species that we tentatively name *Tarsius syricha roseolovirus 1* (TsRV1), or Tarsier Herpes EVE (THEVE) when referring to the paleovirus. THEVE clearly demonstrates the endogenisation of herpesviruses, and has extended the host range of Roseoloviruses from being ape-exclusive to include tarsiers. The only Roseoloviruses identified to date infect chimpanzees [62] and humans, suggesting that the lineage may be associated with many more primates. Previous studies have shown that the topology and relative branch lengths of Betaherpesviruses are largely congruent with that of their hosts which is strongly suggestive of co-divergence [2,28]. They are thought to have arisen by co-speciation with Euarchontoglires hosts, but the lineage containing Roseoloviruses deviates from this pattern and must have arisen by cross species transmission [3,4]. There is not enough information to conclude whether single or multiple transmission events occurred, but if the lineage arose from a single spill over, then THEVE suggest that Roseoloviruses have been infecting primates since the *Tarsiformes* divergence approximately 81 mya [22], and our estimate for the integration date of THEVE between 55-75mya is consistent with this scenario.

Studies of ciHHV6 have confirmed that the telomeres are the chromosomal target for integration, and Marek's disease virus is also known to integrate into the telomere of chicken chromosomes, albeit not heritably (reviewed in [10]). As well as the presumed dependence of integration on homologous recombination of viral and host TMRs, the preference for telomeric regions is likely to have evolved because the large genomes of herpesviruses would not be tolerated elsewhere in the chromosome. Indications that HHV6 may integrate into some chromosomes preferentially suggests that even the

telomeres of some chromosomes are too crucial to withstand such a drastic integration [15,63]. Moreover, although the precise telomeric integration site is not specific in HHV6, it is unlikely to be evolutionarily stable at the extreme end of the telomere, which is gradually shortened at each cell division [64]. This raises the possibility that the region between the sub-telomere and chromosomal side of telomeres contains an underappreciated diversity of EVEs that would be very difficult to study. THEVE offers a way to study the long-term effects of ciHHV6 on both telomere biology and the general physiological and evolutionary impact of acquiring ~150Kb of viral genomic material. It is also an alternative investigative avenue in which to fully characterise the Roseolovirus integration model. The mechanism of HHV6 integration has yet to be confirmed, and although a number of models have been proposed, it is difficult to assert with any confidence which (if any) are applicable to THEVE [10,11,14,63,65]. There is strong evidence to suggest that the HHV6 capacity to integrate is brokered by its captured copy of *rep* [63]. It has been demonstrated that HHV6 can rescue replication of Adeno-Associated Virus with defective *rep* [66]. We cannot know the exact process that lead to herpesvirus *rep* acquisition, but if the only selective advantage were conferring the ability to integrate, it would mean that the mechanism of *rep* transfer to Roseoloviruses is also the reason for its evolutionary maintenance, and in the case of THEVE, the eventual genomic fossilization. Very little is known about how non-retroviral EVEs are formed, and in this case it would be interesting for future work to test the idea that the endogenisation of a parvovirus into a herpesvirus is in turn responsible for the endogenisation of the herpesvirus itself. Interestingly, although THEVE resolves basally to HHV6 and 7, the gene is absent in HHV7, which in fact has not been shown to be capable of genomic integration. A phylogeny of the gene comparing the copies of endogenous and exogenous parvoviruses robustly places the THEVE and HHV6 *rep* as sister taxa with 98% bootstrap support, which is consistent with a common origin. We propose that the combined evidence of positional orthology and co-linearity of the *rep* gene in THEVE and HHV6, along with monophyly of the genes in the *rep* tree, suggest the subsequent loss of the gene from HHV7.

The data presented also includes the description of sequences that appear to be from exogenous viral contaminants of their host genome projects. In doing so, we have described a number of virulence genes shared between these viruses and pathogenic human herpesviruses. Our results indicate that the PanHVLs almost certainly represent the previously recognised *Pan paniscus Lymphocryptovirus 1* (PpanLCV1), for which only two genes had been previously sequenced. The DmadHVL sequences are part of a previously undescribed *Rhadinovirus* of aye-ayes, provisionally named *Daubentonia*

Madagascariensis Rhadinovirus 1 (DmRV1). For both viruses, the assembly status of the genomes meant that the viral sequences were composed of discontiguous stretches homologous to reference viruses. Interestingly, in the case of the largest PpanHVL contig that spanned over 40 Kb, the homologous region of HHV4/EBV is flanked by internal repeat regions; a genome assembler would have not been able to reconstruct the junction easily (figure 5). Nonetheless, neither set of sequences contains disruptive mutations in coding sequences, and can be easily aligned to closely related and well-characterised herpesvirus genomes. The exogenous status of DmRV1 is further supported by the lack of any detectable viral sequences in the recently released re-sequencing project of 11 aye-aye individuals [67]. Both BLAST and read-mapping techniques were employed to search for HVL sequences in the other genomes, with no reportable matches. This also suggests that the virus is variably present in the population, and given that aye-ayes are a near-threatened species, understanding the factors that influence this will be informative from a conservation standpoint.

The techniques used to describe the exogenous viruses demonstrate the possibility of circumventing some of the challenges that clinical viral genomics is encumbered by. The relative paucity of viral nucleic acid compared to that of the host is a particular obstacle, and viral enrichment steps risk introducing biases [18,68]. Doing this in the absence of host genome information is especially important due to the perpetual threat of emerging diseases by zoonosis [69], and the primate viruses we describe belong to lineages that include human pathogens (EBV, HHV6 and HHV7, HHV8). The herpesvirus capacity to cross species has been demonstrated experimentally [70], and such a phenomenon has already been observed in the case of the *Macacine herpesvirus 1,* an Alphaherpesvirus of macaques that is deadly to humans [71]. The PpanLCV1 gene set is identical to that of EBV, meaning the overall pathology will likely be similar given that chimpanzees are physiologically comparable to humans. This makes it a substantial zoonotic risk, but also means PpanLCV1 an ideal system for comparative molecular analyses that may contribute to functional studies of EBV.

Across Rhadinoviruses, the genomic regions that are most divergent (both genetically and in terms of gene-content) were also the regions that we could not recover for DmRV1. Namely, these are the terminal regions, as well as regions which are consistently gene-sparse in each species. These sequences are notoriously difficult to identify, needing dedicated treatment in genome characterisation, as was done, for example, in the recently described elephant endotheliotropic herpesviruses [72]. Nevertheless, the genomic

arrangement closely resembles *Ateline herpesvirus 3* and *Saimiriine herpesvrius 2* (figure 3), and the vast majority of ORFs are characteristic of Rhadinoviruses, which allows us to speculate on its likely pathogenic mechanisms. The presence of an ORF similar to the RTA gene tells us that like most herpesviruses, DmRV1 is capable of switching between latent and lytic replication, which means that infected animals are probably largely asymptomatic. Conversely, the LANA-like ORF on contig 15 indicates that the virus may be capable of upregulating cellular proliferation, potentially promoting tumorogenisis in certain circumstances. The Rhadinovirus lineage as a whole probably arose by cross species transmission [2], followed by multiple subsequent transmission events, raising the question of why Rhadinoviruses are particularly amenable to host switching. It would be important for future work to address this, and DmRV1 will be key to understanding the origins and cross-species transferability of Rahdinoviruses. Although the genome most closely resembles Rhadinoviruses overall, we have described a number of similarities to other groups, such as the partial Ov5 gene and captured bcl2 homologue. Combined with the phylogenetic placement as the most basal member of the group this suggests that DmRV1 still maintains features of the transitional ancestor that probably began infecting primates in the late Cretaceous.

The techniques employed in this work were developed for the study of endogenous viral elements and have proved capable of identifying valuable viral genetic data embedded in online databases. Our objective was to address the lack of herpesviruses in the genomic fossil record, and so we incorporated metagenomic approaches to over come bioinformatic hurdles such as their large genomes. In addition to describing the first Herpes EVE, the serendipitous identification of two probable exogenous viruses reveals a wider benefit of this multidisciplinary perspective. By establishing this proof-of-principle on both a paleovirus and accidentally sequenced viruses, there are promising practical applications in other areas of virology and general genomics that will certainly have far reaching consequences.

Materials and Methods

All computational analyses were conducted on an Apple MacPro running MAC OSX 10.7.5 with 12-core Intel Xenon processor and 64GB RAM. Automation of tasks was achieved by a variety of custom scripts written in Python. The overall strategy involved searching for amino acid similarity in a 6-frame translation of a nucleotide sequence database [7,73]. This is because a combination of high sequence divergence and frame-shifting mutations, often render the nucleotide sequences of EVEs undetectably dissimilar

in a single-frame search. The ideal query protein represents the viral diversity well, but is simultaneously evolutionarily conserved enough to detect distant sequences. Poor sequence conservation across the order *Herpesvirales* means that there are no genes or products that captures sequence diversity in this way, and to compensate by using a large query set would be computationally expensive. Instead, 33 separate conserved protein clusters from the *Alloherpesviridae* and *Herpesviridae* families were used to generate consensus sequences and used to query the NCBI databases WGS and RefSeq. BLAST alignments from aye-aye (*Daubentonia madagascariensis*), bonobo (*Pan paniscus*) and Tarsus (*Tarsius syrichta*) indicated that similarity from multiple viral genes appeared syntenically on single contigs, motivating further investigation (table S1).

Sequence Collation

The genome records for the three primate species under investigation are at various levels of assembly completion (table 1). Consequently, the exercise of characterising viral sequences embedded within them is complicated by their fragmented state. Establishing a biological explanation for the presence of these HVLs requires an assessment of their identity, i.e., whether or not these are endogenous to the host genome or merely a contaminant. Evidence that would suggest that these are EVEs, includes finding the 5' and/or 3' flanking host genome regions, which would show that viral genes are integrated. Furthermore, the presence of mutations that result in frame shifts and stop codons of genes would indicated that the HVLs had been subject to random genetic drift as would be expected from EVEs. Herpesviruses are known to integrate chromosomally, however (reviewed in [10]), and some EVEs have been captured by the host genome to provide some function [24,74,75]. This means that the absence of inactivating mutations would not rule out endogenisation, nor would it necessarily be confirmed by finding flanking regions. Rather, a combination of different lines of evidence together builds a case for establishing the nature of these HVLs.

Database Mining

Locally implemented NCBI BLAST+ was used to run tBLASTn search of the genomic databases wgs and RefSeq. The set of queries used were consensus sequences obtained from 33 different 'protein clusters' of *Herpesviridae* and *Alloherpesviridae*, which were conserved to at least the sub family level, and represented a broad functional range (table S3). We performed a qualitative assessment of the resulting alignments, in order to identify host species that might warrant further investigation. The bonobo, aye-aye and Tarsier were shortlisted for the second search phase. The best matching viral groups to

the species belonged to the Lymphocryptovirus, Rhadinovirus, and Roseolovirus genera respectively, as determined by a reciprocal BLAST search of the contigs against a viral database of protein products. Delimiting the searchable database to only these species, allowed a more exhaustive search, using all protein sequences from the best viral groups. The resulting list of sequence records from each genome was then subjected to a final round of shortlisting. BLAST hits from each genome to viral proteins under 50 amino acids long or with an e-value score below 1E-10 were discarded from further investigation. Finally, a BLASTx search of these contigs against NCBI protein database was performed to identify false positive hits, which can result from host homologues of viral proteins (table S1).

Contig visualisation

The resulting set of BLAST hits were converted from XML to genbank format using a custom python script. These hits were then merged into the genome records of the viruses from which the protein queries were extracted, and used to generate a graphic representation of the host contigs using the reference viral genome as a guide. The map in figure 5 was created using a trial version of MacVector, while those in figure 1C, 2 and 3C were drawn using Pixelmator v2.2.2. Given the high sequence similarity to HHV4, it was possible to align the PpanHVLs to the reference genome (figure 5). As for the TsyrHVLs, it was not possible to predict the ORFs due to the abundance of stops, and their genetic distance to the reference Roseoloviruses meant that an alignment would not be informative in terms of gene mapping. Instead, we used BLASTx results to evaluate the regions of similarity in the TsyrHVLs to known coding sequences.

PCR amplification

DNA from *Carlito syrichta* (syn. *Tarsier syrichta*) kindly provided by Dr. Christian Roos from the Gene Bank of Primates at the German Primate Centre in Göttingen, Germany. The sample was obtained from the muscle tissue of a specimen from Frankfurt Zoo. A first round of PCR amplification was performed using *T. syrichta* genomic DNA extracted from muscle. Two reactions with a final volume of 50ul included 25ul RedTaq Ready mix (sigma-aldrich), 1.5ul 10uM forward and reverse primer, 20ul DNase/RNase free water, and two different concentrations of template. In one reaction, 2ul of template at ~15ng/ul and 1ul of ~46ng/ul in the other. Primers used were F4:GGCAGAAGATGGCACCTCC with R4:CCAACCCATTGAACTGCTAAAATTCC. Cycling parameters were: 2 minutes pre-denaturation at 94°C followed by 40 cycles of 40 seconds denaturation at 94°C, 40 second annealing at 55°C and 2 minute extension at 72°C, and a final extension at 72°C

for 7 minutes. PCR purification was performed using QIAquick PCR purification kit by Qiagen, and although no bands were visible by agarose gel electrophoresis, DNA concentration of elution was performed using a nanodrop spectrophotometer indicating 6 and 8.3 ng/ul was present in the samples. In the second round of amplification, 2ul of each of the first round amplicons were used in a reaction mix with the same ratio of reagents. Primer pairs used in each reaction were:
F2:GGGATCACTGAATAAAGTGCGATGG, with
R4:CCAACCCATTGAACTGCTAAAATTCC and
F3:GGCATTTAAATTGACGAGTCCAAACGG with R4. Cycling parameters for second round were: 2 minutes pre-denaturation at 94°C followed by 20 cycles of 40 seconds denaturation at 94°C, 40 second annealing at 55°C and 2 minute extension at 72°C. We then ran a further 20 cycles with annealing temperature dropped to 50°C, and a final extension at 72°C for 7 minutes. Amplicons were visualised by 1% agarose gel electrophoresis (Figure 2A). Purified products were diluted to a concentration of 20ng/ul and sent for Sanger sequencing by a commercial service (Source Bioscience Oxford). Sanger method cannot achieve good quality sequencing beyond ~700bp, so we used internal primers for the sequencing reaction as well as those used for amplification. The primers sequences mapped in figure 2 are:
F27:GGTAAGTAGAGGTATAGCGGTAGTGTG,
R27:CTTAGCCCATGTTTTACACCTTTAAG
comp_R27:CTTAGCCCATGTTTTACACCTTTAAG,
in_aluF:TAGGCATTGCCAAAAGAACA, R3:GGATTCAATGAAGGAGAATGG,
R18:AACCTGGCTGCTAGCATCTG. The fragments sequenced from within the unique viral region were amplified using the cycling parameters as the first round of junction PRC, and the primers used on contig ABRT02259801.1 are F1:TGAGGGTTGCAGGTTGGTTT,
R1:CCCGATATTCATCAGGCCCC; F2:TTCTGCAGTAAGGGCGAGTG,
R2:TTCCTTGCTTCGGCATCCAT, F3:GATCTGCAACGGACAGGTGA,
R3:AGCACACTGTCAAACCCACA, F4:CTTCCGTGTGAACTCCTGCT,
R4:CTTTCAGCCACACAACTGCC, F5:GGCTTGGCCTCCTGATTTCT,
R5:GCCCTTTAGCAGCTGTACCA. Primer pairs for fragments on contig ABRT02337366.1 are F6:CTAGCTGGGTGCACAGTCTC,
R6:GACCCAAGAGAAGACGAGCC; F7:TTGGAATGCCATCTGTCGCT,
R7:TTGGGAGGAAGCTGAACCAC. Sequence for all amplified fragments can be found in supporting information S1, and have been assigned the Genbank accession ID KJ410306.

Alignment and Phylogenetic Reconstruction

Amino acid alignments were initially generated using the Mega 5.0 implementation of Muscle, and thereafter adjusted manually, including the removal of highly divergent (therefore unalignable) blocks. Tests for the best-fit evolutionary model were performed in Prottest 3.0, and chosen according to the Akaike information criterion. Initially, trees were estimated in PhyML, but comparison with RaxML indicated that the latter software was capable of finding more likely tree. The final phylogenies were estimated by maximum likelihood in RaxML and statistical support for the trees was calculated via 1000 nonparametric bootstrap replicates. The phylogenetic trees in figures 1A, 3A and 4C represent a 639 amino acid alignment of a conserved region in DNA polymerase for 144 Herpesviridae taxa (including the HVLs), and the best model was LG+G. To more accurately examine the relationships within each subfamily, taxonomic coverage was sacrificed in exchange for viruses for which a larger number of genes had already been sequenced. A 5095 long concatenated amino acid alignment of DNA polymerase, glycoprotein B, Helicase/Primase, ssDNA binding protein, transport protein and major capsid protein was used to generate the gammaherpesvirus phylogeny of 18 taxa (figure 4B and 4B). Best models for each partition were LG+G+F but fixed frequency estimates for glycoprotein B and DNA polymerase. For the Betaherpesvirus tree in figure 1B, we used a concatenated alignment for 16 taxa of terminase, large tegument, uracil-DNA glycosylase, kinase, capsid protein and helicase. The best model in each case was LG+G with empirical frequency estimates for terminase, large tegument, and helicase. Lastly, a phylogeny that exploited more gammaherpesvirus sequences from the database was also constructed. This tree did not compromise as much taxonomic coverage as for DNA polymerase (73 taxa), but using two genes in the alignment allowed for better-supported nodes than in the initial tree intended for quick taxonomic placement. In this case, we used a concatenation of Glycoprotein B and DNA polymerase genes in an 876 amino acid alignment with LG+G as the best model in both cases. The phylogeny of captured parvovirus *rep* genes (figure S2) in herpesviruses was reconstructed from a 281 amino acid alignment of the *rep* NS1 domain from 20 taxa, with model LG+G+F.

References

1. Lavergne A, de Thoisy B, Pouliquen J-F, Ruiz-García M, Lacoste V (2011) Partial molecular characterisation of New World non-human primate lymphocryptoviruses. Infect Genet Evol 11: 1782–1789. doi:10.1016/j.meegid.2011.07.017.
2. McGeoch DJ, Gatherer D, Dolan A (2005) On phylogenetic relationships among major lineages of the Gammaherpesvirinae. J Gen Virol 86: 307–316. doi:10.1099/vir.0.80588-0.
3. McGeoch DJ, Rixon FJ, Davison AJ (2006) Topics in herpesvirus genomics and evolution. Virus Res 117: 90–104. doi:10.1016/j.virusres.2006.01.002.
4. McGeoch DJ, Davison AJ, Dolan A, Gatherer D, Sevilla-Reyes EE (2008) Molecular Evolution of the

Herpesvirales. In: Domingo E, Holland JJ, editors. Origin and Evolution of Viruses. Academic Press. pp. 447–475. doi:10.1016/B978-0-12-374153-0.00020-5.
5. Fields BN, Knipe DM, Howley PM (2007) Fields Virology. 5th ed. Philadelphia: Lippincott Williams and Wilkins.
6. Tischer BK, Osterrieder N (2010) Herpesviruses--a zoonotic threat? Vet Microbiol 140: 266–270. doi:10.1016/j.vetmic.2009.06.020.
7. Katzourakis A, Gifford RJ (2010) Endogenous viral elements in animal genomes. PLoS Genet 6: e1001191.
8. Feschotte C, Gilbert C (2012) Endogenous viruses: insights into viral evolution and impact on host biology. Nat Rev Genet 13: 283–296. doi:10.1038/nrg3199.
9. Hurley EA, Agger S, McNeil JA, Lawrence JB, Calendar A, et al. (1991) When Epstein-Barr virus persistently infects B-cell lines, it frequently integrates. J Virol 65: 1245–1254.
10. Morissette G, Flamand L (2010) Herpesviruses and chromosomal integration. J Virol 84: 12100–12109. doi:10.1128/JVI.01169-10.
11. Kaufer BB, Jarosinski KW, Osterrieder N (2011) Herpesvirus telomeric repeats facilitate genomic integration into host telomeres and mobilization of viral DNA during reactivation. J Exp Med 208: 605–615. doi:10.1084/jem.20101402.
12. Delecluse HJ, Hammerschmidt W (1993) Status of Marek's disease virus in established lymphoma cell lines: herpesvirus integration is common. J Virol 67: 82–92.
13. Delecluse HJ, Schüller S, Hammerschmidt W (1993) Latent Marek's disease virus can be activated from its chromosomally integrated state in herpesvirus-transformed lymphoma cells. EMBO J 12: 3277–3286.
14. Arbuckle JH, Pantry SN, Medveczky MM, Prichett J, Loomis KS, et al. (2013) Mapping the telomere integrated genome of human herpesvirus 6A and 6B. Virology 442: 3–11. doi:10.1016/j.virol.2013.03.030.
15. Arbuckle JH, Medveczky MM, Luka J, Hadley SH, Luegmayr A, et al. (2010) The latent human herpesvirus-6A genome specifically integrates in telomeres of human chromosomes in vivo and in vitro. PNAS 107: 5563–5568. doi:10.1073/pnas.0913586107.
16. Pellett PE, Ablashi D V., Ambros PF, Agut H, Caserta MT, et al. (2012) Chromosomally integrated human herpesvirus 6: Questions and answers. Rev Med Virol 22: 144–155. doi:10.1002/rmv.715.
17. Delwart E (2013) A Roadmap to the Human Virome. PLoS Pathog 9: e1003146. doi:10.1371/journal.ppat.1003146.
18. Wilkie GS, Davison AJ, Watson M, Kerr K, Sanderson S, et al. (2013) Complete Genome Sequences of Elephant Endotheliotropic Herpesviruses 1A and 1B Determined Directly from Fatal Cases. J Virol 87: 6700–6712. doi:10.1128/JVI.00655-13.
19. Ehlers B, Ochs A, Leendertz F, Goltz M, Boesch C, et al. (2003) Novel simian homologues of Epstein-Barr virus. J Virol 77: 10695–10699.
20. Kumar S, Subramanian S (2002) Mutation rates in mammalian genomes. PNAS 99: 803–808. doi:10.1073/pnas.022629899.
21. Pace JK, Gilbert C, Clark MS, Feschotte C (2008) Repeated horizontal transfer of a DNA transposon in mammals and other tetrapods. PNAS 105: 17023–17028. doi:10.1073/pnas.0806548105.
22. Perelman P, Johnson WE, Roos C, Seuánez HN, Horvath JE, et al. (2011) A molecular phylogeny of living primates. PLoS Genet 7: e1001342. doi:10.1371/journal.pgen.1001342.
23. Holmes EC (2011) The Evolution of Endogenous Viral Elements. Cell Host Microbe 10: 368–377. doi:10.1016/j.chom.2011.09.002.
24. Aswad A, Katzourakis A (2012) Paleovirology and virally derived immunity. Trends Ecol Evol 27: 627–636. doi:10.1016/j.tree.2012.07.007.
25. Tadagaki K, Nakano K, Yamanishi K (2005) Human herpesvirus 7 open reading frames U12 and U51 encode functional beta-chemokine receptors. J Virol 79: 7068–7076. doi:10.1128/JVI.79.11.7068-7076.2005.
26. Thomson BJ, Efstathiou S, Honess RW (1991) Acquisition of the human adeno-associated virus type-2 rep gene by human herpesvirus type-6. Nature 351: 78–80. doi:10.1038/351078a0.
27. McGeoch DJ, Dolan A, Ralph AC (2000) Toward a Comprehensive Phylogeny for Mammalian and Avian Herpesviruses. J Virol 74: 10401–10406. doi:10.1128/JVI.74.22.10401-10406.2000.
28. Ehlers B, Dural G, Yasmum N, Lembo T, de Thoisy B, et al. (2008) Novel mammalian herpesviruses and lineages within the Gammaherpesvirinae: cospeciation and interspecies transfer. J Virol 82: 3509–3516. doi:10.1128/JVI.02646-07.
29. McGeoch DJ (2001) Molecular evolution of the gamma-Herpesvirinae. Philos Trans R Soc Lond B Biol Sci 356: 421–435. doi:10.1098/rstb.2000.0775.
30. Gerner CS, Dolan A, McGeoch DJ (2004) Phylogenetic relationships in the Lymphocryptovirus genus of the Gammaherpesvirinae. Virus Res 99: 187–192. doi:10.1016/j.virusres.2003.10.011.
31. Davison AJ, Eberle R, Ehlers B, Hayward GS, McGeoch DJ, et al. (2009) The order Herpesvirales. Arch Virol 154: 171–177. doi:10.1007/s00705-008-0278-4.
32. Ehlers B, Spieß K, Leendertz F, Peeters M, Boesch C, et al. (2010) Lymphocryptovirus phylogeny and the origins of Epstein-Barr virus. J Gen Virol 91: 630–642. doi:10.1099/vir.0.017251-0.
33. Crawford DH (2001) Biology and disease associations of Epstein-Barr virus. Philos Trans R Soc Lond B

Biol Sci 356: 461–473. doi:10.1098/rstb.2000.0783.
34. Parkin DM (2011) Cancers attributable to infection in the UK in 2010. Br J Cancer 105 Suppl: S49–S56. doi:10.1038/bjc.2011.484.
35. Ramasubramanyan S, Kanhere A, Osborn K, Flower K, Jenner RG, et al. (2012) Genome-wide analyses of Zta binding to the Epstein-Barr virus genome reveals interactions in both early and late lytic cycles and an epigenetic switch leading to an altered binding profile. J Virol 86: 12494–12502. doi:10.1128/JVI.01705-12.
36. Fujii K, Yokoyama N, Kiyono T, Kuzushima K, Homma M, et al. (2000) The Epstein-Barr Virus Pol Catalytic Subunit Physically Interacts with the BBLF4-BSLF1-BBLF2/3 Complex. J Virol 74: 2550–2557. doi:10.1128/JVI.74.6.2550-2557.2000.
37. El-Guindy A, Ghiassi-Nejad M, Golden S, Delecluse H-J, Miller G (2013) Essential role of Rta in lytic DNA replication of Epstein-Barr virus. J Virol 87: 208–223. doi:10.1128/JVI.01995-12.
38. Chiu Y-F, Sugden B, Chang P-J, Chen L-W, Lin Y-J, et al. (2012) Characterization and intracellular trafficking of Epstein-Barr virus BBLF1, a protein involved in virion maturation. J Virol 86: 9647–9655. doi:10.1128/JVI.01126-12.
39. Ressing ME, Horst D, Griffin BD, Tellam J, Zuo J, et al. (2008) Epstein-Barr virus evasion of CD8(+) and CD4(+) T cell immunity via concerted actions of multiple gene products. Semin Cancer Biol 18: 397–408. doi:10.1016/j.semcancer.2008.10.008.
40. Cohen JI, Lekstrom K (1999) Epstein-Barr virus BARF1 protein is dispensable for B-cell transformation and inhibits alpha interferon secretion from mononuclear cells. J Virol 73: 7627–7632.
41. Hislop AD, Ressing ME, van Leeuwen D, Pudney VA, Horst D, et al. (2007) A CD8+ T cell immune evasion protein specific to Epstein-Barr virus and its close relatives in Old World primates. J Exp Med 204: 1863–1873. doi:10.1084/jem.20070256.
42. Kvansakul M, Wei AH, Fletcher JI, Willis SN, Chen L, et al. (2010) Structural basis for apoptosis inhibition by Epstein-Barr virus BHRF1. PLoS Pathog 6: e1001236. doi:10.1371/journal.ppat.1001236.
43. Lake CM, Hutt-Fletcher LM (2004) The Epstein-Barr virus BFRF1 and BFLF2 proteins interact and coexpression alters their cellular localization. Virology 320: 99–106. doi:10.1016/j.virol.2003.11.018.
44. Gonnella R, Farina A, Santarelli R, Raffa S, Feederle R, et al. (2005) Characterization and intracellular localization of the Epstein-Barr virus protein BFLF2: interactions with BFRF1 and with the nuclear lamina. J Virol 79: 3713–3727. doi:10.1128/JVI.79.6.3713-3727.2005.
45. Pavlova S, Feederle R, Gärtner K, Fuchs W, Granzow H, et al. (2013) An Epstein-Barr virus mutant produces immunogenic defective particles devoid of viral DNA. J Virol 87: 2011–2022. doi:10.1128/JVI.02533-12.
46. Whitehurst CB, Vaziri C, Shackelford J, Pagano JS (2012) Epstein-Barr virus BPLF1 deubiquitinates PCNA and attenuates polymerase η recruitment to DNA damage sites. J Virol 86: 8097–8106. doi:10.1128/JVI.00588-12.
47. Massa M, Mazzoli F, Pignatti P, De Benedetti F, Passalia M, et al. (2002) Proinflammatory responses to self HLA epitopes are triggered by molecular mimicry to Epstein-Barr virus proteins in oligoarticular juvenile idiopathic arthritis. Arthritis Rheum 46: 2721–2729. doi:10.1002/art.10564.
48. Rezaee SAR, Cunningham C, Davison AJ, Blackbourn DJ (2006) Kaposi's sarcoma-associated herpesvirus immune modulation: an overview. J Gen Virol 87: 1781–1804. doi:10.1099/vir.0.81919-0.
49. Kelly GL, Strasser A (2011) The essential role of evasion from cell death in cancer. Adv Cancer Res 111: 39–96. doi:10.1016/B978-0-12-385524-4.00002-7.
50. Davison AJ, Stow ND (2005) New genes from old: redeployment of dUTPase by herpesviruses. J Virol 79: 12880–12892. doi:10.1128/JVI.79.20.12880-12892.2005.
51. Hart J, Ackermann M, Jayawardane G, Russell G, Haig DM, et al. (2007) Complete sequence and analysis of the ovine herpesvirus 2 genome. J Gen Virol 88: 28–39. doi:10.1099/vir.0.82284-0.
52. Ahuja SK, Murphy PM (1993) Molecular piracy of mammalian interleukin-8 receptor type B by herpesvirus saimiri. J Biol Chem 268: 20691–20694.
53. Glaunsinger B, Ganem D (2004) Lytic KSHV Infection Inhibits Host Gene Expression by Accelerating Global mRNA Turnover. Mol Cell 13: 713–723. doi:10.1016/S1097-2765(04)00091-7.
54. Glaunsinger B, Chavez L, Ganem D (n.d.) The Exonuclease and Host Shutoff Functions of the SOX Protein of Kaposi's Sarcoma-Associated Herpesvirus Are Genetically Separable. J Virol.
55. Zhu FX, King SM, Smith EJ, Levy DE, Yuan Y (2002) A Kaposi's sarcoma-associated herpesviral protein inhibits virus-mediated induction of type I interferon by blocking IRF-7 phosphorylation and nuclear accumulation. PNAS 99: 5573–5578. doi:10.1073/pnas.082420599.
56. Damania B, Jeong JH, Bowser BS, DeWire SM, Staudt MR, et al. (2004) Comparison of the Rta/Orf50 Transactivator Proteins of Gamma-2-Herpesviruses. J Virol 78: 5491–5499. doi:10.1128/JVI.78.10.5491-5499.2004.
57. De León Vázquez E, Kaye KM (2011) The internal Kaposi's sarcoma-associated herpesvirus LANA regions exert a critical role on episome persistence. J Virol 85: 7622–7633. doi:10.1128/JVI.00304-11.
58. Kaul R, Verma SC, Robertson ES (2007) Protein complexes associated with the Kaposi's sarcoma-associated herpesvirus-encoded LANA. Virology 364: 317–329. doi:10.1016/j.virol.2007.03.010.
59. Watanabe T, Sugaya M, Atkins AM, Aquilino EA, Yang A, et al. (2003) Kaposi's sarcoma-associated herpesvirus latency-associated nuclear antigen prolongs the life span of primary human umbilical vein

endothelial cells. J Virol 77: 6188–6196.
60. Fujimuro M, Hayward SD (2003) The latency-associated nuclear antigen of Kaposi's sarcoma-associated herpesvirus manipulates the activity of glycogen synthase kinase-3beta. J Virol 77: 8019–8030.
61. Cai Q, Xiao B, Si H, Cervini A, Gao J, et al. (2012) Kaposi's sarcoma herpesvirus upregulates Aurora A expression to promote p53 phosphorylation and ubiquitylation. PLoS Pathog 8: e1002566. doi:10.1371/journal.ppat.1002566.
62. Lacoste V, Verschoor EJ, Nerrienet E, Gessain A (2005) A novel homologue of Human herpesvirus 6 in chimpanzees. J Gen Virol 86: 2135–2140. doi:10.1099/vir.0.81034-0.
63. Arbuckle JH, Medveczky PG (2011) The molecular biology of human herpesvirus-6 latency and telomere integration. Microbes Infect 13: 731–741.
64. Riethman H (2008) Human telomere structure and biology. Annu Rev Genomics Hum Genet 9: 1–19. doi:10.1146/annurev.genom.8.021506.172017.
65. Huang Y, Hidalgo-Bravo A, Zhang E, Cotton VE, Mendez-Bermudez A, et al. (2013) Human telomeres that carry an integrated copy of human herpesvirus 6 are often short and unstable, facilitating release of the viral genome from the chromosome. Nucleic Acids Res: gkt840 – . doi:10.1093/nar/gkt840.
66. Thomson BJ, Weindler FW, Gray D, Schwaab V, Heilbronn R (1994) Human Herpesvirus 6 (HHV-6) Is a Helper Virus for Adeno-Associated Virus Type 2 (AAV-2) and the AAV-2 rep Gene Homologue in HHV-6 Can Mediate AAV-2 DNA Replication and Regulate Gene Expression. Virology 204: 304–311.
67. Perry GH, Louis EE, Ratan A, Bedoya-Reina OC, Burhans RC, et al. (2013) Aye-aye population genomic analyses highlight an important center of endemism in northern Madagascar. PNAS 110: 5823–5828. doi:10.1073/pnas.1211990110.
68. Renzette N, Bhattacharjee B, Jensen JD, Gibson L, Kowalik TF (2011) Extensive genome-wide variability of human cytomegalovirus in congenitally infected infants. PLoS Pathog 7: e1001344. doi:10.1371/journal.ppat.1001344.
69. Christou L (2011) The global burden of bacterial and viral zoonotic infections. Clin Microbiol Infect 17: 326–330. doi:10.1111/j.1469-0691.2010.03441.x.
70. Chang H, Wachtman LM, Pearson CB, Lee J-S, Lee H-R, et al. (2009) Non-human primate model of Kaposi's sarcoma-associated herpesvirus infection. PLoS Pathog 5: e1000606. doi:10.1371/journal.ppat.1000606.
71. Huff JL, Barry PA (2003) B-virus (Cercopithecine herpesvirus 1) infection in humans and macaques: potential for zoonotic disease. Emerg Infect Dis 9: 246–250. doi:10.3201/eid0902.020272.
72. Wilkie GS, Davison AJ, Watson M, Kerr K, Sanderson S, et al. (2013) Complete genome sequences of elephant endotheliotropic herpesviruses 1A and 1B determined directly from fatal cases. J Virol: JVI.00655–13 – . doi:10.1128/JVI.00655-13.
73. Horie M, Honda T, Suzuki Y, Kobayashi Y, Daito T, et al. (2010) Endogenous non-retroviral RNA virus elements in mammalian genomes. Nature 463: 84–87. doi:10.1038/nature08695.
74. 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. doi:10.1371/journal.ppat.1001030.
75. Cornelis G, Heidmann O, Bernard-Stoecklin S, Reynaud K, Véron G, et al. (2012) Ancestral capture of syncytin-Car1, a fusogenic endogenous retroviral envelope gene involved in placentation and conserved in Carnivora. PNAS 109: E432–E441. doi:10.1073/pnas.1115346109.

Chapter Two: Unusual large dsDNA viruses discovered among genome data of 15 different fish

Abstract

Viruses pervasively infect all cellular life and fulfill a crucial ecosystem role as agents of death in every habitable environment on earth. Disease-causing viruses often result from cross-species transmissions such as HIV and Ebola that are asymptomatic in the original host. Despite this, our knowledge of viral biodiversity is biased towards clinical and veterinary samples, leaving us vulnerable to emerging zoonotic infections. In this study we reveal a new family of dsDNA virus discovered in publically available host genome data. These novel viruses are associated with at least 15 different fish, including a full-length genome of a new salmon virus and partial genomes in other important fish species. Our study demonstrates that publically accessible databases of host genomes are a source of novel viral diversity that can be mined bioinformatically under an evolutionary framework. This approach also enabled us to simultaneously identify sequences that likely represent ancient endogenous viral elements. Moreover, approximately half of the open reading frames we predict in the extant viruses do not exhibit similarity to any known genes. These results clearly demonstrate the utility of mining non-viral genome data for its viral content, allowing us to characterize both extinct and extant diversity. The identification of novel dsDNA viruses in this study will be important to basic virology research and studies on the origin of life, and the approach in general will have crucial applications in clinical, veterinary and plant epidemiology.

Author Summary

This study describes the discovery and characterization of a previously unknown group of viruses, solely through bioinformatic means. The sequences we identify are associated with a diverse range of 15 species of fish, several of which are economically important. Among the fragments we describe is 200,000 base-pair sequence that represents the whole genome of a virus associated with the Atlantic salmon, meaning this would be among the largest viruses that we know of. As well as features that suggest distant ancestry to herpesviruses, our genomic analysis reveals many predicted open reading frames that have no detectable similarity to any known genes. Our results highlight the value of exploiting vast quantities of next-generation sequencing data for viral research, using techniques from evolutionary biology and genomics. At a larger scale, our techniques can be used to identify and characterize viruses that could pose a threat to animal and human health, particularly in the case of cross-species transmissible viruses that my be asymptomatic in the reservoir host.

Introduction

Viruses are extremely diverse and abundant parasites that infect all life on earth including other viruses [1]. Unlike cells, viral genomes can be composed of either RNA or DNA (both single or double stranded) and reproduce via a diverse set of strategies that require hijacking cellular replicative machinery. This dependence on their hosts is a unifying feature of viruses that allows them to maintain minimal genomes with as few as two genes. For example, the smallest virus we know of is the *Porcine circovirus 1* which has a 1.7Kb genome enclosed in a 17nm capsid [2]. Indeed, their size was the criterion by which viruses were first identified, when the tobacco mosaic virus was discovered using a Chamberland filter over 120 years ago. However, while they are generally much smaller than cells, recently (accidentally) discovered 'giant viruses' have surpassed the smallest eukaryotes, both in terms of physical size and genome length [3].

Viral identification has traditionally involved low throughput techniques such as filtration, microscopy and cell culture that introduces severe biases, including the fact that viruses larger than porcelain filter pores were not known to exist until 2004 [4].The discovery of such 'giant viruses' is part of a wider trend of novel approaches to viral discovery in the genomic age [5]. Among the drivers of this methodological shift is the recognition of the important ecosystem roles of viruses [6], which drives the need to study (and therefore sequence) the total viral content of environmental samples. Moreover, the demand for high-through put techniques also came in the form of ambitious exploratory projects such as mapping the human virome [7]. The introduction of next generation sequencing (NGS) has facilitated the rise of culture-independent metagenomic techniques that are unencumbered by the lack of universal genetic markers, and immune to many of the biases in older approaches [8]. NGS has also resulted in the widespread availability of animal and plant genomes, that has in turn lead to the establishment of paleovirology, a field that studies the remnants of viruses that have integrated into the genomes of their hosts.

In as much as it aims to uncover viral sequences from large genomic databases, paleovirology shares similar methods to viral discovery through metagenomics. Although far more viruses are identified through metagenomics [9], the systematic investigation of host genomes in paleovirology has revealed many novel viral lineages that often represent long extinct viruses [10–16]. Not only do such studies reveal ancient groups that could have contemporary relatives, they have involved the development of methods to mine genomic data under an evolutionary framework. Moreover, while most endogenous

viral elements (EVEs) are non-functional ‘fossils’ of viruses past, there is increasing recognition of their active influence on host genomes [17–19]. Therefore, alongside environmental sampling and the large-scale virome projects, paleovirology is an important and complementary part of the emergent metagenomic approach to virology, offering solutions to an overlapping set of scientific questions with transferable techniques.

It has previously been suggested that the deployment of a paleovirological approach to viral discovery could be particularly useful in the identification of novel herpesviruses [20]. This proposal supports the findings of our previous study that simultaneously identified paleoviral data and novel herpesviruses among primate host genomes [12]. This raised the intriguing prospect of reapplying our approach to investigate unexplored areas of herpesvirus diversity, where the identification of novel viruses is the main goal rather than a secondary result. Our previous investigation uncovered both known and new viruses from *Herpesviridae* [12], which is the most well-studied of the three families in the order *Herpesvirales*. Compared to hundreds of *Herpesviridae,* only 14 highly diverse members of the fish-infecting *Alloherpesviridae* have been identified thus far [21]. This is despite the fact that viral diseases are a major threat to the aquaculture industries that are addressing the rising global demand for seafood as an alternative to unsustainable fishing practices [22].

In this study, we aimed to describe novel viruses by mining fish genomes for similarity to alloherpesviruses. The analyses in this study reveal virus-like sequences among the genome records of 15 species of fish from 9 different teleost families. These viral sequences share phylogenetic and genomic similarity to alloherpesviruses, and include among them a full-length viral genome associated with the Atlantic salmon (*Salmo salar*). The preliminary characterization of these viruses also indicates genetic similarity to genes from at least five other viral families, as well as dozens of predicted open reading frames (ORFs) with no detectable similarity to any known protein-coding genes. Finally, there is also evidence to suggest that at least some of the sequences represent EVEs that have heritably integrated into the host genome. In addition to recognizing the importance of identifying these viral sequences specifically, we also conclude that the results strongly advocate in favor of broader and wider scale of genomic database mining.

Results

Database mining and phylogenetic reconstruction

Virus-like sequences were identified in 15 different species in a search of 54 teleost fish genomes for similarity to the DNA polymerase of *Ictalurid Herpesvirus 1* (IcHV1). Using tBLASTn, we initially considered the top 21 hits that aligned to at least 60% of the query, shortlisting only the top scoring sequences for each species, with the exception of two distinct strains of *Oryzias latipes* that both contained two distinct hits (table S1). The average amino-acid identity was 27% between IcHV1 and these 21 hits. We then used 3 sequences from the 11 different species in a BLASTn to reveal similar sequences in an additional 4 species of fish that were an average of 60% identical at the nucleotide level (table S1). After aligning the shortlisted contigs and delineating the DNA polymerase gene, we performed a BLASTn search against each individual fish genome to retrieved all contigs in the database that contain the DNA polymerase sequence under investigation (table S2).

Phylogenetic reconstruction revealed that the sequences discovered in fish genome records are a monophyletic group related to other dsDNA viruses. Based on midpoint rooting, the fish sequences are most closely related to the *Alloherpesviridae,* with a posterior probability of 100% (fig. 1A). Considering the patristic distances between the various dsDNA viral families represented in the tree, the sequences are unlikely to be are part of the *Alloherpesviridae*, but rather represent a distinct clade of previously undescribed viruses. We also investigated the relationships within this clade, using a longer alignment that only considered the new sequences (fig. 1B). This is because the alignment used to estimate the tree in figure 1A was limited to a small conserved region and does not offer statistically robust information on the relationships among these sequences. In contrast, the tree reconstructed from a longer region of DNA polymerase offers robust support for all nodes in the tree (fig. 1B). Unlike other dsDNA viruses such as the *Herpesviridae,* the topology revealed is only somewhat congruent to that of their hosts. For example, the sequences from Salmonidae species and Cichlidae are both monophyletic groups, although the latter clade also contains sequences identified in genomes of Gobidae and Cyprinodontiformes fish (fig. 1B). On the onther hand, there is a broad topological incongruence to the host tree, such as the fact that Salmonidae and Esociformes fish are basal to the Cichlidae [23], but this is not the case with their associated viral sequences (fig. 1B).

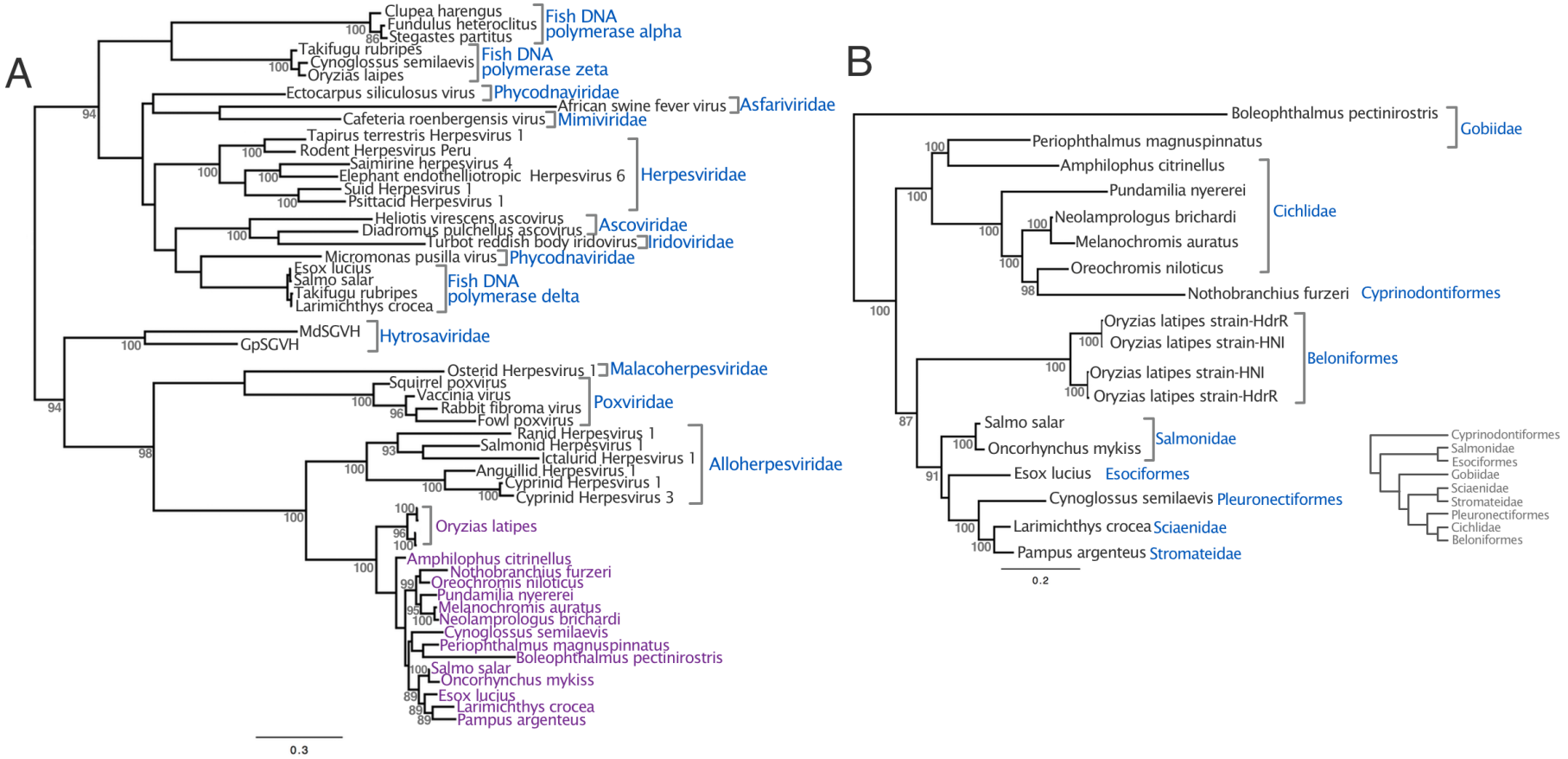


Figure 1 Phylogenetic analysis of viral-like sequences in fish

Both panels depict midpoint rooted Bayesian phylogenetics trees reconstructed from an alignment of DNA polymerase. The branch lengths represent the number of substitutions per site and the numbers at each node represent posterior probabilities above 85%. Panel A: As well as the sequences under investigation (annotated in purple) the 233 amino acid alignment included viruses representing eight dsDNA virus families, as well as delta, zeta and epsilon fish DNA polymerases. Panel B: An extended 2,904-nucleotide alignment of the new sequences without other viral groups intended to obtain more robust support for the topology within the clade. The clades are annotated according to the family of the fish species in whose genome the viral-like data was identified.

Comparative analysis and sequence characterization

The synteny among contigs identified in all 15 species of fish were investigated using the genome aligner MAUVE [24] (fig. 2). Although the length of the contigs found varied greatly, the analysis revealed sizable co-linear blocks among the viral sequences, particularly near DNA polymerase. However, the MAUVE alignment also showed that the contigs are highly rearranged relative to one another, and the length of co-linear blocks is highly variable. We also identified repetitive elements (LINEs, SINEs and endogenous retroviruses, hereafter referred to collectively as REs) in the contigs of three of the species (*Esox Lucius*, *Larimichthys crocea* and *Oncorhynchus mykiss*). In the case of *Esox Lucius* and *Larimichthys crocea,* we found contigs that did not contain the same profile of REs. For the pair of *Larimichthys crocea* sequences, the RE is visibly interfering with the colinearity between the contigs (fig. 2).

ORF prediction using GeneMarkS was performed on the DNA polymerase-containing contigs from four species with the largest contigs (one each from *Salmoniformes*, and *Cyprinodontiformes* and two *Perciformes*) (fig. 3, table S3-S6). The majority of predicted ORFs are most similar to genes in *Alloherpesviridae*. However, we also identified ORFs with similarity to other dsDNA viral groups, with irridoviral genes being the second most common top hit (fig. 3). Across the four annotated contigs, the ORFs are 1,165nt long on average with a median length of 834nt. The longest ORF is 5,916nt, although no similarity could be detected to any known gene product using BLAST, which was also the case for many of the predicted ORFs. The longest recognizable ORFs were the DNA polymerase-like sequences in *Nothobranchius furzeri* and *Amphilophus citrinellus* (4,923nt and 4,845nt, respectively). The contig with the highest number of predicted genes (115) was identified in the *Salmo salar* genome record, which also contains a flanking pair of inverted repeat regions ~194Kbp apart (fig. 3). This genomic structure is also observed in herpesviruses, which are also between 100 and 200Kbp, suggesting that the contig represents a near-complete viral genome; although all the gene-containing regions appear are present, one of the repetitive flanks is truncated by the end of the contig, making it half as long as the other (fig. 3).

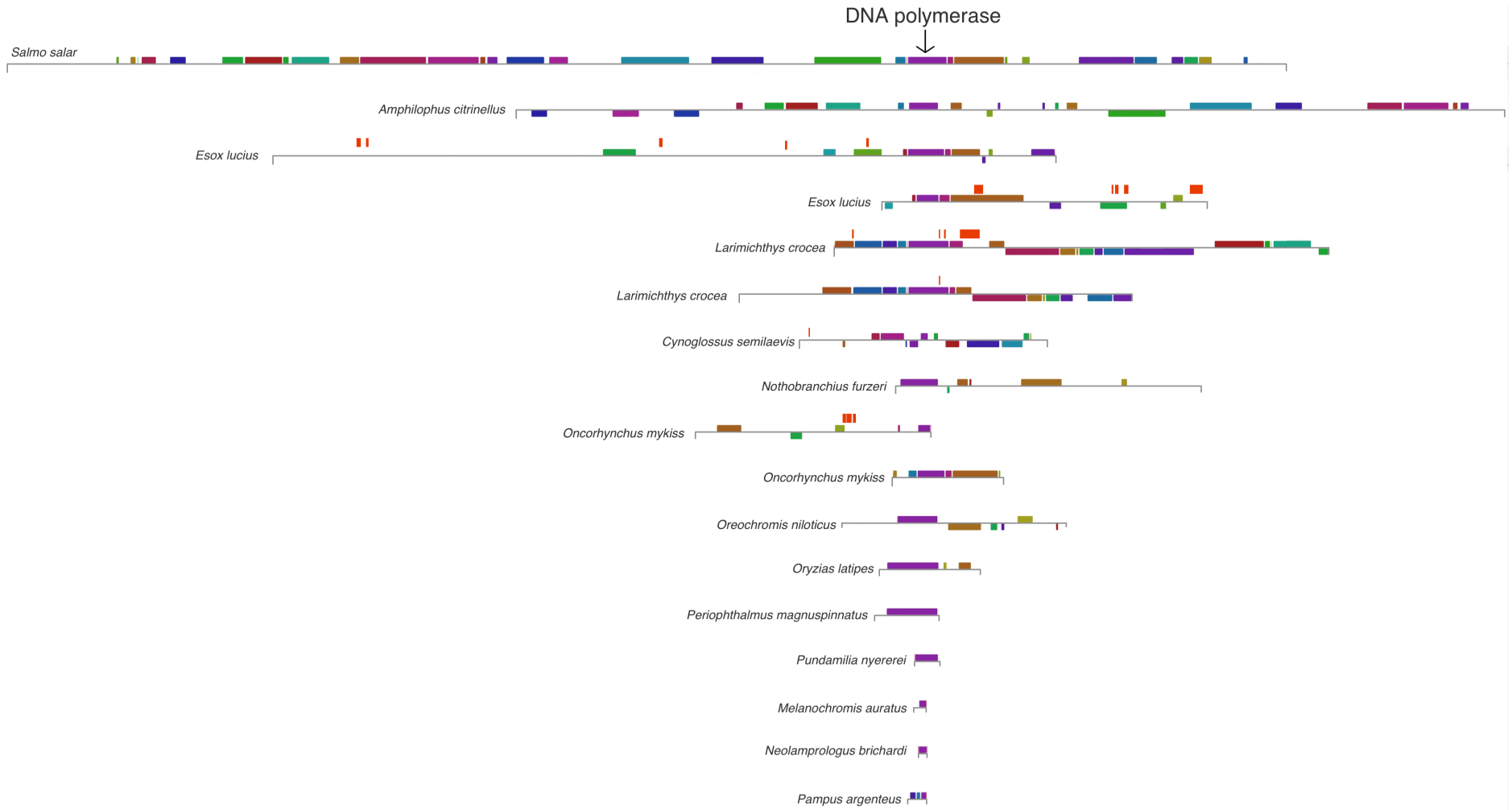


Figure 2 Annotation of viral-like contigs identified in fish genomes

A schematic diagram depicting the blocks of co-linear sequence similarity among the viral contigs, which are drawn relative to the Salmo salar sequence. Homologous blocks across the different sequences are indicated by the same color, and shown below the line if a co-linear block is found in reverse orientation. The alignment is centered at the midpoint of DNA polymerase, which is located in slightly different places for each sequence within the co-linear block. Repetitive elements (not simple repeats) are indicated as red blocks above the representation of each contig.

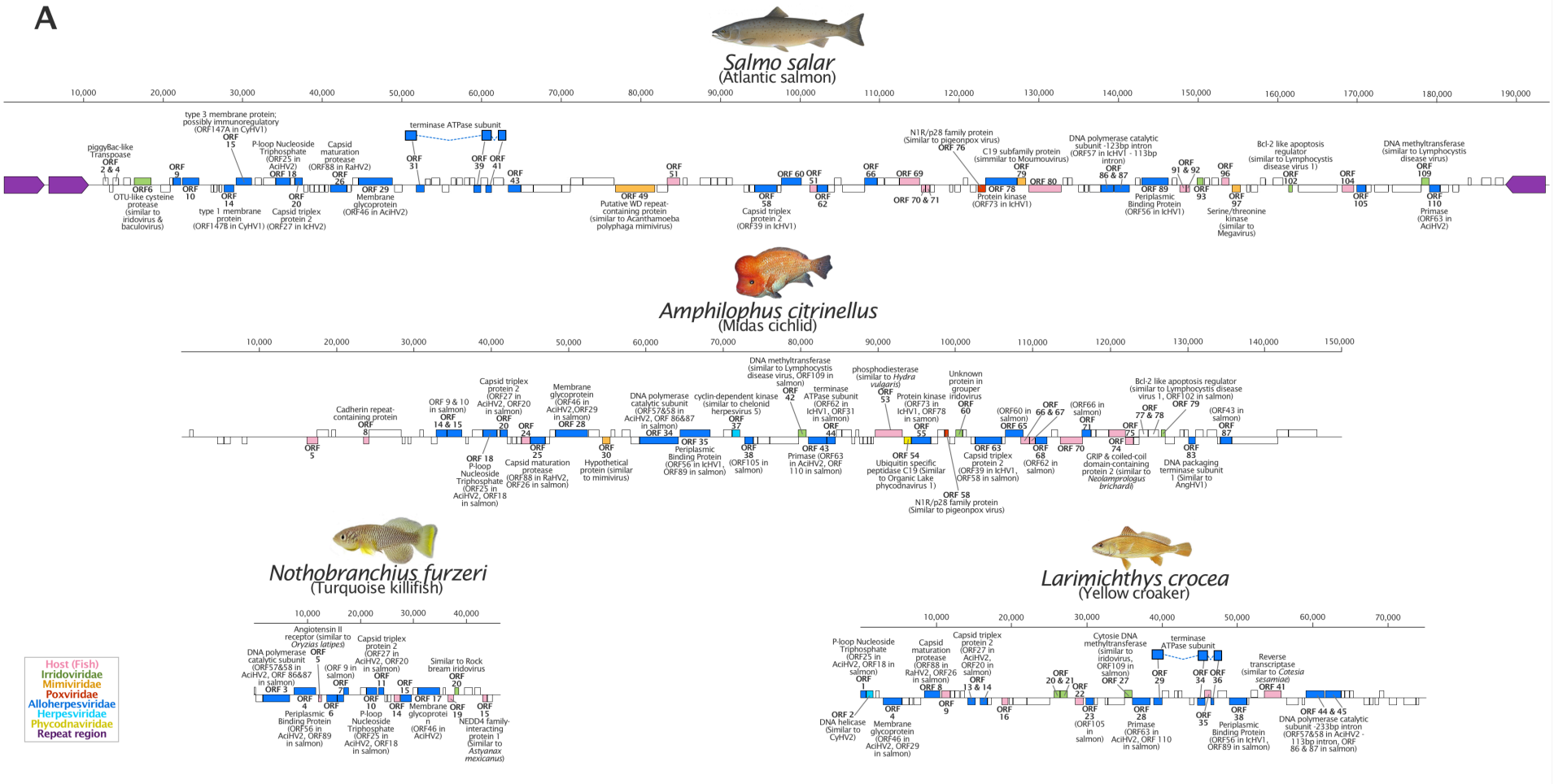


Figure 3 Analysis of synteny for the viral-like contigs identified in fish

Four of the viral sequences detected in fish genomes with predicted open reading frames (ORFs) represented as boxes. ORFs in the forward and reverse orientation are indicated above and below the line, respectively. ORFs without any detectable similarity to known proteins are not labeled, and those with similarity to unnamed proteins are only indicated by their ORF ID. Boxes are colored according to the taxonomic group of the most similar protein.

Discussion

The results of this investigation reveal viral-like sequences identified in open-access genome data of 15 species of teleost fish. This was achieved using common bioinformatic tools and evolutionary techniques in a methodological approach that was developed for the discovery and characterization endogenous viral elements. Our data conclusively demonstrate the utility of this approach to viral discovery, which has uncovered a previously unknown viral family. Through phylogenetic reconstruction, we show that this new group is related to alloherpesviruses and other large dsDNA viruses. Crucially, our results include a large contig identified in the *S. salar* genome data that appears to represent the full-length of the viral coding region, with a high depth of read-coverage of differently sized libraries (fig. 4). This will be instrumental for future characterization of this unusual clade of viruses, several of which were identified in genome data of fish that have substantial conservation, research and commercial interest.

It should be noted that this characterization of these sequences as most closely related to alloherpesviruses would be invalidated if the true root of the tree is at the branch leading to the new clade, to the alloherepsviruses or between the two. We opted to root the tree at the midpoint as a conservative solution to the fact that the phylogenetic history of large dsDNA viruses has proved extremely difficult to reconstruct with accuracy, due to the complex history of their core genes (e.g. gene loss/gain) [25]. Moreover, it is not clear what the evolutionary relationship of large dsDNA viral polymerase is with similar hosts genes, which could be due to any number of independent acquisition events [26]. Nonetheless, we think that the alloherpesviruses are probably the most closely related group, since our preliminary genomic characterization shows that most predicted ORFs are most similar to alloherpesvirus genes (fig. 3).

More importantly that determining which is most closely related group to the new viruses, the results indicate we can rule out that these sequences are simply alloherpesviruses. First, there is a large patristic distance between the identified sequences and alloherpesviruses, which is visually comparable to the distance between other dsDNA families (fig. 1). Moreover, most of the similarity to alloherpesvirus genes was only partial and shared a low sequence identity. For example, in the *S. salar* contig, the average identity to alloherpesviruses proteins is 30%, with average gene coverage of 63% (table S1). Furthermore, as well as similarity to other viral groups (9 ORFs were similar to non-alloherpesviruses), 64/115 of the predicted *S. salar* ORFs did not exhibit any similarity to known genes. Interestingly, a similarly high proportion of unknown genes was observed in

the initial sequencing of the giant mimivirus genome, in which only 10% of the ORFs were identifiable [4].

The topology of the tree without other viral groups indicates that a cross-species transmission of these viruses may have occurred between *Periophthalmus magnuspinnatus* and *Amphilophus citrinellus*. This is because they form a monophyletic group with strong support despite being associated to different families of fish, although we cannot determine the direction of transfer (fig. 1B). Similarly, the tree topology suggests that the virus may have transmitted from cichlids to the African Killifish *Nothobranchius furzeri*, which is placed among a clade of sequences associated to cichlid species. More specifically, it is tempting to speculate that the transmission event occurred from the Nile tilapia *Oreochromis niloticus*, since the two sequences group with 98% probability. Moreover, both species are cultivated by humans for food and as a research model (tilapia and killifish, respectively), which could have offered more opportunity for transmission. The tree of these fish sequences also revealed that the two pairs of sequences identified in two *Oryzias latipes* strains do not group together, suggesting that each genome contains two different strains of the virus (fig. 1B).

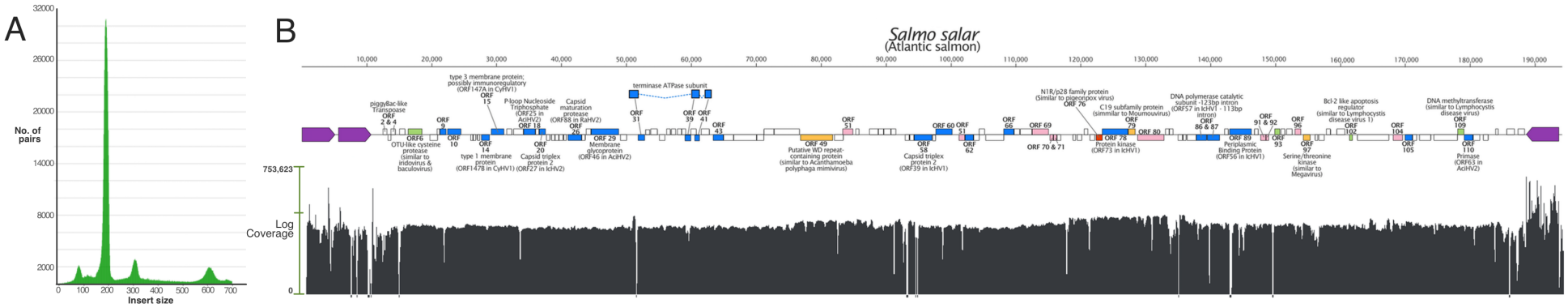


Figure 4 read-mapping analysis of *Salmo salar contig*

(A) The graph shown indicates the number of read pairs for each insert size. The peaks at 180, 300 and 600 correspond to the known sizes of libraries used in the sequencing project. Only high quality, well-aligned and paired reads were included in the coverage count. (B) a coverage graph across the length of the salmon contig. The X-axis represents the log coverage, with a horizontal bar indicating the mean at ~3000X. Short regions of zero coverage are indicated by a dash along the x axis, all of which are bridged by read pairs indicating that the contig is not erroneously assembled.

We have previously identified three viruses from primate genome data that belong to the *Herpesviridae* [12], where the data revealed an endogenous betaherpesvirus found in the Philippine tarsier. The evidence in support of this came from PCR confirmation of the host-virus junction, and that the genome contained many nonsense mutations as well as a number of host-specific mobile elements. Among the fish sequences in this study, we find a similar set of evidence that at least some of the sequences might be endogenous. For instance, the partial viral genome identified in the *Amphilophus citrinellus* genome (fig. 3) represents the last 150Kb of a 722Kb contig, although we cannot rule out that this is not an assembly error without identification and characterization of the junction. In *Esox luicus, Larimichthys crocea, Cynoglussus semilaevis* and *Oncorhynchus mykiss,* we have identified a number of mobile elements that indicate these sequences could be endogenous (fig. 2). Unlike the endogenous tarsier herpesvirus, in-frame stop codons or frame shifting indels are more difficult to determine because these sequences are not easily comparable to known genes. Nevertheless, if we consider the alignment of over 1,000 amino acids used to reconstruct the fish phylogeny (fig. 1B), there are a number frameshifts and stop codons in exactly the same species that contain mobile elements. Although not conclusive, the combination of both findings strongly suggests that these four species of fish are harboring endogenous representatives of this new viral clade. If confirmed, it would mean that through our evolutionarily informed data mining approach, these sequences represent the first group of viruses that were simultaneously identified along with their paleovirus counterparts.

The chief drawback to relying exclusively on draft assembly data is the certainty with which we can characterize the viral genomes. This is because we have to rely on imperfect genome assembly methods that can erroneously group reads into artificial contigs. One solution to this drawback would be to perform supplementary experiments by cell culture or targeted sequencing, but this introduces limitations in the form of additional time and cost. In order to achieve a higher level of confidence bioinformatically, a supplementary investigation into the raw sequence reads would be beneficial, with particular attention paid to repetitive regions within the virus. It would be important to ensure that there is high-coverage across such regions, represented by a combination paired-end reads of short and long insert sizes [27]. For example in the case of the largest contig identified within the salmon genome data, read mapping indicates a mean coverage of ~3000X, and there is a mixture of insert-sizes represented that bridge the

small number of short ambiguities (fig. 4). In other cases where there is much lower coverage or if major ambiguities persist, this technique could be bolstered with additional evidence from diagnostic targeted PCR sequencing to conclusively confirm 'suspicious' junctions as necessary.

The innovative sequencing technology that have enabled high-throughput viral discovery through metagenomics is beginning to reveal the vastness of unexplored viral biodiversity. In addition, it has also resulted in the growth of whole genome databases of viral plant and animal hosts, which gave rise to the large-scale exploration of ancient viruses through paleovirology. The methodology and objectives of both paleovirology and discovery through metagenomic are highly complementary to one another, which is an observation that has only recently been recognized [20]. We first recognized the value of this union of perspectives and techniques in the discovery of novel herpesvirus data, where we revealed sequences that were readily incorporated into existing taxonomy due to their indisputable homology to known viruses [12]. In this paper however, the technique has successfully demonstrated the capacity to identify an entirely new viral lineage, which like the recently described giant viruses, contain many unrecognizable predicted ORFs. In the case of these fish viruses, there is great downstream value of discovering novel diversity for epidemiological surveillance and disease management of both wild and farmed populations. Although the results of this analysis undoubtedly require further scrutiny, they exhibit the success of this approach that can be applied to a vast array of different context in basic science and industry.

Materials & Methods

Genome database mining

DNA polymerase protein sequences of four alloherpesviruses and two malacoherpesviruses were used as a query set against the NCBI nr, WGS, HTGS, EST & TSA. The search scope was limited to the 54 species of fish for which there are assembled genome sequence data. We retrieved sequences that aligned to at least 60% of the query (ictalurid orf57), choosing a single sequence per species, except for *Oryzias latipes,* for which there were two different strains. For the second round of searching, we used three representatives from the results of the previous round as search queries, guided by a preliminary phylogeny to avoid choosing redundantly similar sequences. Although all the full-length hits belonged to species identified in the previous round, all three of these searches revealed four additional species that were at least 50% identical (*Pampus argenteus, Cynoglossus semilaevis, Melanochromis auratus*, and

Neolamprologus brichardi). The Salmo salar accession that was eventually used (Table S1) does not correspond to the contig identified through BLAST search in this study. Rather, this contig was first identified in a similar but preliminary search conducted several years ago against an earlier draft of the genome record.

Alignment & Phylogenetics

The objective of the phylogeny was to determine the relationship of these viral DNA polymerase-like sequences to other viruses. To this end, we aligned representative polymerase sequences from a diverse range of dsDNA virus families, as well as host polymerases that also exhibit similarity to dsDNA virus polymerases. Because of this broad taxonomic strategy, the alignment was limited to the conserved catalytic domain. For the most divergent sequences, we only aligned the most conserved region of this domain. We used a Bayesian phylogenetic approach to reconstruct the tree using MrBayes [28] (10 million generations; 25%burn-in), and implemented the best evolutionary model according to the corrected Akaike Information criterion in ProtTest3 [29] (WAG+G).

Contig annotation, synteny mapping & visualization

Similarly to other studies on alloherpesvirus genomes [30], ORFs were identified using GeneMarkS [31]. The output GFF files were converted into Genbank records and visualized using MAUVE [24] with further graphical editing in Pixelmator 3.3. MAUVE was also used for a whole contig alignment of all of the species represented in the phylogeny, to identify 'locally co-linear blocks' with the *Salmo salar* contig as the reference. Repetitive elements in each contig were identified using RepeatMasker, excluding the reporting of simple repeats and low-complexity regions. The results were then visualized in IGV and transposed onto the ORF maps. The amino acid sequence for each of the predicted ORFs was used as a query for a BLASTp search against a database of all dsDNA viral proteins in the NCBI nr database. Only results with an e-value better than $1x10^{-4}$ were considered, and a table was constructed detailing the results for the top-scoring hit of each ORF. For ORFs that did not exhibit similarity to any dsDNA virus proteins, we performed a second more exhaustive search of the rest of the nr database. Read mapping for the salmon contig was performed using the BWA mem algorithm using the Salmon contig as a referenc sequence for 101 mate-paired libraries of the salmon genome project (NCBI Accession: PRJNA72713). The mapped reads were strictly filtered for erroneous alignment using Bamtools filter, leaving only those reads that mapped in a proper pair with an edit distance of less than 5.

Acknowledgments

We would like to thank Pakorn Aiewsakun and Tara Hurst for helpful comments on the manuscript. AK is funded by the Royal Society.

References

1. La Scola B, Desnues C, Pagnier I, Robert C, Barrassi L, Fournous G, et al. The virophage as a unique parasite of the giant mimivirus. Nature. Macmillan Publishers Limited. All rights reserved; 2008;455: 100–4. doi:10.1038/nature07218
2. Finsterbusch T, Mankertz A. Porcine circoviruses--small but powerful. Virus Res. 2009;143: 177–83. doi:10.1016/j.virusres.2009.02.009
3. Philippe N, Legendre M, Doutre G, Couté Y. Pandoraviruses: amoeba viruses with genomes up to 2.5 Mb reaching that of parasitic eukaryotes. Science (80-). 2013; Available: http://www.sciencemag.org/content/341/6143/281.short
4. Raoult D, Audic S, Robert C, Abergel C. The 1.2-megabase genome sequence of Mimivirus. Science (80-). 2004; Available: http://www.sciencemag.org/content/306/5700/1344.short
5. Roux S, Hallam SJ, Woyke T, Sullivan MB. Viral dark matter and virus-host interactions resolved from publicly available microbial genomes. Elife. eLife Sciences Publications Limited; 2015;4: e08490. doi:10.7554/eLife.08490
6. Suttle CA. Marine viruses--major players in the global ecosystem. Nat Rev Microbiol. 2007;5: 801–12. doi:10.1038/nrmicro1750
7. Delwart E. A Roadmap to the Human Virome. Racaniello V, editor. PLoS Pathog. Public Library of Science; 2013;9: e1003146. doi:10.1371/journal.ppat.1003146
8. Mokili JL, Rohwer F, Dutilh BE. Metagenomics and future perspectives in virus discovery. Curr Opin Virol. 2012;2: 63–77. doi:10.1016/j.coviro.2011.12.004
9. Edwards RA, Rohwer F. Viral metagenomics. Nat Rev Microbiol. 2005;3: 504–10. doi:10.1038/nrmicro1163
10. Katzourakis A, Gifford RRJ. Endogenous viral elements in animal genomes. PLoS Genet. 2010;6: e1001191. Available: http://dx.plos.org/10.1371/journal.pgen.1001191.g007
11. Katzourakis A, Tristem M, Pybus OG, Gifford RJ. Discovery and analysis of the first endogenous lentivirus. Proc Natl Acad Sci U S A. 2007;104: 6261–5. doi:10.1073/pnas.0700471104
12. Aswad A, Katzourakis A. The first endogenous herpesvirus, identified in the tarsier genome, and novel sequences from primate rhadinoviruses and lymphocryptoviruses. Feschotte C, editor. PLoS Genet. Public Library of Science; 2014;10: e1004332. doi:10.1371/journal.pgen.1004332
13. Taylor DJ, Leach RW, Bruenn J. Filoviruses are ancient and integrated into mammalian genomes. BMC Evol Biol. 2010;10: 193. doi:10.1186/1471-2148-10-193
14. Belyi VA, Levine AJ, Skalka AM. Unexpected inheritance: multiple integrations of ancient bornavirus and ebolavirus/marburgvirus sequences in vertebrate genomes. Buchmeier MJ, editor. PLoS Pathog. Public Library of Science; 2010;6: e1001030. doi:10.1371/journal.ppat.1001030
15. Gilbert C, Feschotte C. Genomic fossils calibrate the long-term evolution of hepadnaviruses. Emerman M, editor. PLoS Biol. Public Library of Science; 2010;8: 12. doi:10.1371/journal.pbio.1000495
16. Belyi VA, Levine AJ, Skalka AM. 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. 2010;84: 12458–62. doi:10.1128/JVI.01789-10
17. Aswad A, Katzourakis A. Paleovirology and virally derived immunity. Trends Ecol Evol. 2012;27: 627–636. doi:10.1016/j.tree.2012.07.007
18. Feschotte C, Gilbert C. Endogenous viruses: insights into viral evolution and impact on host biology. Nat Rev Genet. Nature Publishing Group, a division of Macmillan Publishers Limited. All Rights Reserved.; 2012;13: 283–296. doi:10.1038/nrg3199
19. Mi S, Lee X, Li X, Veldman GM, Finnerty H, Racie L, et al. Syncytin is a captive retroviral envelope protein involved in human placental morphogenesis. Nature. 2000;403: 785–789.
20. Houldcroft CJ, Breuer J. Tales from the crypt and coral reef: the successes and challenges of identifying new herpesviruses using metagenomics. Front Microbiol. 2015;6: 188. doi:10.3389/fmicb.2015.00188
21. Davison AJ, Eberle R, Ehlers B, Hayward GS, McGeoch DJ, Minson AC, et al. The order Herpesvirales. Arch Virol. 2009;154: 171–177. doi:10.1007/s00705-008-0278-4
22. Walker PJ, Winton JR. Emerging viral diseases of fish and shrimp. Vet Res. 41: 51. doi:10.1051/vetres/2010022
23. Near TJ, Eytan RI, Dornburg A, Kuhn KL, Moore JA, Davis MP, et al. Resolution of ray-finned fish phylogeny and timing of diversification. Proc Natl Acad Sci U S A. 2012;109: 13698–703. doi:10.1073/pnas.1206625109
24. Darling ACE, Mau B, Blattner FR, Perna NT. Mauve: multiple alignment of conserved genomic sequence

with rearrangements. Genome Res. 2004;14: 1394–403. doi:10.1101/gr.2289704
25. Yutin N, Koonin E. Hidden evolutionary complexity of Nucleo-Cytoplasmic Large DNA viruses of eukaryotes. Virol J. 2012; Available: http://www.biomedcentral.com/content/pdf/1743-422X-9-161.pdf
26. Iyer LM, Aravind L, Koonin E V. Common origin of four diverse families of large eukaryotic DNA viruses. J Virol. 2001;75: 11720–34. doi:10.1128/JVI.75.23.11720-11734.2001
27. Treangen TJ, Salzberg SL. Repetitive DNA and next-generation sequencing: computational challenges and solutions. Nat Rev Genet. Nature Publishing Group, a division of Macmillan Publishers Limited. All Rights Reserved.; 2012;13: 36–46. doi:10.1038/nrg3117
28. Ronquist F, Huelsenbeck JP. MrBayes 3: Bayesian phylogenetic inference under mixed models. Bioinformatics. 2003;19: 1572–1574. doi:10.1093/bioinformatics/btg180
29. Darriba D, Taboada GL, Doallo R, Posada D. ProtTest 3: fast selection of best-fit models of protein evolution. Bioinformatics. 2011;27: 1164–5. doi:10.1093/bioinformatics/btr088
30. van Beurden SJ, Bossers A, Voorbergen-Laarman MHA, Haenen OLM, Peters S, Abma-Henkens MHC, et al. Complete genome sequence and taxonomic position of anguillid herpesvirus 1. J Gen Virol. Society for General Microbiology; 2010;91: 880–7. doi:10.1099/vir.0.016261-0
31. Besemer J. GeneMarkS: a self-training method for prediction of gene starts in microbial genomes. Implications for finding sequence motifs in regulatory regions. Nucleic Acids Res. 2001;29: 2607–2618. doi:10.1093/nar/29.12.2607

Chapter Three: Paleovirology and Virally Derived Immunity

Abstract

Paleovirology, the study of viruses on evolutionary timescales, can exploit information from genomic features known as endogenous viral elements (EVEs), the result of heritable horizontal gene transfer from viruses to hosts. The availability of genomic data has increased opportunities to study EVEs, and bioinformatics techniques have been crucial to cataloguing their diversity and taxonomic coverage. Recent advances have shown that a number of EVEs have been co-opted as cellular genes, often as inhibitors of viral infection. These genes are an intriguing strategy in virus-host evolutionary battles: genetic material transferred from virus to host, then used by the host against the virus. In this review, we consider the genes and processes involved in EVE derived immunity, assess the factors that lead to their emergence, and outline how future work will benefit from incorporating evolutionary approaches.

Endogenous Viral Elements are relics that inform natural history

Viruses are probably the most abundant and diverse biological entity on earth [1]. They infect organisms and hijack the necessary molecular machinery needed for reproduction. Viruses typically spread between individuals, and occasionally, horizontal gene transfer can occur by the chromosomal integration of viral genes in the host germline (Box 1). These integrations, which are known as endogenous viral elements (EVEs), are revealed in genomic data using bioinformatics techniques [2], and are often described as the viral 'fossil record' [3–5]. The discovery and analysis of EVEs in host genomes has resulted in the emerging field of paleovirology. EVEs reveal the ancient history of viruses, which is otherwise extremely difficult to interpret: viruses are not geologically preserved and inferences from the genetics of extant viruses are also fraught with inaccuracies and limitations. This is largely due to rapid mutation rates that erode evolutionary signal. EVEs provide a convenient solution, as is exemplified by the identification of lentiviral EVEs in Lagomorpha and Mustelidae, revising the age of the Lentiviruses by at least 12 million years [6,7]. Paleovirology is pivotal for the appreciation of virus-host evolution and we focus here on the significance of EVE-derived immunity genes, proposing that they might be the products of a particular evolutionary mechanism.

The persistent and influential evolutionary consequences of EVEs

The vast majority of EVEs are passive relics, subject to the forces of genetic drift, accumulating frame shifts and stop codons [4]. However, some EVEs have maintained open reading frames (ORFs), exhibit low ratios of nonsynonymous to synonymous mutations (dN/dS) or have fewer stop codons than expected under neutrality, indicating a degree of purifying selection [8,9]. Virally derived genes have been discovered in several species, in many cases functioning as promoters and transcriptional regulators [10–12]. For example, the long terminal repeat regions of retroviruses often contain regulatory signals, and there are instances of their co-option in great apes [11]. The focus of this review, is to consider a subset of functional EVEs that act as part of the intrinsic immune system (Box 2) [5], representing an unusual immunity strategy: the use of the viruses against themselves.

EVE-derived immunity genes (EDI) have only been confirmed in a handful of cases (Table 1). Their numbers will no doubt increase as some EVEs have already been shown to encode uninterrupted ORFs, or are even expressed [8,13]. The presence of EDIs and the relative abundance and diversity of EVEs among hosts [4] require evolutionary explanations and phylogenetic analysis as well as making use of genomic survey data. EDIs are the focus of this review, defined as genes with viral ancestry acting as direct

inhibitors of viral infection. The molecular mechanisms of EDIs will be described to appreciate the link between their function and natural history. We will then discuss the causes and consequences of EDIs, and consider their broader evolutionary significance. There is much to be gained by intersecting what we understand about genetic evolution and immunity, with particular perspectives from paleovirology [14,15].

Box 1 - What is Paleovirology?

Viruses sometimes integrate into the genomes of their host; whether as part of their reproductive cycle, as a host evasion tactic, or by chance. In an infected individual, when this occurs in germline cells, the virus is inherited vertically in the population as part of the repertoire of host genes [8]. Known as endogenous viral elements (EVEs), such genomic features would be eliminated from the population by selection unless they pose negligible deleterious, or beneficial effects [2,4]. Paleovirology involves the study of EVEs; they have been found in all genomes that have been checked and viruses of all reproductive modes have the ability to endogenise. While we have known about EVEs for several decades, the increasing availability of whole-genome data has stimulated a more systematic approach to their discovery and analysis. For an unknown proportion of EVEs, the probability and time to fixation are not governed by neutral drift, but rather a product of their co-option by the host: they provide a selective advantage to the individuals carrying them [9,32]. For EVEs conferring antiviral properties, the extent of the benefit conferred will depend on the presence and prevalence of the viral infection that they protect against. Co-option of EVEs by hosts is also known to occur for non-antiviral purposes, most notably illustrated by Syncytin [81,86,87]. Originating from an endogenised retroviral envelope gene (*env*), Syncytin played a pivotal role in the evolution of mammalian placentation and has independently re-emerged at least seven times in multiple lineages [81,83]. This suggests a non-random pattern to their co-option and raises questions about the forces that drive endogenisation and co-option. As well as understanding co-option, the utility of paleovirology in deciphering the long-range evolutionary history of viruses is an invaluable tool. Anciently integrated EVEs serve as a record of viral biology, preserved by the slower host mutation rate that they evolve at because of endogenisation. The early evolution of a viral group within certain hosts can be confirmed by the discovery of paleoviruses. Co-evolution of foamy viruses and mammals began over 100 million years ago, as confirmed with the discovery of endogenous variants in Xenathrans [84]. Indeed, dating the age of EVEs and using them for the calibration of phylogenies reveals that long-term substitution rate of viruses is much slower than previously thought [88,89].

Viruses and hosts ordinarily compete by perpetually adapting and counter-adapting their molecular arsenal. EDIs however, are an example of evolutionary opportunism, representing the occasional evolutionary shortcut. Distinct host phenotypes arise via the acquisition of viral genes, which exist both as part of a viral and anti-viral mechanism [16]. The biological utility of such a gene is redefined while nonetheless maintaining molecular identity and function, and is now being used by the host after taking ‘ownership’ of the gene. For instance, a receptor binding protein might be used in viruses to facilitate cellular entry whereas a host could co-opt it to block entry. While under viral ownership, the receptor-binding gene already encodes for a protein capable of binding to cellular receptors - the host is bypassing the need for gradual molecular counter-adaptation by receiving a gene that has already undergone the necessary evolution to function effectively.

Molecular strategies of EDIs can be categorised

Cellular defense strategies involve interference by the host with the virus at cell entry or exit, or by disrupting viral replication in the cytoplasm (figure 1). EDIs that function at each

stage can be categorised; Cell entry inhibition, intracellular disruption and immune tolerisation. Independently acquired EDIs in several species perform almost identical functions and by appreciating the various examples within this framework, we can more easily discern their evolutionary implications.

Resisting viral infection by blocking cell entry

Eliminating a virus's ability to cross the cell membrane is one way to counteract infection. EDIs in mice (*Mus musculus*), sheep (*Ovis aries*), chickens (*Gallus gallus domesticus*) and cats (*Felis cattus*) are known to achieve this by saturating receptors that mediate viral entry. In mice, susceptibility to a subgroup of Murine Leukaemia Virus (MuLV) dubbed the 'friend strain' is genetically determined [17,18]. One of these genes, known as *Fv4*, is a defective endogenous retrovirus (ERV), composed of a fully intact envelope gene (*env*), that shares ~90% protein similarity with an exogenous MuLV [19]. The non-infectious, defective *Fv4 env* proteins bind to the same cell surface receptors used by MuLV, leaving them unavailable for invading virions [20]. *Rmcf* and *Rmcf2* are also EVE-derived mouse genes, offering resistance to a different strain of the virus by the same but independently derived mechanism. [21,22]. For sheep, EVEs related to the Jaggsiekte Sheep Retrovirus (JSRV) provide resistance to infection by saturating the Hyal-2 entry receptors [23]. *env*-mediated receptor blocking EDIs are also found in chickens and cats, offering resistance to Avian Leukosis Virus (ALV) [24] and Feline leukemia virus (FeLV), respectively [25].

EDI	Viral Origin/Host Genome	Function(s)	Reference
Fv4	Retroviral *env*/mouse	Receptor interference	[20]
enJSRV	Betaretroviral *gag;env*/sheep	Heteromultimerisation; Receptor interference	[30]
enFeLV	Retroviral *env*/cat	Receptor interference	[25]
enALV	Avian leukosis virus *env;whole*/chicken	Receptor interference, immune tolerisation	[32]
IRIS	*env*/fruit fly	*Unknown*	[9]
Rmcf, 2	*env*/Mouse	Receptor interference	[21,22]
Fv1	Retroviral *gag*/mouse	Capsid interaction	[16]
***Gypsy*-like LTR retrotransposons**	Retroviral *gag*/vertebrates	*Unknown*	[40]
enMMTV	Mouse Mammary tumor virus *sag*/mouse	T-cell deletion	[36,37]
EBLN	Bornavirus-like NP/mammalian	*Unknown*	[13]
NIRVs	Ebolavirus-like *VP-35;NP*/Bats	*Unknown*	[44]
CGIN1	Retroviral RNase/mammalian	Unconfirmed ubiquitination mechanism	[61]
APOBEC3	Retroviral regulatory element	Up regulation of viral inhibition by APOBEC	[62]
IAPV	Picornavirus	*Unknown*	[75]

Table 1

Endogenous gag genes disrupt viral replication after cell entry

Fv1 is another gene that contributes to MuLV resistance by a different mechanism [26–28]. Although the precise biology is poorly understood, we know that *Fv1* is derived from a

retroviral *gag* (group specific antigen gene), and belongs to ERV-L, a family of mammalian ERVs distantly related to MuLV [16]. At some stage after cell entry but before nuclear integration, the restrictive effect of *Fv1* occurs by interaction with the viral capsid protein. *Fv1* resistance depends on one of three allele variants carried by the mouse, each offering resistance to different strains of the virus (*Fv1*^{b}*, Fv1*^{n} *and Fv1*^{NR}*)* [29]. As well as acting by receptor blocking, an endogenous-JSRV in sheep also restricts reproduction of JSRV viruses. At least two of the 27 known copies of endogenous-JSRV with the relevant polymorphisms inhibit cell exit of an invading JSRV, despite the fact that intact virions are visible by microscopy [30]. Whereas Fv1 has its effect before genomic integration, the association of the endogenous Gag protein (capsid) with JSRV interferes with the cellular trafficking of assembled virions. Indeed, capsid proteins are commonly involved in the evolution of EDI acquisition, having been co-opted independently in several species including some fungi [31,32].

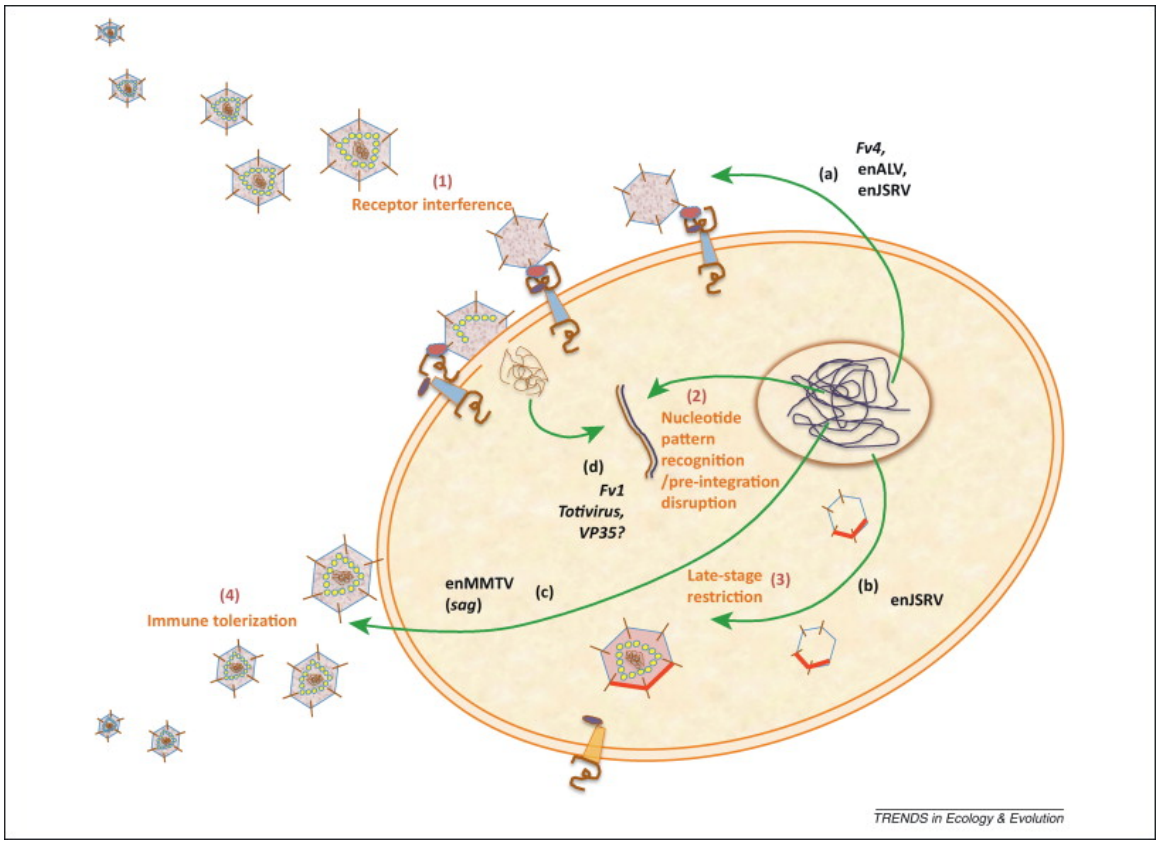


Figure 1 EDI systems and molecular mechanisms.
The intrinsic immune system is composed of constitutively expressed antiviral genes that have evolved to disrupt viruses at different stages of replication. Although viruses have a diverse range of replication strategies, they share generic features to which an immune system can be applied. Antiviral genes of viral origin are key members of the intrinsic immune system and their mode of action is a reflection of their evolutionary viral past. (1/a) A virus needs to gain cellular entry in order to hijack the necessary reproductive machinery, and to do so often requires receptor-mediated recognition and cell membrane translocation. Receptor blocking is an EDI mechanism whereby defective viral particles or proteins are produced by the cell and saturate entry receptors, thereby inhibiting infection. (2/d) Intracellular disruption by EDIs can occur before integration as with *Fv1*, which interacts with incoming viral capsids. Also, disruption by recognising and interacting with foreign RNA is an existing defense system known as RNA interference (RNAi) and similar mechanisms have been proposed for experimentally unconfirmed EDIs. (3/b) EDI Heteromultimerisation can occur when the cell expresses defective viral components, associating with the assembling virus and causing structural or functional deficits. This is known to disrupt trafficking ability or cell exit, often achieved using endogenous capsid proteins expressed by *gag* genes. (4/c) Immune tolerisation can also be induced by EDIs. Expressing viable viruses at early stages in development for example, eradicates subsets of T-cells that are used by a virus to spread.

Anticipatory strategies offered by EDIs

Immunological tolerance can be defined as a process by which the immune system does not attack an antigen. EVE-induced immunological tolerance is known to interfere with infection outcomes in mice and chickens. The expression of an endogenous-ALV [33] in chickens reduces non-neoplastic symptoms in chickens infected with the exogenous strain RAV-1, and chickens lacking endogenous-ALV are highly susceptible to infection by exogenous ALV- J [34], a strain that probably emerged by recombining with another endogenous-ALV, EAV-HP [35]. In principle, any EVE encoded peptide could tolerise the

host against infections with related exogenous viruses if expressed during thymic development, as demonstrated with an endogenous env protein that tolerises mice against MuLV [36].

Box 2 –Intrinsic Immunity is a cellular system of antiviral defense

The immune system is a layered set of defense structures that an organism uses to protect itself against disease. If a virus manages to evade mechanical barriers such as skin and/or exoskeleton, as well as humoral (e.g. antibody secretion) and cell-mediated responses (e.g. t-lymphocytes), it is still challenged by intracellular defenses. Intrinsic immunity is a term used for a set of constitutively expressed inhibitors of viral pathogens [90]. Such inhibitors are sometimes referred to as restriction factors, and act to disrupt viral infection by a variety of strategies. TRIM5α, for example, recognises viral capsids soon after cell entry and mediates their disruption [91]. APOBEC is restriction factor family that interferes with virus replication by introducing cytidine-to-uracil mutations in the minus strand of retroviral DNA [92]; several human and macaque APOBEC3 families have the ability to disrupt HIV infection, but only in strains lacking the necessary Vif accessory protein [93]. Vif is a counter-adaptation in the virus to the host's intrinsic defense, and acts by encouraging the ubiquitination and degradation of APOBEC3G [94]. SAMHD1 is a tissue specific HIV restriction factor functioning in dendritic and myeloid cells, the effect of which is counteracted by the viral auxiliary protein Vpx [95]. This kind of evolutionary response puts viruses at an advantage over their hosts; their faster mutation rates resulting from lower-fidelity replication mean that beneficial substitutions can accumulate more quickly. Their ability to adapt rapidly to host evasion tactics, however, is buffered by the fact that the molecular conflict between viruses and hosts occurs at multiple interfaces simultaneously (e.g. cell surface entry receptors, intracellular recognition mechanisms, etc.). This means that for viruses to successfully infect a cell they need to have the optimum combination of the compatible components that will allow it to reach and enter the cell, evade detection, and/or avoid antiviral effects. This encourages a perpetual arms-race because a catalogue of alleles exist in the intrinsic immune system, effectively resulting in more genotype combinations than there are alleles, to each of which a virus needs to evolve compatibility at the same time [14]. Virus-derived antiviral factors in the intrinsic immune system are the main subject of this review, and while several molecular mechanisms exist, intrinsic immune genes of viral origin are a significant contributor to intrinsic immunity, representing a strategy in the evolution of host defense.

Expression of an EDI during development can also benefit the host via an anticipatory mechanism other than immunological tolerance. In mice, B-cells infected with Mouse Mammary Tumour Virus (MMTV) express a superantigen (*sag*) which induces cytokine secretion from T-cell [37]. This promotes cell division of infected cells, and benefits the virus by acting as a dispersal mechanism. Endogenous-MMTV, however, also expresses *sag* during host development; this results in deletion of the specific *sag* responsive T-cells [38,39], thereby reducing the ability of MMTV to spread. It is interesting that such strategies can exist, given that tolerisation or the deletion of a large proportion of T-cells leaves an individual vulnerable to other pathogens. Infection probability is heterogeneous in time and space, and the success of such a costly strategy will depend on the cost-benefit balance, which will be influenced by ecological factors.

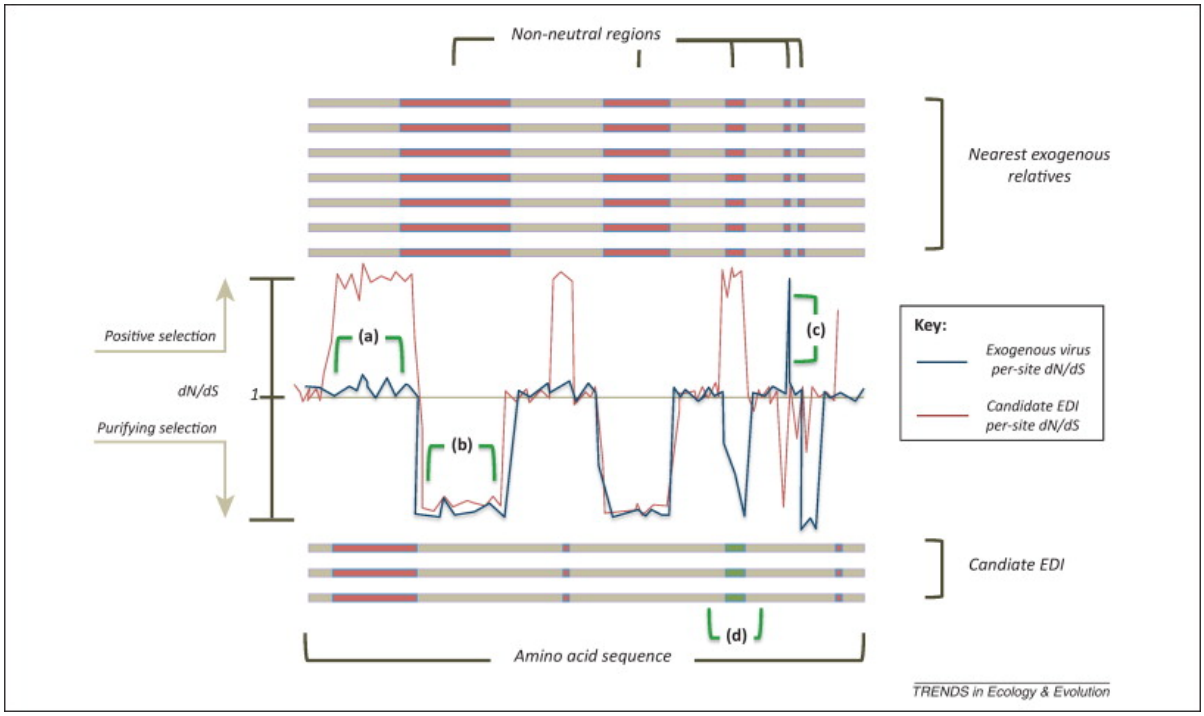


Figure 2 Interpreting selection patterns differences between EDIs and viruses.
By considering the molecular requirements of transitioning from a viral gene to a host defence gene, a set of selection patterns can be expected of a protein undertaking such an evolutionary trajectory. Ratios of non-synonymous to synonymous changes can be used as a measure of selection direction (positive or negative), although understanding this requires information about biological context and interpretation. Comparing selection dynamics between an EDI and its exogenous counterpart is a useful approach that might suggest particular changes in function. Although this is a hypothetical illustration, it is useful to consider various results that would be expected if viral genes diversified in this way. (a) This pattern would be expected if the region were a neutrally evolving site in an active virus, but is undergoing recent adaptive change in the EVE. (b) Conserved regions in the exogenous virus may also appear to be subjected to similarly high levels of purifying selection in a co-opted EVE, but with a different sequence identity, despite an initial sweep of directional selection for amino acids that were selected for different functional reasons. This highlights the need to consider amino acid properties when drawing conclusions from potentially deceptive selection patterns. (c) Positively evolving residues in a virus sequence may be expected, for example, if determinants of tropism need to account for variation in the host population. (d) A positively selected site in an EVE might be the result of selection for diversity. (c) and (d) are the kind of virus-host adaptive conflict that are perpetuated by the evolutionary arms race.

EDI mechanisms show independently derived endogenous viruses, in different species performing the same function (receptor blocking in sheep and mice). Furthermore different viral genes within one host can function by distinct mechanisms (Fv1 and Fv4), as well as similar viral genes performing different functions in different hosts (gag interference by Fv1 and enJSRV). EDIs that involve anticipatory strategies also differ greatly in the benefits that they confer – ALV-E functions by reducing a debilitating immune reaction (although increasing viral replication), while MMTV functions via a pre-emptive strategy, removing the cells required for viral spread.

EDI evolution: Interpreting selection patterns by direct and indirect paleovirology

Direct paleovirology involves examining viral integrations within host genomes in order to deduce the evolutionary history of EVEs and EDIs [3,8,13]. This is achieved by querying genomes with viral proteins using sequence similarity tools such as BLAST, and interpreting the evolution of identified EVEs with reference to known related exogenous viruses [3,8,40–43]. Determining the age of EVE integration can be achieved using several approaches in order to provides a minimum bound for the expected date of integration, and an expectation of the amount of divergence that would have been accrued under neutrality. Indeed, some non-functional EVEs have not been subject to neutral evolution, as evidenced by fewer stop codons than would be expected, or by low dN/dS ratios [44]. For example, *Fv1* integrated prior to the expansion of *Mus* but after the *Mus-Rattus* split, providing an estimated date for *Fv1* of 7 million years [28], and Filovirus-like genes in bats (*Myotis* spp.) are thought to have been maintained throughout the radiation of the genus [44]. Furthermore, selection analyses performed with a panel of fruit fly (*Drosophila*) species suggest that retroviral *iris* genes may have been domesticated by their host [9], illustrating how direct paleovirology can also be used for discovering potential EDIs.

Indirect paleovirology can be used to describe wider aspects of virus-host evolutionary history. Inferring the existence of extinct viruses might be achieved by considering the evolutionary effects imposed on the antiviral host genes that antagonised them [15]. For example, the differential selection patterns observed for primate antiviral genes such as TRIM5 [45] and APOBEC3H [46] may have resulted from episodes of positive selection driven by different viruses at different times [15]. This approach can be used with EDIs to greater effect, as they also contain information about the ancestral virus population – the EDIs origins are rooted in the viral population that drove its evolution. This overlapping natural history enables us to interpret selection patterns, but we are limited to comparable domains; EDIs will loose by deletion aspects of its viral origin that are not useful to cell defence. Both *Fv1* and *Fv4* for example, have lost their polymerase gene, *Fv4* encodes a defective *env*, while *Fv1* has lost *env* but maintains *gag* [16,47]. Nonetheless, we can compare the effects of selection between exogenous viruses and surviving domains of the EDI sequence. This will allow us to examine how exaptation events influenced the trajectory of extant virus evolution, as well as partially deduce the drivers of the exaptation itself (figure 2).

Thoughts on the fate of EDIs: Are they predestined for failure?

EDIs are part of the complex co-evolutionary dynamics affecting viruses and host. By virtue of their short generation times and low-fidelity replication, viruses have high substitution rates that allow them to respond relatively rapidly to host adaptation. Because slow-evolving hosts cannot compete by adaptation, a multifaceted immune strategy compensates for this rate disparity [48]. Adaptive immunity is a rapid response to viral threats, and can in part ameliorate this disparity, but the first line of defence is the intrinsic immune system. These fast evolving genes counter-adapt to viral evasion and manage to compete only because there are many alleles that are variable in the population. This makes it difficult for viruses to fully optimise their molecular arsenal, despite their substitution rate advantage [49]. Furthermore, evidence suggests that these genes resample a limited set of alleles in a 'rock-paper-scissors' model, suggesting a degree of cyclicality to the arms-race [14]. EDIs are a peculiar evolutionary shortcut in this repertoire of strategies because their makeshift origins might come with fitness costs. Immune tolerisation exposes hosts to a multitude of pathogens [50], and there is certainly a cost associated with receptor blocking. For example, the Hyal-2 receptor that is blocked by endogenous-JSRV is important in a myriad of functions [51] and mice missing the receptor display a range of abnormalities [52].

Adopting an EDI therefore will be more likely when the selective cost of a viral infection is higher than the cost imposed by an EDI. This will depend on a combination of permissive circumstances involving genomics, population structure and disease ecology. The fitness advantage gained by EDIs will depend on a multitude of features such as genomic compensatory mechanisms that minimise its negative fitness costs. For example, detrimental side-effects of Hyal-2 blockage are probably offset by the redundancy of its function, which overlaps with similar receptors [52]. Moreover, the features of an acquiescent host population will be closely linked to the patterns of viral infection, for example, a chronically infectious virus in a transmissive population with a high, but non-lethal fitness cost would more amenable to EDI population spread. The frequency of this fixation will be dependant on the probability of EVE integration, which in turn depends on the frequency of infection and the first individual surviving the integration event. If the new EVE is neutral, it might drift in the population before gaining the necessary beneficial mutations for EDI function, or else be beneficial at the point of integration, which would result in a far quicker spread within the host population.

If an EDI succeeds against a viral threat or if the virus counteradapts, the advantage that maintained it will be lost, allowing negative selection to drive its extinction (figure 3).

Moreover, the rate of EDI acquisition probably far exceeds their population fixation rate, and exactly how transient an EDI is depends on the delicate cost-benefit balance. Population structures of hosts that influence viral epidemiology would result in localized EDI acquisitions that never reach species-wide fixation. This could lead to complex macroevolutionary patterns – for example, if heterogeneous EDI acquisition and extinction occurred in different sub-populations, coupled with heterogeneous epidemics, then intermittent contact between partially isolated subgroups could cause oscillating rates of EDI acquisition, caused by variable evolutionary pressure. Cyclical dynamics might also emerge, if the viral threat returns to a subpopulation soon enough after an EDI has begun to degrade and the returning pressure encourages restoration of the EDI. Alternatively, an extinct EDI might be then co-opted by the host for a different function, which raises the question of whether we could consider EDI acquisition as a mechanism of HGT in its own right – the beneficial effects allow these foreign sequences to rapidly expand in the population, introducing genetic novelty that can be exploited.

EVEs with a history of purifying selection support the idea of transience, exemplified by primate endogenous Bornavirus-like N viruses (EBLN). These EVEs exhibit evidence for a history of purifying selection that has since relaxed; the Orang-utan (*Pongo abelii*), Marmoset (*Callithrix jacchus*) and Macaque (*Macaca mulatta*) EBLNs have all acquired stop codons, whereas open reading frames are maintained in Humans (*Homo sapiens*), Gorillas (*Gorilla gorilla*), and Chimpanzees (*Pan troglodytes*) [8].

EDIs are an extreme example of virus-host evolutionary entanglement

We can consider the degree of evolutionary entanglement between hosts, viruses, and EDIs to lie on a continuum. On one end of the continuum lie viruses and hosts that are engaged in gradual adaptation and counteradaptation. EDIs lie at the other extreme of this continuum, and are indicative of a much more entangled relationship; a virus has effectively become a host gene. Between the two exists a range of intermediate interactions that includes superinfection exclusion. Some viruses result in resistance to secondary infection by the same or closely related viruses, by a variety of mechanisms functioning at all stages of the viral replication cycle [53–57]. For example, rat astroglia cells infected with Borna disease virus [58] resist infection by other strains. This relationship is neither based on independent rapid molecular adaptation, nor on permanent genetic transfer. It could certainly be an evolutionary intermediate to becoming a receptor blocking EDI as previously implied [59]. Other virus-host evolutionary relationships also lie along this continuum. For example, EDIs can emerge from EVE-host

gene fusion in the same way that ordinary gene fusion events between host genes or domains can give rise to novel proteins [60]. A mammalian gene named CGIN1 (cousin of gypsy integrase-1) may have arisen by the fusion of a host gene with retroviral genes to regulate viral protein ubiquitination [61]. We know of EVE that have been co-opted for regulatory functions [12], and when this happens as part of immune regulation we could consider them to be a kind of EDI. For example, in mice, APOBEC3 is a restriction factor containing a retroviral regulatory element insertion, responsible for up-regulating expression [62]. Such examples lie close to EDIs on the continuum, but involve host genes as part of their emergence. Completely usurping a viral gene, and using components of viral genes to complement host strategies in a molecular partnership are distinct events. Other forms of virus-host interaction might arguably fall along this continuum also, although they are speculative and not emphasised in this review (Box 3).

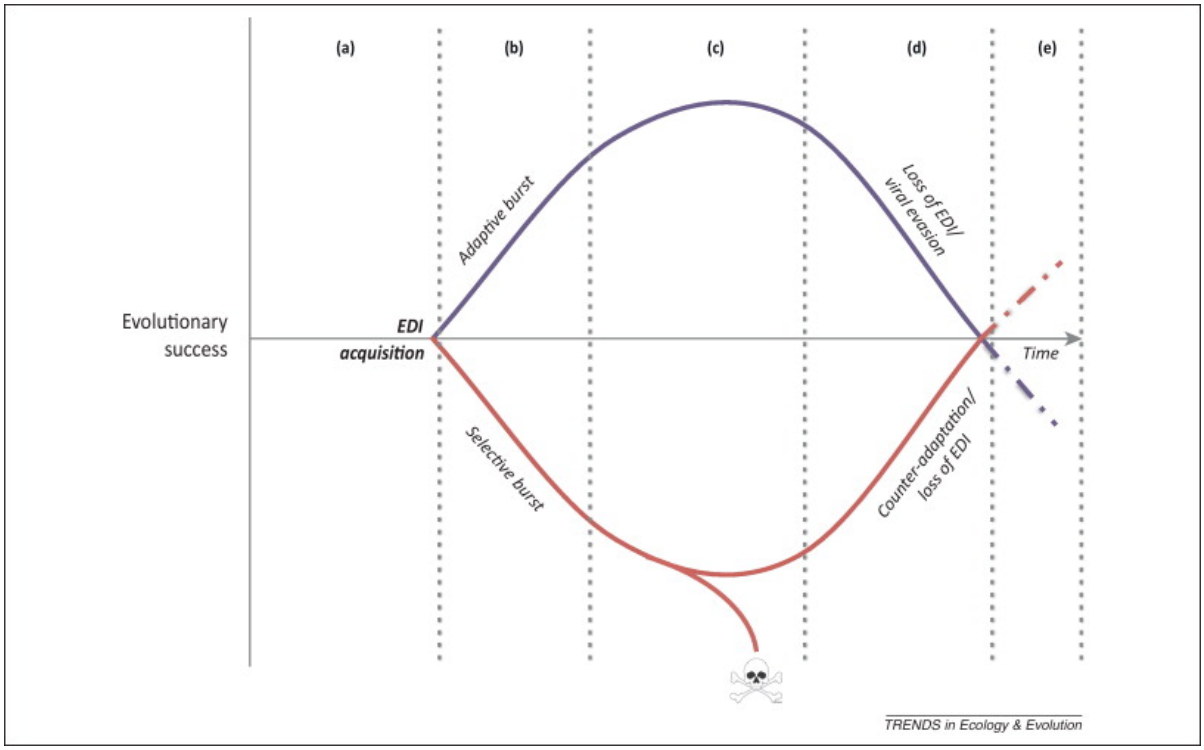


Figure 3 A conceptual model describing EDI evolutionary dynamics.
Purple curve corresponds to the host while red corresponds to the virus. (a) The acquisition of an EDI boosts the host's ability to evade viral infection (b). This is mirrored in the virus population with a sudden decline of the vulnerable viral strain. (c) Two scenarios are possible; either the virus population is decimated beyond recovery, or evasion occurs because viral strains immune to the EDI emerge and expand in the population. (d) The decline of the host ability to resist is inevitable due to the two scenarios described in (c), which encourages a recovery of viral infection success. Once an EDI is redundant, any negative fitness effects it causes will promote accelerated molecular degradation, eventually reverting to a non-functional EVE. (e) Several outcomes are now possible. 1. The virus is so successful and results in host extinction, which would explain the unusual abundance of EVEs, and is supported by EVEs with a history of purifying selection. 2. Ordinary arms-race is restored with a baseline likelihood of EDI appearance. 3. Returning evolutionary pressure encourages either novel EDI acquisition, or the revival of the former EDI. 4. Secondary co-option occurs to fulfil a different non-immunity related host function, which could also happen in (d).

The big picture: EDI-like evolutionary mechanisms are ubiquitous in nature

Describing these different co-evolutionary mechanisms should not lead to conclusions of a virus host mutualism as some have proposed EDIs to be in support of [63]. Exaptation means an EDI cannot be considered as viral, but rather a macro-mutation resulting in neofunctionalisation – they are selfish elements, with allegiances to neither virus nor host. This view is supported by the fact that viruses also partake in gene capture as part of the evolutionary arms-race. The emergence of the ALV-J virus by recombination with an endogenised ALV [34], means that viruses also exploit the same evolutionary shortcut as their hosts. Several large viruses of metazoans encode captured genes with immunomodulatory function [64]. A gene in Ovine herpesvirus-2 was shown to be a member of the interleukin-10 family [65], also present in other viruses [66]. Interleukins natively perform immunostimulatory and immunosuppressive functions, but their viral counterparts are often modified to only improve viruses' ability to cause persistent infections.

Box 3- Immunomodulatory EVEs may share EDI evolutionary Dynamics

There are EVEs that have been co-opted for immunomodulatory functions, which may share some of the evolutionary dynamics as EDIs but do not fall into the categories we define. EVEs might for instance be co-opted to defend against entities other than viruses. Syncytin functions by re-purposing the fusogenic properties of the *env* transmembrane (TM) domain, originally intended as a means of cell surface to virion envelope fusion during cell entry. Syncytin uses this capability in the formation of the epithelial layer of placental villi, known as the syncytiotrophoblast, by the cell-cell fusion of cytotrophoblast cells during embryonic development. TM domains sometimes also contain immunosuppressive domains, an attractive ability for a foetus needing protection from maternal immune attack. The repeated co-option of Syncitin (see box 1) [81,83] has prompted speculation that this may reflect new or improved opportunities for maternal exploitation by the foetus. Might some Syncytins provide foetal protection from maternal immunity? This idea is at least partly supported by reports of positive selection in some Syncytin genes [81].The repeat exaptation could reflect something akin to arms-race dynamics, driven by variable periods of high and low maternal-paternal conflict, in the same way that the heterogeneity of EDI acquisition and extinction might result in complex evolutionary dynamics. Variation in ecological factors relevant to the foetus, such as lack of paternal investment or multiple paternity might be encouraging variation in foetal-maternal conflict, be responsible for this repeated co-option [83].

An EVE could provide protection to the host cell from autoimmune attack. An "integration hypothesis" has been proposed to explain why natural simian immunodeficiency virus (SIV) hosts (e.g. sooty mangabeys) remain healthy despite high viral load [96]. A hypothetical EVE might tolerise the immune system inhibiting chronic immune hyperactivation during infection, which damages the immune system. While there is no evidence of such a mechanism, it is true that sooty mangabeys do not exhibit many signs of hyperactivation [97]. The issue of SIV pathogenesis is controversial, however, and there could be other mechanisms involved [98].

It is interesting to consider a co-opted EVE offering protection from other selfish genetic elements. A set of orthologous *gypsy*-like LTR-retrotransposons are present in the genomes of sheep, rat, mice and humans, exhibiting strong evidence for exaptation [99]. It was speculated that gypsy-like retrotransposons are unusually scarce in mammalian genomes [40] and that they function as inhibitors of intact *gypsy* LTR-retrotransposons.

The EDI mechanism is different to other forms of HGT in that arms-race dynamics are at play between the donor and recipient. HGT occurs most extensively in prokaryotes and is a major force in their evolution [67]. They participate in a molecular strategy known as CRISPR (Clustered Regularly Interspaced Short Palindromic Repeats, reviewed in [68]) that is in some ways similar to the EDI phenomenon. CRISPR systems are a pre-existing adaptable mechanism of defence, compared to transient and narrow range EDIs. They are an example of viruses being used against viruses, extending this kind of HGT to all domains of life.

Finding EDIs: informed bioinformatics strategies and perspectives from ecology

Considering plausible biological scenarios will allow targeted approaches for tackling specific paleovirological questions and EVE research with potential for EDI discovery already exists. Mononegavirales-like EVEs have endogenised in a variety of mammalian species [8] and though the sampling is far from exhaustive, a notable bias towards their endogenisation exists [4]. The co-option of filovirus-like EVE with VP35-like (Viral Protein-35) sequences has been inferred in several species of bat (with positional orthologs in rodents) [44]. VP35-like proteins might be a counter-adaptive response to VP35 immunosuppressive activity, which natively functions to supress RNA silencing and antagonism of the interferon response [69]. Nearly 200 Rhabdovirus fossils have been identified in the mosquito (*Aedes aegypti*) and the tick (*Ixodes scapularis*), with evidence of recent purifying selection [70]. Potential EDIs are also apparent in plants [71], and in fungi totivirus-like genes are undergoing significant levels of purifying selection [32]. The fungus *Debaryomces hansenii* contains a totivirus-like polymerase and has maintained the sites necessary for its catalytic activity. It was proposed that antiviral defence occurred via RNAi; by expressing viral-specific transcripts, the EVE recognises and disrupts incoming viruses [72].

If, as we propose, EDIs are transient, examining genomes of single individuals will fail to identify the majority of EDIs that might exist in the population, because we predict that in most cases, an EDIs lifetime will be shorter than the time taken for it reach population fixation. Thus we would largely find either old EDIs, or EVEs that no longer encode a function. Screening multiple individuals will reveal unfixed EDIs, and their distribution and abundance will allow quantification of their transience. A koala (*Phascolarctos cinereus*) ERV maintaining high levels of genome activity, is variably prevalent in the population and absent in some populations [73]. An endogenous Israeli acute paralysis virus (IAPV) in

honeybees (*Apis mellifera*) is found in a third of tested populations, and bees carrying the insertion are resistant to infection [74]. It would be revealing to examine the effects this EDI has on population structure, both in the wild and experimentally, by tracking allele frequency turnover and molecular changes over time.

The benefits or costs of EDIs are likely to be heterogeneous and dependent on pleiotropic effects and environmental factors. The benefit conferred by active EDIs must be higher than the negative fitness effect as previously discussed with endogenous-JSRV [51,52]. EDI emergence depends on the prevalence of infection because an EDI's co-option is contingent on a high enough viral threat. Furthermore, because compensatory mechanisms are themselves variable, an EDI might impose a selective influence beyond the involved genes. The human CCR5Δ32 allele imparts resistance against HIV-1 [75] and displays spatially heterogeneous frequencies [76]. HIV emerged too recently to be responsible for purifying selection in CCR5Δ32, yet the allele is maintained in Europeans. This frequency difference could be attributed to changes in cost-benefit ratios of ancestral and modern conflicts [76,77]. The evolutionary reach of EDI influence is likely to be similarly complex and can only be elucidated if ecological variables are properly appreciated. Former EDIs are ideal candidates for subsequent co-option, perhaps to perform other, non-antiviral functions (figure 3e). enJSRV is already known to regulate aspects of growth and differentiation in preimplantation-stage embryos, which exemplifies the multifunctional potential of EVEs [78].

Finding EDIs will elucidate the nature and validity of the phenomena

There are not many known EDIs, and those that are confirmed are poorly understood. The limitations of BLAST have not yet been reached, partly because it is a powerful tool, but also because the growth rate of genomic data is far outstripping our ability to analyse it. Alternative techniques will be needed to reveal older EDIs that have been obscured by nucleotide divergence. Indeed, some approaches employ model-based techniques such as amino acid profiling [79] and degenerate protein motif searching that accounts for frame shifts [80]. Evidence of purifying selection, such as low dN/dS or intact ORFs can be a test for exaptation candidates. Such EVEs then need to demonstrate their expressiblility, e.g. by transcriptomic screening. *in silico* functional analysis (e.g. protein prediction and modelling) and *in vitro* experiments are also crucial. The Syncytin gene of carnivores (Syncytin-Car1) was initially pursued after *in silico* screening with retroviral *env* probes and confirmed by RT-PCR and in-situ hybridization that showed localised placental expression [81]. 'Reverse paleovirology' is an alternative approach - look for EDIs by

examining cellular expression profiles under relevant conditions such as an interferon response, using a candidate-gene approach. Similarly to *Fv1* and enJSRV, *gag*-like genes have been co-opted in humans to fulfil several other functions such apoptosis and ubiquitin-ligase binding [82], which is further evidence of their propensity for co-option.

Though there are exceptions [14,15,83,84], there has been little emphasis on the evolutionary significance of virally derived antiviral genes. Because EDI dynamics can result in the rapid and frequent introduction of novel genetic material, it might be that they serve to increase the speed of viral to host LGT; it will be interesting to determine if exaptation of an EDI facilitates subsequent exaptation of the same gene to fulfil another function, and whether EDIs result from EVEs that had drifted for a while or were immediately beneficial. Elucidating the validity of this hypothesis will necessitate intra-species screening for viral elements, which will also provide insights into the nature and validity of EDI transience, as well as characterising the population genetic and biogeographic factors that determine EDI acquisition likelihood. Examining whether any true bias exists for certain genes, and whether EDI presence corresponds to any co-evolutionary patterns between viral groups and their hosts, will reveal some of the conditions necessary for EDI acquisition. Whether the acquisition bias of some genes such as those that encode for capsids or nucleoproteins is the result of functional flexibility, and whether the overrepresentation of retroviruses is purely a by-product of their replication, are outstanding questions that will need a fuller understanding of the precise molecular biology of EDIs. Whatever the technique, identifying EDIs is a prerequisite step to addressing all of these points, as well as validating their significance and clarifying the particulars of this kind of virus-host HGT.

References

1 Kristensen, D. M. *et al.* (2010) New dimensions of the virus world discovered through metagenomics. *Trends in microbiology*. 18, 11-9

2 Feschotte, C. and Gilbert, C. (2012) Endogenous viruses: insights into viral evolution and impact on host biology. *Nature Reviews Genetics*. 13, 283-296

3 Patel, M. R. *et al.* (2011) Paleovirology—ghosts and gifts of viruses past. *Current Opinion in Virology*. 1, 304-309

4 Holmes, E. C. (2011) The Evolution of Endogenous Viral Elements. *Cell Host & Microbe*. 10, 368-377

5 Stoye, J. P. (2012) Studies of endogenous retroviruses reveal a continuing evolutionary saga. *Nature reviews. Microbiology*. 10, 395-406

6 Han, G.-Z. and Worobey, M. (2012) Endogenous lentiviral elements in the weasel family (Mustelidae). *Molecular Biology and Evolution*. DOI: 10.1093/molbev/mss126

7 Keckesova, Z. *et al.* (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

8 Katzourakis, A. and Gifford, R. J. (2010) Endogenous viral elements in animal genomes. *PLoS Genetics*. 6, e1001191

9 Malik, H. S. and Henikoff, S. (2005) Positive selection of Iris, a retroviral envelope-derived host gene in Drosophila melanogaster. *PLoS genetics*. 1, e44

10 Wang, T. *et al.* (2007) Species-specific endogenous retroviruses shape the transcriptional network of the human tumor suppressor protein p53. *Proceedings of the National Academy of Sciences of the United States of America*. 104, 18613-8

11 Beyer, U. *et al.* (2011) Endogenous retrovirus drives hitherto unknown proapoptotic p63 isoforms in the male germ line of humans and great apes. *Proceedings of the National Academy of Sciences of the United States of America*. 108, 3624-9
12 Cohen, C. J. *et al.* (2009) Endogenous retroviral LTRs as promoters for human genes: a critical assessment. *Gene*. 448, 105-14
13 Horie, M. *et al.* (2010) Endogenous non-retroviral RNA virus elements in mammalian genomes. *Nature*. 463, 84-7
14 Meyerson, N. R. N. R. and Sawyer, S. L. S. L. (2011) Two-stepping through time: mammals and viruses. *Trends in microbiology*. 19, 286-94
15 Emerman, M. and Malik, H. S. (2010) Paleovirology—modern consequences of ancient viruses. *PLoS biology*. 8, e1000301
16 Best, S. *et al.* (1996) Positional cloning of the mouse retrovirus restriction gene Fv1. *Nature*. 382, 826-9
17 Han, G.-Z. and Worobey, M. (2012) An Endogenous Foamy-like Viral Element in the Coelacanth Genome. *PLoS Pathogens*. 8, e1002790
18 ODAKA, T. *et al.* (1962) Inheritance of susceptibility to Friend mouse leukemia virus. *The Japanese journal of experimental medicine*. 32, 405-13
19 Masuda, M. and Yoshikura, H. (1990) Construction and characterization of the recombinant Moloney murine leukemia viruses bearing the mouse Fv-4 env gene. *Journal of virology*. 64, 1033-43
20 Taylor, G. M. *et al.* (2001) Fv-4: identification of the defect in Env and the mechanism of resistance to ecotropic murine leukemia virus. *Journal of virology*. 75, 11244-8
21 Wu, T. *et al.* (2005) Rmcf2, a xenotropic provirus in the Asian mouse species Mus castaneus, blocks infection by polytropic mouse gammaretroviruses. *Journal of virology*. 79, 9677-84
22 Jung, Y. T. *et al.* (2002) Characterization of a polytropic murine leukemia virus proviral sequence associated with the virus resistance gene Rmcf of DBA/2 mice. *Journal of virology*. 76, 8218-24
23 Spencer, T. E. *et al.* (2003) Receptor Usage and Fetal Expression of Ovine Endogenous Betaretroviruses: Implications for Coevolution of Endogenous and Exogenous Retroviruses. *Journal of Virology*. 77, 749-753
24 Tikhonenko, A. T. and Lomovskaya, O. L. (1990) Avian endogenous provirus (ev-3) env gene sequencing: implication for pathogenic retrovirus origination. *Virus genes*. 3, 251-8
25 McDougall, A. S. *et al.* (1994) Defective endogenous proviruses are expressed in feline lymphoid cells: evidence for a role in natural resistance to subgroup B feline leukemia viruses. *J. Virol.* 68, 2151-2160
26 Hartley, J. W. *et al.* (1970) Host-range restrictions of murine leukemia viruses in mouse embryo cell cultures. *Journal of virology*. 5, 221-5
27 Pincus, T. *et al.* (1971) A major genetic locus affecting resistance to infection with murine leukemia viruses. I. Tissue culture studies of naturally occurring viruses. *The Journal of experimental medicine*. 133, 1219-33
28 Yan, Y. *et al.* (2009) Origin, antiviral function and evidence for positive selection of the gammaretrovirus restriction gene Fv1 in the genus Mus. *Proceedings of the National Academy of Sciences of the United States of America*. 106, 3259-63
29 Stoye, J. P. (1998) Fv1, the mouse retrovirus resistance gene. *Revue scientifique et technique (International Office of Epizootics)*. 17, 269-77
30 Mura, M. *et al.* (2004) Late viral interference induced by transdominant Gag of an endogenous retrovirus. *Proceedings of the National Academy of Sciences of the United States of America*. 101, 11117-22
31 Valle, R. P. and Wickner, R. B. (1993) Elimination of L-A double-stranded RNA virus of Saccharomyces cerevisiae by expression of gag and gag-pol from an L-A cDNA clone. *Journal of virology*. 67, 2764-71
32 Taylor, D. J. and Bruenn, J. (2009) The evolution of novel fungal genes from non-retroviral RNA viruses. *BMC biology*. 7, 88
33 Crittenden, L. B. *et al.* (1984) Influence of endogenous viral (ev) gene expression and strain of exogenous avian leukosis virus (ALV) on mortality and ALV infection and shedding in chickens. *Avian diseases*. 28, 1037-56
34 Silva, R. F. *et al.* (2000) Hypervariability in the envelope genes of subgroup J avian leukosis viruses obtained from different farms in the United States. *Virology*. 272, 106-11
35 Sacco, M. A. *et al.* (2004) Assessing the roles of endogenous retrovirus EAV-HP in avian leukosis virus subgroup J emergence and tolerance. *Journal of virology*. 78, 10525-35
36 Young, G. R. *et al.* (2012) Negative Selection by an Endogenous Retrovirus Promotes a Higher-Avidity CD4(+) T Cell Response to Retroviral Infection. *PLoS pathogens*. 8, e1002709
37 Acha-Orbea, H. and MacDonald, H. R. (1995) Superantigens of mouse mammary tumor virus. *Annual review of immunology*. 13, 459-86
38 Barnett, A. *et al.* (1999) Expression of Mouse Mammary Tumor Virus Superantigen mRNA in the Thymus Correlates with Kinetics of Self-Reactive T-Cell Loss. *J. Virol.* 73, 6634-6645
39 Bhadra, S. *et al.* (2009) BALB/Mtv-null mice responding to strong mouse mammary tumor virus superantigens restrict mammary tumorigenesis. *Journal of virology*. 83, 484-8
40 Miller, K. *et al.* (1999) Identification of Multiple Gypsy LTR-Retrotransposon Lineages in Vertebrate Genomes. *Journal of Molecular Evolution*. 49, 358-366

41 Kapoor, A. *et al.* (2010) Discovery and characterization of mammalian endogenous parvoviruses. *Journal of virology*. 84, 12628-35
42 Cui, J. and Holmes, E. C. (2012) Endogenous lentiviruses in the ferret genome. *Journal of virology*. 86, 3383-5
43 van der Kuyl, A. C. (2011) Characterization of a full-length endogenous beta-retrovirus, EqERV-beta1, in the genome of the horse (Equus caballus). *Viruses*. 3, 620-8
44 Taylor, D. J. *et al.* (2011) Evolutionary maintenance of filovirus-like genes in bat genomes. *BMC evolutionary biology*. 11, 336
45 Sawyer, S. L. *et al.* (2005) Positive selection of primate TRIM5alpha identifies a critical species-specific retroviral restriction domain. *Proceedings of the National Academy of Sciences of the United States of America*. 102, 2832-7
46 OhAinle, M. *et al.* (2008) Antiretroelement activity of APOBEC3H was lost twice in recent human evolution. *Cell host & microbe*. 4, 249-59
47 Ikeda, H. *et al.* (1985) Characterization of a molecularly cloned retroviral sequence associated with Fv-4 resistance. *J. Virol.* 55, 768-777
48 Trowsdale, J. and Parham, P. (2004) Mini-review: defense strategies and immunity-related genes. *European journal of immunology*. 34, 7-17
49 Gilman, R. T. *et al.* (2012) Coevolution in multidimensional trait space favours escape from parasites and pathogens. *Nature*. 483, 328-330
50 Bhadra, S. *et al.* (2006) Endogenous MMTV proviruses induce susceptibility to both viral and bacterial pathogens. *PLoS pathogens*. 2, e128
51 Lepperdinger, G. *et al.* (2001) Hyal2 — less active, but more versatile? *Matrix Biology*. 20, 509-514
52 Jadin, L. *et al.* (2008) Skeletal and hematological anomalies in HYAL2-deficient mice: a second type of mucopolysaccharidosis IX? *FASEB journal : official publication of the Federation of American Societies for Experimental Biology*. 22, 4316-26
53 Folimonova, S. Y. (2012) Superinfection Exclusion Is an Active Virus-Controlled Function that Requires a Specific Viral Protein. *Journal of virology*. DOI: 10.1128/JVI.00310-12
54 Ehrengruber, M. U. and Goldin, A. L. (2007) Semliki Forest virus vectors with mutations in the nonstructural protein 2 gene permit extended superinfection of neuronal and non-neuronal cells. *Journal of neurovirology*. 13, 353-63
55 Claus, C. *et al.* (2007) Rubella virus-induced superinfection exclusion studied in cells with persisting replicons. *The Journal of general virology*. 88, 2769-73
56 Walters, K.-A. *et al.* (2004) Superinfection exclusion in duck hepatitis B virus infection is mediated by the large surface antigen. *Journal of virology*. 78, 7925-37
57 Tscherne, D. M. *et al.* (2007) Superinfection exclusion in cells infected with hepatitis C virus. *Journal of virology*. 81, 3693-703
58 Geib, T. *et al.* (2003) Selective Virus Resistance Conferred by Expression of Borna Disease Virus Nucleocapsid Components. *Journal of Virology*. 77, 4283-4290
59 Horie, M. and Tomonaga, K. (2011) Non-Retroviral Fossils in Vertebrate Genomes. *Viruses*. 3, 1836-1848
60 Innan, H. and Kondrashov, F. (2010) The evolution of gene duplications: classifying and distinguishing between models. *Nature reviews. Genetics*. 11, 97-108
61 Marco, A. and Marín, I. (2009) CGIN1: a retroviral contribution to mammalian genomes. *Molecular biology and evolution*. 26, 2167-70
62 Sanville, B. *et al.* (2010) Adaptive evolution of Mus Apobec3 includes retroviral insertion and positive selection at two clusters of residues flanking the substrate groove. *PLoS pathogens*. 6, e1000974
63 Roossinck, M. J. (2011) The good viruses: viral mutualistic symbioses. *Nature reviews. Microbiology*. 9, 99-108
64 Filée, J. *et al.* (2008) Phylogenetic evidence for extensive lateral acquisition of cellular genes by Nucleocytoplasmic large DNA viruses. *BMC evolutionary biology*. 8, 320
65 Jayawardane, G. *et al.* (2008) A captured viral interleukin 10 gene with cellular exon structure. *The Journal of general virology*. 89, 2447-55
66 Slobedman, B. *et al.* Virus-Encoded Homologs of Cellular Interleukin-10 and Their Control of Host Immune Function. *Journal of virology*. 83, 9618-9629
67 Frost, L. S. *et al.* (2005) Mobile genetic elements: the agents of open source evolution. *Nature reviews. Microbiology*. 3, 722-32
68 van der Oost, J. *et al.* (2009) CRISPR-based adaptive and heritable immunity in prokaryotes. *Trends in biochemical sciences*. 34, 401-7
69 Fabozzi, G. *et al.* (2011) Ebolavirus proteins suppress the effects of small interfering RNA by direct interaction with the mammalian RNA interference pathway. *Journal of virology*. 85, 2512-23
70 Fort, P. *et al.* (2011) Fossil Rhabdoviral Sequences Integrated into Arthropod Genomes: Ontogeny, Evolution, and Potential Functionality. *Molecular biology and evolution*. 29, 381-90
71 Bock, R. (2010) The give-and-take of DNA: horizontal gene transfer in plants. *Trends in plant science*. 15, 11-22
72 Koonin, E. V. (2010) Taming of the shrewd: novel eukaryotic genes from RNA viruses. *BMC biology*. 8, 2

73 Tarlinton, R. E. *et al.* (2006) Retroviral invasion of the koala genome. *Nature*. 442, 79-81
74 Maori, E. *et al.* (2007) Reciprocal sequence exchange between non-retro viruses and hosts leading to the appearance of new host phenotypes. *Virology*. 362, 342-9
75 Lusso, P. (2006) HIV and the chemokine system: 10 years later. *The EMBO journal*. 25, 447-56
76 Balanovsky, O. *et al.* (2005) Is spatial distribution of the HIV-1-resistant CCR5Delta32 allele formed by ecological factors? *Journal of physiological anthropology and applied human science*. 24, 375-82
77 Faure, E. and Royer-Carenzi, M. (2008) Is the European spatial distribution of the HIV-1-resistant CCR5-Delta32 allele formed by a breakdown of the pathocenosis due to the historical Roman expansion? *Infection, genetics and evolution : journal of molecular epidemiology and evolutionary genetics in infectious diseases*. 8, 864-74
78 Dunlap, K. A. *et al.* (2006) Endogenous retroviruses regulate periimplantation placental growth and differentiation. *Proceedings of the National Academy of Sciences of the United States of America*. 103, 14390-5
79 Llorens, C. *et al.* (2011) The Gypsy Database (GyDB) of mobile genetic elements: release 2.0. *Nucleic acids research*. 39, D70-4
80 Bénit, L. *et al.* (2001) Identification, phylogeny, and evolution of retroviral elements based on their envelope genes. *Journal of virology*. 75, 11709-19
81 Cornelis, G. *et al.* (2012) Ancestral capture of syncytin-Car1, a fusogenic endogenous retroviral envelope gene involved in placentation and conserved in Carnivora. *Proceedings of the National Academy of Sciences of the United States of America*. 109, E432-41
82 Campillos, M. *et al.* (2006) Computational characterization of multiple Gag-like human proteins. *Trends in genetics : TIG*. 22, 585-9
83 Malik, H. S. (2012) Retroviruses push the envelope for mammalian placentation. *Proceedings of the National Academy of Sciences of the United States of America*. 109, 2184-5
84 Katzourakis, A. *et al.* (2009) Paleovirology reveals the macroevolution of complex retroviruses. *Retrovirology*. 6, O1
85 Gould, S. and Vrba, E. (1982) Exaptation-A Missing Term in the Science of Form. *Paleobiology*. 8, 4 - 15
86 Dupressoir, A. *et al.* (2005) Syncytin-A and syncytin-B, two fusogenic placenta-specific murine envelope genes of retroviral origin conserved in Muridae. *Proceedings of the National Academy of Sciences of the United States of America*. 102, 725-30
87 Mi, S. *et al.* (2000) Syncytin is a captive retroviral envelope protein involved in human placental morphogenesis. *Nature*. 403, 785-9
88 Gilbert, C. and Feschotte, C. (2010) Genomic fossils calibrate the long-term evolution of hepadnaviruses. *PLoS biology*. 8, 12
89 Lefeuvre, P. *et al.* (2011) Evolutionary time-scale of the begomoviruses: evidence from integrated sequences in the Nicotiana genome. *PloS one*. 6, e19193
90 Hatziioannou, T. and Bieniasz, P. D. (2011) Antiretroviral restriction factors. *Current Opinion in Virology*. 1, 526-532
91 Nisole, S. *et al.* (2005) TRIM family proteins: retroviral restriction and antiviral defence. *Nature reviews. Microbiology*. 3, 799-808
92 Malim, M. H. (2009) APOBEC proteins and intrinsic resistance to HIV-1 infection. *Philosophical transactions of the Royal Society of London. Series B, Biological sciences*. 364, 675-87
93 Hultquist, J. F. *et al.* (2011) Human and rhesus APOBEC3D, APOBEC3F, APOBEC3G, and APOBEC3H demonstrate a conserved capacity to restrict Vif-deficient HIV-1. *Journal of virology*. 85, 11220-34
94 Mehle, A. *et al.* (2004) Vif overcomes the innate antiviral activity of APOBEC3G by promoting its degradation in the ubiquitin-proteasome pathway. *The Journal of biological chemistry*. 279, 7792-8
95 Laguette, N. *et al.* (2011) SAMHD1 is the dendritic- and myeloid-cell-specific HIV-1 restriction factor counteracted by Vpx. *Nature*. 474, 654-7
96 Rey-Cuillé, M. A. *et al.* (1998) Simian immunodeficiency virus replicates to high levels in sooty mangabeys without inducing disease. *Journal of virology*. 72, 3872-86
97 Müller, V. and De Boer, R. J. (2006) The integration hypothesis: an evolutionary pathway to benign SIV infection. *PLoS pathogens*. 2, e15
98 Borghans, J. A. M. *et al.* (2005) Limited role for the thymus in SIV pathogenesis. *European journal of immunology*. 35, 42-5
99 Lynch, C. and Tristem, M. (2003) A Co-opted gypsy-type LTR-Retrotransposon Is Conserved in the Genomes of Humans, Sheep, Mice, and Rats. *Current Biology*. 13, 1518-1523

Chapter Four: Host genes evolve under stronger purifying selection in viral genomes

Abstract

Large DNA viruses such as herpesviruses are known to frequently capture host genes, including those that are involved in immunomodulation. The genomes of all herpesviruses contain host-derived chemokines, cytokines and other such immune-system derivatives, interspersed among highly conserved 'core-blocks' of herpesviral genes. Whereas functional studies of such genes reveal their important roles in helping evade the host immune system and improving infectivity, the underlying dynamics of their evolutionary history remain unclear. This is particularly important because host immune genes evolve rapidly due to selection pressure imposed by pathogens, including the viruses that eventually acquire them. In this study, we reveal that host-derived genes a group of herpesviruses are evolving under stronger purifying selection when compared to their host counterparts. We also develop a new implementation of existing maximum likelihood selection analysis to identify selection dynamics at fine detail, allowing us to evaluate the sequences at the level of specific protein domains and less. Our data reveal a previously unknown evolutionary mechanism of the arms race between viruses and hosts. The systematic approach we developed includes a new technique of selection analysis that is widely applicable to a variety of contexts. Furthermore, this pattern is likely to be identified in many other viruses, including those with significant clinical and economic importance. By characterizing such features of the arms-race between viruses and hosts, we will be in a better position to inform epidemiological management and the development of medical interventions.

Introduction

The *Herpesviridae* are a well-described and diverse group of large dsDNA viruses. Their genomes are between 100-200Kbp, often encoding over 100 open reading frames (ORFs), including some with sequence similarity to cellular genes [1]. Although the gene repertoire of different herpesvirus lineages can differ greatly, there are conserved regions that contain a set of core genes that are found across all three *Herpesviridae* subfamilies [2]. Lineage-specific genes are interspersed between these core blocks, including those that derive from the host cell by horizontal gene transfer (HGT). Viruses acquire host genes that increase their fitness, and they are often those involved in nucleic acid metabolism and replication, or gene that evade host defenses [3]. Cellular immunomodulatory genes are frequently identified in herpesviruses, such as cytokines, chemokines and interferon regulatory factors [4,5]. This evolutionary mechanism is so successful that some genes such as interleukin 10 have been repeatedly captured independently in different lineages [6].

Cellular gene capture occurs as part of the evolutionary ‘arms-race’ between viruses and their hosts, where each side is under constant selective pressure to evolve a defensive (host) or evasive (virus) advantage. Because of their large population sizes and high mutation rates, viruses have an advantage over hosts that take longer to adapt. In herpesviruses and indeed other large dsDNA viruses, this advantage is not as clear, since they often have much slower evolutionary rates than other viruses [7]. In their vertebrate hosts, the disadvantage of a slower evolutionary rate is compensated for by complex, multi-layered immunity that includes an adaptive system. It is difficult to study the long-term dynamics of the evolutionary arms-race using contemporary sequence data because the high evolutionary rate of viruses ‘overwrites’ genetic patterns that could have been used to reconstruct past evolutionary processes.

Instead, long term virus evolution can be examined through endogenous viral elements (EVEs), which are genomic features that result from the occasional integration of viral genetic material in the host germline [8–11]. EVEs are usually non-functional relics that are most often used for the purposes of evolutionary inference [11,12]. However, there are a number of examples where they play an active role in the host genome, including as antiviral defense genes [13]. We can investigate the dynamics of this kind of HGT through paleovirology, which synthesizes genomic techniques with information from the reconstruction of EVE natural history through phylogenetics and evolutionary analyses. For instance, mining host genome data to characterize the viral fossil record has revealed

a number of viral elements with intact genes that could represent EVE-derived immunity genes (EDIs) [14,15]. The capture of these EDIs is a mechanism employed by hosts as part of the evolutionary arms race to defend against viruses, just as viruses capture cellular genes to evade host immunity.

In order to explore the evolutionary dynamics of gene capture by viruses, it is important to make use of both comparative viral genomics and phylogenetics reconstruction of genes that exhibit similarity to host counterparts [3]. With the recent abundance of viral genomic data, we can now take a genome-wide approach to studying host-derived genes in the same way that EDIs can be revealed and investigated through paleovirology. In a sense, captured genes are the conceptual counterpart to EVEs; a kind of 'endogenous host element' in viral genomes. In this study, we set out to investigate virally captured host genes in rhadinoviruses, a group of well-described herpesviruses that infect a broad range of mammals including humans. Rhadinoviruses are an ideal study group to systematically reconstruct the evolutionary history of captured genes, since there are seven well-annotated genomes that contain many non-core genes. Moreover, the genus includes *Human Herpesvirus 8* (HHV8) that is the etiological agent of Kaposi's Sarcoma, for which ORFs with sequence similarity to human genes have already been identified [5].

The results of this study conclusively demonstrate the mammalian origin of captured rhadinoviral genes, identifying the specific donor group for seven of them. We investigate the history of evolutionary pressure imposed on captured genes, and develop an implementation of existing maximum likelihood selection analyses to reveal a detailed picture of the dynamics within genes. Our study reveals the remarkable finding that host-derived genes in viruses are under stronger purifying selection in viruses compared to hosts, and we interpret this change as a means to maintain the original function of the gene product. This is further reflected in the fact that amino acid sequence similarity, although low, is maintained despite a dramatic shift in nucleotide composition away from that of the original gene. Moreover, we identify evidence of this functional preservation in the predicted proteins in that they are structurally congruent to their cellular predecessors despite low sequence similarity. Our study reveals that a shift towards stronger purifying selection is a common evolutionary mechanism underlying the capture of cellular genes in rhadinoviruses, demonstrating the advantage of investigating virus-host HGT through comparative genomics under and evolutionary framework.

Results

We constructed a whole genome alignment of seven rhadinoviruses using Mauve [16] to structure the investigation according to the locus of each putative gene capture (figure 1). We focused on genes outside the conserved core blocks, which we know to be shared across all herpesviruses, and arbitrarily designated each of the variable regions A-C (figure 1). Each gene was matched to its likely homolog in other rhadinoviruses by searching for sequence similarity using BLASTp with a .a minimum alignment score threshold of 50. Similar rhadinoviral genes were considered syntenic if they shared similar flanking genes in the equivalent region of the genome. BLAST was also used for each gene to identify similarity to host sequences, which is an indication that the viral sequence could have been derived by HGT. After excluding genes that were too divergent to align to similar host genes with confidence, we shortlisted 13 genes for further phylogenetic and evolutionary analysis.

Major Histocompatibility Complex, Interleukin 6 and Interleukin 17

The first three genes in RHVP, designated as R1, R2 and R3, show sequence similarity to major histocompatibility complex (MHC) genes (figure 1), and were aligned in the initial characterization of the viral genome with human HLA-A2 and mouse H-2K^{b} [17]. In a re-analysis of these genes, we have determined that each of the genes were similar to different regions of the MHC gene in a co-linear manner. This lead us to the conclusion that they are likely derived from the capture of a single gene, and it was possible to align all three ORFs to a range of MHC homologs. We reconstructed a phylogeny from a diverse set of mammalian hosts, revealing that that RHVP R1-3 are most closely related to the mouse and rat homologs. Along with the fact that RHVP was isolated from the pygmy rice rat (*Oligoryzomys microtis*), the phylogenetic similarity to *Mus musculus* and *Rattus norvegicus* is consistent with an acquisition from a rodent host (figure 2).

Interleukins (ILs) are a large group of cytokines, of which genes from several different families have been captured by rhadinoviruses. IL-17 is a family of 6 types of interleukin involved in the inflammatory response (IL-17A to F) [18]. SaHV2 acquired a copy of IL-17A, and the fact that this gene is not present in AtHV3 suggests that the capture event occurred after their speciation. Phylogenetic analysis is consistent with this hypothesis (figure 3), since despite the long branch of SaHV2 IL-17A, the gene groups robustly with the IL-17genes of squirrel monkey (*S. boliviensis*) and marmoset (*C. jacchus*), which speciated after the divergence of their common ancestor from spider monkeys (*Ateles sp.*). This could mean that the gene was acquired by the SaHV2 ancestor after the squirrel

monkey-spider monkey split, but before the squirrel monkey-marmoset split at ~20 million years [19]. This result suggests that the capture event may have occurred in a small 3 million year window, when new world monkeys were undergoing a major radiation that separated the major clades (*Atelidae*, *Cebidae* & *Pithicidae*) [20].

The genomes of HHV8 and MaHV5 contain homologs of IL-6, which is a secreted cytokine known to have a variety of functions, ranging from roles in immunomodulation to oncogenesis [21]. The phylogenetic analysis of these genes revealed the sequences are embedded within a clade that includes mouse, rat and tree shrew (figure 3). This is surprising given that both HHV8 and MaHV5 are primate-infecting viruses, which suggests that the gene may have been acquired from a non-primate host during cross-species transmission. Alternatively, the results could be false, and the topology is an artifact of the virus' long branch.

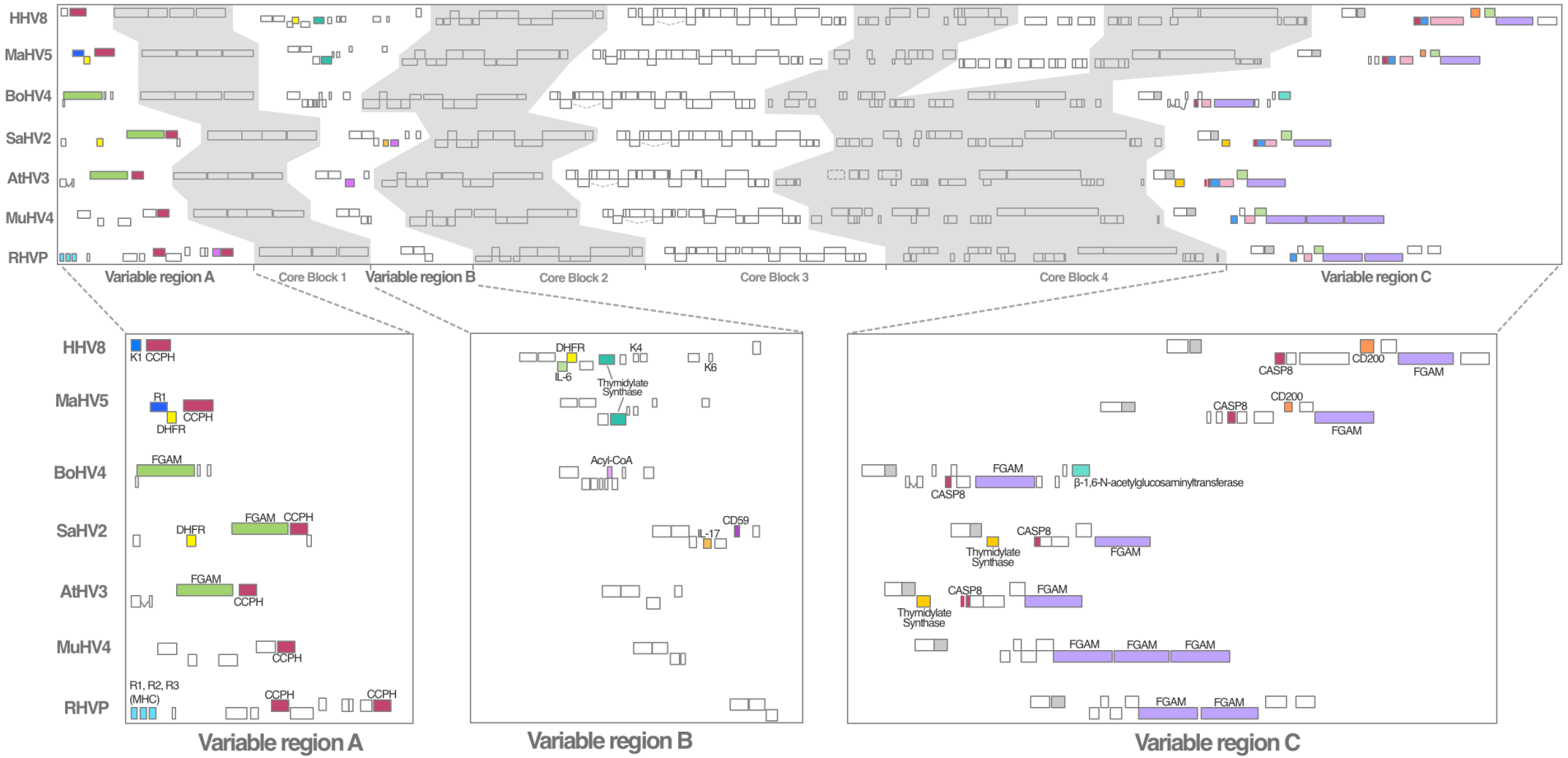


Figure 1 Whole genome alignment of Rhadinoviruses

The figure depicts a scaled cartoon of an alignment of seven rhadinoviral genomes, where boxes represent ORFs according to the genbank annotation. Both lineage-specific (variable) and core herpesvirus blocks are indicated by the labels below the alignment, and are highlighted in grey or while in alternating fashion. The region in HHV8 and MaHV5 corresponding to core block 4 includes a stretch of lineage specific genes that is not part of the set of conserved genes. Horizontally acquired genes investigated in this analysis are indicated in colour and annotated in the zoomed-in panels below the alignment. Virus abbreviations are: HHV8: Human Herpesvirus 8, MaHV5: Macacine Herpesvirus 5, BoHV4: Bovine Herpesvirus 4, AtHV3: Ateline Herpesvirus 3, MuHV4: Murine Herpesvirus 4, RHVP: Rodent Herpesvirus Peru.

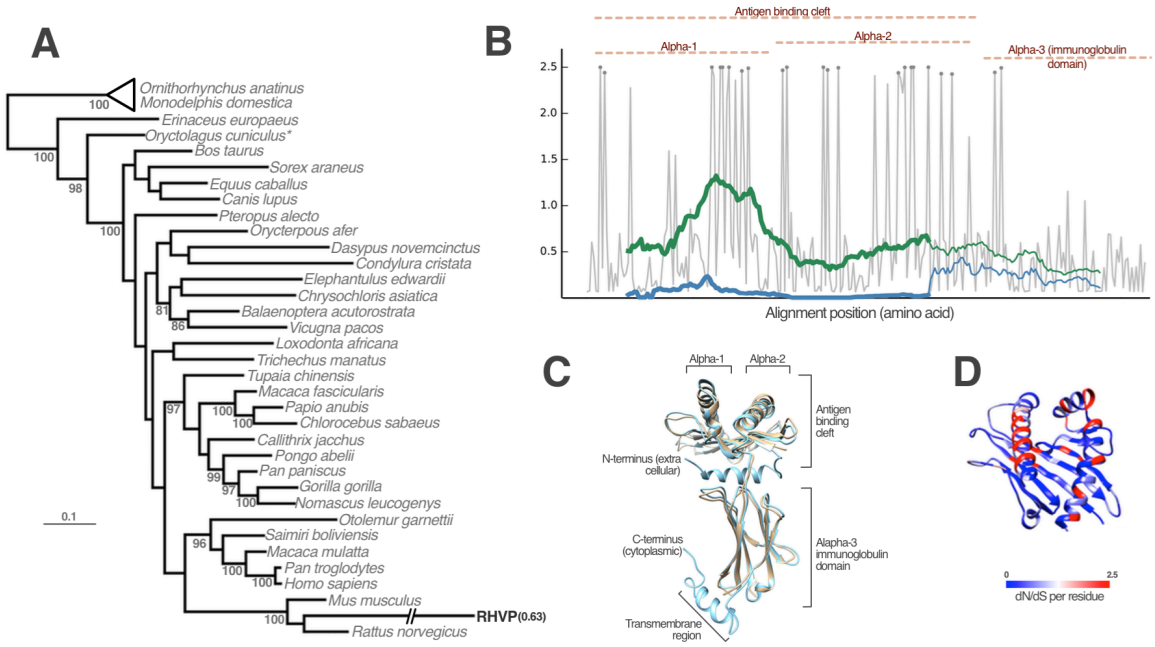


Figure 2 Capture of a Major histocompatibility complex (MHC) gene by RHVP
Panel A: Bayesian phylogenetic reconstruction of mammalian MHC class I genes with the RHVP homolog. The RHVP branch length has been truncated to fit in the figure and the total branch length is indicated in brackets. The scale bar represents 0.1 substitutions per site. The numbers at each node represent posterior probability and values below 80 are not shown. This tree shows that the RHVP homolog and the genes of *M. musculus and R. norvegicus* group together with 100% probability. Panel B: The graph depicts a collation of a number of selection analysis results. The grey line indicates the values of dN/dS obtained for each residue. Sites shown to be under significant positive selection according to the BEB analysis are indicated by a grey dot along the line. The green (host) and blue (virus) lines represent the results of the sliding window branch test analysis. The thick line indicates where the branch test rejected the null hypothesis of a single dN/dS according to a likelihood ratio test. Panel C: Structural model of the MHC homolog with the best matching template in the database. There is a helix at both termini that could not be modelled suggesting that they could have been lost in the viral copy (the corresponding region is not homologous). This suggests that the viral copy does not function as a transmembrane protein since the C-terminus helix is needed for this. Panel D: The same structural model as in panel C, but only showing the antigen binding cleft coloured according to the dN/dS values for each residue. Nearly all of the positively evolving sites are found in this region.

Dihydrofolate reductase, CD200 immunoglobulin & Chemokine C-C motif ligand 3 (CCL3)

Several of the host-derived genes investigated were found in multiple viruses, suggesting that they were acquired before the viruses diverged. The genomes of MaHV5 and HHV8 also include ORFs with similarity to the small cellular cytokine CCL3 that plays a role in the inflammatory response. Three such genes were identified in HHV8, and phylogenetic reconstruction places the clade of viral genes within primates with over 80% probability, but their exact position is not well supported (figure 4). The fact that the HHV8 genes group together to the exclusion of the MaHV5 copy with 97% probability indicates that they were duplicated within the herpesvirus after capture. The two viruses also encode a gene with similarity to CD200, a membrane glycoprotein of the immunoglobulin superfamily also known as OX-2. Experimental evidence has shown that purified HHV8 OX-2 stimulates the production of inflammatory cytokines in myeloid-lineage cells [22].

Phylogenetic reconstruction demonstrates conclusively the gene is primate derived with high posterior probability (figure 5A). The topology indicates that the viral gene is most similar to new world monkey host genes (*Saimiri boliviensis* and *Callithrix jacchus*), although the extremely long branches could have influenced the placement.

Three rhadinoviruses contain a homolog of dihydrofolate reductase (DHFR), which is a universal cellular enzyme that catalyses the conversion of dihydrofolate to tetrahydrofolate [23]. Tetrahydrofolate is required for purine and thymidylic acid synthesis, as well as synthesis of amino acids such as glycine and cysteine [23]. Although the gene is situated in variable region A for MaHV5 and SaHV2, the HHV8 DHFR homolog is located downstream of this in variable region B. If all three genes originate from the same capture event, then the DHFR in HHV8 must have been moved through illegitimate recombination between variable region A and B. However, there is a 23 amino acid long carboxyl-terminus region in HHV8 DHFR that is absent in MaHV5 and SaHV2, a difference that has been pointed out as supportive of a separate capture scenario to explain the distinct locus [24]. Although the tree in this study shows that the genes cluster together with high posterior probability, it is not possible to rule out a separate capture since intermediate lineages may not have been sampled (figure 5).

CD59 glycoprotein & Core 2 β-1,6-N-acetylglucosaminyltransferase-mucin (C2GnT-M)

The BoHV4 Bo17 gene has been previously identified as a host-derived C2GnT-M gene that is involved in glycan synthesis. It has already been established that the gene is likely derived from an ancestor of the African buffalo (*Syncerus caffer*) [25], which has since been shown to be the most likely natural host of BoHV4 that subsequently transmitted to cattle multiple times [26]. The tree reconstructed in our analysis is consistent with these previous findings, placing the BoHV4 C2GnT-M gene next to the African buffalo host gene with a high posterior probability (figure 4).

SaHV2 encodes a homolog of the mammalian cell-surface glycoprotein CD59, first identified during the genome-sequencing project of the virus in 1992. The results of our phylogenetic analysis demonstrate that the gene was acquired through HGT, since the viral homolog resolves next to *Saimiri boliviensis* with a high posterior probability to the exclusion of *Callithrix jacchus* CD59 (figure 4). We can therefore conclude with confidence that the gene was captured after the *Callithrix-Saimiri* divergence (~20mya [19]) from either the current host (*Saimiri*) or a closely related monkey that may be susceptible to

infection. This is also supported by the fact that the gene is absent in AtHV3, the most closely related virus to SaHV2.

Phosphoribosylformylglycinamidine synthase & Complement Control Protein

Although this study is not focused on genes that were likely captured before the emergence of rhadinoviruses, we examined the evolution of two such genes due to their extensive post capture duplication and rearrangement in rhadinoviruses. Phosphoribosylformylglycinamidine synthase (FGAMS) is a purine synthesis enzyme, which was captured early in the evolution of gammaherpesviruses, being shared by all the viruses in the group. Among Rhadinoviruses however, a number of duplications, deletions and translocations have occurred, though have not been fully investigated. We conducted a phylogenetic reconstruction of available gammaherpes FGAMS genes to trace the series of events that lead to the current state of rhadinovirus FGAMS (figure 6A). Consistent with previous observations, the tree shows that after the ancestral capture, an initial duplication event occurred in the ancestor of macaviruses, percaviruses and rhadinoviruses. This second copy appears to have been lost in the MaHV5-HHV8 ancestor, but duplicated a second time in the ancestor of RoHV, MuHV4, WMHV. These rodent viruses also appear to have lost the original FGAMS, since their copies are more closely related to the second copy, despite being in a different position in the genome. This suggests that a translocation event may have occurred at the time of this secondary event. This second duplicate was itself duplicated in the MuHV4-WMHV ancestor, since both genomes contain a third copy that is most closely related to it.

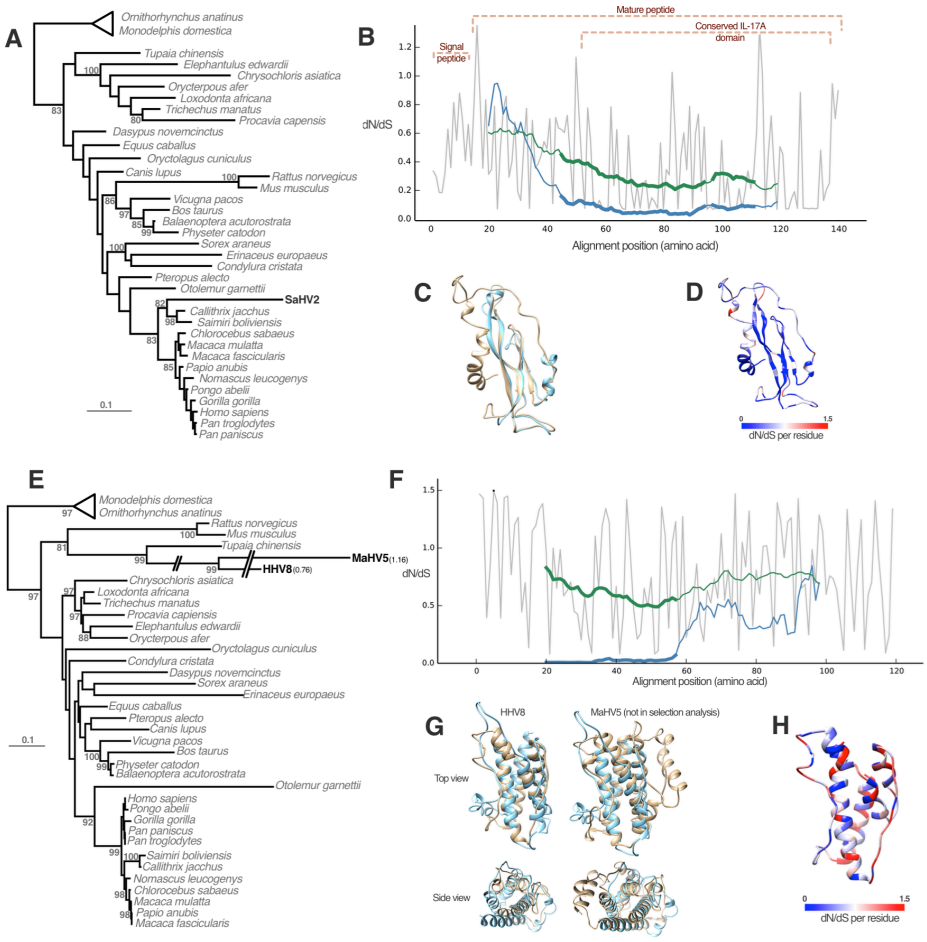


Figure 3 Host-derived Interleukin 6 & 17 in MaHV5, HHV8 and SaHV2

Panel A: Bayesian phylogenetic reconstruction of mammalian interleukin 17 and a homologous sequence in SaHV2. The scale bar represents 0.1 substitutions per site. The numbers at each node represent posterior probability and values below 80 are not shown. The tree shows that the viral gene groups most closely to the new world monkeys *C. jacchus and S. boliviensis*, the latter being the virus' current host. Panel B: The graph depicts a collation of a number of selection analysis results. The grey line indicates the values of dN/dS obtained for each residue. No sites under significant positive selection according to the BEB analysis were identified. The green (host) and blue (virus) lines represent the results of the sliding window branch test analysis. The thick line indicates where the branch test rejected the null hypothesis of a single dN/dS according to a likelihood ratio test. The conserved domains and other regions in interleukin 17 are annotated in brown. Panel C: Structural model of the interleukin 17 homolog with the best matching template in the database, showing precise overlap for most of the structure. Panel D: The same structural model as in panel C without the template model, and coloured according to the dN/dS values for each residue. Panel E: Bayesian phylogenetic reconstruction of mammalian interleukin 6 and homologous sequences in HHV8 and MaHV5. The scale bar represents 0.1 substitutions per site. The numbers at each node represent posterior probability and values below 80 are not shown. The tree shows that the viral genes group most closely to the tree shrew (*T. chinensis*) with 99% probability despite belonging to viruses that infect primates. This could either indicate that a cross species transfer had occurred in the history of the virus, or that long-branch attraction has resulted in this unusual topology. Panel F: The graph depicts a collation of a number of selection analysis results, but these did not include the MaHV5 gene since it could not be aligned for the entire length (219 gaps in a 384 nucleotide alignment). The grey line indicates the values of dN/dS obtained for each residue. Sites shown to be under significant positive selection according to the BEB analysis are indicated by a grey dot along the line. The green (host) and blue (virus) lines represent the results of the sliding window branch test analysis. The thick line indicates where the branch test rejected the null hypothesis of a single dN/dS according to a likelihood ratio test. Panel G: Structural model of both interleukin 6 homologs with the best matching template in the database. Note that as well as sequence divergence, the MaHV5 structure is also more divergent than the HHV8 counterpart. Panel H: The structural model of HHV8 as in panel G without the template model, and coloured according to the dN/dS values for each residue. Residues with a dN/dS above 1 are distributed throughout the structure with no discernable pattern.

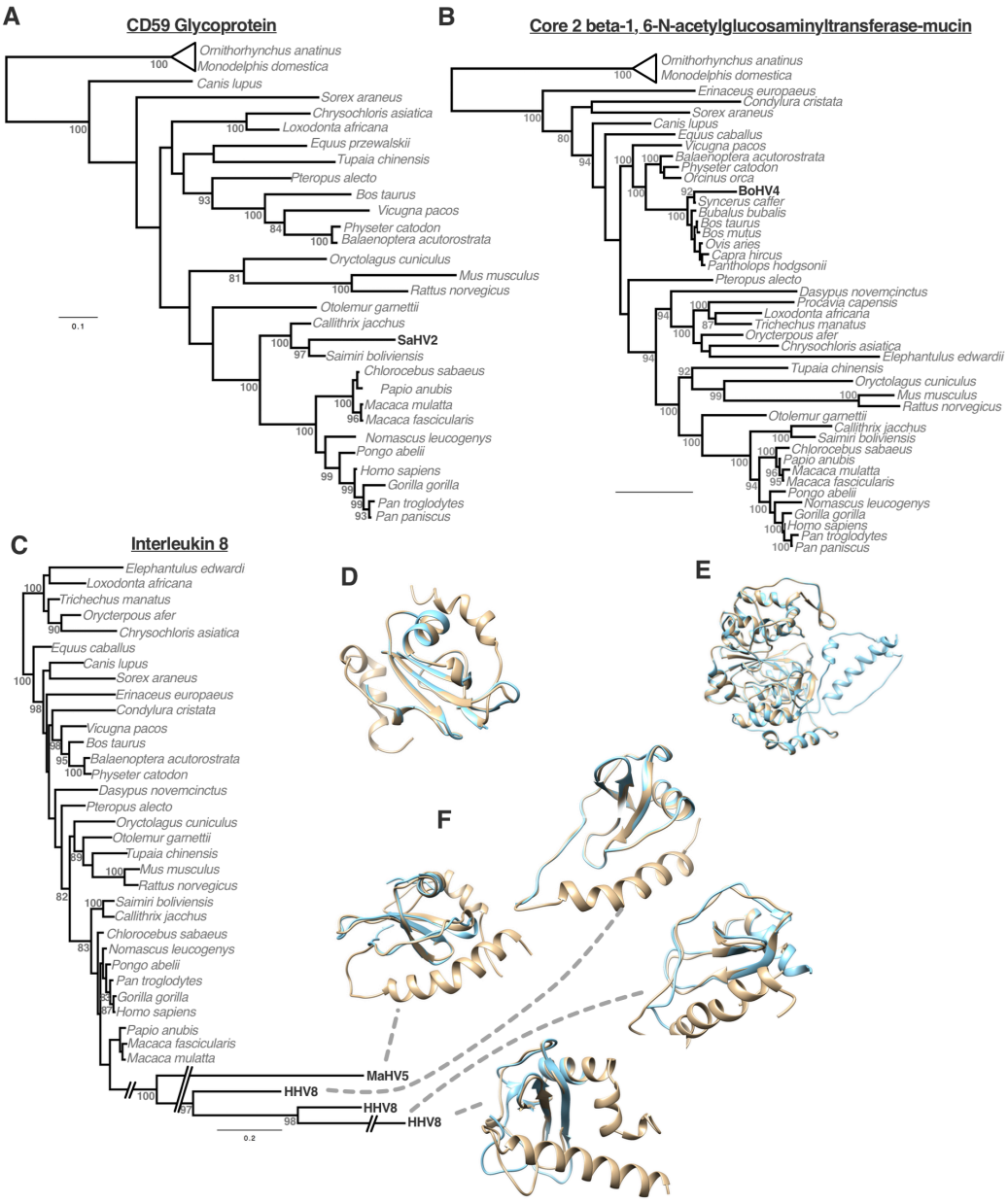


Figure 4 Phylogenetics & structural models of captured genes in HHV8, BoHV4 & SaHV2. This figure contains the genes from our main analysis for which a sliding window analysis was not conducted because of a short alignment length in the case of CD59 and CCL3 and because the overall branch test was not significant for the C2GnT-M. Panel A: Bayesian phylogenetic reconstruction of mammalian CD59 glycoprotein genes and homologous sequences in SaHV2. The scale bar represents 0.1 substitutions per site. The numbers at each node represent posterior probability and values below 80 are not shown. The tree shows that the viral gene groups with *S. boliviensis* with 97% probability, indicating that the gene is most likely derived from this monkey, which is the virus' current host. The relatively short branch length is also an indication that the capture has occurred recently. Panel B: Bayesian phylogenetic reconstruction of mammalian C2GnT-M genes with a homolog from BoHV4. The scale bar represents 0.1 substitutions per site. The numbers at each node represent posterior probability and values below 80 are not shown. This tree confirms previous studies that have reconstructed trees with comparable topology. This has been previously interpreted as evidence that the origin of the gene is *S. caffer.* . Panel C: Bayesian phylogenetic reconstruction of mammalian CCl3 genes with homologous genes from MaHV5 and three copies in HHV8. The scale bar represents 0.1 substitutions per site. The numbers at each node represent posterior probability and values below 80 are not shown. Panels D, E & F: Structural models of CD59, C2GnT-M and CCL3 with the best matching database templates, respectively.

A homolog of the host gene that encodes the complement control protein (CCP) is found in all rhadinoviruses except BoHV4. As with several genes of the complement system (a subset of innate immunity), CCP contains modular sushi repeats, which are conserved ~60 amino acid domains that exist in a range of complement and adhesion proteins [27]. In rhadinoviruses, the number of sushi domains is variable in the different homologs (between 4 and 8). The rhadinoviral homologs are most closely related to the conserved mammalian complement component 4 binding protein, but could not be aligned with confidence due to low sequence identity. Nonetheless, we hypothesized that CCP genes are orthologs because they are all found at the same locus and are partially syntenic (only SaHV2 and AtHV3 have the same gene on both side) (figure 6B). HHV8, MuHV4 and SaHV2 each have one gene with similarity to CCP containing 4 sushi domains, but none of the most closely related viruses to each of these viruses possess a CCP ortholog with the same number of sushi domains. Based on a high posterior probability grouping the genes in the phylogeny, we infer that they originate from a single capture. Inferring from the domain organization of a similar CCP host gene (CD46), we hypothesize that the ancestral gene likely contained 4 sushi domains. On this basis, the AtHV3 appears to have lost a sushi domain since it diverged from SaHV2. In MaHV5, an apparent duplication has resulted in a gene that is twice as long as the HHV8 copy (with 8 sushi domains). A similar duplication seems to have occurred in RHVP compared to MuHV4, but in this case they are two separate ORFs separated by several genes and each missing two sushi domains.

To investigate the possible evolutionary history of duplication and loss of CCP genes, we conducted a co-phylogeny analysis using JANE 4.0. This allowed us to reconstruct the possible series of events based on the respective phylogenetic relationships among the genes and the relative topology of the gene tree to the virus tree. This revealed one most parsimonious solution when HGT between the viruses is permitted, and one solution where no HGT is allowed. In both instances, we used the default cost settings for duplication, loss and failure to diverge. Under the latter scenario, only SaHV2 and AtHV3 have retained the ancestrally captured gene, which was replaced with subsequent duplicates in all other viruses. The first duplication gave rise to the 5' end of the MaHV4 CCP as well as the copy in RHVP and MuHV4. This copy underwent a further two rounds of duplication leading to the HHV8 copy and the 3' end of the MaHV5 gene. In RHVP, a lineage-specific duplication gave rise to the second ORF, after which both genes lost 2

sushi domains. Alternatively, under a model where HGT is permitted, the analysis reveals that the ancestral copy gave rise to the contemporary gene in SaHV2, AtHV3, HHV8 and the 5' end of the MaHV5 gene. The 3' end in MaHV5 CCP was the result of a duplication and transfer of a copy from HHV8, which itself also transferred to the MuHV4-RHVP ancestor. Thereafter, the extra copy in RHVP originates from a similar duplication and transfer from MuHV4.

Other genes

Some non-core genes exhibited similarity to cellular genes but were too divergent to determine the host group of origin with statistical confidence or else could not be confidently aligned to begin with. For example, HHV8 and MaHV5 share a number of interferon-like genes that were probably acquired after their divergence from other rhadinoviruses. Interestingly, these genes are located within a single stretch amidst core block 4 that is otherwise syntentic in nearly all herpesviruses (figure 1). There are also genes that could be aligned to host homologs, but phylogenetic analysis was inconclusive. For example, the BoHV4 genome includes a small hypothetical open reading frame with similarity to Acyl-CoA thioesterase, which is a family of cellular genes that encode enzymes involved in metabolism (figure S1C). Similarly, all of the rhadinoviruses apart from MuHV4 and RHVP encode a CASP8-like gene, which in hosts is a protease involved in programed cell death. Although the phylogeny was inconclusive with regard to the particular host lineage of origin, it revealed that the gene does group within mammalian homologs with 93% probability (figure S1A).

HHV8, MaHV5, SaHV2 and AtHV3 all possess another nucleotide metabolism gene - thymidilate synthase (TS). It is in the same location for MAHV5 and HHV8 but situated near the right terminal of the genome for SaHV2 & AtHV3 (figure 1). Although the phylogeny exhibited poor support for most nodes, the tree reveals that the pair of TS genes in SaHV2 and AtHV3 have much shorter branches than the HHV8/MaHV5 genes, suggesting that they originate from a distinct and more recent capture (figure S1D). The fact that the pairs of TS genes are at opposite locations of their respective genomes also supports this scenario.

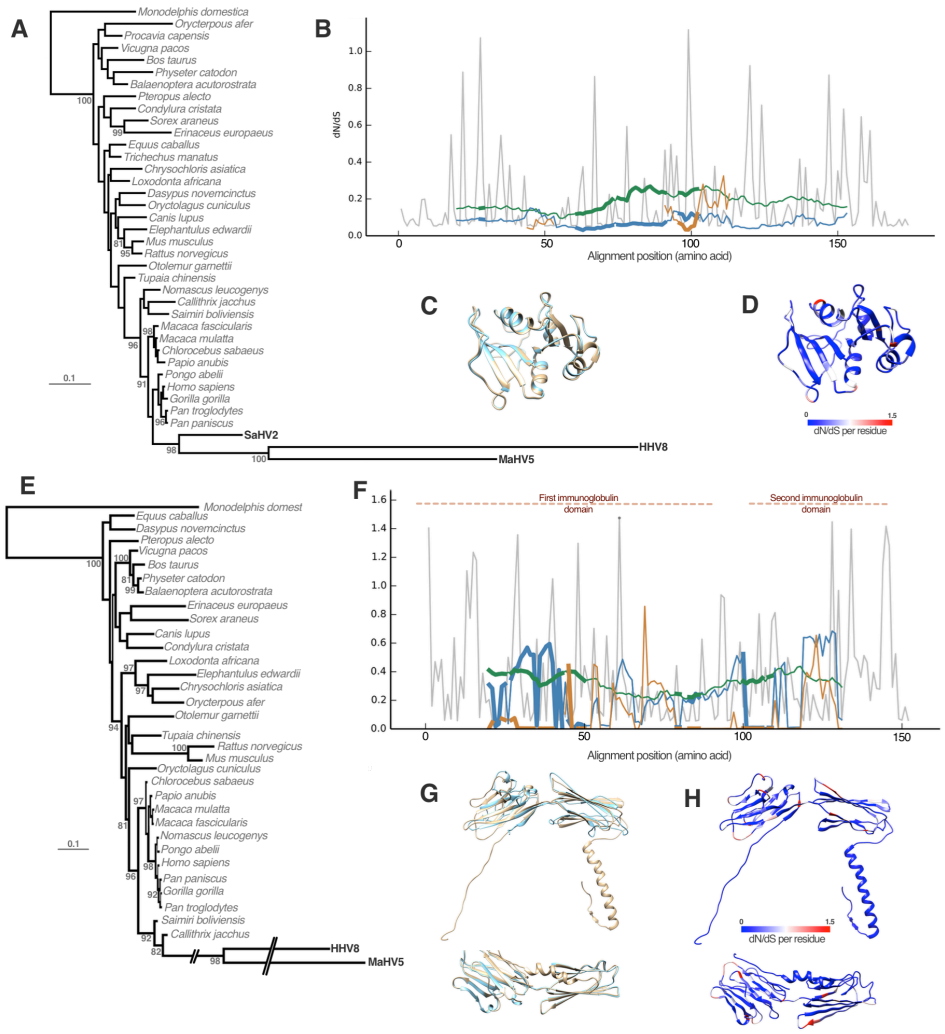


Figure 5 Host-derived DHFR and CD200 genes in SaHV2, HHV8 & MaHV5

Panel A: Bayesian phylogenetic reconstruction of mammalian DHFR and homologous sequences in SaHV2, HHV8 and MaHV5. The scale bar represents 0.1 substitutions per site. The numbers at each node represent posterior probability and values below 80 are not shown. The tree shows that the viral gene groups within primates with 91% probability and is probably most closely related to great ape genes as there is a posterior probability of 96% that the virus and the clade of great apes are sister clades. Panel B: The graph depicts a collation of a number of selection analysis results. The grey line indicates the values of dN/dS obtained for each residue. No sites under significant positive selection according to the BEB analysis were identified. The green (host), blue (virus) and orange (viral stem branch) lines represent the results of the sliding window branch test analysis. The thick line indicates where the branch test rejected the null hypothesis of a single dN/dS according to a likelihood ratio test. dN/dS estimates of infinity for the viral stem branch are not shown for clarity. Panel C: Structural model of the DHFR homolog with the best matching template in the database, showing precise overlap for the entire structure. Panel D: The same structural model as in panel C without the template, and coloured according to the dN/dS values for each residue. Panel E: Bayesian phylogenetic reconstruction of mammalian CD200 genes and homologous sequences in HHV8 and MaHV5. The scale bar represents 0.1 substitutions per site. The numbers at each node represent posterior probability and values below 80 are not shown. The viral genes are most closely related to new world monkey host genes from *C. jacchus and S. boliviensis* with 92% probability, and more closely related to *C. jacchus* with 82% probability. Panel F: The graph depicts a collation of a number of selection analysis results. The grey line indicates the values of dN/dS obtained for each residue. One site under significant positive selection according to the BEB analysis is indicated by a grey dot. The green (host), blue (virus) and orange (viral stem branch) lines represent the results of the sliding window branch test analysis. The thick line indicates where the branch test rejected the null hypothesis of a single dN/dS according to a likelihood ratio test. dN/dS estimates of infinity for the viral stem branch are not shown for clarity. Panel G: Structural model of HHV8 (upper) and MaHV5 (lower) homologs with the best matching template in the database. Panel H: The structural model of HHV8 and MaHV5 as in panel G without the template model, and coloured according to the dN/dS values for each residue.

The leftmost HHV8 gene K1, encodes a transformation-associated glycoprotein that performs roles in signal transduction, possibly involved in immunomodulation and perpetuating cell survival [28]. Based on a BLAST search against herpesviruses, K1 does not appear to have homologs in other genomes, except for very weak sequence similarity to MaHV5 R1 (28% amino acid identity, ~50% query coverage). While we cannot identify host sequences with good similarity to K1, the R1 gene of MaHV5 is most similar to the N-terminus of the low affinity receptor for the FC fragment of immunoglobulin G in primates (31% amino acid identity to *Callithrix jacchus, 41% coverage*). It is possible that this gene is predecessor to the viral ORF, which is is consistent with functional studies of K1, which is distantly similar to R1. It has been shown that K1 can cause the down regulation of the B cell antigen receptor complex (BCR), which is known to occur via the interaction of the N-terminus with the u chain of BCR [29]. This would explain why sequence conservation has been maintained at the N-terminus.

We were unable to robustly estimate a phylogeny of the R1 gene with host homologs, probably due to the effects of long-branch attraction. Nonetheless, the MaHV5 gene groups with placental mammals with high posterior probability (figure S1). K1 and R1 are situated in the same genomic position and both contain immunoglobulin-like domains. This synteny could be interpreted as evidence that the genes are highly divergent orthologs that were acquired from the ancestral primate host, but we cannot rule out the scenario that they are independent acquisitions. The latter possibility is supported by their extreme divergence from one another compared to the rest of the genomes (the viruses are otherwise easily aligned to one another). Even among HHV8 K1 genes, the sequence diversity is extremely high among different strains suggests that there is a mechanism of diversifying selection [30], and the same evolutionary pressure can be invoked to explain the repeated capture of the same gene.

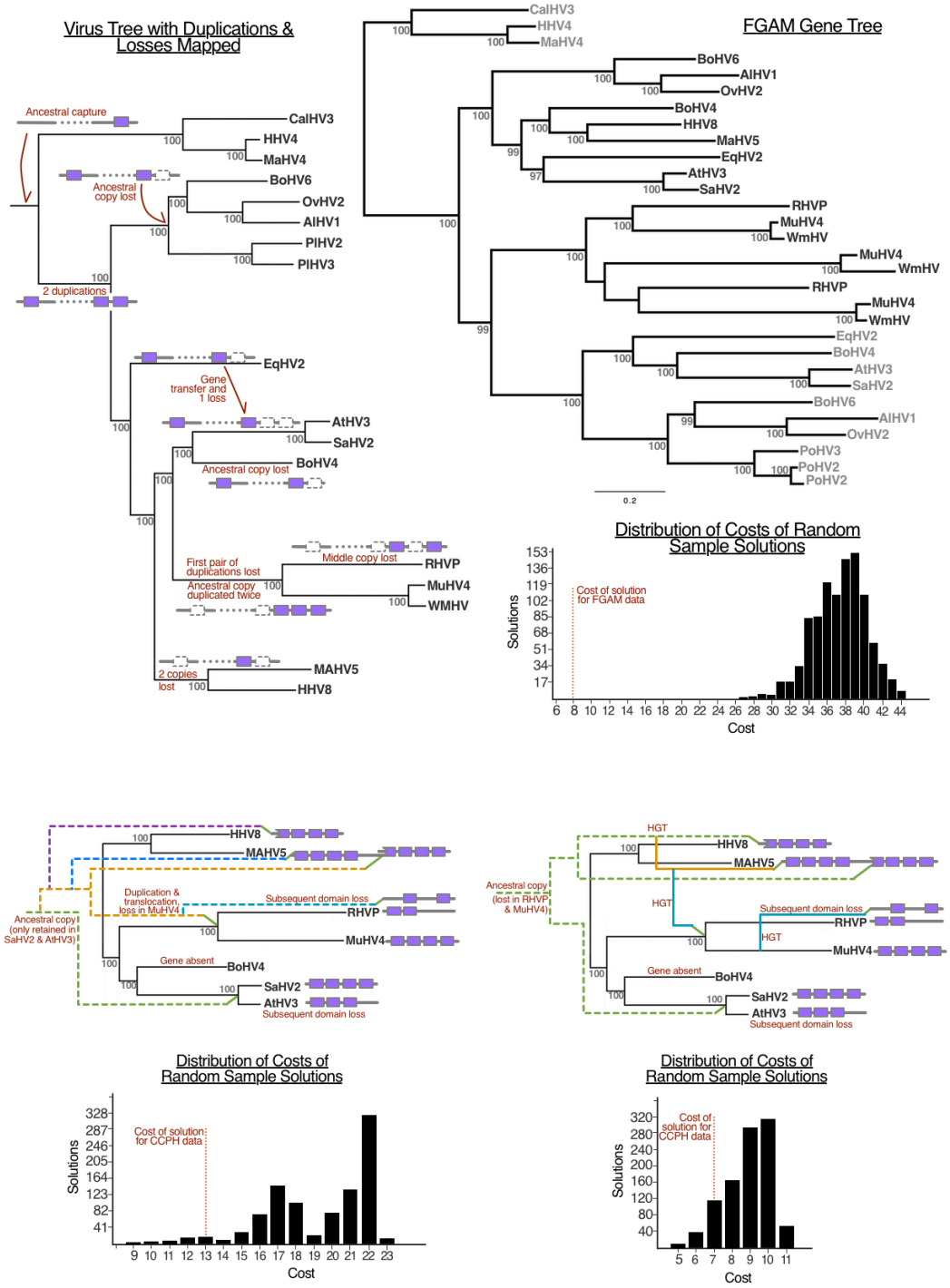


Figure 6 Co-phylogeny analysis for FGAM and CCCP

Panel A: The results of the co-phylogeny analysis for FGAM synthase homologs in rhadinoviruses. Although the viral sequences are too divergent to align with host genes, a series of duplications, losses and rearrangements have occurred in the clade. The viral tree is shown with reconstructed events superimposed at each node. The cartoons represent a simplified viral genome with FGAM genes at either end. Dotted lines indicate a loss event, and the duplications can be deduced based on a comparison of a node with the previous node. Panel B: Phylogenetic tree of FGAM genes from rhadinoviruses. Dark taxon labels indicate 3' end loci. Scale bar represents substitutions per site. Panel C: A histogram of the cost of a random sample solution, showing that the co-phylogeny reconstruction for FGAM is well below the random distribution. Panel D & E: Results of the co-phylogeny analysis for CCP homologs superimposed on the virus tree as a cladogram (the branch lengths of coloured dotted lines is arbitrary). Each of the SUSHI domains within CCP viral homologs is represented as a cartoon next to each taxon label. The best (only) solution is show with (Panel E) and without (Panel D) horizontal gene transfer allowed. The cost for both of the CCP solutions lies within the distribution of costs for random samples.

Post-capture evolutionary changes

Analysis of Selection

We determined that the selection pressure imposed on viral genes was significantly different to that of their cellular homolog (likelihood ratio test, $p<=0.05$) in seven of the eight genes tested (C2GnT-M was the exception). Using the branch-model implemented in codeml, we found that the viral genes exhibit lower values of dN/dS compared to their cellular counterparts (table 1). For genes where multiple viruses contained a homolog (DHFR, CD200 and CCL3), we find that the drop in dN/dS is evident whether the model tested implemented separate estimates for the viral stem and crown group branches or whether both were designated a single estimate.

We also identified between 1-19 sites under positive selection in 6/8 of the genes tested (table 1). Since the test of positive selection considers the whole alignment for all taxa, the results primarily reflect the evolutionary history of the host genes, since they represent the majority of sequences in the data. This indicates that the genes are evolving adaptively in host genomes, and are likely to be an underestimate of the true number of positively selected sites due to the limitations of selection analyses (see methods). We also tested whether the overall dN/dS reduction in viral branches according to the branch model could be due to an artifact of these positively evolving sites in the host that were biasing the dN/dS estimates. We therefore implemented a modification to the branch-model where these sites are excluded, and find that the drop in dN/dS is still significant albeit with a slightly lower magnitude (table 2). Interestingly in the case of the MHC homologue in RHVP, we also found a lower average dN/dS in viruses for an alignment of the sites under positive selection (0.0049 compared to 0.042 in the hosts), with a much higher fold-drop compared to a test of the rest of the alignment (470 fold compared to 9.5 fold in hosts).

We implemented a sliding-window modification of the branch model, and find that the difference in dN/dS between cellular and viral counterparts is non-uniform along the length of the genes. In the case of MHC, the test reveals that the viral homolog is only significantly lower in the regions that correspond to the alpha-1 and alpha-2 domains (figure 2). Together these form the antigen binding cleft, which is where the majority of positively selected sites were detected. Moreover, the viral dN/dS values within the alpha-1/2 region are lower than the values of the alpha-3 region where the virus and cell dN/dS cannot be statistically distinguished (figure 2).

In Il-17A, we find that the significant difference in dN/dS is localized to the conserved Il-17 domain and signal peptide, and for both IL-6 and DHFR, the drop in dN/dS was limited to roughly half of the gene (figure 3 & 4, respectively). Since a homolog of DHFR is found in more than one rhadinovirus, the sliding window branch test also revealed a small stretch of significantly different dN/dS for host branches, viral tips and the viral stem branch (figure 5). For the CD200 homolog, we find that several regions across the gene that exhibit a significant difference in dN/dS between host and viral branches. Similarly to DHFR, the 3-model branch-test also revealed multiple regions with a different dN/dS in the viral stem branch (figure 5). Unlike the rest of the genes analysed however, the CD200 analysis shows that the differences in dN/dS are not contiguous, but rather interspersed by regions where a significant difference could not be detected between virus and host values. Furthermore, the sliding window test also revealed that the viruses are evolving under comparatively higher dN/dS in the virus, with peaks at 3 different regions (figure 5).

Gene	Site model M7 & M8	Model	Free parameters	lnL	2ΔlnL	p-value (better than 1 ratio)	omega virus	omega host	omega virus tip	Δω significantly better than ω0	fold change
Interleukin 17A	no result	ω0	1	-5054.408722							
		ω0,ωSaHV2	2	-5049.903145	9.011154	0.002683369	0.1726	0.3715	n/a	-0.1989	2.152375435
Dihydrofolate reductase	0	ω0	1	-5294.084281							
		ω0,ωiv	1	-5294.049062	0.070438	0.790700117					
		ω0,ωiv=ωHHV8=ωMaHV5=ωSaHV2	2	-5290.308975	7.550612	0.00599897	0.0871	0.1708	n/a	-0.0837	1.960964409
		ω0,ωiv,ωHHV8=ωMaHV5=ωSaHV2	3	-5289.108978	9.950606	0.006906426	∞	0.1694	0.0735	n/a	n/a
glucosaminyl (N-acetyl) transferase 3	5	ω0	1	-13058.34234							
		ω0,ωBoHV4	2	-13058.34176	0.001158	0.972853707					
MHC class II (mammals)	22	ω0	1	-11175.03628							
		ω0,ωRHVP	2	-11137.71358	74.645394	0	0.0506	0.5285	n/a	-0.4779	10.44466403
Interleukin 6	1	ω0	1	-5632.996227							
		ω0,ωiv	2	-5626.185996	13.620462	0.000223734	0.0101	0.7086	n/a	-0.6985	70.15841584
CD59 glycoprotein	8	ω0	1	-3437.224162							
		ω0,ωSaHV2	2	-3418.956735	36.534854	0.000000001	0.0682	0.7418	n/a	-0.6736	10.87683284
CD200 immunoglobulin	1	ω0	1	-5270.40937							
		ω0,ωiv	2	-5265.533087	9.752566	0.001790726	0.0035	0.3212	n/a	-0.3177	91.77142857
		ω0,ωiv=ωHHV8=ωMaHV5	2	-5266.501778	7.815184	0.005180907	0.0163	0.3214	n/a	-0.3051	19.71779141
		ω0,ωiv,ωHHV8=ωMaHV5	3	-5264.964464	10.889812	0.004318246	0.0034	0.3225	0.0329	n/a	n/a
CCL3 chemokine	1	ω0	1	-2001.226622							
		ω0,ωiv	2	-1998.834746	4.783752	0.028729459	0.0025	0.3801	n/a	-0.3776	152.04
		ω0,ωiv=ωHHV8=ωMaHV5	2	-1981.355092	39.743060	0	0.008	0.4829	n/a	-0.4749	60.3625
		ω0,ωiv,ωHHV8=ωMaHV5	3	-1981.150834	40.151576	0.000000002	0.0026	0.4833	0.008	n/a	n/a

Table 2 A collection of the results of the selection analyses conducted for the eight main genes investigated in this study.
The number of positively evolving sites detected according to the Bayes empirical Bayes procedure is indicated under the column heading 'Site model M7 & M8'. The details and results for the branch-model, including each of the models tested for each gene are also shown. Models are represented as the combination of different dN/dS estimates that were obtained for branches in the tree, separated by a comma. ω0 represents the null model or 'background' estimate in the models with more than one estimate. ωv is used when an estimate of the viral internal branch was separately obtained, while ω(virus abbreviation) is used to denote an estimate for a specific tip branch. An equals symbol '=' is used to indicate where multiple branches are assigned the same dN/dS estimate.

Branch test with positives removed										
Gene	**Model**	**Free parameters**	**lnL**	**2ΔlnL**	**p-value (better than 1 ratio)**	**omega virus**	**omega host**	**omega virus tip**	**Δω significantly better than**	**fold change**
glucosaminyl (N-acetyl) transferase 3	ω0	1	-12530.23401							
	ω0,ωBoHV4	2	-12530.23368	0.000652	0.979628797					
MHC class II	ω0	1	-8829.756359							
	ω0,ωRHVP (no positives)	2	-8802.133267	55.246184	0	0.0422	0.4022	n/a	-0.36	9.530805687
Interleukin 6	ω0	1	-5546.811709							
	ω0,ωiv	2	-5540.241282	13.140854	0.000288926	0.0101	0.6965	n/a	-0.6864	68.96039604
CD59 glycoprotein	ω0	1	-2894.257134							
	ω0,ωSaHV2 (no positives)	2	-2880.878985	26.756298	0.000254573	0.0749	0.608	n/a	-0.5331	8.117489987
CD200 immunoglobulin	ω0	1	-5176.010121							
	ω0,ωiv	2	-5171.379368	9.261506	0.002340205	0.0035	0.317	n/a	-0.3135	90.57142857
	ω0,ωiv=ωHHV8=ωMaHV5	2	-5172.8689	6.282442	0.012193996	0.0241	0.3164	n/a	-0.2923	13.12863071
	ω0,ωiv,ωHHV8=ωMaHV5	3	-5171.340174	9.339894	0.009372766	0.0034	0.3176	0.0376	-0.3142	93.41176471
CCL3 chemokine	ω0	1	-1911.393951							
	ω0,ωiv	2	-1909.447286	3.893330	0.048478214	0.0025	0.3523	n/a	-0.3498	140.92
	ω0,ωiv=ωHHV8=ωMaHV5 (no positive)	2	-1892.910528	36.966846	0.000000001	0.0078	0.4535	n/a	-0.4457	58.14102564
	ω0,ωiv,ωHHV8=ωMaHV5	3	-1892.714923	37.358056	0.000000008	0.0028	0.4539	0.0078	-0.4511	162.1071429
Branch test of positives only										
Gene	**Model**	**Free parameters**	**lnL**	**2ΔlnL**	**p-value**	**omega virus**	**omega host**	**omega virus**	**Δω**	**fold change**
MHC class II positives only	ω0	1	-2315.339982							
	ω0,ωRHVP (positives only)	2	-2311.963402	6.753160	0.00935818	0.4363	1.8886		-1.4523	4.328672931
CD59 glycoprotein	ω0	1	-521.232785							
	ω0,ωSaHV2 (positives only)	2	-516.396918	9.671734	0.001871246	0.0001	3.7312		-3.7311	37312

Table 3 A collection of the results of the modified selection analyses conducted

These results are for the six genes that exhibited evidence for positive selection at specific sites according to the M7 M8 comparison. The same branch models were run as in the main selection analysis, but in this instance either the positively selected sites were removed or the test was run on an alignment of only these sites. The details and results for the branch-model, including each of the models tested for each gene are also shown. Models are represented as the combination of different dN/dS estimates that were obtained for branches in the tree, separated by a comma. ω0 represents the null model or 'background' estimate in the models with more than one estimate. ωv is used when an estimate of the viral internal branch was separately obtained, while ω(virus abbreviation) is used to denote an estimate for a specific tip branch. An equals symbol '=' is used to indicate where multiple branches are assigned the same dN/dS estimate.

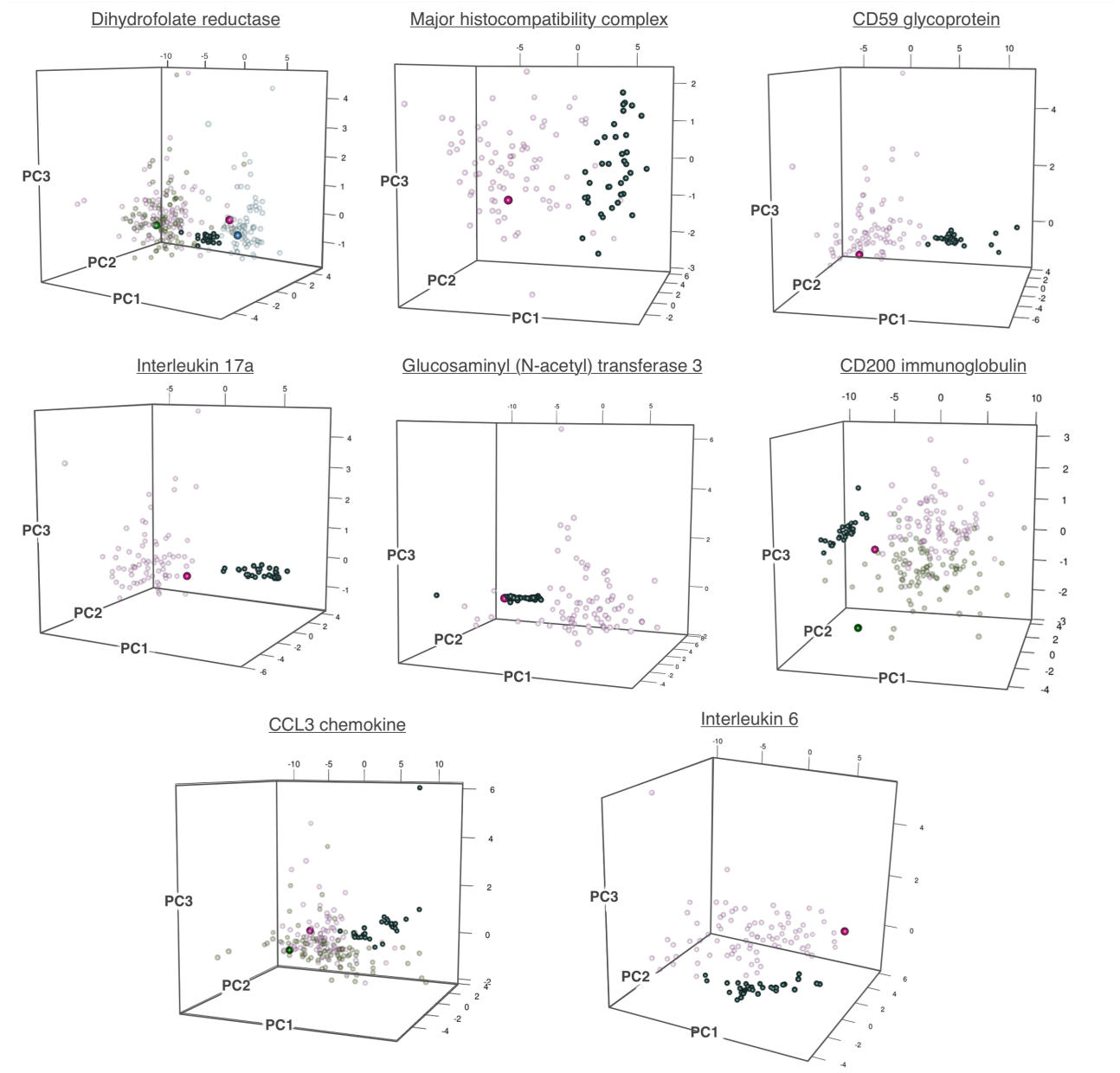


Figure 7 Principal components analysis of sequence composition, length and CpG bias
Principal components analysis of the 8 main genes investigated in this study. The variables used were all four mononucleotide and sixteen dinucleotide frequencies, in addition to CpG bias and gene length. For each gene, the scores for the first 3 principal components (accounting for most of the variance) are plotted for the host genes, the viral homolog and the rest of the viral genes in the relevant genome. In all of the genes except for C2GnT-M, there is a clear distinction in variable space between the viral homologs and the host genes from which they originate. See table S2 for a detailed breakdown of the variable loadings each principal component.

Sequence composition & Structural modeling

We further investigated the differences between viral genes and their host homologs by comparing nucleotide frequencies (mono- & dinucleotide), CpG bias and gene length in a principal components analysis. As was the case for the branch-site analysis, the results showed that there was a substantial difference for all the genes except for the C2GnT-M

homologue (figure 7). We also included all the genes in each viral genome in the analysis. This revealed that the sequence composition of the captured viral genes was moving away from that of the host from which they were derived towards the characteristics of the virus genome (figure 7). In contrast to this, structural modeling of the viral homologs shows that the sequences are compatible with the structure of their host homolog, suggesting that they maintain a similar function. However, structural modeling revealed some distinct deviations from the host gene that are suggestive of functional divergence. Namely, the MHC transmembrane helix is absent in the viral copy (figure 2), and there are extended N- and C-terminal regions in the viral version of CD200 (figure 5).

Discussion

We set out to investigate the history and evolutionary dynamics of host-derived genes in rhadinoviruses, and our results are consistent with the observation that large DNA viruses frequently acquire new genes through HGT from their host. For the first time, our analyses reveal the host origin of rhadinoviral genes through statistically robust phylogenetic reconstruction. By undertaking a systematic approach, we identify a pattern of evolutionary changes that occur in captured genes. We provide evidence for a shift in nucleotide composition away from the host of origin for these genes, despite phylogenetic and structural similarity. Through maximum likelihood dN/dS analysis, we reveal the remarkable finding that hosts exert purifying selection on host-derived genes in viruses that are acquired as part of the evolutionary arms race between them.

The arms-race between hosts and viruses describes the general dynamic of rapid defensive and evasive adaptation. However, the high evolutionary rate and large population size of viruses puts them at an advantage over hosts if point mutation was the only adaptive mechanism [31,32]. Although they may have a small effect individually, a combination of adaptations can act together at different evolutionary timescales to mitigate the slow rate of host evolution [33]. At the ecological scale, multi-layered and adaptive immune systems can partially compensate for the rate disparity between viruses and hosts - for example, the evolution of rearrangeable immunoglobulin segments allows for an effectively infinite number of receptors for antigen recognition [34]. At the evolutionary level, it has been proposed that hosts can accelerate the adaptation to new viral threats by switching between a small number of genotypes that only require one or very few amino acid changes [35].

Sequence similarity to cellular counterparts has already been established for most of the genes in our study [36,37], but the evolutionary relationship had yet to be examined in great detail. In this study, we have conclusively demonstrated their evolutionary origins as host genes through systematic phylogenetic reconstruction (figure 2-5, S1). The viral genes in our trees exhibit long branches, most likely due to the high evolutionary rate of viruses, but nonetheless maintain enough phylogenetic information to group more closely to some mammalian genes than others. Moreover, our results show that the nucleotide composition of these genes has shifted away from their original hosts, bearing more resemblance to other viral genes (figure 7). Despite this compositional shift and high evolutionary rate, we have identified the host group of origin for 8 host-derived genes, For example, the trees in figure 5 both indicate that the cellular homologs originate from primates. We have also been able to demonstrate more specifically that the CD59 homolog in SaHV2 was likely derived from *Saimiri* sp., since the virus and host genes group together with 97% probability (figure 4A).

The capture of cellular genes is among the many adaptations employed by viruses as part of the arms-race, particularly when they are directly involved in immunomodulation. Conversely, hosts are also known to capture viral genes as part of the arms race when EVEs are repurposed to function as anti-viral immune genes [13,38]. Both host-derived viral genes and virally-derived host genes are an 'evolutionary short-cut', circumventing the need for incremental adaptation by capturing a gene already adapted for a particular purpose [13]. For viruses, this is consistent with the fact that captured genes are primarily found in dsDNA viruses that have the lowest mutation rates among viruses [39] and therefore the smallest rate advantage. Moreover, whereas this evolutionary strategy appears to be rare in hosts, the evidence suggests that this is a highly recurring event for viruses, as evidenced by the high number of lineage-specific cellular genes [3]. However, unlike viral genes that are re-purposed to fulfill an anti-viral role in hosts, host-derived immunomodulatory genes are often performing the function that they evolved for. For instance, the retroviral *gag* gene encodes the viral capsid protein, but in mice, an endogenous retroviral *gag* has evolved to restrict infection murine leukemia virus [13,40]. However in viruses, many captured immunomodulatory genes have been shown to maintain at least parts of their original function [4]. This means that although both strategies represent 'evolutionary shortcuts', the host is taking advantage of a novel function that has not evolved within the evolutionary arms-race and is therefore entirely unanticipated by the virus [13]. The viral 'evolutionary shortcut' succeeds for precisely the

opposite reason, by exploiting the fact that host immunomodulatory genes are optimized for many viruses through a range of different arms-races.

This functional conservation is further supported by the fact that the structural modeling revealed a general compatibility with host homologs for all of the genes that we analyzed. The simplification of host-derived genes has previously been observed, particularly in multi domain proteins [41]. Our findings are consistent with this, and despite a number of such modifications (e.g. the missing transmembrane domain in the RHVP MHC homolog, figure 2), the conserved domains are nonetheless maintained, which indicates that the structural similarity is not simply due to recent acquisition. This is especially remarkable given that the genes are evolving between 1-2 orders of magnitude faster in viruses compared to host (the mammalian average rate is ~2.2×10^{-9} [42] compared to dsDNA viruses rates that are approximately 1×10^{-7} [7,43]). Furthermore, the results of our selection analysis reveal a pattern of stronger purifying selection in the viral genes compared to the host counterpart, (figures 2, 3, 5) despite the large rate disparity.

The selection tests we conducted and developed demonstrate that the viral strategy simultaneously involves the preservation of function and change in selection pressure. This result is seemingly counterintuitive given that host immune genes are relatively fast evolving (due to the pressure of the arms-race), and evolve at an uncharacteristically high evolutionary rate compared to other host genes [44,45]. This 'evolutionary tempo' is an adaptive mechanism that evolved in an environment with multiple threats from a variety of viruses. Apart from the relatively higher average dN/dS compared to the viral homologs, we also identify this aspect of the 'host-mode' of adaptive evolution. Specifically, we observed between 1 and 22 positively evolving sites in all of the immunomodulatory genes tested (table 1). Conversely, each of these molecular adversaries on the viral side has a much narrower target that they need to evade in that viruses usually only infect a single species (or very closely related species).

Through the development of our sliding-window approach to the codeml branch model, we are able to dissect the dynamics of the dramatic shift in selection pressure of captured genes. The adaptive nature of this slow-down is most clear in the Il-17 and MHC homologs (figure 1 & 3, respectively). In these cases, we are able to overlay basic information pertaining to the functional structure of the gene products in order to contrast this against the results of the branch test of selection. In both of these genes, the statistically significant drop of dN/dS in the virus' gene compared to host counterparts is

specific to the functional domains. For the homolog of MHC, these domains are also where the vast majority of the 22 positively selected sites were detected, which represent the positive selection undergone by the host genes (figure 1, table 1). This is because the codeml site test of positive selection considers the whole alignment that is predominantly composed of host data.

The shift in selection dynamics of captured genes is especially interesting in the case of the MHC homolog, which had the highest number of adaptive sites in our study that were specifically targeted in the viral 'evolutionary slow-down'. Indeed vertebrate MHCs are the most rapidly evolving vertebrate genes, and there are multiple hypotheses that together explain their paradoxical levels of genetic diversity [46,47]. The MHC molecule functions by binding to antigens for presentation to T-cells that will result in an immune response. The high diversity of MHCs is needed to recognize a wide range of antigens, driven primarily by a selection pressure from pathogens, since hosts with the ability to recognize a diverse set of foreign antigens will have an evolutionary advantage [47]. Among the many evasive strategies employed by viruses is the down regulation of host MHC molecules to avoid detection by T-cells, but this leaves the infected cell vulnerable to lysis by natural killer (NK) cells that respond to changes in MHC presentation. It is therefore plausible that RHVP captured an MHC homologue in order to disguise infected cells as healthy using a 'decoy' MHC, which is a known strategy that has been demonstrated in other viruses [48,49]. Our sliding-window technique is also useful in revealing the hidden structure of the shift in selection pressure where the gene is not as structured like in interleukin 6 (figure 3), where the significant difference in dN/dS is localized to the first half of the gene. This can be used to inform future functional characterization to pinpoint the specific functional effects that primarily drove the acquisition of the gene, since IL6 has been implicated in a wide range of different processes [21].

By analyzing the evolutionary history of host-derived viral genes systematically, the shift in selection we detected reveals an underlying adaptive mechanism. The stronger purifying selection in viruses explains why there is a high turnover of such genes that are repeatedly captured by the same virus, as well as independently acquired in different lineages. For example in our dataset, all of the immunomodulatory cellular genes that we analyzed are limited to one or two viruses, which is an indication of specialized capture in response to a specific host [3,36]. This is consistent with our findings that viruses exert strong purifying selection on such genes in order to maintain their function, which also explains their frequent loss. This is because the strategy of exploiting a specific 'version'

of the captured homolog will loose efficacy once the fitness landscape shifts away, or if the virus switches to a new host. In these events, it is simple for the virus to re-capture an 'updated' host homolog, to which the host target has not been actively evolving to evade. A good example of this phenomenon is Interleukin-10, which has been repeatedly captured by a variety of different DNA viruses, independently from a range of donor hosts [6].

In contrast to such short-lived gene captures, we also analysed homologs of genes for FGAM-synthase that are present in all of the rhadinoviruses, indicating that they are anciently captured and retained through speciation of the virus rather that lost and replaced. FGAM is an enzyme involved in purine synthesis and are therefore not captured for a transiently effective purpose such as immune evasion of a particular host with a specific immune strategy. Moreover, the genes are found in orthologous positions, and our analysis reveals that they have undergone a number of duplications, rearrangements and losses (figure 6). Similarly, we investigated the evolution of host-derived genes that encode homologs of complement control protein (CCP) that are also in all of the viruses except BoHV4 and are capable of evading the complement system [50]. These CCP homologs are therefore an example of an immunomodulatory gene that is not evolving according to the capture-and-replace mechanism we have described. This could be because it is a regulator of innate immunity that does not evolve in an arms-race framework, therefore remaining functionally useful for a longer time. In HHV8, a host-derived Interleukin 8 gene appears to have duplicated twice since capture, indicating that like those ancient genes such as FGAM-synthase, this can also occur at shorter evolutionary timescales for lineage specific genes (figure 4). Moreover, these homologues appear to have gained some distinct functions, but nonetheless functionally overlap in their ability to target the same receptors on Th2 cells, as well as all sharing the capacity to recruit $CD4^+$ and $CD25^+$ T-cells to down-modulate the immune response [37].

Although occurring at short evolutionary timescales, our data suggest that the evolutionary patterns that we have detected do not appear immediately after capture. For instance, if we consider the homologous genes for DHFR in SaHV2, HHV8 and MaHV5 as well as the CD200 homolog in HHV8 and MaHV5 (figure 5), our sliding window branch test analysis revealed that the significant drop in dN/dS was mainly observed for the model that distinguished the viral branches from the rest of the tree. The models where the internal viral branch was designated a separate dN/dS value only shows a significantly lower estimate across very short regions, suggesting that it takes time for the genes to exhibit

the effects of the shifting selection pressure. Compared to the results of the selection analysis of other genes, the pattern of high purifying selection is not as clear in CD200 (figure 5). Moreover, the sequence composition analysis shows that the viral homolog is somewhere between host and virus composition (figure 7). This suggests that the CD200 is an example of a host-derived gene with intermediate patterns of selection and compositional change. C2GnT-M is an example of an early-stage capture, where the selection shift has not yet left a detectable pattern in the sequence data, and the nucleotide composition is still more host-like than virus-like (figure 7). Nonetheless, experimental evidence has shown that BoHV4 can rescue enzymatic activity in cells missing the C2GnT-M gene [51], which offers us conclusive evidence of functional conservation.

Several of the genes studied in this analysis have been previously recognized as probable captured genes, but this is the first example of a systematic investigation of their evolutionary history that conclusively demonstrates their cellular origin through phylogenetic analysis. Unlike other examples of HGT, this kind of gene capture is a natural experiment that allows us study the dynamics of protein evolution either side of a molecular arms-race. This is because we can control for the identity of the protein, while exploring the effect of being in a different genome. Indeed, by reconstruction the history of these genes for a clade of herpesviruses, we revealed a remarkable pattern in the evolution of selection pressure that could not have otherwise been observed.

Methods

Sequence collation, orthology detection & alignment

Full genome alignments for seven rhadinoviruses were downloaded from NCBI in genbank format and aligned using ProgressiveMauve [16] (figure 1). A table of all coding sequence accession IDs from each virus was constructed whereby each row represents a group of genes that probably represent orthologs. Orthology was determined using a BLASTp search of each protein sequence against all other rhadinovirus proteins, in combination with an assessment of their synteny from the whole genome alignments (table S1). Each row of likely orthologs were then used as a BLASTp and tBLASTn query (or group of queries) against the NCBI sequence databases nr, nt, RefSeq and wgs with viruses excluded (except wgs, which does not contain viruses) in order to identify potential homologs in non-viral species. Any results above a BLAST score of 40 were manually examined for their potential to be aligned with the viral sequences. We then shortlisted

genes that exhibited at least 30% identity and could be confidently aligned for further investigation.

Taxon choice & phylogenetic reconstruction

We performed a phylogenetic analysis for each of the shortlisted viral genes to assess their relationship to host homologs. We constructed alignments consisting of a diverse range of mammalian species (table S2). Our taxon choice included more primate representatives than any other order since 4/7 of the viruses infect primate hosts and we wished to evaluate the genes' relationship in this part of the tree in finer detail. All phylogentic reconstruction was performed in parallelized MrBayes [52,53]. We ran two independent MCMC chains for 10 million generations to ensure convergence. Due to the high divergence of viral homologs, we favoured amino acid alignments to mitigate the effect of highly divergent viral sequences, but nucleotide versions were also used in the instances where poor statistical support was exhibited in the amino acid tree (posterior probability below 85% for the nodes relevant to our investigation). The best evolutionary model for each alignment was determined using JModelTest 2 [54] or ProtTest 3 [55] according to the corrected Akaike information criterion. The tree shown in figure 2 was reconstructed from a 266 amino acid alignment of MHC class 1 genes. The RHVP homolog was derived from the 3 separate predicted ORFs R1, R2 and R3 (RV138-140 in table S1). In figure 3, the panel A tree was reconstructed using a 144 amino acid alignment of 36 mammalian IL17 genes and a homolog in SaHV2 (RV134 in table S1). The tree in figure 5A was reconstructed from a 176 amino acid alignment of DHFR for 35 mammalian species and homologs from SaHV2, MaHV5 and HHV8 (RV10 in table S1). The CD59 tree in figure 4A was reconstructed from a 100 amino acid alignment of 29 mammalian genes and the SaHV homolog (RV136 in table S1). The figure 4B tree was reconstructed from a 431 amino acid alignment of 45 mammalian C2GnT-M genes and the BoHV4 homolog (RV130 in table S1). Additional taxa that are closely related to *Bos Taurus* were added in this case to evaluate the relationship of the viral gene in finer detail. The best model for all five of these amino acid based trees was JTT+G. The tree shown in figure 4C was constructed using a 93 nucleotide alignment of 35 mammalian CCL3 genes, an MaHV5 homolog and 3 homologs identified in HHV8 (RV13, RV14, RV17 in table S1). The CASP8 tree shown in figure S1 was reconstructed from a nucleotide alignment of 36 mammalian genes and their homologs in BoHV4, SaHV2, AtHV3, HHV8, and MaHV5 (RV79 in table S1). GTR+G was the best evolutionary model for both of these trees. For a number of genes, neither amino acid nor conventional nucleotide models succeeded in reconstructing reliable trees.. We therefore implemented the SRD06 model (separate

HKY+G for codon 1+2 and 3) in the case of the Il-6 tree (figure 3E), CD200 tree (figure 5E), R1 and Acyl-CoA trees (figure S1). SRD06 is thought to be more biologically realistic and has been shown to frequently outperform other approaches [56].

Selection Analysis

We assessed the effect of viral capture on selection pressure using a maximum likelihood approach implemented in CODEML. We tested for individual amino acid sites under positive selection in the alignment for eight genes where the placement of the viral branch was statistically robust. At least one positively selected site was identified in all but 2 of the gene tested (table 1).. For each of the 8 genes, we tested whether our results may have been biased by the fact that we chose to include many primate sequences that are very similar, which could have obscured positively evolving sites (table S3). we repeated the analysis with fewer primate sequences and found that this had no influence on the results except for CD59, where fewer sites were detected (table S3). We also implemented the 'branch-site test' to detected positively evolving sites along the viral branch (or internal viral branch in the trees with multiple viral tips). The goal was to identify the sites that underwent rapid evolutionary change immediately after being captured, but the analysis revealed no such sites. It should be noted that a major limitation in this test is the need for a much large number of sequences than we have attempted to analyze, in addition to requiring a data set of sequences that are neither too similar nor too divergent [57]. In our case, the size of the data set is probably the main reason why it has failed to detect any sites.

We used the branch model test implemented in CODEML, where an *a priori* model of selection is compared against a null hypothesis. For each gene, we tested the null hypothesis that dN/dS is the same across the tree (table 1; model=ω0), and an alternative model where the viral branch dN/dS is estimated separately. For three of the genes, there are more than one 'viral branch' for genes that are found in multiple viruses (homologs of CD200 & CCL3 and DHFR). We therefore tested three alternative models, allocating either separate dN/dS for the branch internal to the viral group, estimating a separate dN/dS for all viral branches or designating separate values for the internal branch and viral tips (table 1). In order to examine the selection dynamics along each gene in finer detail, we developed and implemented a sliding-window modification of the branch test. In our modification, rather than comparing the overall dN/dS values for the branches under consideration, we analyse subsections of the alignment, plotting the dN/dS values along with whether or not a significant difference is exhibited. A custom script was written to run

multiple instances of CODEML for a window of 120 base pairs with a step size of 3 both the null and alternative hypotheses were run for each window, and a likelihood ratio test was used to determine significance according to a chi2 test with a p-value threshold of 0.05. Note that these p-values are indicative of the magnitude of the drop in dN/dS, and have not been adjusted for multiple testing or pseudo-replication resulting from the sliding windows. Window sizes of 30, 60, 90 and 150 were also attempted but these resulted in poorer results – shorter windows produced fewer stretches of statistically significant differences in dN/dS and a larger window resulted in lower resolution for shorter genes.

Sequence composition analysis, structural modeling & co-phylogeny analysis

For each alignment in our main analysis, we measured the gene lengths, CpG bias, all four mononucleotide and 16 dinucleotide frequencies using a custom script. These variables were used in a principal components analysis to identify whether a pattern could be detected that distinguished viral genes from their host homologs. We examined the first three principal components in all genes analysed (figure 7), which together account for >90% of the variance in the data in all cases. In addition to the viral gene under investigation and its homologs in host species, we also included the rest of the genes in the viral genome to which the gene belongs to assess their location in the plot relative to the host gene and viral homolog.

We used the Phyer2 web portal to model the tertiary structure of viral homologs to evaluate the potential functional significance of any sequence divergence since HGT [58]. We generated an overlap of the model generated with the database template that is most likely a homolog of the query in the UCSF Chimera package [59] (figure 2-5). In addition, we visually mapped the results of the selection analysis on each residue as a heat map representing dN/dS estimates. The dN/dS values are represented along a colour spectrum of blue to red in equallysized incriments of 0.25, where the darkest blue is equivalent to 0 and the darkest red is the highest estimated value.. It should be noted that such modeling approaches infer structure from sequence homology, and should be interpreted as a rough guide for how compatible the viral homolog is to a reference structure. A crucial limitation to this approach is interpreting whether differences between similar sequences would influence structure, and if so, in what way this would happen.

For the FGAM-synthase and CCP homologs, we conducted a co-phylogeny analysis using JANE 4.0, which is an event-based method of reconciling tree topologies of hosts and parasites [60]. The method works by considering five possible evolutionary events: co-

speciation, duplication, loss, host switching and failure to diverge, and reconstructs possible evolutionary histories that could explain the data with minimum a minimum cost of events. In our analysis, we used the default cost-scheme of 0,1,2,1,1 for co-speciation, duplication, loss, host switching and failure to diverge, respectively, and ran an additional analysis that disallowed host switching. Note that in the context of our investigation, the host switching parameter is the equivalent of horizontal gene transfer, since we are comparing the congruence of tree topology between the viral phylogeny and a gene tree rather than host-parasite. We calculated the distribution of costs for random sample solutions for 1000 samples hat included the original problem instance, using tip mapping as the randomization method.

References

1. McGeoch DJ, Davison AJ, Dolan A, Gatherer D, Sevilla-Reyes EE (2008) Molecular Evolution of the Herpesvirales. In: Domingo E, Holland JJ, editors. Origin and Evolution of Viruses. Academic Press. pp. 447–475. doi:10.1016/B978-0-12-374153-0.00020-5.
2. Fields BN, Knipe DM, Howley PM (2007) Fields Virology. 5th ed. Philadelphia: Lippincott Williams and Wilkins.
3. Shackelton LA, Holmes EC (2004) The evolution of large DNA viruses: combining genomic information of viruses and their hosts. Trends Microbiol 12: 458–465.
4. Alcami A (2003) Viral mimicry of cytokines, chemokines and their receptors. Nat Rev Immunol 3: 36–50. doi:10.1038/nri980.
5. Holzerlandt R, Orengo C, Kellam P, Albà MM (2002) Identification of new herpesvirus gene homologs in the human genome. Genome Res 12: 1739–1748. doi:10.1101/gr.334302.
6. Ouyang P, Rakus K, van Beurden SJ, Westphal AH, Davison AJ, et al. (2014) IL-10 encoded by viruses: a remarkable example of independent acquisition of a cellular gene by viruses and its subsequent evolution in the viral genome. J Gen Virol 95: 245–262. doi:10.1099/vir.0.058966-0.
7. Duffy S, Shackelton LA, Holmes EC (2008) Rates of evolutionary change in viruses: patterns and determinants. Nat Rev Genet 9: 267–276. doi:10.1038/nrg2323.
8. Feschotte C, Gilbert C (2012) Endogenous viruses: insights into viral evolution and impact on host biology. Nat Rev Genet 13: 283–296. doi:10.1038/nrg3199.
9. Holmes EC (2011) The Evolution of Endogenous Viral Elements. Cell Host Microbe 10: 368–377. doi:10.1016/j.chom.2011.09.002.
10. Katzourakis A, Gifford RJ (2010) Endogenous viral elements in animal genomes. PLoS Genet 6: e1001191.
11. Katzourakis A (2013) Paleovirology: inferring viral evolution from host genome sequence data. Phil Trans R Soc B 368: 20120493 – . doi:10.1098/rstb.2012.0493.
12. Patel MR, Emerman M, Malik HS (2011) Paleovirology—ghosts and gifts of viruses past. Curr Opin Virol 1: 304–309. doi:10.1016/j.coviro.2011.06.007.
13. Aswad A, Katzourakis A (2012) Paleovirology and virally derived immunity. Trends Ecol Evol 27: 627–636. doi:10.1016/j.tree.2012.07.007.
14. Katzourakis A, Gifford RJ (2010) Endogenous viral elements in animal genomes. PLoS Genet 6. doi:10.1371/journal.pgen.1001191.
15. 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. doi:10.1098/rstb.2012.0499.
16. Darling ACE, Mau B, Blattner FR, Perna NT (2004) Mauve: multiple alignment of conserved genomic sequence with rearrangements. Genome Res 14: 1394–1403. doi:10.1101/gr.2289704.
17. Loh J, Zhao G, Nelson CA, Coder P, Droit L, et al. (2011) Identification and sequencing of a novel rodent gammaherpesvirus that establishes acute and latent infection in laboratory mice. J Virol 85: 2642–2656. doi:10.1128/JVI.01661-10.
18. Gu C, Wu L, Li X (2013) IL-17 family: cytokines, receptors and signaling. Cytokine 64: 477–485. doi:10.1016/j.cyto.2013.07.022.
19. Perelman P, Johnson WE, Roos C, Seuánez HN, Horvath JE, et al. (2011) A molecular phylogeny of living primates. PLoS Genet 7: e1001342. doi:10.1371/journal.pgen.1001342.
20. 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. doi:10.1016/j.ympev.2005.11.015.
21. Kishimoto T (2010) IL-6: from its discovery to clinical applications. Int Immunol 22: 347–352. doi:10.1093/intimm/dxq030.
22. Chung Y-H, Means RE, Choi J-K, Lee B-S, Jung JU (2002) Kaposi's Sarcoma-Associated Herpesvirus OX2 Glycoprotein Activates Myeloid-Lineage Cells To Induce Inflammatory Cytokine Production. J Virol 76: 4688–4698. doi:10.1128/JVI.76.10.4688-4698.2002.
23. Berg J, Tymoczko J, Stryer L (2002) Biochemistry. 5th ed. New York: W.H. Freeman & Co Ltd. 1050 p.
24. Cinquina CC, Grogan E, Sun R, Lin SF, Beardsley GP, et al. (2000) Dihydrofolate reductase from Kaposi's sarcoma-associated herpesvirus. Virology 268: 201–217. doi:10.1006/viro.1999.0165.
25. Markine-Goriaynoff N, Georgin J-P, Goltz M, Zimmermann W, Broll H, et al. (2003) The core 2 beta-1,6-N-acetylglucosaminyltransferase-mucin encoded by bovine herpesvirus 4 was acquired from an ancestor of the African buffalo. J Virol 77: 1784–1792.
26. Dewals B, Thirion M, Markine-Goriaynoff N, Gillet L, de Fays K, et al. (2006) Evolution of Bovine herpesvirus 4: recombination and transmission between African buffalo and cattle. J Gen Virol 87: 1509–1519. doi:10.1099/vir.0.81757-0.
27. Kirkitadze MD, Barlow PN (2001) Structure and flexibility of the multiple domain proteins that regulate complement activation. Immunol Rev 180: 146–161. doi:10.1034/j.1600-065X.2001.1800113.x.
28. Rezaee SAR, Cunningham C, Davison AJ, Blackbourn DJ (2006) Kaposi's sarcoma-associated herpesvirus immune modulation: an overview. J Gen Virol 87: 1781–1804. doi:10.1099/vir.0.81919-0.
29. Lee B-S (2000) Inhibition of Intracellular Transport of B Cell Antigen Receptor Complexes by Kaposi's Sarcoma-associated Herpesvirus K1. J Exp Med 192: 11–22. doi:10.1084/jem.192.1.11.
30. Zong JC, Ciufo DM, Alcendor DJ, Wan X, Nicholas J, et al. (1999) High-level variability in the ORF-K1 membrane protein gene at the left end of the Kaposi's sarcoma-associated herpesvirus genome defines four major virus subtypes and multiple variants or clades in different human populations. J Virol 73: 4156–4170.
31. Daugherty MD, Malik HS (2012) Rules of engagement: molecular insights from host-virus arms races. Annu Rev Genet 46: 677–700. doi:10.1146/annurev-genet-110711-155522.
32. Duggal NK, Emerman M (2012) Evolutionary conflicts between viruses and restriction factors shape immunity. Nat Rev Immunol 12: 687–695. doi:10.1038/nri3295.
33. Trowsdale J, Parham P (2004) Mini-review: defense strategies and immunity-related genes. Eur J Immunol 34: 7–17. doi:10.1002/eji.200324693.
34. Pancer Z, Cooper MD (2006) The evolution of adaptive immunity. Annu Rev Immunol 24: 497–518. doi:10.1146/annurev.immunol.24.021605.090542.
35. Meyerson NRNR, Sawyer SLSL (2011) Two-stepping through time: mammals and viruses. Trends Microbiol 19: 286–294. doi:10.1016/j.tim.2011.03.006.
36. Holzerlandt R, Orengo C, Kellam P, Albà MM (2002) Identification of new herpesvirus gene homologs in the human genome. Genome Res 12: 1739–1748. doi:10.1101/gr.334302.
37. Nicholas J (2005) Review: Human Gammaherpesvirus Cytokines and Chemokine Receptors. J Interf Cytokine Res 25: 373–383. doi:10.1089/jir.2005.25.373.
38. Feschotte C, Gilbert C (2012) Endogenous viruses: insights into viral evolution and impact on host biology. Nat Rev Genet.
39. Holmes EC (2011) What does virus evolution tell us about virus origins? J Virol 85: 5247–5251. doi:10.1128/JVI.02203-10.
40. Stoye JP (1998) Fv1, the mouse retrovirus resistance gene. Rev Sci Tech 17: 269–277.
41. Rappoport N, Linial M (2012) Viral proteins acquired from a host converge to simplified domain architectures. PLoS Comput Biol 8: e1002364. doi:10.1371/journal.pcbi.1002364.
42. Kumar S, Subramanian S (2002) Mutation rates in mammalian genomes. PNAS 99: 803–808. doi:10.1073/pnas.022629899.
43. Sanjuán R (2012) From molecular genetics to phylodynamics: evolutionary relevance of mutation rates across viruses. PLoS Pathog 8: e1002685. doi:10.1371/journal.ppat.1002685.
44. Nielsen R, Hellmann I, Hubisz M, Bustamante C, Clark AG (2007) Recent and ongoing selection in the human genome. Nat Rev Genet 8: 857–868. doi:10.1038/nrg2187.
45. Sabeti PC, Schaffner SF, Fry B, Lohmueller J, Varilly P, et al. (2006) Positive natural selection in the human lineage. Science 312: 1614–1620. doi:10.1126/science.1124309.
46. Bernatchez L, Landry C (2003) MHC studies in nonmodel vertebrates: what have we learned about natural selection in 15 years? J Evol Biol 16: 363–377. doi:10.1046/j.1420-9101.2003.00531.x.
47. Spurgin LG, Richardson DS (2010) How pathogens drive genetic diversity: MHC, mechanisms and misunderstandings. Proc Biol Sci 277: 979–988. doi:10.1098/rspb.2009.2084.
48. Babić M, Krmpotić A, Jonjić S (2011) All is fair in virus-host interactions: NK cells and cytomegalovirus. Trends Mol Med 17: 677–685. doi:10.1016/j.molmed.2011.07.003.
49. Pyzik M, Dumaine A, Dumaine AA, Charbonneau B, Fodil-Cornu N, et al. (2014) Viral MHC class I-like molecule allows evasion of NK cell effector responses in vivo. J Immunol 193: 6061–6069. doi:10.4049/jimmunol.1401386.
50. Lambris JD, Ricklin D, Geisbrecht B V (2008) Complement evasion by human pathogens. Nat Rev

Microbiol 6: 132–142. doi:10.1038/nrmicro1824.
51. Vanderplasschen A, Markine-Goriaynoff N, Lomonte P, Suzuki M, Hiraoka N, et al. (2000) A multipotential beta -1,6-N-acetylglucosaminyl-transferase is encoded by bovine herpesvirus type 4. PNAS 97: 5756–5761. doi:10.1073/pnas.100058897.
52. Ronquist F, Huelsenbeck JP (2003) MrBayes 3: Bayesian phylogenetic inference under mixed models. Bioinformatics 19: 1572–1574. doi:10.1093/bioinformatics/btg180.
53. Altekar G, Dwarkadas S, Huelsenbeck JP, Ronquist F (2004) Parallel Metropolis coupled Markov chain Monte Carlo for Bayesian phylogenetic inference. Bioinformatics 20: 407–415. doi:10.1093/bioinformatics/btg427.
54. Darriba D, Taboada GL, Doallo R, Posada D (2012) jModelTest 2: more models, new heuristics and parallel computing. Nat Methods 9: 772. doi:10.1038/nmeth.2109.
55. Darriba D, Taboada GL, Doallo R, Posada D (2011) ProtTest 3: fast selection of best-fit models of protein evolution. Bioinformatics 27: 1164–1165. doi:10.1093/bioinformatics/btr088.
56. Shapiro B, Rambaut A, Drummond AJ (2006) Choosing appropriate substitution models for the phylogenetic analysis of protein-coding sequences. Mol Biol Evol 23: 7–9. doi:10.1093/molbev/msj021.
57. Yang Z, dos Reis M (2011) Statistical properties of the branch-site test of positive selection. Mol Biol Evol 28: 1217–1228. doi:10.1093/molbev/msq303.
58. Kelley LA, Mezulis S, Yates CM, Wass MN, Sternberg MJE (2015) The Phyre2 web portal for protein modeling, prediction and analysis. Nat Protoc 10: 845–858. doi:10.1038/nprot.2015.053.
59. Pettersen EF, Goddard TD, Huang CC, Couch GS, Greenblatt DM, et al. (2004) UCSF Chimera--a visualization system for exploratory research and analysis. J Comput Chem 25: 1605–1612. doi:10.1002/jcc.20084.
60. Conow C, Fielder D, Ovadia Y, Libeskind-Hadas R (2010) Jane: a new tool for the cophylogeny reconstruction problem. Algorithms Mol Biol 5: 16. doi:10.1186/1748-7188-5-16.

Chapter Five: Convergent Capture of Retroviral Superantigens by Mammalian Herpesviruses

Abstract

Horizontal gene transfer from retroviruses to mammals is well documented and extensive, but is rare between unrelated viruses with distinct genome types. Three herpesviruses encode a gene with similarity to a retroviral superantigen gene (*sag*) of the unrelated mouse mammary tumor virus (MMTV). We uncover ancient retroviral *sags* in over 20 mammals to reconstruct their shared history with herpesviral *sags*, revealing that the acquisition is a convergent evolutionary event. A retrovirus circulating in South American primates over 10 million years ago was the source of *sag* in two monkey herpesviruses, and a different retrovirus was the source of *sag* in a Peruvian rodent herpesvirus. We further show through a time-scaled phylogenetic analysis that a cross-species transmission of a monkey herpesviruses occurred after the acquisition of *sag*. These results reveal that a diverse range of ancient *sag*-containing retroviruses independently donated *sag* twice from two separate lineages that are distinct from MMTV.

Introduction

Superantigens (*sag*) are unique virulence genes capable of eliciting a major immune response by non-specific T-cell stimulation [1,2]. Ordinarily, antigenic processing and presentation by major histocompatibility complexes (MHCs) is required before an immunogenic molecule can provoke T-cell recognition. *Sags* are unusual in their ability to cross-link antigen presenting cells and T-cells by directly binding to MHC class II and T-cell receptors (TCRs) simultaneously [1]. Their affinity to a diverse range of TCRs results in large numbers of activated T-cells, causing a cytokine storm that can lead to potentially fatal toxic shock syndrome. *Sags* have also been implicated in a spectrum of unrelated syndromes, including sepsis, Kawasaki's disease, Scarlet Fever and autoimmune disorders [3]. Unrelated to bacterial *sag*, viral *sags* are also known, most notably in the Mouse Mammary Tumor Virus (MMTV) [4], a retrovirus with an RNA genome. The integrated proviral genome encodes the main retroviral genes *gag*, *pro*, *pol* and *env*, as well as flanking long terminal repeats (LTR) that are created during reverse transcription. LTRs are an integral part of retrovirus replication, but are ordinarily non-coding. In MMTV, the majority of the *sag* gene is found within the U3 region of the LTR [4]. Multiple *sag* mRNA promoters have been described that give rise to different transcripts beginning in the U3/R junction in the 5'LTR [5–7], the *env* gene [8] and an internal 3' LTR U3 in *sag* [9].

The *Herpesvirales* are a large, well-characterized and extremely diverse order known to infect amphibians, reptiles, fish, birds and mammals [10]. Humans are susceptible to eight different herpesviruses that are responsible for a number of diseases. These include, but are not limited to, oral and genital lesions, varicella (chicken pox), Burkitt's lymphoma and Kaposi's sarcoma [11]. Herpesviruses have large (over 100Kb) dsDNA genomes normally flanked by repetitive regions and blocks of conserved coding regions that vary between subfamilies. Gene content among herpesviruses is highly comparable, with the majority of lineage-specific genes existing towards the termini [10]. Members of the genus *Rhadinovirus* have a notably high number of host-derived genes, particularly in the region between the gene for dUTPase and the minor capsid such as the cellular genes CD59 that in hosts is involved in complement regulation, and the proinflamatory cytokine IL-17. Also in this region, *sag*–like genes have been identified in *Saimirine Herpesvirus 2* (SaHV2) and its sister taxon *Ateline Herpesvirus 3* (AtHV3) that infect squirrel and spider monkeys, respectively [12,13]. A similar gene is also found in the genome of *Rodent Herpesvirus peru* (RHVP) [14], a rhadinovirus of the pygmy rice rat (*Oligoryzomys microtis*).

We do not know the source of *sag*-like genes in these herpesviruses, nor do we know their evolutionary relationship to MMTV *sag*. However, because MMTV and herpesvirus *sags* are known to exhibit sequence similarity, we hypothesized that they are homologous as a result of HGT from *sag*-containing retroviruses. Although horizontal gene transfer (HGT) is prevalent among viruses in general, it is predominantly a phenomenon limited to those with comparable genomes. Virus to virus HGT between unrelated groups, such as between RNA and DNA viruses, is rare and not fully understood, but there have been an increasing number of documented cases in recent years [15,16].

Since MMTV is the only extant *sag*-containing retrovirus identified to date, we set out to investigate if ancient retroviruses related to MMTV may have possessed *sag*. Viruses of all genome types and replication strategies are known to occasionally integrate into the germline genome of their host and can thereafter be inherited as part of the host gene repertoire [17–19]. These endogenous retroviruses (ERVs) can eventually reach population fixation, and drift at the host neutral evolutionary rate. ERVs remain recognizable for millions of years after endogenisation, and can be used as a 'fossil record' of past infections to understand viral evolution [17,18]. Our strategy was to use ERVs to determine the number of times *sag* was acquired by herpesviruses, and from how many retroviral donor lineages. We also aimed to characterize the relationship between retroviral *sags*, MMTV *sag* and herpesvirus *sags,* as well as the pattern of selective pressures involved in their respective evolutionary histories. To tackle these goals, we collected ancient retroviral data by mining ERVs from mammalian genomes. We then use these ERVs to reconstruct the history of this unique virulence gene, by synthesizing evidence from phylogenetics and bioinformatic techniques under an evolutionary framework.

To date, *sag*-containing ERVs have only been reported in a small number of genomes (e.g. rat and deer mouse [20]), but our systematic search reveals that there are numerous *sag*-like sequences in at least 20 different mammals. Through a series of phylogenetic analyses, we demonstrate that herpesvirus *sags* originate from a large, previously undescribed clade of betaretroviruses. Through an analysis of selection we demonstrate that these ancient *sag*-like genes were functional. The trees show that at least two HGT events occurred from different retroviral lineages that are both distinct from MMTV, and we provide estimates for the timing of these events. Our results show that although gene transfer between unrelated viruses is rare, it can occur as part of the evolutionary arms-race in a convergent evolutionary manner.

Results

Characterization of the herpesvirus sag loci

In order to determine if herpesvirus and MMTV *sags* share an evolutionary history, we needed to first establish the evolutionary history of the gene among the herpesviruses. We conducted an analysis of the herpesvirus genomes to determine whether the three herpesvirus *sag* loci are syntenic, which would strongly suggest that their similarity is the result of common ancestry. It is straightforward to rule this out for RHVP *sag,* which is situated at a different locus to the other two herpesviruses (Fig. 1). The region between the dUTPase and minor capsid genes in SaHV2 and AtHV3 both share *sag* and bcl-2 in a collinear manner despite the presence of ORF12, Il-17 and CD59 in SaHV2 (Fig. 1). While ORF12 is absent in AtHV3, sequence similarity to SaHV2 at that locus is maintained, indicating that it was a shared gene in their ancestor that later degraded in AtHV3 (Fig. 1). This is not the case for IL-17 and CD59 where no sequence similarity is evident, demonstrating that they were acquired after the divergence of SaHV2 and AtHV3. Indeed, phylogenetic evidence shows the SaHV2 genes are closely related to the squirrel monkey homologs, which is the current host of SaHV2. Thus, these results show that as well as *sag,* the region between dUTPase and minor capsid in the SaHV2-AtHV3 ancestor included ORF12 and bcl-2. Whereas bcl-2 persisted in both viruses after speciation, ORF12 was only maintained in SaHV2, which thereafter also acquired IL-17 and CD59 (Fig. 1).

Survey and alignment construction of sag-containing ERVs

We surveyed mammalian genomes to search for evidence that retroviruses with *sag* have existed in the past. We used BLAST with the herpesviral *sag* open reading frames and a representative MMTV *sag* transcript sequence as queries to search for similarity in NCBI genome databases (Fig. 1, Supplementary Fig. 1). To discard false positives and sequences that are too divergent to be aligned, we applied stringent shortlisting criteria to the results (see materials and methods). The search revealed many ERVs with *sag* similarity, representing sequences from the genomes of a diverse set of 20 species of mammal (Fig. 2). All of these *sags* were betaretroviral in origin and closely related to MMTV. We also searched for betaretrovirus ERVs using MMTV polymerase as the query, in order to identify those closely related to MMTV that did not show *sag* similarity. To investigate the evolution of herpesvirus *sag* acquisition, we used the collected sequences to conduct a series of phylogenetic analyses that included data from both sets of searches.

We employed a phylogenetic strategy that would allow us to both investigate *sag*-less sequences and also mitigate the constraints of low sequence similarity among various *sags* (only a short region (~80aa) could be aligned across all species (Supplementary Note 1)). Conserved regions of all four retroviral genes plus *sag* were concatenated in a 'master' alignment of 138 sequences including the three herpesvirus *sag*s, MMTV, related ERVs with or without *sag*, and a number of other retroviruses such as JSRV, ALV and HERVK that are known outgroups to MMTV (Supplementary Fig. 2). Because some sequences did not include all genes (Supplementary Table 1), the proportion of gaps in the master alignment was prohibitive for meaningful phylogenetic reconstruction. For this reason, we constructed a concatenated alignment of all genes excluding many sequences that were predominantly gapped (Supplementary Fig. 2) except for those of the monkey *sags* that are the most closely related to SaHV2 and AtHV3 *sag* (Fig. 3). The subset of these sequences that also contained an intact polymerase were additionally aligned to the outgroup sequences to identify the correct root of the main concatenated tree (Supplementary Fig. 3). We also reconstructed a tree from the *sag*-only alignment to evaluate the placement of sequences omitted from the main tree, and a *pol*-only tree to assess the relationship of *sag*-containing ERVs to closely related non-*sag* ERV.

Reconstruction of the evolutionary history of sag

The concatenated alignment revealed that there are two major lineages of *sag* - one belonging to retroviruses that primarily infect rodents (designated 'clade A'), and a second mainly primate-infecting 'clade B' (Fig. 3). Importantly, the *sags* of primate-infecting herpesviruses (SaHV2 and AtHV3) are most closely related to primate retroviral *sags*. We identified three independent lineages in the genome of *Callithrix jacchus* and *Aotus nancymaae,* which are not orthologous despite the hosts having only recently speciated (~20 mya). Together, they form a robust monophyletic group that is consistent with a single capture event of *sag* from a new world monkey retrovirus to the ancestor of the two new world monkey herpesviruses. The RHVP *sag* does not cluster with the monkey herpesvirus *sags,* indicating a separate gene-capture event. Instead, RHVP *sag* is more closely related to rodent retroviruses, resolving as a sister taxon to *J. jaculus* and *P. maniculatus*. Both herpesvirus *sag* lineages resolve within a clade of retroviruses to the exclusion of other retroviral sequences with high posterior probability, indicating that the direction of transfer was from retrovirus to herpesvirus. Furthermore, the tree shows with 100% probability that both of the herpesvirus *sag* lineages are independent from MMTV *sag*, which is the only other extant *sag*-containing virus.

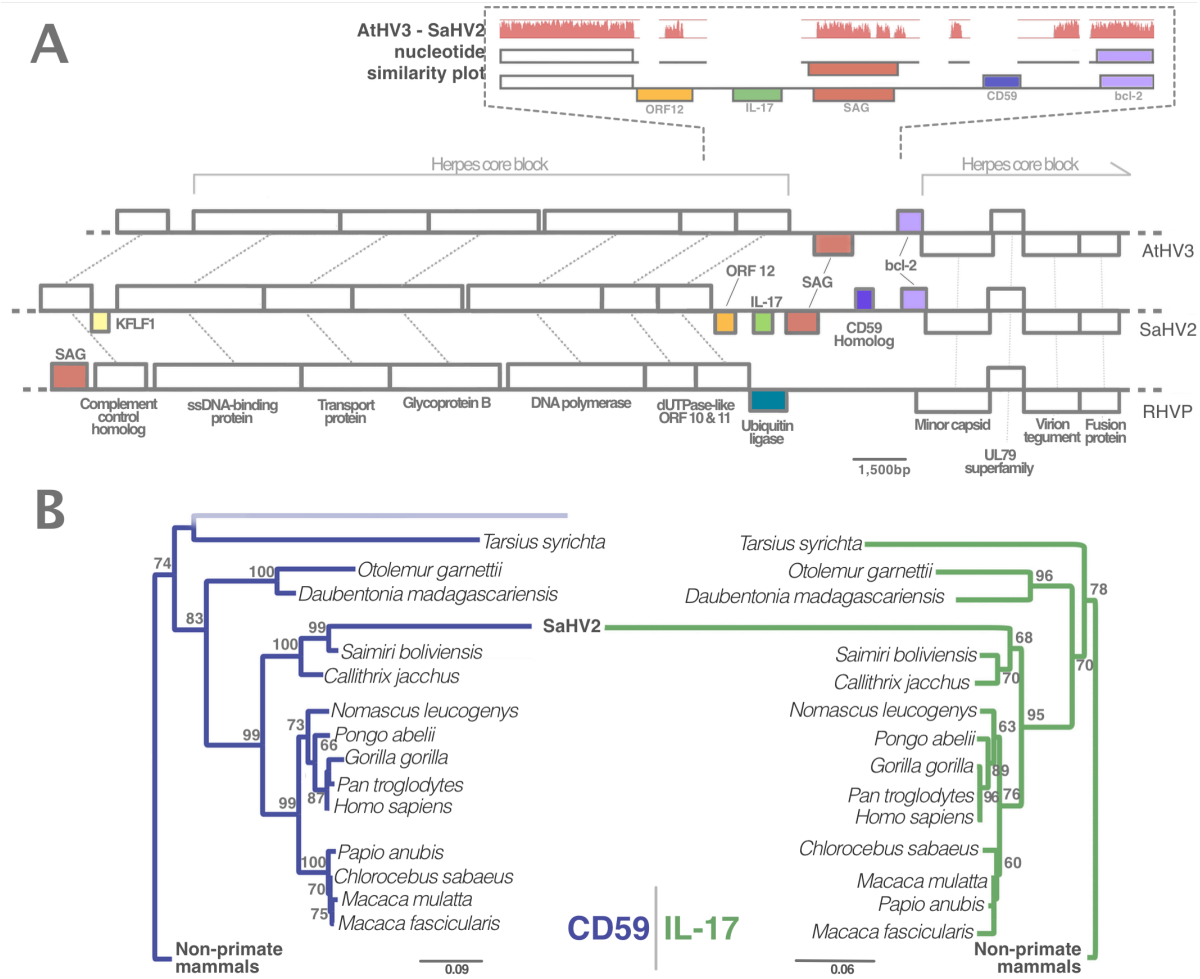


Figure 1 The herpesviral *sag* locus

A: Cartoon of a partial genome alignment of Saimrine Herpesvirus 2 (SaHV2), Ateline Herpesvirus 3 (AtHV3) and Rodent Herpesvirus Peru (RHVP), showing conserved gene blocks in blank boxes and lineage-specific genes in colour. The figure also shows an alignment for SaHV2 and AtHV3 of the region between ORF11 and the minor capsid gene. The red graph depicts the nucleotide-level similarity shared by the viruses. **B:** Maximum-likelihood phylogenies of host IL-17 and CD59 from primates including a viral homolog in SaHV2. SaHV2 independently acquired these genes after diverging from AtHV3. CD59 in particular, appears to have been captured from the squirrel monkeys, which is the virus's current host. Only the primate tips have been made visible. The scale bar represents the number of nucleotide substitutions per site. The results of 1,000 non-parametric bootstrap replicates are indicated at each node.

Not all of the sag genes identified included full-length ERVs (Supplementary table 1). Most were found as 'solo-LTRs', which are the result of recombination of LTRs that results in the deletion of the rest of the ERV. 490/527 of the shortlisted *sag*-containing contigs did not return BLAST hits to other retroviral genes (using MMTV *gag, pol* and *env* as queries), indicating that they were solo-LTRs. Therefore the separate *sag* alignment allowed the inclusion of more retroviral sequences, some of which were excluded from the concatenated alignment due to the absence of any other genes. Because it is based on a much smaller alignment of a highly divergent gene, the topology in this tree is less reliable and exhibits poor support overall. Nevertheless, the *sag*-only tree does recover the distinction between the predominantly rodent- and primate-infecting viruses, albeit with lower support for deep nodes (clade A and B, Fig. 4). The *sag*-only tree shows evidence

for multiple invasions by MMTV-like retroviruses of the same host species; there are three independent lineages of ERVs in *Rattus norvegicus*, four lineages in *Mus musculus*, and two separate *Dipodomys ordii* clades (Fig. 4). The *sag*-only tree also shows a larger subset of the ~300 *Pteropus sp. sag*-like BLAST hits (Fig. 2) that indicate a history of intra-genomic proliferation. As in the case of the multi-lineage *sag*s of mouse, rat and kangaroo rat, there are two distinct lineages in *C. jacchus* as well as a lineage in *A nancymaae* (Fig. 4).

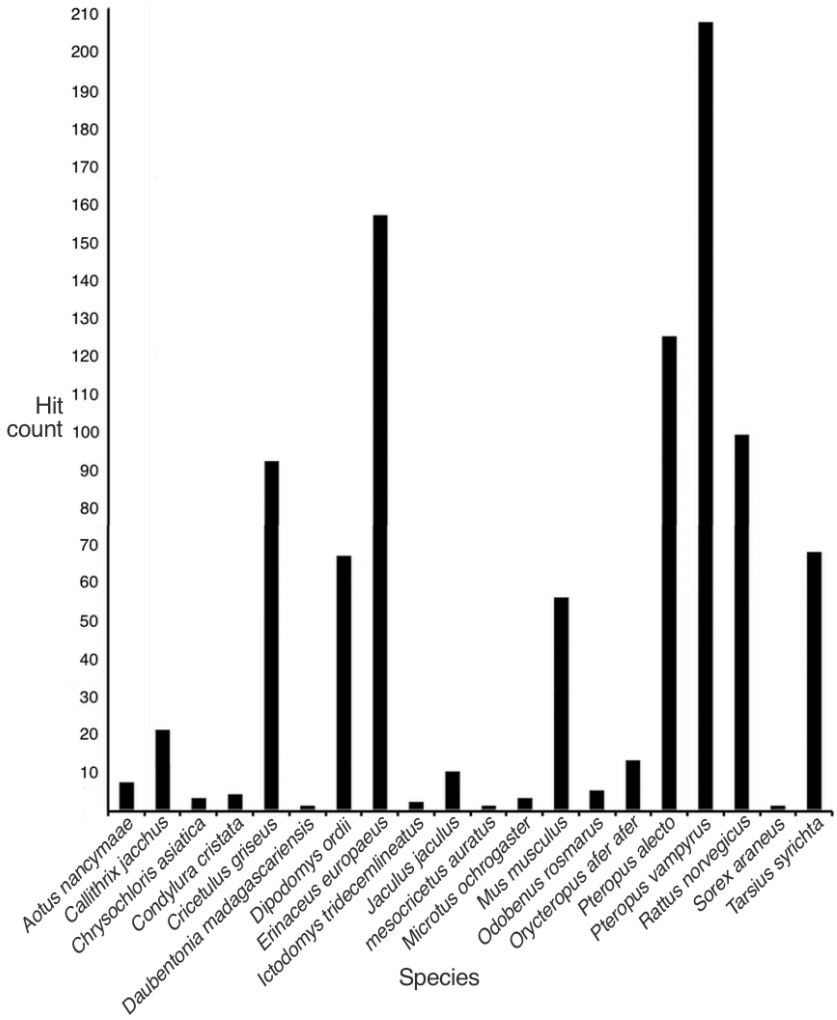


Figure 2 Summary of BLAST *sag* hits in mammalian genomes
The histogram shows the number of tBLASTn hits per mammalian genome to any of the four viral *sag*s. A conserved 100 amino acid region of the *sag* sequences of Saimrine Herpesvirus 2 (SaHV2), Ateline Herpesvirus 3 (AtHV3), Rodent Herpesvirus Peru (RHVP) and Mouse mammary tumor virus (MMTV) was used as a query. The mouse, rat, Chinese hamster and hedgehog data represent multiple individual genomes. The data was obtained from the NCBI WGS database.

The polymerase (*pol*) tree (Fig. 4) allows us to reliably identify the phylogenetic placement of sag-containing ERVs identified in this study with respect to other retroviruses. It was possible to include ERVs that were identified as being closely related to MMTV but did not contain recognizable *sag* according to our similarity cut-off. This tree also shows that none

of the polymerase sequences from *sag*-containing ERVs are more closely related to a non-betaretrovirus (e.g. in the case that a recombination event occurred). The inclusion of known outgroup sequences in this tree confirms the distinction between clades A and B from which the separate herpesvirus *sag* acquisitions originate. Interestingly, the *pol* tree reveals that there is a diverse group of mammalian ERVs closely related to the *sag*-containing viruses. Most of these viruses did not contain detectable *sag* sequences according to our similarity threshold for consideration (Fig. 4), except for a lineage of *Tarsius syrichta* ERVs, which resolve at the root of the main phylogeny.

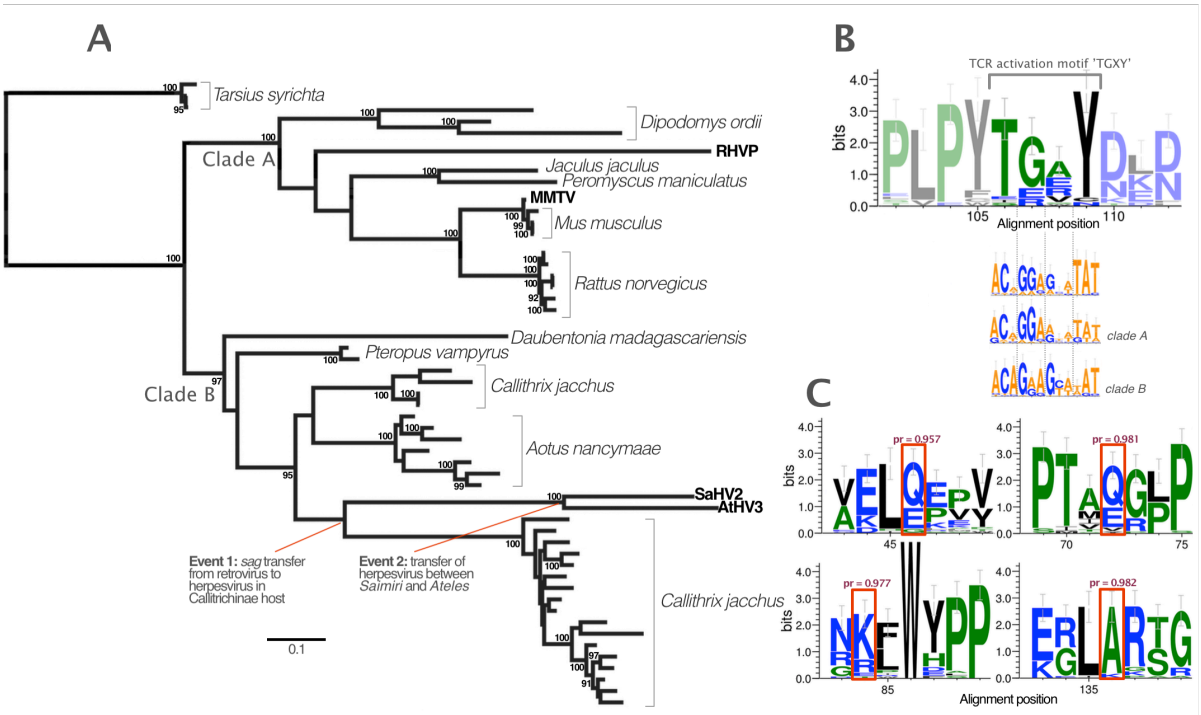


Figure 3 Evolutionary history and selection analysis of *sags*
A: A phylogeny of *gag, pro, pol, env* and *sag* genes. Clade-A shows the rodent ERVs, and Rodent Herpesvirus Peru (RHVP), while clade-B is the predominantly primate clade that also contains Ateline Herpesvirus 3 (AtHV3) and Saimirine Herpesvirus 2 (AtHV2). The rooting was determined by reconstructing a tree from an alignment *pol* from sequences included in this tree as well as 11 outgroup sequences. The scale bar shown represents the number of amino acid substitutions per site and posterior probability values above 90% are indicated at each node. **B:** WebLogo of a region of the alignment used to reconstruct the tree in panel-A, including the underlying codons on the first line. The region shown is the crucial TCR activation motif. The subsequent lines show the codons for the same region using only clade-A or clade-B sequences, including a larger taxon set than the representative subset used for the tree. **C:** Results of the Branch-Site test of positive selection, assessing the likelihood of positive selection at individual sites of the branch leading to SaHV2/AtHV3. The WebLogo was generated using the expanded clade-A alignment. The highlighted sites have a dN/dS>1 with a significant posterior probability, denoted as pr.

Reconstructing the sag evolutionary timeline

Since retroviral LTRs are identical at the time of integration, their divergence from one another is used as a measure of their age when combined with estimates of the host neutral rate. We calculated the age of the ERVs based on the average mammalian rate [21] of 2.2×10^{-9} per site per year, revealing that sequences in clade B are relatively young.

The rat, mouse and deer mouse ERVs are between 1.7 and 6.2mya, whereas the marmoset and tarsier ERVs are estimated to be between 14 and 19mya. Using the mouse evolutionary rate ($4.5x10^{-9}$ per site per year) [22], age estimates for clade B are calculated to be between 0.8 and 3mya. However, both sets of estimates should be interpreted as conservative minimum bounds for the integration date of the exogenous virus (Fig. 5), because they are averages over time and across the genome. Moreover, apart from mouse, a species-specific evolutionary rate has not been obtained for the other hosts.

We wanted to test whether the divergence of the herpesvirus is similar to their respective host divergence dates (i.e. the age of the common ancestors of SaHV2-AtHV3 and RHVP-MuHV4 should be consistent with the age of the *Saimiri-Ateles* and *Oligoryzomys-Mus* divergences, respectively). We therefore reconstructed a time-calibrated phylogeny of rhadinoviruses using a relaxed log-normal molecular clock implemented in BEAST. We calibrated at previously identified co-diverging nodes (using host divergence dates [23]) that are outside the clades of interest, therefore enabling us to date the sag-containing herpesvirus splits without assuming co-divergence *a priori.* This revealed that the SaHV2-AtHV3 split occurred ~10.6mya (95% HPD: 8.4-13), while that of RHVP and its closest relative MuHV4 diverged 37.5mya (95% HPD: 32.7-42.6) (Fig. 5).

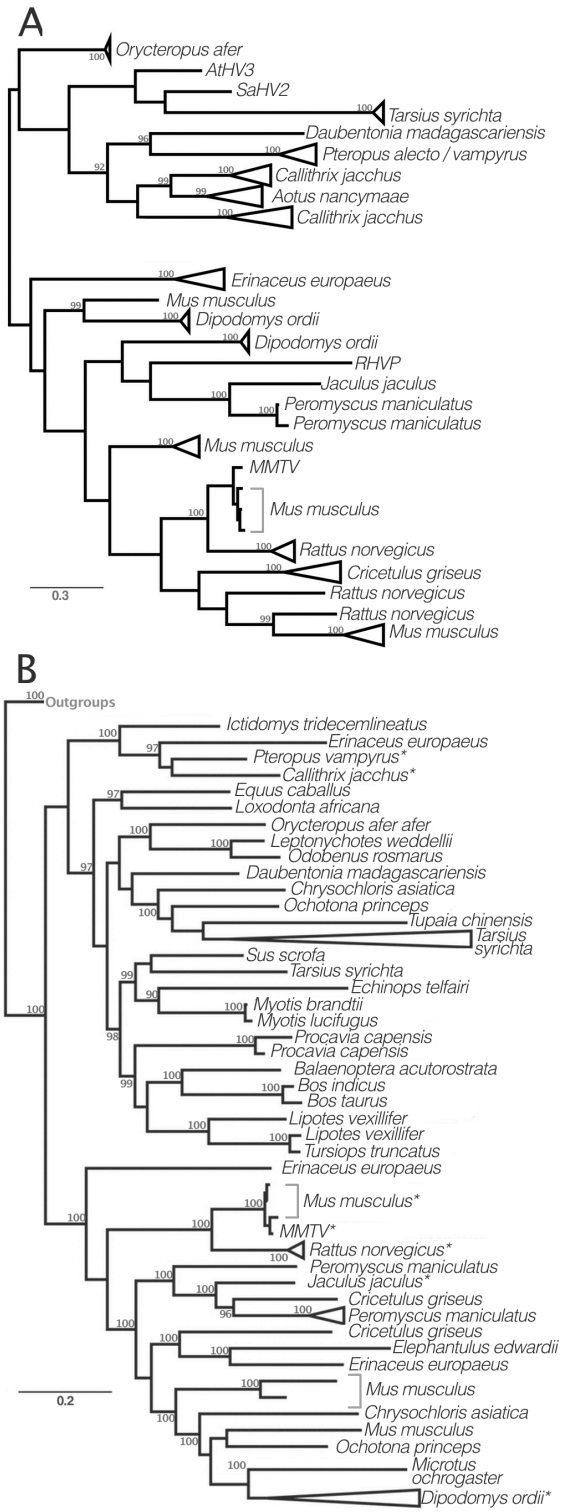


Figure 4 Phylogenetic tree of MMTV-like polymerase and sag sequences
Panel A: The Bayesian phylogenetic tree shown was reconstructed using only the *sag* partition of the alignment. Albeit with lower support for deep nodes, the *sag*-only tree recovers the distinction between predominantly rodent- and primate-infecting viruses. The tree reveals that there are three independent lineages of rat ERVs, four lineages of mouse, and two separate kangaroo rat clades. The scale bar shown represents amino acid substitutions per site and posterior probability values above 90% are indicated at each node. Panel B: The Bayesian phylogenetic tree shown was reconstructed using only the reverse-transcriptase and integrase (RT & INT) partition of the alignment. Both clade-A and B are observed as with the Fig. 3 and *sag*-only tree in the upper panel. In addition, the RT-INT tree shows that clade-A is closely related to a large and diverse group of *sag*-less mammalian ERVs. Starred tips indicate a *pol* of an ERV also included in the phylogenetic analysis of a concatenated alignment of *gag, pol, env* and *sag*. The scale bar shown represents amino acid substitutions per site and posterior probability values above 90% are indicated at each node.

Selection analysis of retroviral and herpesviral sags

The results from the phylogenetic analysis support the hypothesis that *sag* was co-opted by herpesviruses from retroviruses. If this is true, the retroviral donors must have possessed functional *sag,* and so we sought evidence for a history of purifying selection in the *sag*-like sequences that we identified. We conducted a maximum likelihood based analysis of the ratio of non-synonymous to synonymous (*dN/dS* or ω) using PAML [24]. We estimated average *dN/dS* along branches in the tree (the so-called branch models) by allowing every lineage to have its own ω (free-ratio test) or fixing the tree to a single estimate (one-ratio test). We also tested a number of two- and multi-ratio tests that did not return a significant improvement of likelihood over the one-ratio test (*see* materials and methods, Supplementary Table 2). Despite the fact that the majority of the sequences are neutrally evolving non-functional ERVs, the one-ratio test revealed ω = 0.289, which is indicative of purifying selection. This means that their endogenisation was relatively recent (which is consistent with the results of the LTR dating) and these sequences were likely functional in the past. This also indicates that the donor was most likely an infectious retrovirus rather than an integrated virus, which would have a dN/dS value much closer to one because of neutral drift in the host genome. Furthermore, we find that a known conserved motif - crucial for T-cell receptor (TCR) activation [1] – is well conserved in the *sag*-like sequences of the ERVs investigated (Fig. 3). In sequences without the exact TGXY motif, conservation is nonetheless maintained at the nucleotide level suggesting that the site has drifted since endogenisation (Fig. 3).

To investigate if adaptive changes occurred in *sag* during capture by herpesviruses, we implemented the branch-site model to test for positive selection at individual sites along the internal monkey herpesvirus branch. This revealed that four amino acids exhibited evidence for positive selection (*p<0.0001*, Fig. 3). For sites 46 and 72, we see either glutamine or glutamic acid in the majority of *sags*, both of which are polar but have changed in monkey herpesviruses to the small, non-polar glycine and valine, respectively (Fig. 3). Interestingly, amino acid 84 is a lysine in most sequences, but has undergone a triple-site change in SaHV2/AtHV3 (AAA-CGC, relative to *C. jacchus*) to encode the similarly positively charged arginine, despite only needing a single site change. This could indicate that this biochemical property is important here, despite being a positively selected site. A triple-site change at residue 136 for AtHV3/SaHV2 encodes lycine rather than alanine as in most *sags.* Positions 84 and 136 are also different in RHVP relative to the phylogenetically closest sequences, but neither change is equivalent to the change in SaHV2 and AtHV3, nor are they biochemically comparable. This could mean that these

sites are under diversifying selection, or else could reflect adaptation to host-specific targets. Due to the limitations of selection analyses, we consider these 4 sites as the minimum number of possible adaptive sites. This is because selection analyses cannot detect selective sweeps (where an adaptive site is adopted across the entire population).

While the principal drivers of these patterns of selection would have taken effect on *sag* in the exogenous virus, they could have continued post-endogenisation, if *sag* was co-opted by the host. For instance, endogenous *sag* expression in mice can be viewed as beneficial to the host since it reduces the spread of exogenous MMTV [25]. To look for post-endogenisation purifying selection, we examined orthologous loci to control for the signal of purifying selection that results from exogenous viral evolution, but did not find any significant evidence of $dN/dS < 1$. We identified orthologous sag sequences in the two bat species as well as in the different strains of mouse by aligning the LTRs and checking to see if the host flanking sequence was identical. There were zero differences between the mouse sequences and none of the 25 orthologous *Pteropus* LTRs exhibited a dN/dS significantly different to 1 according to the likelihood ratio test with a p-value threshold 0.05 (Supplementary table 3). Nonetheless, it is noteworthy that a search of the NCBI expression databases EST and TSA reveals hits to *sag* in the mouse (120), deer mouse (12), rat (3) and Chinese hamster (3) with an e-value threshold of 0.00001, a minimum alignment length of 60 amino acids with no more than 20 mismatches.

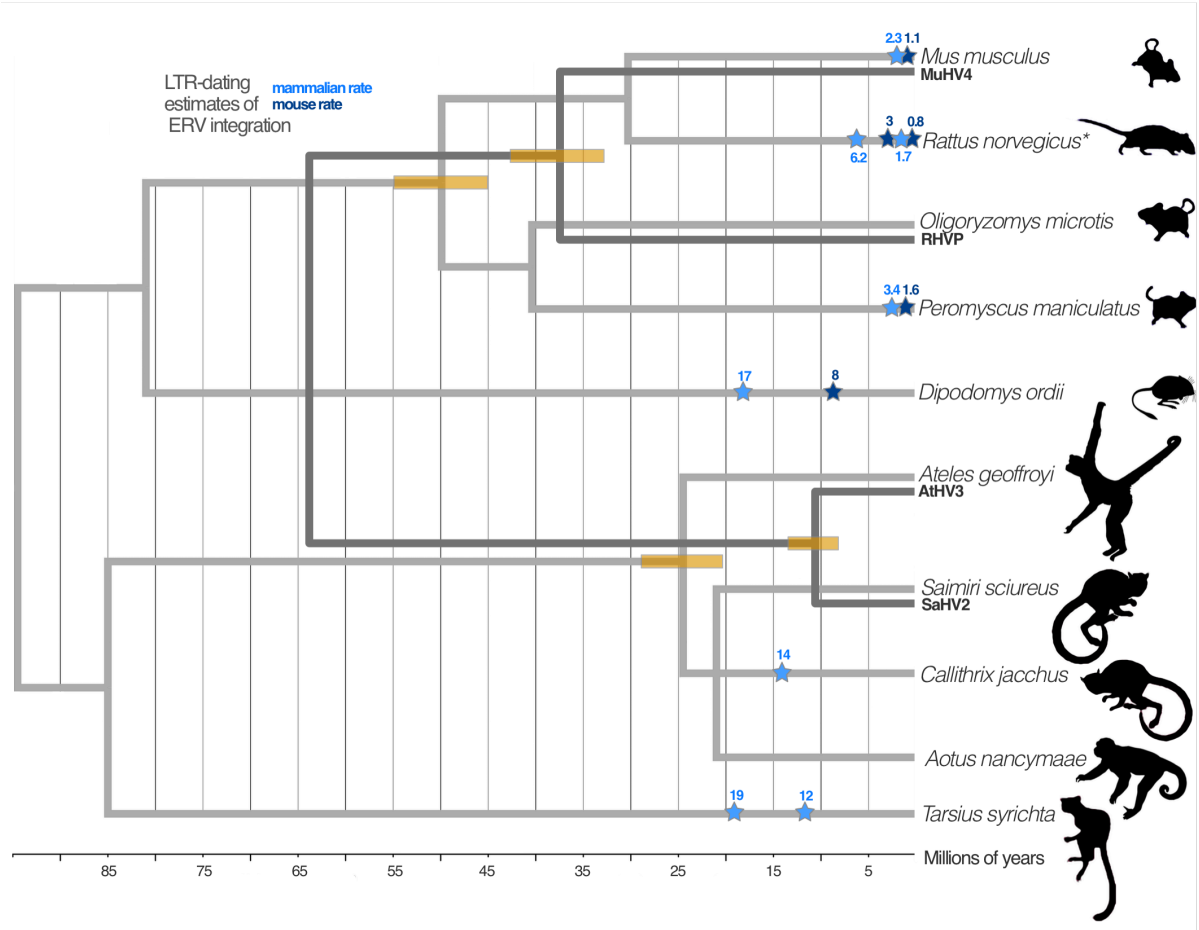


Figure 5 Evolutionary scenario of herpesvirus and retrovirus *sag*
The figure depicts time-calibrated phylogenies of Herpesviruses (dark grey), and mammalian hosts (light grey). Both trees are shown on the same axis, and have been pruned down to the taxa relevant to this study. Each of the hosts for which *sag*-containing ERVs could be dated through LTR divergence are marked by stars that indicate the age of estimate obtained using different neutral rates of evolution. Only the highest and lowest values for *Rattus norvegicus* are shown for clarity (denoted by an asterix). The herpesvirus branches are drawn alongside the relevant host branch, in order to evaluate whether their respective divergence times are consistent with the null hypothesis of co-speciation. Yellow bars indicate the upper and lower bound age estimates. We cannot confidently reject a scenario of co-speciation for RHVP and MuHV4 since there is a known increase in the rate of evolution along this herpesvirus lineage, and the divergence date estimates are close to those of their hosts. The split between SaHV2 and AtHV3, however, is far younger then that of their hosts with no discernable rate increase. This discrepancy is consistent with cross species transmission of the herpesviruses rather than co-speciation. The divergence date of the two monkey viruses therefore is a minimum bound for the acquisition of *sag*. This date (~10mya) is also similar to the integration date estimate of a closely related *sag* in the *Callithrix jacchus* genome (~14mya), which supports the existence of an exogenous retroviruses circulating in new world monkeys, with a maximum bound at the root of host clade ~25 million years ago.

Discussion

We set out to reconstruct the unusual evolutionary history of viral *sags*, which exhibit sequence similarity despite belonging to unrelated viruses. In this study, we uncovered an unexpected diversity and richness of *sags* in the retroviral fossil record, offering us the necessary evolutionary context in which to reconstruct the shared history of herpesvirus and MMTV *sag.* The results conclusively demonstrate that the herpesviral *sag*s were the result of convergent HGT from ancient retroviruses related to MMTV. Our analyses reveal

that this HGT occurred on at least two different occasions into the same herpesvirus genus, from two independent retroviral donors that are both distinct from MMTV.

The multiple origins of herpesvirus *sags* are consistent with the both the polyphyly of *sag*-containing herpesviruses (Supplementary Fig. 4), as well as the fact that the RHVP *sag* locus is not orthologous to the other herpesvirus *sags*. This allows us to conclusively rule out a single-capture scenario in the ancestor of the three rhadinoviruses, all of which infect South American hosts. Indeed, a scenario where *sag* was transferred from retroviruses once a does not explain why *sag*-containing herpesviruses are not monophyletic, and would also require multiple losses of *sag* in other herpesviruses, and at least one recombination event to explain why it is in a different location in RHVP.

We are able to reconstruct a timeline of events in the evolutionary history of *sag* by combining information from the sag phylogeny with host and herpesvirus speciation dates, as well as the dates of ERV integration (Fig. 5). It should be noted that this should only be considered an estimate of the timeframe due to the inaccuracies of molecular clock dating. For example, the evolutionary rate can vary between different locations in the genome [26], and is known to correlate with a number of other variables such as body size, lifespan, generation time and fecundity [27], as well as varying between different lineages for specific genes [28]. Such variation could be particularly influential for the LTR dating, because we could be applying estimates based on averages to specific loci in different species (e.g. using the mouse average rate for specific mouse loci or for loci in other rodents). However, this variation in evolutionary rate is less problematic for both the host tree dates obtained from [23] and our time-scaled herpesvirus tree, which is calibrated using these dates. This is because of the observation that evolutionary rates are relatively homogenous across mammals when the estimates are obtained by averaging over many genes simultaneously [28].With these caveats in mind, the lower bound estimate for the mouse-pygmy rice rat split is 45.1my, which is close to the upper bound of the 95% confidence interval for the MuHV4-RHVP split (42.6mya) obtained from our dating analysis. However, although the dates do not overlap exactly, this can be explained by the increased rate of evolution in the MuHV4-RHVP lineage that would result in an underestimated date. This observation has been previously made [29] and is confirmed in our analysis by the long branches of MuHV4 and RHVP. Another reason for the discrepancy in ages is that the viral speciation could have occurred slightly after that of the hosts, since it would have continued to cross-infect for some time. We therefore cannot reject the hypothesis that RHVP and MuHV4 co-speciated with their hosts. We can

therefore use the speciation of mouse and pygmy rice rat as an upper bound for the acquisition of *sag,* since the gene is absent in MuHV4.

On the other hand, the AtHV3-SaHV2 common ancestor is estimated to be much younger than the split of their hosts at 10.6mya compared to 24.5mya. Moreover, we did not detect an elevated evolutionary rate that could explain the difference, nor have any other previously published studies [29,30]. This means that one of the two viruses arose from a cross-species transmission several million years after the hosts diverged. Because the *sag* gene in both viruses are at the same locus (Fig. 1) and are phylogenetically most similar to each other, they most likely originated from a single acquisition in their common ancestor. However, since the genes are located in a gene-capture hotspot, we cannot rule out the possibility that they appear syntenic by coincidence, and that their phylogenetic similarity is due to convergent evolution. To examine the latter possibility, we re-estimated the phylogeny after removing the amino-acid positions that we identified as positively selected in herpesviruses (Fig. 3), in case these sites led them to appear artificially more similar (Supplementary Fig. 5). Nonetheless, the viruses still resolve as a monophyletic group, which is consistent with the single-capture hypothesis.

It will be interesting to see if future studies identify retroviral lineages that split the monophyly of the herpesvirus *sags*, which would mean that the convergent capture of *sag* occurs more often than our findings demonstrate. Whether there were one or two HGT events, multiple independent integrations of non-orthologous ERVs in both *Aotus and Calithrix* species indicate that they had a single infectious retroviral ancestor, which is also ancestral to the SaHV2/AtHV3 clade (Fig. 3). This means that the retroviral donor to the monkey herpesviruses was also infectious, which is consistent with the finding of low dN/dS values along the monkey herpesvirus *sag* internal branch. Moreover, the fact that *sag* is absent from any Old World primate ERVs suggests that the retrovirus crossed into New world monkeys after their arrival in South America in the late Eocene/Early Oligocene [31]. Together, these data support the existence of an infectious retrovirus circulating among and transmitting between monkeys in South America, leaving endogenous counterparts in multiple host and herpesvirus genomes.

The search for past evidence of *sag* in the retroviral fossil record revealed at least 20 separate lineages of ERV from 8 mammalian orders with *sag* similarity in their LTR (Fig. 4). While our shortlisting criteria are a conservative estimate of the past diversity of *sag*-containing retroviruses, the number of taxa included in the trees and the long branches

between clades is strong evidence of the diverse and widespread nature of *sags* in ancient betaretroviruses. The vast majority of ERVs will eventually succumb to deletion because of illegitimate recombination between the LTRs. For example, the number of solo-LTRs in the human genome far outnumbers the count of full length ERVs [32]. Being in the LTR, *sag* is therefore more likely to be preserved in the fossil record, and this is reflected in the results of our search. The LTRs of retroviruses tend to be poorly conserved beyond closely related clades [33], since there is no selective pressure to maintain a particular sequence. However, the fact that the LTRs encode *sag* in MMTV-like betaretroviruses means that they are conserved across a more diverse range of retroviruses. Therefore, as well as revealing the history of herpesvirus *sag* acquisition, the *sag* tree also represents an analysis of LTR evolution in deep time that is based on a sequence alignment, rather than the alternative methods that are necessary to circumvent the difficulty of aligning LTRs [34].

Herpesviruses typically infect hosts indefinitely, and move from latent to lytic replication during periods of immunosuppression. A host expressing *sag* mRNA is more likely to be in an immune-compromised state that is the ideal environment for *sag* gene-transfer in an activated herpesvirus. For the gene transfer to take place a herpesvirus would have likely infected a host with a high retroviral load, which could consequently lead to the accidental integration of the retrovirus followed by the formation of a solo LTR. Experimental evidence establishing the ability of retroviral integration into herpesviruses was demonstrated over a decade ago, where *Reticulendotheliosis virus* (REV) and Marek's Disease Virus (MDV) were co-infected both *in vitro* and *in vivo* [35]. More recently, a number of studies have shown that the naturally occurring REV LTR in field strains of MDV directly influence MDV expression and pathogenicity [36]. The phenomenon is not limited to herpesvirus as recipients, nor are retroviruses their only donor. The Fowlpox Virus is also known to have recombined with REV [37], and the small single-stranded DNA adeno-associated virus was the source of gene captured by *Human Herpesvirus 6* [38]. A common biological mechanism in these examples is that the recipients are large DNA viruses, while the donors are small viruses with an established DNA integration mechanism. The high chance of illegitimate recombination between LTRs, leading to the loss of all but a solo-LTR, explains the absence of the rest of the virus in herpesviruses, and is the mechanism that has been demonstrated in the case of MDV [39]. Alternatively, although we identified a low dN/dS value along the internal herpesvirus branch, it is possible that the acquisition did not occur during co-infection by retroviral integration followed by solo-LTR formation. It could be that a host-integrated retroviruses expressed

sag mRNA that was then reverse-transcribed and integrated by recombining with a LINE or SINE, although no evidence of any such element was detected using RepeatMasker, nor did we identify a poly-A tail that is characteristic of mRNA.

The convergent acquisition of the same gene by herpesviruses of the same genus suggests a shared evolutionary mechanism. MMTV-like retroviruses have donated *sag* to retroviruses, which implicitly reveals a history of retrovirus-herpesvirus co-infection in two different mammalian orders, The immunomodulatory benefit offered by *sag* explains its tendency to be captured as part of the perpetual arms-race that viruses and hosts are engaged in; herpesviruses tend to capture sequences with immunomodulatory functions [40]. In the case of the monkey herpesviruses, the gene was maintained after speciation, which reflects the selective advantage it may have offered at a time when new world monkeys were rapidly diversifying in South America.

The finding that *sag* transfer between unrelated viruses was convergent could indicate that rhadinoviruses and retroviruses are particularly amenable to co-infection, which would explain the recurrence of an otherwise extremely rare event. In the case of *Human Herpesvirus 8*, the etiological agent of Kaposi's sarcoma, symptoms are almost exclusive to HIV patients, which is also a retrovirus. The data have also revealed that *sag* genes are widespread in ancient mammalian betaretroviruses, suggesting that many mammals are potential reservoirs of hitherto unknown *sag*-containing viruses, potentially offering immediate applications in epidemiological surveillance. Furthermore, our findings could facilitate future surveillance efforts in anticipate similar gene acquisitions that might drastically change the recipient virus' pathology. Indeed, we do not fully understand what drives the differences in pathogenicity of many herpesviruses, both in the course of a lifetime infection and in different individuals.

Host and virus evolutionary rates differ dramatically, an imbalance favoring viruses that quickly adapt to host defenses. One compensation mechanism employed by hosts is the occasional evolutionary shortcut in the form of viral gene capture; bypassing the need for slow molecular adaptation by re-purposing viral genes that have already undergone the necessary evolution [25]. Interestingly, the endogenous counterpart of MMTV in the mouse offers an advantage against exogenous MMTV. Expression of *sag* during thymic development results in deletion of T-cells that are *sag*-responsive, reducing the viruses capacity to proliferate [41]. Viral gene capture by hosts may facilitate viral endogenisation in general, but is especially likely in the case of *sag* [25]. This is because the endogenous

expression of *sag* is itself a rapid solution to the problems it causes, and the selective pressure imposed on *sag*-less individuals mitigates the negative effect of host *sag* expression. The herpesvirus capture of *sag* is exploiting the same evolutionary advantage as the host for the opposite reason – as a countermeasure to host immunity. Moreover, the fact that *sag* acquisition is convergent indicates a common evolutionary mechanism, which is distinct from that involved in host gene capture. Herpesviruses are exploiting the product of the evolutionary arms race in which retroviruses developed *sag*. In this way, the virus receives a gene with the specific function of counteracting host defenses, but without incurring the cost of incremental molecular evolution.

Methods

Sequence collection & phylogenetic reconstruction.

We conducted similarity searches by tBLASTn, using MMTV amino acid sequences of sag and pol as a query against the WGS database (Genbank Accession NP_056884.1 and NP_056880.1, respectively). 19/20 of the genomes identified in this study were obtained from the database as it was in November 2013, with the later addition of the ERVs in the *Aotus nancymaae* genome that was released during manuscript preparation. The search using MMTV pol retrieved many distantly related ERVs. In order to delimit the sequences to those closely related to MMTV, a neighbor-joining tree was reconstructed using the reverse transcriptase domain of 4750 sequences. We discarded all but a subset of ~100 sequences from the MMTV-like clade and three known outgroups (Avian Leukosis Virus, Human ERV-K108 and JSRV). We also included 8 ERV sequences that were closely related to these outgroups: *Rattus norvegicus, Pteropus alecto* x2*, Ovis aries, Aotus nancymaae, Nomascus leucgenys and Mus musculus*. We manually curated the shortlist by re-examining WGS, using query sequences distributed across the tree. These searches were also done against individual species, as the BLAST scores vary according to the size of the target database, and erroneously excluded ERVs were reintroduced into the list. Each gene was aligned automatically with MUSCLE with manual adjustment and only conserved regions were used for further analysis. The *sag* sequences used were shortlisted according to another round of BLAST searching using the conserved region from each of MMTV, SaHV2, AtHV3 and RHVP as a query. We shortlisted *sag* sequences that aligned to at least 80% of any of the queries with at least 45% identical positions and no more than 45 mismatches. The short list was then reduced to a representative subset using a guide tree, discarding highly similar sequences and favoring contigs that contained the whole ERV. Bayesian phylogenetic analysis was performed using parallel MrBayes 3.2.2. For each tree, two independent MCMC chains were run for 50 million

generations with a 25% burn-in. The tree depicted in Fig. 3 was reconstructed from a concatenated amino acid alignment of *gag, pro, pol, env* and *sag*. We could not include the outgroup set in this concatenated alignment due to the absence of any retroviral genes in solo-LTRs and herpesvirus *sags.* Therefore in order to determine the rooting of the tree, we reconstructed a diagnostic phylogeny from an alignment of the outgroup with a subset of the concatenated alignment that excluded the solo-LTRs and herpesvirus sags. This tree revealed the *Tarsius syrichta* ERVs as the most basal sequences with a posterior probability of 100, allowing us to correctly root the Fig. 3 tree. Prottest 3.4 was used to determine the best evolutionary model, which was JTT+G for *gag*, env and *sag*; RTREV+G for pro and WAG+G for *pol*.

LTR dating & BEAST analysis

The date of integration was estimated for paired LTRs from ERVs included in the tree reconstructed from the concatenated alignment. The LTRs were identified using BLASTn to search one half of the ERV against the other. The ML distance of aligned LTRs was obtained using BASEML [24], and dates were calculated by dividing half the genetic distance by the host neutral rate of evolution. We reconstructed a maximum clade credibility Bayesian phylogenetic tree in BEAST from a concatenated alignment of 6 core genes (terminase, large tegument, uracil-DNA glycosylase, kinase, capsid protein and helicase) using the GTR+Γ substitution model and a relaxed uncorrelated lognormal molecular clock model. As in previous studies [29], we calibrated the tree using the divergence date of *Ovine Herpesvirus 2* from *Porcine Lymphotropic Herpesvirus 2*, their divergence from *Equine Herpesvirus 2*, and the split between Old World primate Lymphocryptoviruses from everything else. Based on the split of their hosts, the respective values in millions of years used were: ruminant-pig divergence (71.6), artiodactyl-perissodactyl divergence (87.3), and primate-ungulate divergence (98.9) [23]. Only a pruned version of the BEAST tree is shown in figure 5; the full tree can be found in Supplementary Fig. 6.

Selection analysis

We analyzed *dN/dS* values along branches in the tree using CODEML [24]. We estimated ω across the whole tree assuming the same value for all branches (one-ratio test) or allowing each branch a separate value (free-ratio test). The free-ratio model explained the data significantly better than the one-ratio (p=0.006 Fig. 3) by a likelihood ratio test where the X^2 value was equal to twice the difference of the log likelihoods with 49 degrees of freedom. The free-ratio ω estimate for the monkey herpesvirus internal branch (0.461)

was higher than the tips (0.238 & 0.307). These values represent an average along the branches, and should be interpreted only as a guide in the exploration of selection patterns in the tree. We were interested whether this reflected an episode of positive selection at the time of *sag* acquisition, and whether a convergent change could be detected in RHVP. To investigate this and address the problem of over-parameterization, a number of models were implemented that group branches into a small number of categories. This included distinguishing herpesviruses from retroviruses, clade-A from clade-B, and tips from internal branches (Supplementary table 2). We also implemented the branch-site test for the internal branch of SaHV2 and AtHV3. Significance was determined by a log likelihood ratio test where the null hypothesis is that dN/dS = 1 ($p<0.0001$). Because selective sweeps cannot be detected, our findings are likely to be an underestimate of the positively evolving sites in *sag.* Nonetheless the Bayes Empirical Bayes BEB analysis indicated 5 sites evolving under positive selection. One of the sites detected is tentative, due to low certainty of the alignment at that position, and the remaining sites were visualized using a WebLogo (Fig. 3). The likelihood scores of estimated pairwise dN/dS values of *Pteropus* orthologs were compared against the score of fixing the dN/dS to one using an LTR.

References

1. Fortin J-S, Genève L, Gauthier C, Shoukry NH, Azar GA, et al. (2014) MMTV superantigens coerce an unconventional topology between the TCR and MHC class II. J Immunol 192: 1896–1906. doi:10.4049/jimmunol.1203130.
2. Fraser JD, Proft T (2008) The bacterial superantigen and superantigen-like proteins. Immunol Rev 225: 226–243. doi:10.1111/j.1600-065X.2008.00681.x.
3. Llewelyn M, Cohen J (2002) Superantigens: microbial agents that corrupt immunity. Lancet Infect Dis 2: 156–162.
4. Acha-Orbea H, MacDonald HR (1995) Superantigens of mouse mammary tumor virus. Annu Rev Immunol 13: 459–486. doi:10.1146/annurev.iy.13.040195.002331.
5. Wheeler DA, Butel JS, Medina D, Cardiff RD, Hager GL (1983) Transcription of mouse mammary tumor virus: identification of a candidate mRNA for the long terminal repeat gene product. J Virol 46: 42–49.
6. van Ooyen AJ, Michalides RJ, Nusse R (1983) Structural analysis of a 1.7-kilobase mouse mammary tumor virus-specific RNA. J Virol 46: 362–370.
7. Jarvis CD, Germain RN, Hager GL, Damschroder M, Matis LA (1994) Tissue-specific expression of messenger RNAs encoding endogenous viral superantigens. J Immunol 152: 1032–1038.
8. Arroyo J, Winchester E, McLellan B, Huber B (1997) Shared promoter elements between a viral superantigen and the major histocompatibility complex class II-associated invariant chain. J Virol 71: 1237–1245.
9. Günzburg WH, Heinemann F, Wintersperger S, Miethke T, Wagner H, et al. (1993) Endogenous superantigen expression controlled by a novel promoter in the MMTV long terminal repeat. Nature 364: 154–158. doi:10.1038/364154a0.
10. McGeoch DJ, Davison AJ, Dolan A, Gatherer D, Sevilla-Reyes EE (2008) Molecular Evolution of the Herpesvirales. In: Domingo E, Holland JJ, editors. Origin and Evolution of Viruses. Academic Press. pp. 447–475. doi:10.1016/B978-0-12-374153-0.00020-5.
11. Fields BN, Knipe DM, Howley PM (2007) Fields Virology. 5th ed. Philadelphia:Lippincott Williams&Wilkins.
12. Albrecht J-C (2000) Primary Structure of the Herpesvirus Ateles Genome. J Virol 74: 1033–1037. doi:10.1128/JVI.74.2.1033-1037.2000.
13. Albrecht JC, Nicholas J, Biller D, Cameron KR, Biesinger B, et al. (1992) Primary structure of the herpesvirus saimiri genome. J Virol 66: 5047–5058.

14. Loh J, Zhao G, Nelson CA, Coder P, Droit L, et al. (2011) Identification and sequencing of a novel rodent gammaherpesvirus that establishes acute and latent infection in laboratory mice. J Virol 85: 2642–2656. doi:10.1128/JVI.01661-10.
15. Krupovic M, Koonin E V (2014) Evolution of eukaryotic single-stranded DNA viruses of the Bidnaviridae family from genes of four other groups of widely different viruses. Sci Rep 4: 5347. doi:10.1038/srep05347.
16. 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. doi:10.1186/1745-6150-7-13.
17. Katzourakis A, Gifford RRJ (2010) Endogenous viral elements in animal genomes. PLoS Genet 6: e1001191.
18. Feschotte C, Gilbert C (2012) Endogenous viruses: insights into viral evolution and impact on host biology. Nat Rev Genet 13: 283–296. doi:10.1038/nrg3199.
19. Johnson WE (2010) Endless forms most viral. PLoS Genet 6: e1001210. doi:10.1371/journal.pgen.1001210.
20. Demogines A, Abraham J, Choe H, Farzan M, Sawyer SL (2013) Dual host-virus arms races shape an essential housekeeping protein. PLoS Biol 11: e1001571. doi:10.1371/journal.pbio.1001571.
21. Kumar S, Subramanian S (2002) Mutation rates in mammalian genomes. PNAS 99: 803–808. doi:10.1073/pnas.022629899.
22. Waterston RH, Lindblad-Toh K, Birney E, Rogers J, Abril JF, et al. (2002) Initial sequencing and comparative analysis of the mouse genome. Nature 420: 520–562. doi:10.1038/nature01262.
23. Bininda-Emonds ORP, Cardillo M, Jones KE, MacPhee RDE, Beck RMD, et al. (2007) The delayed rise of present-day mammals. Nature 446: 507–512. doi:10.1038/nature05634.
24. Yang Z (2007) PAML 4: phylogenetic analysis by maximum likelihood. Mol Biol Evol 24: 1586–1591. doi:10.1093/molbev/msm088.
25. Aswad A, Katzourakis A (2012) Paleovirology and virally derived immunity. Trends Ecol Evol 27: 627–636. doi:10.1016/j.tree.2012.07.007.
26. Hodgkinson A, Eyre-Walker A (2011) Variation in the mutation rate across mammalian genomes. Nat Rev Genet 12: 756–766. doi:10.1038/nrg3098.
27. Welch JJ, Bininda-Emonds ORP, Bromham L (2008) Correlates of substitution rate variation in mammalian protein-coding sequences. BMC Evol Biol 8: 53. doi:10.1186/1471-2148-8-53.
28. Bininda-Emonds ORP (2007) Fast genes and slow clades: comparative rates of molecular evolution in mammals. Evol Bioinform Online 3: 59–85.
29. McGeoch DJ, Gatherer D, Dolan A (2005) On phylogenetic relationships among major lineages of the Gammaherpesvirinae. J Gen Virol 86: 307–316. doi:10.1099/vir.0.80588-0.
30. Ehlers B, Dural G, Yasmum N, Lembo T, de Thoisy B, et al. (2008) Novel mammalian herpesviruses and lineages within the Gammaherpesvirinae: cospeciation and interspecies transfer. J Virol 82: 3509–3516. doi:10.1128/JVI.02646-07.
31. Poux C, Chevret P, Huchon D, de Jong WW, Douzery EJP (2006) Arrival and diversification of caviomorph rodents and platyrrhine primates in South America. Syst Biol 55: 228–244. doi:10.1080/10635150500481390.
32. Sverdlov ED (1998) Perpetually mobile footprints of ancient infections in human genome. FEBS Lett 428:1–6.
33. Benachenhou F, Jern P, Oja M, Sperber G, Blikstad V, et al. (2009) Evolutionary Conservation of Orthoretroviral Long Terminal Repeats (LTRs) and ab initio Detection of Single LTRs in Genomic Data. PLoS One 4: e5179. doi:10.1371/journal.pone.0005179.
34. Benachenhou F, Sperber GO, Bongcam-Rudloff E, Andersson G, Boeke JD, et al. (2013) Conserved structure and inferred evolutionary history of long terminal repeats (LTRs). Mob DNA 4: 5. doi:10.1186/1759-8753-4-5.
35. Isfort R, Jones D, Kost R, Witter R, Kung HJ (1992) Retrovirus insertion into herpesvirus in vitro and in vivo. PNAS 89: 991–995. doi:10.1073/pnas.89.3.991.
36. Sun A-J, Xu X-Y, Petherbridge L, Zhao Y-G, Nair V, et al. (2010) Functional evaluation of the role of reticuloendotheliosis virus long terminal repeat (LTR) integrated into the genome of a field strain of Marek's disease virus. Virology 397: 270–276. doi:10.1016/j.virol.2009.11.017.
37. Hertig C, Coupar BE, Gould AR, Boyle DB (1997) Field and vaccine strains of fowlpox virus carry integrated sequences from the avian retrovirus, reticuloendotheliosis virus. Virology 235: 367–376.
38. Thomson BJ, Efstathiou S, Honess RW (1991) Acquisition of the human adeno-associated virus type-2 rep gene by human herpesvirus type-6. Nature 351: 78–80. doi:10.1038/351078a0.
39. Jones D, Isfort R, Witter R, Kost R, Kung HJ (1993) Retroviral insertions into a herpesvirus are clustered at the junctions of the short repeat and short unique sequences. PNAS 90: 3855–3859. doi:10.1073/pnas.90.9.3855.
40. Shackelton LA, Holmes EC (2004) The evolution of large DNA viruses: combining genomic information of viruses and their hosts. Trends Microbiol 12: 458–465.
41. Bhadra S, Lozano M, Dudley J (2009) BALB/Mtv-null mice responding to strong mouse mammary tumor virus superantigens restrict mammary tumorigenesis. J Virol.

Discussion: State of the art and future directions

The studies undertaken throughout this thesis are an investigation of horizontal gene transfer between viruses and their hosts. By drawing from the principles of paleovirology that were being established at the start of this DPhil, I developed additional methods and techniques to study EVEs and address some of the outstanding questions at the time. For the final pages of this thesis, I will review the various results presented in previous chapters, highlighting the questions that motivated each study. In addition, I will discuss the challenges faced in each chapter, including those that were not overcome and how they influence the interpretation of results. The various chapters broadly belong to two interconnected themes. I will first discuss the chapters that together represent the 'direct paleovirology' aspect of this thesis, highlighting the contributions I have made to the field both in terms of results but also novel perspectives. Second, I will address discuss the ideas and results of the last three chapters as part of the general theme horizontal gene transfer, and how the concepts elucidated through the results and discussions suggest that EVEs and host derived genes can be considered as two sides of the same biological strategy within an evolutionary arms race. For both themes I will also highlight the various future directions that are clear avenues of further research.

Chapter one and two: The discovery of herpesvirus EVEs reveal the interdisciplinary potential of paleovirology

I set out to determine whether vertebrate genomes could accommodate endogenous viruses as large as herpesviruses, while simultaneously aiming to address the technical challenges that had discouraged their investigation in the first place. For instance, viruses over 100Kbps were not investigated in the 2010 survey that identified EVEs representing all main genome types and replication strategies [1]. Unlike most viruses, large dsDNA virus genomes can contain over 100 genes, many of which are lineage specific including those that are derived from hosts [2]. Therefore rather than a search strategy that considered all of the virus' sequence data, I needed to design an appropriate method that would still be sensitive enough to detect an EVE while avoiding false positives with similar host genes. This particular problem was overcome by performing reciprocal BLAST searching to exclude obvious false positives, as well as phylogenetic analyses that could rapidly demonstrate whether a potential 'hit' was more closely related to hosts or viruses.

The herpesvirus genes used to search host genomes were selected based on their high levels of conservation across a diverse range of herpesviruses, in addition to being

suitable for phylogenetic reconstruction as demonstrated by their use in the existing literature [2–5]. Moreover, they were also found at distinct locations in the genome, which is an advantage when searching against draft assemblies with contigs that are likely to be smaller than the size of the full viral genome (i.e. the lower the N50 statistic, the higher the chance that the virus would not be found intact in a single contig). It became apparent that in addition to modifying existing paleovirology techniques, the task of finding a herpesvirus EVE necessitated the incorporation of methods in genomics to circumvent the problem of a low N50. Developing a targeted contig-extension algorithm similar to those used in scaffolding software partially solved this problem in the case of the novel aye-aye virus. Rather than attempting to improve the assembly overall, my technique allowed a more efficient targeting of only the viral genome that we expected [3].

As well as significantly contributing to paleovirology by identifying the largest vertebrate EVE in a primate, the study also revealed the utility of paleovirology techniques to discover novel viruses. In addition to the tarsier herpesvirus EVE, the investigation also uncovered an entirely novel aye-aye virus related to the human Kaposi's sarcoma virus, as well as characterized novel sequences of the chimpanzee equivalent of Epstein-Barr virus [3]. For the first time, I argued that online genome databases could be viewed as a metagenomic resource for viral discovery that includes more biodiversity than the sum of its contributing host species. Rather than treating viral reads in genome data as contaminant, we should be opportunistically characterizing them as potential novel sequences, and this was subsequently supported in a review by Houldcroft and Breuer that discussed approaches to herpesvirus discovery [6].

The motivation behind the work in chapter two came from the perspective that was formulated in chapter one, where novel viruses that are not EVEs were identified. However, both the chimpanzee and aye-aye viruses that were characterized are very closely related to existing gammaherpsviruses. In chapter two I endeavored to test the efficacy of this approach by considering the parts of the herpesvirus phylogeny that are far less characterized, with an aim to identify more genetically diverse viruses. For this reason, I reapplied the same technique using query genes from malacoherpesviruses, which are the most poorly understood group of the *Herpesvirales*. Rather than revealing novel malacoherpesvirus diversity, the analyses revealed an entirely new lineage of dsDNA virus in fish. Moreover, their genomes contained strikingly diverse genes that only shared very low amino acid sequence identify to known genes, as well as many genes

that could not be identified at all. This is reminiscent of the giant viruses that contain many unknown genes, which are also dsDNA viruses ([7] and Appendix two).

The primary limitation evident in both chapter one and two is that only partial viral genomes are identified, even after the improvements made to contig length through the targeted scaffolding technique [3]. Some further improvements could conceivably be made to the bioinformatic pipeline that may improve the output, but ultimately the approach would always benefit from a supplementary sequencing component. Even in the ideal case of the novel salmon virus identified in chapter two, where the full length of the viral genome was identified, confirmatory PCR is still warranted. For instance, although it was outside the scope of this study, it would have been useful to sequence the regions of the contig with the lowest read coverage to serve as additional evidence for the structure of the genome. This is even more important in chapter two than chapter one, since none of the viruses identified exhibit a comparable genomic organization to known viruses.

Chapter three, four and five: EVEs as a special case of horizontal gene transfer between different viruses and hosts

The other main theme of this thesis is the study of gene exchange between viruses and their hosts. This began with chapter three, where I reviewed and discussed how some EVEs have been shown to perform beneficial functions in their hosts, most often as anti-viral immune genes [8]. However, rather than simply viewing these as rare accidents that evolved from EVE 'junk DNA', I developed an evolutionary framework to study EVE-derived immunity genes (EDIs) as examples of gene flow from viruses to hosts. In particular, I argued that this was an adaptive response that may have recurred many times, and could even explain the initial capture of currently non-functional EVEs that had since lost the selective pressure required to maintain them. The ideas that were developed in this chapter considered the capture of viral genes by hosts as part of the molecular arms-race between viruses and hosts.

Crucial to the concepts I used in chapter three was the idea that EDIs are one side of a genetic exchange between viruses and hosts. With this perspective in mind, the last two chapters of my thesis departed from the way that existing paleovirology had been approached and instead applied the same techniques to investigate host-derived genes in viruses. Unlike EDIs, such host-derived genes are more widely studied and particularly common in large dsDNA viruses. However, in chapter four I set out to systematically reconstruct the evolutionary history of captured genes in a group of herpesviruses. This

included an in-depth phylogenetic analysis to determine the host group of origin, as well as a series of selection tests to understand how such genes evolved. Through these analyses, I uncovered the remarkable finding that host-derived genes undergo an evolutionary shift towards lower dN/dS values compared to the host counterparts, indicating stronger purifying selection in the virus. This indicates there is a common evolutionary mechanism that preserves the function of these genes, suggesting that such genes are repeatedly captured by the same adaptive mechanism as part of the evolutionary arms-race. The main limitation in this study was that while the analyses were comprehensive, they remained limited to the seven genomes of the genus *Rhadinovirus.* Given the opportunity and time to expand the study, it would have been advantageous to include a large sample of species, and to compare the processes we described across a wider variety of host genes. Furthermore, it would be crucial in future work to compare the same host derived genes that are found in different viruses. For instance, members of gene families that are frequently captured by herpesviruses (such as interleukin 10) are also captured by poxviruses [9]. Finally, one of the main discussion points drawn from the results of chapter four was that the evolutionary patterns we revealed could be the driving force behind the high turnover of captured genes, as indicated by the fact that many host-derived genes are lineage specific. In order to test this idea, another avenue for future research would be to examine whole-genome data for a large number of samples for the same virus, to estimate the frequency of host-gene uptake at the population level.

The study described in chapter five illustrates the utility of paleovirology to investigate the evolution of gene transfer in the context of the evolutionary arms-race. However, while chapters three and four were dedicated to the exchange of genes between virus and host, chapter five deals with HGT between different viruses in their shared host. Not only is this phenomenon extremely rare, this study revealed that the *sag* gene had transferred into rhadinoviruses from different retroviruses on at least two occasions. This indicates that there is an underlying mechanism at play, and it was possible to explore the dynamics of this evolutionary process using the theoretical framework developed in chapter three. Moreover, this chapter is a good display of the synthesis between ‘direct’ and ‘indirect’ paleovirology, since it involved both mining mammalian ERVs and evolutionary analyses in order to disentangle the relationship between extant herpesviral and retroviral *sag.*

Although it was outside the scope of this particular endeavor, given the opportunity to work on chapter five further, it would have been interesting to pursue at least two other avenues of enquiry. First, because the *sag* gene is located in the LTR of the retroviruses,

it presents an unprecedented opportunity to study a large group of ERVs to a greater degree of accuracy than has been possible to date. This is because distantly related retroviruses share very little similarity in their LTRs, which precludes the alignment of the vast majority of ERVs that are found as solo-LTRs. However, because of the purifying selection imposed on the *sag* gene, this large and diverse group of betaretroviruses can be studied in-depth using alignment-based phylogenetic analysis, rather than alternative techniques that less powerful/accurate [10]. This would allow us to investigate the evolutionary history of a large group of viruses in more detail than ever before. One example of how this could be used is to map the history of cross species transmission for a viral group, reaching further back in time than would be possible if we were to depend on scarce ERV genes of closely related viral lineages. The second major question that was beyond the scope of this project is the question of the function of *sag.* Although selection analysis has clearly indicated that the *sag* sequences in ERVs of these mammals were functional, I did not fully characterize the sequences to identify whether any of the genes are intact ORFs, nor did I screen species of interest by PCR such as addition new world monkeys. Moreover, it will be particularly interesting to design experiments that test for differential expression of active *sag* genes, as a preliminary indication of function. It is a high research priority to identify potentially functional *sags* that may be acting as EDIs in their host, as is the case in the endogenous MMTV of mouse that I reviewed in chapter three [8].

The causes and consequences of EVEs

Because I aimed to address broad evolutionary questions in this doctorate, I did not focus on determining the biological mechanisms involved in EVE formation. Indeed this avenue of inquiry has yet to be fully examined by others in the field, and represents a major direction for future work. Nonetheless, It has been less then a decade since non-retroviral EVEs have been studied under the same evolutionary framework – i.e. paleovirology – that also links them to the large body of research into endogenous retroviruses. The presence of ERVs in host genomes is relatively well understood in that we know the biological mechanisms responsible for their integration and subsequent intragenomic proliferation, but it is less clear how other viruses integrate [12]. EVE formation involves two steps; integration into the germline followed by vertical transmission to the progeny. For some EVEs we have been able to infer the mechanism of germline integration, either by extrapolating from the known properties of related exogenous viruses, or observations that indicate a plausible mechanism of integration. For example, we know that a number of herpesviruses have been shown to somatically integrate into the genome, including

human roseoloviruses [11]. Given the close genomic and phylogenetic similarity between roseoloviruses and the endogenous tarsier herpesvirus, we can infer that its integration involved homologous recombination with host telomeric repeats, which have been identified in the tarsier herpesvirus EVE described in chapter one [3]. In the case of EVEs related to RNA viruses, it has been speculated that their conversion to dsDNA (a prerequisite of integration) could be achieved by reverse-transcriptase that is abundant in host genomes due to the presence of ERVs and other retroelements [12]. Indeed it has been shown that cDNA of bornaviruses can be detected in infected cell, and endogenous bornavirus EVEs exhibit features such as poly-A tails and target sited duplications that suggest LINE-mediated reverse transcription and integration [13].

The herpesvirus and bornavirus examples are instances where we have inferred integration events by observing the outcome, but integration itself has also been experimentally demonstrated. Cell culture experiments showed that while the majority of detectable bornaviruses were in the cytosol, a proportion of integrated elements can be detected ~30 days after infection [13]. The group 5 *Lymphocytic choriomeningitis virus* can also become integrated into the genome after reverse transcription and recombination with a mouse ERV [14]. Interestingly, the ERVs in this case belonged to the IAP family of *env*-less ERVs that are particularly prolific in the mouse genome [15], which might positively influence the probability of recombination since there are more opportunities to recombine because they are expressed more frequently. We have also observed the process of vertical transmission of an integrated virus, which is the second step in the formation of an EVE. Both *Human Herpesvirus 6* and the koala retroviral EVE KoRV are unfixed in the host population, and the inheritance of the same EVE locus from parent to offspring has been demonstrated in genealogical studies [16,17]. While none of these examples demonstrates the *in vivo* integration followed by proof of vertical inheritance, synthesizing the evidence for both steps in the endogenisation process allows us to support hypotheses about EVE formation. Of particular relevance to this topic is the work that is outlined in chapter one. For both the tarsier EVE that I describe and the integrated human herpesvirus 6, the hypothesized mechanism of integration is via the *rep* gene, which is itself likely derived from the integration of an adeno-associated virus [3,18,19]. Two highly informative experiments that could be pursued are to test the integration capacity of HHV6 without *rep,* and conversely, to test whether *rep* donates the ability to integrate to a virus that otherwise is not known to do so.

Also in this thesis, I was interested in exploring the evolutionary dynamics of gene transfer as part of the arms-race between viruses and their hosts. However, I did not directly seek evidence of EVEs that are either beneficial or harmful to their host. The harmful effect of any single integration that disrupts a host gene or important regulatory region are comparable for retroviral and non-retroviral EVEs [12]. However, active ERVs impose an additional fitness cost to the host because they can continue to produce pathogenic viruses, which presents the harm of additional integrations that result from its proliferation. Furthermore, inactivated ERV families contribute to non-homologous recombination that can result in major genomic rearrangements. On the other hand, we know of multiple instances where ERVs have played a role beneficial to the host. Syncitins are a group of related genes in mammalian genomes that are derived from the *env* gene of ERVs, and are a crucial component of placental development. Indeed in chapter three, I reviewed and discussed a number of ERV-derived immunity genes (EDIs) that have been described, and there are co-opted ERV regulatory sequences that contribute to host genome function [8]. To date, only one non-retroviral EVE has been shown to be beneficial to its host, but this is sure to change given the numerous EVEs with intact open reading frames and/or evidence of expression. Since the beginning of this project, progress has already been made in this respect. At their initial discovery, evidence of purifying selection in EBLNs suggested that they might have been providing the host with a beneficial function, perhaps as an EDI [8,20,21]. It was recently shown that, *in vitro,* an EBLN in the genome of the thirteen-lined ground squirrel (*Ictidomys tridecemlineatus*) was capable of both reducing *Borna disease virus* replication as well as inhibiting new infections [22]. Other EVEs are on a similar route that will potentially lead to the characterization of novel functional genes. For example, the dependovirus identified in the genome of the degu (*Octodon degus*) is differentially expressed, which indicates a potentially tissue-specific function [23]. Were it within the scope of this thesis, I would have been keen to look for evidence that known EVEs with intact ORFs are beneficial in some way, and this would have linked well thematically with chapter three.

Current and foreseeable challenges in paleovirology

While the precise definition and remit of paleovirology continues to adjust in these early years, at its core is the study of virus evolution within a framework that considers the deep history of EVEs. Across all of the different viral groups, the recurring finding is that viral groups are a lot older than has been inferred by extrapolation from contemporary sequences. This discrepancy between short and long term viral evolutionary rates is an unusual phenomenon that is an area of ongoing research (see [24] for a review).

However, despite the prefix that implies a limitation of the scope to ancient EVEs, the methodology and objectives to studying recently integrated viruses largely overlap, since both broadly involve a sequence similarity search followed by phylogenetic analysis and genomic characterization. In addition to ERVs, a number of non-retroviral EVEs had been described before they were recognized as part of a broader genomic fossil record, but it was not until the advent of cheap, rapid genome sequencing that they could be studied systematically. Efficient data management is one of the primary difficulties that paleovirologists currently face, a challenge that will grow alongside the rapid and voluminous generation of genomic data.

Another issue that is symptomatic of new disciplines is a general lack of standardization across several aspects of EVE research. At the most basic level, nomenclature is often inconsistent, with various studies referring to endogenous viruses differently. Although 'EVE' has been featured in a number of high impact publications since its inception, EVE sub-groups have also been designated as descriptors that sometimes overlap, such as 'non-retroviral integrated RNA viruses' (NIRVs) [25], 'non-retroviral RNA virus sequences' (NRVSs) [26] or 'endogenous plant pararetroviruses' (EPRVs) [27]. Moreover, classification within the ICTV requires a virus species to exhibit multiple distinguishable features or phenotypes that set it apart from other viruses [28], which can be difficult if an EVE is only represented by a sequence. Similarly, viruses at higher taxonomic levels (e.g. genus) are typically grouped according to a set of distinguishing phenotypic criteria, such as morphology or pathogenicity. This means that even if phylogenetic evidence strongly supports the inclusion of an EVE within an established group, the rules that define that group cannot be evaluated for an EVE. An exception to this might be some of the more recent integrations, if an EVE is still capable of producing viral particles.

More importantly, there is a lot of variation in the standards and criteria for authenticating and describing an EVE. The evidence used to support the identity of a viral sequence as genuinely endogenous range from as little as sequence similarity that is not scrutinized further, to fully characterized genomic loci with robust phylogenetic reconstruction and additional sequencing. Furthermore, there are variations in what we think of as a single EVE, for example in the case of those that proliferate intragenomically after the initial integration, or those that undergo duplication events. This is additionally problematic in cases where a single integration is distributed across different loci as in the case of PDVs [29] or hepadnaviruses [20,30,31].

The size of genomic databases has already reached a point where the routine and efficient searching of all known viral proteins against a six-frame translation of all genomes is not feasible on standard desktop computers, requiring instead a computer cluster or super-computer. Even using a high-performance workstation, the search for EVEs is usually limited in scope, either by reducing the number of query sequences or target hosts, or both. This means that our knowledge of the EVE repertoire of host genomes is not usually complete, nor is the full distribution and abundance of EVEs related to particular viral groups. In addition to the taxonomic and nomenclature inconsistencies and variation in methodological rigor, this incompleteness of the genomic fossil record impedes empirical lines of enquiry that will help us understand aspects of viral evolution and history of host interaction. For example in the case of *Parvoviridae*-like EVEs, there have already been over a hundred described loci, the vast majority of which were identified nearly five years before the writing this chapter [20,32]. The number of publically available host genomes has more than doubled since then, suggesting that a comparable number of new *Parvoviridae* EVEs are yet to be described. This means that it will be possible to investigate factors that influence their distribution in host species. Unlike ERVs, the fact that EVEs do not proliferate will allow us for the first time to dissociate the effect of endogenisation on host genomics, ecology and evolution from subsequent intra-genomic proliferation. The answers to such questions as well as new questions will emerge from paleovirology as it establishes norms and standard practices, as well as community resources such as curated, publically available non-redundant databases. To date progress has been made in answering such questions for ERVs, such as the finding that mammalian body size negatively correlates with the total number of ERVs [33], or that they positively correlate with hosts that reproduce via internal fertilization [34].

Conclusion

In addition to the many advances that arise from each of the results in all of the chapters, this thesis has also contributed two major step-changes for the field. First, the synthesis of techniques and perspectives from paleovirology and metagenomics allowed me to demonstrate that large herpesviruses can stably endogenise in primate genomes for millions of years. The methods that were developed to identify the first herpesvirus EVE also revealed the utility of paleovirology for the identification of novel viruses. The identification of a new family of fish viruses demonstrates that we can indeed use data mining and evolutionary reconstruction for future discoveries, particularly as public genome databases continue to grow. With the discovery of giant viruses, research into the

systematics of dsDNA viruses has been reinvigorated [35–37], and this has consequences for the study of the origin of life and our definition of viruses [38–40]. Appendix 1 discusses the potential contribution that paleovirology will bring to giant virus research [7]. This is especially the case for disentangling the relationship between giant viruses, their hosts and the viruses that parasitize the infection itself. This will require in-depth phylogenetic reconstruction and evolutionary analyses of the many horizontally transferred genes and genetic elements that have seemingly been exchanged multiple times, in multiple directions, between several entities [41–44]. The theoretical and methodological approaches developed and used throughout this thesis are ideally suited to address these questions. More generally, as paleovirology continues to transition from its infancy, the scale and scope of the discipline will undoubtedly increase to encompass a wider variety of questions with a wider range of applications in clinical medicine and evolutionary biology. This will in turn be greatly significant to our understanding of viruses, gene exchange, the origin of cellular life and much more.

References

1. Katzourakis A, Gifford RJ (2010) Endogenous viral elements in animal genomes. PLoS Genet 6: e1001191.
2. McGeoch DJ, Davison AJ, Dolan A, Gatherer D, Sevilla-Reyes EE (2008) Molecular Evolution of the Herpesvirales. In: Domingo E, Holland JJ, editors. Origin and Evolution of Viruses. Academic Press. pp. 447–475. doi:10.1016/B978-0-12-374153-0.00020-5.
3. 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. doi:10.1371/journal.pgen.1004332.
4. Ehlers B, Spieß K, Leendertz F, Peeters M, Boesch C, et al. (2010) Lymphocryptovirus phylogeny and the origins of Epstein-Barr virus. J Gen Virol 91: 630–642. doi:10.1099/vir.0.017251-0.
5. McGeoch DJ, Gatherer D, Dolan A (2005) On phylogenetic relationships among major lineages of the Gammaherpesvirinae. J Gen Virol 86: 307–316. doi:10.1099/vir.0.80588-0.
6. Houldcroft CJ, Breuer J (2015) Tales from the crypt and coral reef: the successes and challenges of identifying new herpesviruses using metagenomics. Front Microbiol 6: 188. doi:10.3389/fmicb.2015.00188.
7. Katzourakis A, Aswad A (2014) The origins of giant viruses, virophages and their relatives in host genomes. BMC Biol 12: 51. doi:10.1186/s12915-014-0051-y.
8. Aswad A, Katzourakis A (2012) Paleovirology and virally derived immunity. Trends Ecol Evol 27: 627–636. doi:10.1016/j.tree.2012.07.007.
9. Shackelton LA, Holmes EC (2004) The evolution of large DNA viruses: combining genomic information of viruses and their hosts. Trends Microbiol 12: 458–465.
10. Benachenhou F, Sperber GO, Bongcam-Rudloff E, Andersson G, Boeke JD, et al. (2013) Conserved structure and inferred evolutionary history of long terminal repeats (LTRs). Mob DNA 4: 5. doi:10.1186/1759-8753-4-5.
11. Morissette G, Flamand L (2010) Herpesviruses and chromosomal integration. J Virol 84: 12100–12109. doi:10.1128/JVI.01169-10.
12. Holmes EC (2011) The Evolution of Endogenous Viral Elements. Cell Host Microbe 10: 368–377. doi:10.1016/j.chom.2011.09.002.
13. Horie M, Honda T, Suzuki Y, Kobayashi Y, Daito T, et al. (2010) Endogenous non-retroviral RNA virus elements in mammalian genomes. Nature 463: 84–87. doi:10.1038/nature08695.
14. Geuking MB, Weber J, Dewannieux M, Gorelik E, Heidmann T, et al. (2009) Recombination of retrotransposon and exogenous RNA virus results in nonretroviral cDNA integration. Science 323: 393–396. doi:10.1126/science.1167375.
15. Magiorkinis G, Gifford RJ, Katzourakis A, De Ranter J, Belshaw R (2012) Env-less endogenous retroviruses are genomic superspreaders. PNAS 109: 7385–7390. doi:10.1073/pnas.1200913109.
16. Ishida Y, Zhao K, Greenwood AD, Roca AL (2014) Proliferation of Endogenous Retroviruses in the Early

Stages of a Host Germ Line Invasion. Mol Biol Evol: msu275 – . doi:10.1093/molbev/msu275.
17. Arbuckle JH, Medveczky MM, Luka J, Hadley SH, Luegmayr A, et al. (2010) The latent human herpesvirus-6A genome specifically integrates in telomeres of human chromosomes in vivo and in vitro. PNAS 107: 5563–5568. doi:10.1073/pnas.0913586107.
18. Arbuckle JH, Medveczky PG (2011) The molecular biology of human herpesvirus-6 latency and telomere integration. Microbes Infect 13: 731–741.
19. Thomson BJ, Weindler FW, Gray D, Schwaab V, Heilbronn R (1994) Human Herpesvirus 6 (HHV-6) Is a Helper Virus for Adeno-Associated Virus Type 2 (AAV-2) and the AAV-2 rep Gene Homologue in HHV-6 Can Mediate AAV-2 DNA Replication and Regulate Gene Expression. Virology 204: 304–311.
20. Katzourakis A, Gifford RRJ (2010) Endogenous viral elements in animal genomes. PLoS Genet 6: e1001191.
21. 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. doi:10.1098/rstb.2012.0499.
22. Fujino K, Horie M, Honda T, Merriman DK, Tomonaga K (2014) Inhibition of Borna disease virus replication by an endogenous bornavirus-like element in the ground squirrel genome. PNAS 111: 13175–13180. doi:10.1073/pnas.1407046111.
23. Arriagada G, Gifford RJ (2014) Parvovirus-derived endogenous viral elements in two South American rodent genomes. J Virol 88: 12158–12162. doi:10.1128/JVI.01173-14.
24. Aiewsakun P, Katzourakis A (2015) Endogenous viruses: Connecting recent and ancient viral evolution. Virology. doi:10.1016/j.virol.2015.02.011.
25. Taylor DJ, Dittmar K, Ballinger MJ, Bruenn J (2011) Evolutionary maintenance of filovirus-like genes in bat genomes. BMC Evol Biol 11: 336. doi:10.1186/1471-2148-11-336.
26. Chiba S, Kondo H, Tani A, Saisho D, Sakamoto W, et al. (2011) Widespread endogenization of genome sequences of non-retroviral RNA viruses into plant genomes. PLoS Pathog 7: e1002146. doi:10.1371/journal.ppat.1002146.
27. Staginnus C, Richert-Pöggeler KR (2006) Endogenous pararetroviruses: two-faced travelers in the plant genome. Trends Plant Sci 11: 485–491. doi:10.1016/j.tplants.2006.08.008.
28. King AMQ, Adams MJ, Lefkowitz EJ, editors (2012) Virus Taxonomy: Ninth Report of the International Committee on Taxonomy of Viruses. San Diego: Elsevier. 1327 p.
29. Herniou EA, Huguet E, Thézé J, Bézier A, Periquet G, et al. (2013) When parasitic wasps hijacked viruses: genomic and functional evolution of polydnaviruses. Philos Trans R Soc Lond B Biol Sci 368: 20130051. doi:10.1098/rstb.2013.0051.
30. Liu W, Pan S, Yang H, Bai W, Shen Z, et al. (2012) The first full-length endogenous hepadnaviruses: identification and analysis. J Virol 86: 9510–9513. doi:10.1128/JVI.01164-12.
31. Gilbert C, Feschotte C (2010) Genomic fossils calibrate the long-term evolution of hepadnaviruses. PLoS Biol.
32. Liu H, Fu Y, Xie J, Cheng J, Ghabrial SA, et al. (2011) Widespread endogenization of densoviruses and parvoviruses in animal and human genomes. J Virol 85: 9863–9876. doi:10.1128/JVI.00828-11.
33. Katzourakis A, Magiorkinis G, Lim AG, Gupta S, Belshaw R, et al. (2014) Larger mammalian body size leads to lower retroviral activity. PLoS Pathog 10: e1004214. doi:10.1371/journal.ppat.1004214.
34. Hayward A, Cornwallis CK, Jern P (2014) Pan-vertebrate comparative genomics unmasks retrovirus macroevolution. PNAS 112: 464–469. doi:10.1073/pnas.1414980112.
35. Koonin E, Yutin N (2010) Origin and evolution of eukaryotic large nucleo-cytoplasmic DNA viruses. Intervirology.
36. Colson P, De Lamballerie X, Yutin N, Asgari S, Bigot Y, et al. (2013) “Megavirales”, a proposed new order for eukaryotic nucleocytoplasmic large DNA viruses. Arch Virol 158: 2517–2521. doi:10.1007/s00705-013-1768-6.
37. Yutin N, Koonin E (2013) Pandoraviruses are highly derived phycodnaviruses. DNA.
38. Koonin E, Senkevich T, Dolja V (2009) Compelling reasons why viruses are relevant for the origin of cells. Nat Rev Microbiol.
39. Forterre P (2006) The origin of viruses and their possible roles in major evolutionary transitions. Virus Res 117: 5–16. doi:10.1016/j.virusres.2006.01.010.
40. Raoult D, Forterre P (2008) Redefining viruses: lessons from Mimivirus. Nat Rev Microbiol.
41. La Scola B, Desnues C, Pagnier I, Robert C, Barrassi L, et al. (2008) The virophage as a unique parasite of the giant mimivirus. Nature 455: 100–104. doi:10.1038/nature07218.
42. Pritham EJ, Putliwala T, Feschotte C (2007) Mavericks, a novel class of giant transposable elements widespread in eukaryotes and related to DNA viruses. Gene 390: 3–17. doi:10.1016/j.gene.2006.08.008.
43. Krupovic M, Bamford DH, Koonin E V (2014) Conservation of major and minor jelly-roll capsid proteins in Polinton (Maverick) transposons suggests that they are bona fide viruses. Biol Direct 9: 6. doi:10.1186/1745-6150-9-6.
44. Blanc G, Gallot-Lavallée L, Maumus F (2015) Provirophages in the Bigelowiella genome bear testimony to past encounters with giant viruses. PNAS 112: E5318–E5326. doi:10.1073/pnas.1506469112.

Appendices

Appendix One: The origins of giant viruses, virophages and their relatives in host genomes

Abstract

Giant viruses have revealed a number of surprises that challenge conventions on what constitutes a virus. The Samba virus newly isolated in Brazil expands the known distribution of giant mimiviruses to a near-global scale. These viruses, together with the transposon-related virophages that infect them, pose a number of questions about their evolutionary origins that need to be considered in the light of the complex entanglement between host, virus and virophage genomes.

The discovery of giant viruses

Giant DNA viruses are double-stranded DNA (dsDNA) viruses that have particle and genome sizes comparable to those of small bacteria, and a number of features that are uncharacteristic of viruses. These include the presence of several genes that are similar to cellular genes such as those involved in DNA repair, translation, protein folding, and polysaccharide synthesis [1]. Acanthamoeba polyphaga mimivirus was the first of the giant DNA viruses to be discovered, initially isolated in the search for the causative agent of pneumonia during a hospital outbreak in Bradford, UK [1]. Since then, related viruses have been identified in a range of environments, including the discovery last year of the morphologically and genetically distinct pandoraviruses, which are even larger than the mimiviruses [2]. More recently, the 30,000-year-old Pithovirus sibericum was unearthed and brought back to life from Siberian permafrost [3].

Many of the giant viruses discovered to date have amoeba hosts and amoeba culture techniques have proved instrumental in identifying these giants, including the discovery last month of Samba virus, a wild mimivirus from the Amazonian Rio Negro [4]. Although slightly larger, Samba virus shares identity across the majority of its genome to the original Bradford mimivirus, further expanding the widespread distribution of these giant viruses. The defining feature of giant viruses is that they are an extreme outlier in terms of genome size: Acanthamoeba polyphaga mimivirus has a 1.2 Mb genome [1], which was double the size of the largest virus known at the time, and pandoravirus genomes reach up to 2.5 Mb [2]. Giant viruses are also extreme outliers in terms of their physical size, being too large to pass through porcelain filters, a criterion historically used to define a virus. As a further challenge to the traditional definition of viruses, giant viruses have several essential protein synthesis genes that have thus far been thought to be exclusive to cellular life [1].

Evolutionary origins of viral giantry

Determining the evolutionary relationships among viruses is crucial to investigating the origins of features such as their size, but is complicated by the absence of universally conserved viral genes. The Baltimore system classifies viruses according to genome type and replication strategy, therefore placing giant viruses among others with dsDNA genomes. They are also considered on the basis of distinguishing biological features to belong within the nucleocytoplasmic large DNA viruses (NCLDVs) alongside viral families such as poxviruses and iridoviruses [5]. While dsDNA viruses in general do not appear to have a single evolutionary origin, the NCLDVs all contain five core genes and tend to

share a suite of 50 or so likely ancestral genes [5] that partition them from other large eukaryotic dsDNA viruses such as nudiviruses, herpesviruses and baculoviruses. NCLDVs do share some genes with these other large DNA viruses, but are additionally distinguished by an either completely or largely cytoplasmic replication cycle [5].

Although the low levels of genetic similarity among NCLDVs complicate the precise phylogenetic placement of giant viruses, the relationships between NCLDV families have been reconstructed using multiple conserved genes [5]. The phylogenetic relationships of DNA polymerase genes from NCLDVs reveal that the Mimiviridae family, one of the larger members of which is the Samba virus [4], groups with the Marseilleviridae and Iridoviridae that are 1.2 Mb, 350 kbp and 200 kbp in size, respectively (Figure 1). Pandoraviruses are most closely related to Emiliana huxleyi virus, which has a genome that is 0.41 Mb in size. Thus, pandoraviruses are derived members of Phycodnaviridae and therefore phylogenetically distinct from the rest of the giant viruses [2,5]. While it has been argued that the large genomes of these giants suggest a large and complex ancestor, the most parsimonious interpretation of the phylogenetic evidence is that viral giantry evolved independently on at least two occasions from ancestors with much smaller genomes [5] (Figure 1).

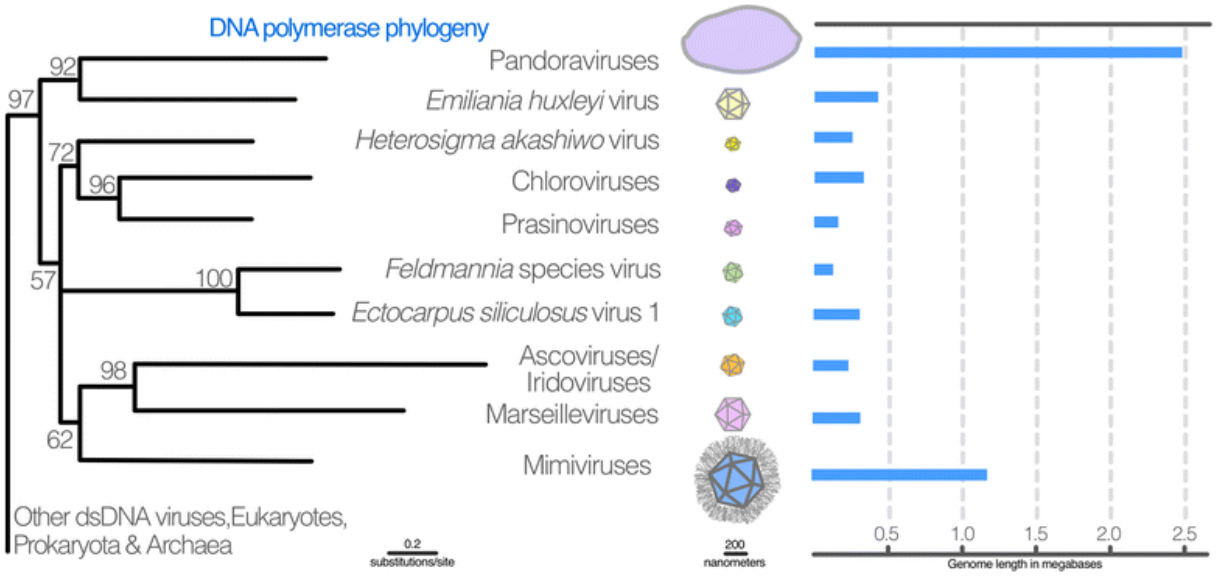


Figure 1 Stylized figure depicting the phylogenetic relationships, genome length and virion size of various nucleocytoplasmic large DNA viruses (NCLDVs).
The maximum likelihood tree shown is a simplified version of the NCLDV subtree for DNA polymerase adapted from [6], where collapsed clades in the original tree are represented by single branches. Numbers at each node are expected likelihood weights from 1,000 rearrangements. The scale throughout the figure is approximate.

The genome sizes of the NCLDVs vary greatly, from the 150 kb genomes of the poxviruses to the 2.5 Mb genomes of pandoraviruses [2,5]. This hints at the possibility that

viruses with intermediate genome sizes may exist. While many lineages may be extinct, it seems likely that at least some will eventually be found through metagenomic sampling. It would be premature to conclude that mimivirus and pandoravirus represent the largest DNA viruses that will be found. Interestingly, the recently uncovered pithoviruses are phylogenetically closer to the mimivirus/marseillevirus group despite a morphological resemblance to pandoraviruses that have ovoid rather than icosahedral morphology [3]. Moreover, of the more than 1,000 pandoravirus genes, 93% are previously unknown to biologists [2]. Many viruses contain some of these so-called orphan genes, but the high percentage of orphans in a single virus highlights how limited our sampling of the diversity of viral genes is.

Giant viruses are susceptible to viral infection by relatives of eukaryotic DNA transposons

Giant viruses reproduce in 'viral factories', which are cytoplasmic compartments of the host cell that can be as large as the nucleus (Figure 2). Accompanying the surprise of discovering giant viruses was the discovery of a group of associated viruses that are not capable of replication in their absence, and instead exploit the viral factory for replication. Other viruses are known to require helper viruses for replication, but these 'virophages' result in the formation of defective mimiviruses, implying a parasitic relationship [6]; they make the giant viruses ill. For example, the infection of Samba virus by its virophage results in a reduction of viral titer of over 80%, as well as partial recovery of the host amoebae [4]. This parasitism is therefore part of a complex relationship between giant viruses, their hosts and the virophages [6]. Virophage genomes are dwarfed by the mimiviruses that they infect, being approximately 20 kbp in length, and have been identified in association with several members of Mimiviridae [6,7]. There are now multiple strains of the first virophage, named Sputnik [6], including the Rio Negro isolate from the Samba virus [4] and more distantly related virophages such as Mavirus, identified from the Cafeteria roenbergensis virus (CroV) [7].

Virophages are related to a class of eukaryotic DNA transposons called Mavericks (or Polintons). These genomic parasites share a set of four core genes [7], as well as widespread conservation of the characteristically viral jelly-roll capsid [8]. Two of these core genes are present in virophages, indicating a close evolutionary relationship, and the Mavirus virophage in particular shares a total of seven homologs with Mavericks [7,8]. This indicates a much closer evolutionary relationship between Mavirus and Maverick transposons than Mavirus has with other virophages like Sputnik, with which it only shares four genes in total [7]. Mavericks are thought to derive from a DNA virus that integrated into the host genome, and the discovery of Mavirus strongly suggests a virophage-like progenitor to Mavericks. Several features indicate that the Mavirus ancestor was also a

virophage rather than an escaped Maverick-like transposon [7], although this has been debated [8]. One such notable feature is the dependence on CroV for replication, as indicated by the high similarity of Mavirus promoters to those of CroV [7]. It is hard to imagine how a DNA transposon's replication strategy would evolve to rely on CroV infection, whereas the post-integration loss of this feature in Mavericks is more readily explained [7]. Interestingly, some Maverick elements are more closely related to some virophages than to other Mavericks [8], suggesting that these integrations are a recurring event. The fact that Mavericks are widespread in the animal kingdom indicates that a number of virophages, and therefore NCLDVs, are yet to be discovered in association with these hosts.

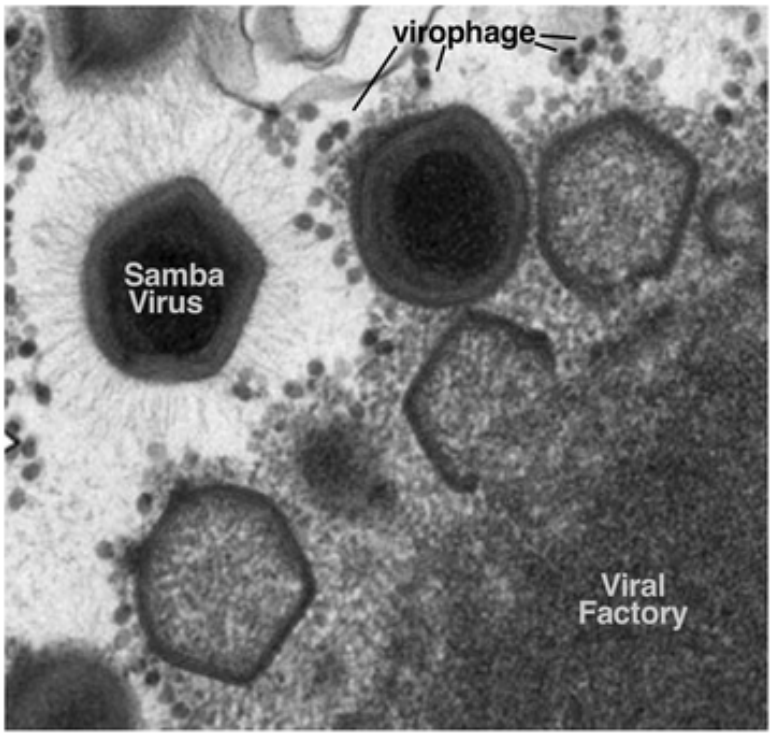


Figure 2. Scanning electron microscope image of Samba virus and Rio Negro virophage adapted from [4].
The figure depicts Samba virus morphogenesis within the viral factory that is formed in Acanthamoeba castellanii, as well as the parasitic virophage particles interspersed among the giant virions.

A question of gene flow and its evolutionary consequences

Gene flow has played a central role in the evolutionary history of virophages. Integrated virophages have been found in a mimivirus genome, and virophage genes also share similarity to genes in other DNA transposons, such as a class of linear plasmids called transpovirons that are also found in mimiviruses [9]. Some virophage genes also show similarity to bacteriophages, cellular genes, and their respective viral hosts [7]. This compound nature of virophage genomes is evidence of extensive horizontal gene transfer, and although the precise details of this gene flow are not fully understood, perspectives from paleovirology - the study of viral remnants, or 'fossils', found in host genomes - may help to clarify them. Analysis of these viral remnants, known as endogenous viral elements (EVEs), has revealed that all viruses can in principle integrate in a heritable fashion into the host genome, thus preserving information from the distant evolutionary

past [10]. Mimivirus EVEs have not been found, and one might suspect that their extraordinarily large genomes mean that they are unlikely to form EVEs. However, we could consider a virophage EVE to exist in the form of Mavericks; in some sense, a mimivirus that donates genes to a subsequently endogenized virophage could be thought of as a 'vicarious EVE'. This flow of genes, from mimivirus to virophage to host genome, is therefore evident in the amoeba genome.

One proposal to explain the endogenization of virophages is that it could have been positively selected for, since the association with a virophage is beneficial to the host cell (owing to its interference with the replication of the large DNA virus) [7]. The survival advantage gained by an integrated virophage could conceivably be the production of virophages as a kind of antiviral response. If the viral threat were lost, then selection for the maintenance of virophage production would be relaxed, explaining the proposed loss of virophage features in Mavericks [7]. Because only a handful of virophages have been described, there are insufficient data to investigate the evolutionary dynamics at play. With improvements in sequencing, bioinformatics and metagenomics, viral discovery is increasing exponentially, and moving beyond studies of immediate medical and economic interest. These advances will generate more data that will be suitable for the study of these evolutionary dynamics.

The entanglement of giant viruses, virophages and their hosts

The discovery of giant viruses has crossed some of the boundaries between viruses and cellular life, although ribosomes remain a distinguishing feature. The conflict between giant viruses and their hosts, with the former also infected by virophages, alongside genomic invasions with related transposons, is reminiscent of Darwin's tangled bank, recapitulated at the microscopic scale in a droplet of water. Elucidating the role of gene flow between these microscopic entities will reveal their evolutionary dynamics and aspects of the origins of viruses and cellular life.

References

1. Raoult D, Audic S, Robert C, Abergel C (2004) The 1.2-megabase genome sequence of Mimivirus. Science (80-).
2. Philippe N, Legendre M, Doutre G, Couté Y (2013) Pandoraviruses: amoeba viruses with genomes up to 2.5 Mb reaching that of parasitic eukaryotes. Science (80-).
3. Legendre M, Bartoli J (2014) Thirty-thousand-year-old distant relative of giant icosahedral DNA viruses with a pandoravirus morphology. Proc ….
4. Campos R, Boratto P (2014) Samba virus: a novel mimivirus from a giant rain forest, the Brazilian Amazon. Virol J.
5. Yutin N, Koonin E (2013) Pandoraviruses are highly derived phycodnaviruses. DNA.
6. Scola B La, Desnues C, Pagnier I, Robert C (2008) The virophage as a unique parasite of the giant

mimivirus. Nature.
7. Fischer M, Suttle C (2011) A virophage at the origin of large DNA transposons. Science (80-) 332: 231–234.
8. Yutin N, Raoult D, Koonin E V (2013) Virophages, polintons, and transpovirons: a complex evolutionary network of diverse selfish genetic elements with different reproduction strategies. Virol J 10: 158. doi:10.1186/1743-422X-10-158.
9. Krupovic M, Bamford DH, Koonin E V (2014) Conservation of major and minor jelly-roll capsid proteins in Polinton (Maverick) transposons suggests that they are bona fide viruses. Biol Direct 9: 6. doi:10.1186/1745-6150-9-6.
10. Katzourakis A (2013) Paleovirology: inferring viral evolution from host genome sequence data. Phil Trans R Soc B 368: 20120493 – . doi:10.1098/rstb.2012.0493.

Appendix Two: The Endogenisation of Retroviruses and their Lasting Influence on Host Evolution

Abstract

Retroviral infection of the host germ line can lead to vertical inheritance of the integrated viral genome. These endogenous retroviruses (ERVs) occupy a significant fraction of vertebrate genomes and can represent the remnants of their exogenous counterparts. ERVs can also be used to study retroviral evolution and aspects of host-virus coevolution that shaped their respective natural histories. Furthermore, ERVs themselves affect host biology at various timescales, both in terms of their direct effect on the genome and the emergence of new viruses, as well as influencing broad patterns of host evolution.

Introduction

Viruses are obligate intracellular parasites that replicate by hijacking their host's molecular machinery. Although the viral life cycle ordinarily takes place in somatic cells, germ line infections can sometimes occur and potentially lead to the inheritance of the virus. Such vertically transmitted viruses are called endogenous viral elements (EVEs; Katzourakis & Gifford 2010) and they can occasionally reach population fixation and persist in the host for millions of years. EVEs contain information about ancient infections that would otherwise not be observable, and this has earned them the metaphorical description of a genomic 'fossil record'. The vast majority of EVEs originate from retroviruses because their life cycle involves obligate genome integration (Figure 1a), and this subset of EVEs are usually referred to separately as endogenous retroviruses (ERVs). The increasing availability of host genomes has revealed the vastness of the viral fossil record, and led to the emergence of paleovirology – a field dedicated to the study of EVEs. While non-retroviral EVEs have only recently been described, ERVs have contributed to studies of the evolutionary history of retroviruses for decades, and have facilitated investigations into the complex evolutionary relationship between retroviruses and their hosts [3,4]

The evolution of retroviruses and their hosts has been shaped by a series of adaptations and counter-adaptations. Adaptations of the host are driven by the selective pressure to restrict viral infections, which in turn inflict pressure on viruses to counter-adapt in order to evade host restriction [5]. Such adaptive changes will recur on both co-evolving sides, predominantly manifesting at the molecular interface between host immunity and viral countermeasures. The consequent evolutionary 'arms-race' can be also observed at various timeframes; from host-virus interactions shortly after the first infection of a host cell, to adaptations in host immunity at deep evolutionary timescales. The detrimental effect of viral integration to their host can be assessed at short timescales, such as the course of a single infection or at the host-population level over a few years or decades. We can also examine the host-virus relationship over evolutionary timescales using ERVs, and we can investigate the lasting effects of the arms race on antiviral genes [6]. The different timescales at which these interactions can be studied, as well as their various consequences can be considered using a paleovirological framework.

This chapter intends to consider examples from the literature that are a cross section of studies on the host-virus relationship as they relate to paleovirology. After a brief description of ERV classification, we discuss the process of ERV formation and the possible events shortly after viral endogenisation. We then review a variety of different

effects that ERVs can have on host biology, before considering how they can be useful tools in the study of both viral evolution and the evolution of host immunity. Finally, we examine how ERV research has begun to tackle much broader questions about the evolutionary consequences of the virus-host arms race that will eventually lead to a better understanding of the ERVs, viruses and their respective pathogenicity.

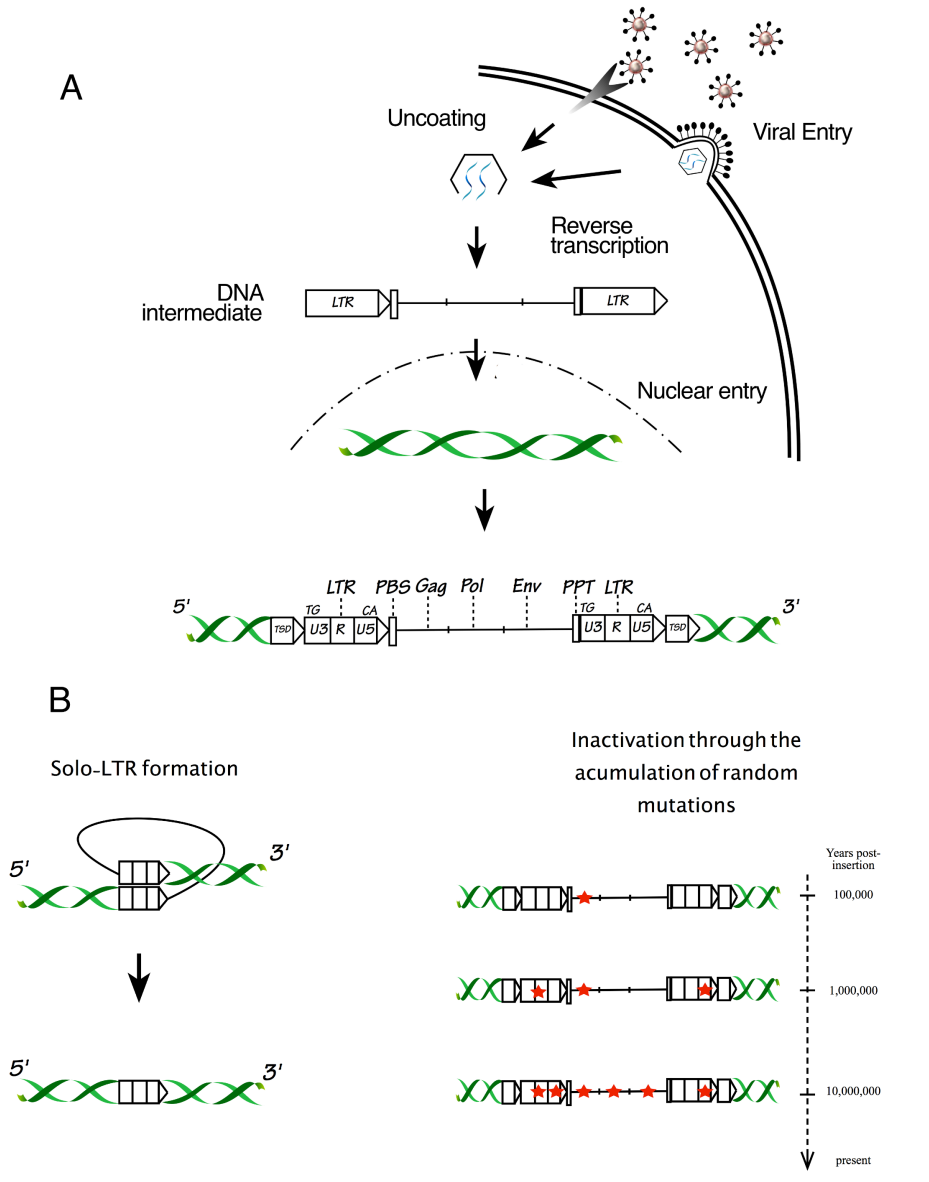


Figure 1. The process of endogenisation and inactivation of ERVs.
Panel A: The diagram depicts a stylised cell infected by retroviruses. The virion achieves cell entry by directly fusing to the cell membrane or via receptor-mediated endocytosis. Once inside the cell, the RNA genome is released from the capsid and is converted to double stranded DNA by reverse transcriptase. This DNA intermediate then enters the nucleus where insertion into the host genome is catalysed by integrase. This process results in the formation of characteristic 'target site duplications' either side of the provirus that are usually 4-6 nucleotides in length. **Panel B:** The two diagrams shown represent the main mechanisms of ERV deletion or inactivation. Because the long terminal repeats that flank retroviral genomes are identical at the time of integration, accidental recombination deletion of the ERV can occur resulting in 'solo-LTR' formation. Alternatively, the accumulation of inactivating non-sense mutations over time will render an ERV incapable of producing viral particles or achieving retro-transposition.

Classification and identification of ERVs

Historically, the classification of ERVs has been complicated by their relationship to both retrotransposons and exogenous retroviruses, both of which have established and distinct classification schemes. Efforts towards a unified system for ERV classification have encountered obstacles when results that use one naming scheme are compared to studies that use another [7,8]. Early systems attempted to classify ERVs based on their primer-binding site. However, numerous ERVs use the same PBS, and with the continuing characterization of novel families, it soon became clear that this scheme could not serve as a unified classification system. Contemporary classification of ERVs is based on reverse transcriptase phylogenies that include endogenous sequences and their exogenous relatives, but maintain widely accepted names based on older schemes. Current exogenous retrovirus taxonomy follows conventions set by the International Committee on Taxonomy of Viruses (ICTV). The family *Retroviridae* is subdivided into two subfamilies, the Orthoretrovirinae and Spumaretrovirinae with the genera Alpharetrovirus, Betaretrovirus, Gammaretrovirus, Deltaretrovirus, Epsilonretrovirus and Lentivirus being members of Orthoretrovirinae and the genus Spumavirus of Spumavirinae. ERVs are classified based on relatedness to exogenous retroviruses; Gammaretroviruses are designated as Class I, while Class II and III ERVs are related to betaretroviruses and spuma-like retroviruses, respectively. Although this phylogenetically informed approach has proved capable of classifying highly degraded ERVs that are distantly related to exogenous retroviruses, the three classes have few distinguishing features, posing limitations on their utility. Recent discoveries place certain ERVs firmly within exogenous genera such as lentiviruses, while certain class I ERVs are in fact more closely related to the epsilonretroviruses than the gammaretroviruses [9]. Furthermore, retroviruses undergo complex patterns of recombination, posing further limitations on efforts to classify them based on a single gene.

From exogenous to endogenous: how an ERV is formed

Germline integration and fixation

ERVs typically comprise 5-10% of vertebrate genomes, and the history of these elements can be understood by considering the evolutionary forces at play, beginning with germline invasion of the host by an exogenous circulating retrovirus. After heritable integration, an ERV will be subjected to genetic drift and its frequency in the population may change by chance due to random sampling. ERVs with a fitness cost will be negatively selected out of the population, and a small proportion of neutral (or nearly neutral) ERVs may

eventually reach fixation. In rare instances where an ERV offers a fitness advantage, its population frequency will increase rapidly.

Intragenomic Proliferation

ERVs will proliferate in the host genome if a viable infectious particle can be produced, resulting in the reinfection of germline cells. This will continue until they are either inactivated by nonsense mutations, or are removed as a result of recombination between their long terminal repeats (LTRs) (Figure1b; Stoye 2001). ERVs that are incapable of producing virions (e.g. because of the inactivation of one of its genes) can nonetheless proliferate by transposition. Intracellular retrotransposition can occur in *cis* or *trans* where either an intact ERV or other viruses provide the enzymatic activity, respectively (Belshaw et al. 2004). ERV proliferation will lead to the emergence of phylogenetically distinct groups called families that are defined as monophyletic lineages derived from distinct germ line invasions [13]. Once inactivated however, ERVs succumb to the host evolutionary rate that is much slower than that of the exogenous virus. Whereas the inactivation of a single ERV usually occurs within a few thousand years, the fossilization of an ERV family can presumably take several millions of years, and have substantially different evolutionary dynamics.

One way to determine the proliferation mechanism that dominated the early history of an ERV family is by interpreting the patterns of selection that can be detected in their sequences. For example, evidence of post-integration purifying selection in the *env* gene would suggest that the ERV family multiplied by reinfection, since the envelope is a necessary component of mature virions. For human ERV (HERV) families with less than 200 copies, almost all were found to have proliferated by re-infection [14]. However, loss of the *env* gene shifts the mode of ERV amplification from reinfection to retrotransposition, leading to far greater proliferation [15]. This is possibly because of the fitness costs to the host associated with reinfection or the expression of *env,* which in turn also determines the survival of the ERV [15].

The influence of ERVs on hosts

The effect of ERVs on genome architecture

ERV integrations affect the host in a number of ways that can be selectively neutral, detrimental or in rare cases advantageous to the host. We can assess the post-integration effects on genome architecture by considering the potential selective pressure they impose on their hosts based on the site of integration. For example, ERV insertions in

genes or other functional regions are rapidly purged from the host population since they usually interrupt essential functions [16]. Integration is likely to be selectively neutral if it occurs in a non-coding region, assuming also that it does not interfere with transcription and can no longer produce infectious particles. However, multiple apparently harmless integrations can have non-neutral downstream consequences by providing recombination hotspots that can significantly alter genome architecture. This is because non-allelic homologous recombination (NAHR) between members of the same ERV family can result in genomic re-arrangements such as duplication, deletion, translocation or inversion [3].

In humans, the HERV-K family in particular has undergone extensive recombination since integration (at least 16% of full-length copies) [17]. Moreover, some HERV-related genomic rearrangements have been associated with pathogenesis. For example, recent studies of the human genome have shown that structural variation may result from NAHR between HERV elements rich in recombination hotspot motifs [18]. For instance, HERV-mediated NAHR has been reported to lead to the deletion of the *eyes absent 1* (EYA1) locus, which can cause branchio-oto-renal syndrome (Sanchez-Valle et al. 2010). Another example is the deletion of the azoospermia factor gene (AZFA or AZF1). The consequent male infertility is the result of a non-reciprocal recombination event of two HERV-I loci on the Y chromosome [21,22].

The effect of ERVs on gene expression

In addition to gene disruption, ERVs can also affect genome's integrity by influencing their transcriptional regulation. Retroviral insertions near a host gene can modify its regulation by offering alternative promoters or enhancers during transcription. This is because retroviral LTRs contain a promoter for RNA polymerase II along with enhancers and transcription factor binding sites. In humans for example, salivary amylase expression was found to be associated with an ERV integration in a tissue-specific manner [23]. The domestication of an LTR has also been found to promote the expression of neuronal apoptosis inhibitory protein [24]. If inserted into introns, ERVs may promote changes in gene transcription by introducing alternative termination sites or novel splice donor and acceptor sites. Such an example is the *slc15a2* locus in mice, where an ERV integration was shown to substantially decrease its transcription as a result of truncated transcripts [25].

ERVs, recombination and the emergence of new viral types

The impact of ERVs extends beyond the direct effect of their integration on host genomes. Indirect consequences include the interaction of ERVs with exogenous viruses or other ERVs that can give rise to novel recombinant viral types. ERVs can be a source of emerging infection if recombination with exogenous viruses leads to the acquisition of novel pathogenicity [26]. This can happen through the replacement of the *env* gene with one of different viral origin, and since *env* plays a key role in determining viral tropism, can facilitate cross-species transmission of the virus. A typical example of *env* acquisition that led to a new viral type is the Feline Leukemia Virus type B (FeLV-B), which emerged after the endogenised FeLV sequences acquired an *env* gene by the exogenous circulating FeLV-A viruses [27,28].

Even more complex patterns of recombination can result from a series of sequence exchanges between viruses have also been described. The RD-114 virus is a product of recombination between the *gag-pol* region of the *Felis catus* endogenous retrovirus (FcEV; gammaretrovirus) and the *env* locus of baboon endogenous retrovirus (BaeV; gammaretrovirus; [29]). BaEV is in turn a chimeric virus formed by the *gag-pol* region of *Papio cynocephalus* endogenous retrovirus (PcEV; gammaretrovirus) and the *env* of a beta-like retrovirus. Furthermore, PcEV is the product of a recombination event between the *gag-pol* region of a betaretrovirus and the *env* region of a gammaretrovirus [29]. RD-114 is therefore the result of recombination between four distinct retroviruses from two different families. Members of the deltaretrovirus genus perhaps represent an entire clade of recombinant viruses in that their *env* gene appears to have been derived from gammaretroviruses [30]. Even though they are phylogenetically distant, gamma- and delta- *env* genes exhibit identical modular organization and they seem to utilize the same receptors in order to achieve cell entry [30].

Recombination between two ERVs may also repair inactivating mutations and give rise to viral sequences with novel pathogenic characteristics, which in turn can lead to new disease phenotypes. Porcine endogenous retroviruses have been shown to spontaneously recombine resulting in ERVs that are capable of infecting human cell-lines, which has serious implications for pig-to-human xenotransplantation [31]. Furthermore, *in vivo* models have linked the emergence of recombinant viruses with potential pathologies, such as that of retrovirus-induced cancers [32]. For example, it has been shown that infectious Murine Leukemia Viruses (MLVs) are produced by recombination between

replication-incompetent ecotropic and non-ecotropic endogenous MLVs in antibody-deficient mice [32].

Occasional beneficial effects of ERVs on hosts

ERVs can have a detrimental effect on their hosts by directly altering either genome architecture or gene expression, or indirectly by contributing to the emergence of new viral types through recombination. However, the flow of viral genomes into that of their hosts can also generate genetic novelty. Rather than being detrimental, ERV-mediated changes to expression can have a positive influence on the host. For example, the antiviral restriction factor APOBEC3G was found to be up regulated by an LTR insertion [33]. As well as the beneficial co-option of LTRs, viral genes can also be exapted by the host, and while this is extremely rare, such examples have been described in literature. Domesticated ERV genes can perform important physiological functions - *syncytin*s being the most notable example. *Syncytins* are *env*-derived genes that contribute to the development of the placenta. Specifically, they facilitate the cell-cell fusion of the syncytial cell layer in the syncytiotrophoblast. Evidence shows that *syncytins* were the result of eight independent domestication events in 6 mammalian orders [34–36].

As well as syncytins, a number of ERV genes have been captured as part of the evolutionary arms races and repurposed to fulfill an antiviral role. A number of these ERV-derived immunity (EDI) genes have been recognized and function by a range of mechanisms (Aswad & Katzourakis 2012). Among them is the Friend-virus-susceptibility-1 (*Fv1*) restriction factor, which has been identified in mice and originates from the *gag* region of an endogenous retrovirus [38]. *Fv1* restricts retroviral infections in the early stages of viral entry by interfering with the viral capsid [38]. As well as *Fv1* other EDIs that prevent viral infections originate from the *env* gene. For example, *FV4* is a defective *env* in mice which when expressed results in the occupation of cell-entry receptors, thereby preventing the binding of incoming viruses [39].

The impact of ERVs on their host can be studied based on the nature of the interaction, i.e the direct effects of integration or the indirect consequences such as their influence on emerging infection. The outcomes of integration can also be categorized based on their selective effect, and both of these consequences of integration can be examined across the different evolutionary timescales at which they occur. For example, the consequences of a retroviral insertion into a host gene will take effect almost instantly, whereas the co-option of a viral gene is likely to take time before the selective pressure from a viral threat

drives the emergence of a novel function. The timescale at which intragenomic proliferation occurs is more continuous, and is likely to be ongoing for millions of years, which in turn will influence the more discrete processes that occur at short and long timescales.

ERVs as tools in the study of the host-virus relationship

ERVs can also be used to study the evolutionary history of host-virus arms races by identifying their influence on both viruses and host immunity. Arms races are a distinct form of natural selection in that the selective pressure on viruses and hosts is highly recurrent (Daugherty & Malik 2012). Moreover, because viruses evolve more rapidly than their hosts, they have an opportunity to quickly evade antiviral defenses in the cell. Stretched over evolutionary timescales, the evidence of this rapid adaptation will erode from the exogenous viral sequences. However, evidence of past evolutionary arms races can be detected in ERV sequences, as can the selective pressure past viruses may have imposed on restriction factors such as APOBEC, TRIM, SAMHD1 and Tetherin. In primates, APOBEC3G-derived mutations have been found in members of the HERV-K (HML2) family, indicating that the hypermutation-induced restriction of retroviruses by APOBEC3G had evolved long ago [40]. This is consistent with the evidence of positive selection on APOBEC3G genes that has persisted for millions of years (Sawyer, Emerman, and Malik, 2004).

Different host lineages will have been subjected to infection by different viruses at different times, and we can consider the variable effect this could have had on their respective restriction factors by examining the signatures of selection in their sequences. TRIM5 inhibits retroviral entry by recognizing the capsid protein and interfering with uncoating (Stremlau et al. 2006). Selection analysis of TRIM5 and other TRIMs such as TRIM22 across multiple primate lineages revealed that positive selection has mainly acted on either TRIM5 or TRIM22 in a given lineage, but normally not on both [43]. A possible explanation for the mutually exclusive selection on TRIM5 and TRIM22 could be that they target different viral strains. This changing selection pressure can also result in the complete loss of a restriction factor. For instance in carnivores, there were at least two independent TRIM5 inactivation events in canines (gene disruption; Sawyer et al. 2007) and felines (gene truncation; McEwan et al. 2009), which is evidence for an unknown past viral threat that had been eradicated without replacement. These findings are consistent with the observation that dogs have an uncharacteristically low ERV copy number [45].

Together, these studies suggest that the host repertoire of anti viral defense genes has a high turnover and is shaped by the specific viral threat(s) at a given time.

The inferences we can make about past viruses using only the information imprinted on host genes is indirect evidence of ancient host-virus arms races. Indeed, such approaches have been dubbed 'indirect paleovirology' [37,46] and allow us to easily and cheaply investigate ancient virus-host arms races retrospectively. Fascinatingly however, ancient molecular battles can be studied by creating chimeric viruses that are capable of infection and contain ERV sequences *in vitro.* For example, the capacity of APOBEC3-A, -B, -F and -G to induce hypermutations in HERVK have been demonstrated in cell cultures infected with artificially reconstructed viruses [47]. Similarly, functional envelopes of endogenous chimpanzee retroviruses 1 and 2 (CERV) were reconstructed to investigate the host range of this now extinct retrovirus, revealing the specific entry receptor it uses as CTR1 [48]. Interestingly, while no copies of CERV can be found in the human lineage, the resurrected CERV was nonetheless capable of infecting human cells, which raises questions about the barrier to endogenisation in humans. A different study that investigate the absence of CERV in humans determined that CERV was insensitive to TRIM5 inhibition, ruling out another explanation for its absence [49]. *Tetherin* is another antiviral factor that was initially characterized as an inhibitor of HIV1, where it prevents release of newly formed virions by 'tethering' them to the cell surface [50]. It has been shown that *tetherin* has a broad viral activity, with the ability to restrict many enveloped viruses, including HERVK [51]. Interestingly however, the most active family of HERVs (HERVK-HML2), were shown to inhibit *tetherin* activity *in vitro* [52]. Furthermore, this anti-*tetherin* activity was also found against the *tetherins* of other primates that also possess HERVK-HML2, suggesting that the property was advantageous to its proliferation in the lineage.

ERVs in the study of viruses and virus-host interaction at evolutionary timescales

The study of ERVs goes beyond that of their causes and consequences; they can be used to understand the evolutionary history of viruses. ERVs are direct evidence of past infections that allow us to reach further back into the evolutionary history of viruses than is possible with contemporary data alone, thereby enabling us to characterize host-virus interactions at evolutionary timescales. Not only can this change an estimate of the age of a viral group, but it can also reveal previously unrecognized hosts. For example, foamy viruses are a group of complex retroviruses that infect a range of mammal hosts, and the identification of an endogenous foamy virus in the two-toed sloth (*Choloepus hoffmanni)*

expanded their host range to include Xenarthrans. When combined with both biogeographic data and phylogenetic evidence of coevolution, this finding suggested that the close association between foamy viruses and their hosts had been maintained for over 100 million years [53]. Foamy virus ERVs that confirm this timeframe were later identified in a number of other species, including the aye-aye (*Daubentonia madagascariensis*) and Cape golden mole (*Chrysochloris asiatica*) [54,55].

Another clear example of what ERVs can reveal about deep virus evolution was demonstrated for lentiviruses, which include HIV and were thought to be only thousands of years old based on molecular clock dating, or at most a few million years old if inferring from co-speciation with their hosts [56]. However, the identification of an endogenous lentivirus (RELIK) in the genome of rabbits allowed us to confidently estimate a minimum age of ~7 million years based on LTR-dating [57]. This date was pushed back even further to 12 million years with the discovery of orthologs in the European hare and weasel family (Keckesova et al. 2009; G.-Z. Han & Worobey 2012). These studies were not only surprising in terms of how old we thought lentiviruses were, but also revealed a much broader host range (at least in the past), which could have implications for the identification of future reservoir species. Furthermore, analysis of RELIK revealed a much simpler genome that lacked contemporary features such as the HIV *vif* accessory gene [57] that defend the virus against host restriction factors [60].

Investigating the lasting effect of antiviral factors on ERVs and demonstrating their efficacy *in vitro* has allowed us to understand the long term dynamics of virus host arms races, while studies of ancient foamy viruses and lentiviruses using ERVs have revealed their date of origin and other aspects of evolution. Furthermore, in addition to identifying previously unknown hosts, we can use ERVs to study the *changes* in host ranges in such large timescale studies of virus-host coevolution. In the case of foamy viruses for example, discordance between the topology of host and viral phylogenies indicates at least two interspecies transfers in their evolutionary history [55]. In gammaretroviruses, a major spillover has been revealed by the phylogenetic reconstruction of Murine Leukemia Virus (MLV)-related viruses in vertebrates, indicating that certain mammalian viral strains emerged from a bird-to-mammal cross-species transmission [61]. Furthermore, phylogenetic analysis of endogenous betaretroviruses in bats indicates that interspecies transfer is particularly prevalent, and it has been suggested that this transmissibility could be the result of adapting to evade host defenses that also act as interspecies barriers [62].

Whether we investigate their obvious pathological consequences, or use them to reveal ancient evolutionary history, ERVs enhance our understanding of viral biology and virus-host interaction that can in turn be applied to clinical intervention and epidemiological strategies. As well as their direct consequences on host genomes and effects on host immunity, we are beginning to build a broader picture of the extent to which ERVs have influenced host evolution. For example, the recent finding that the genomes of larger mammals harbor fewer ERVs raises questions about the mechanisms employed to control their endogenisation and proliferation [63]. This observation has been linked to Peto's paradox, which is that larger mammals do not exhibit higher rates of cancer despite a greater risk due to their size. It could be that the same mechanisms that control cancer emergence in larger hosts are also responsible for controlling ERVs, which themselves impose a tumorigenic risk [63]. Answers to such fundamental questions are now being sought through the study of ERVs, which also continue to inform us about the evolutionary processes of viruses and their interaction with host immunity. In parallel to the continued effort to understand the influence of ERVs on host genomes, pathogenesis and broad evolutionary patterns, the study of non-retroviral endogenous viruses are rapidly opening new avenues of enquiry [3,37,64]. As more non-retroviral EVEs are described, we will be able to both re-apply curiosities that emerged out of the study of retrovirus, as well as ask new questions that are unique to each non-retroviral group.

References

1. Katzourakis A, Gifford RRJ (2010) Endogenous viral elements in animal genomes. PLoS Genet 6: e1001191.
2. Holmes EC (2011) The Evolution of Endogenous Viral Elements. Cell Host Microbe 10: 368–377. doi:10.1016/j.chom.2011.09.002.
3. Feschotte C, Gilbert C (2012) Endogenous viruses: insights into viral evolution and impact on host biology. Nat Rev Genet 13: 283–296. doi:10.1038/nrg3199.
4. Weiss R a, Stoye JP (2013) Virology. Our viral inheritance. Science 340: 820–821. doi:10.1126/science.1235148.
5. Daugherty MD, Malik HS (2012) Rules of engagement: molecular insights from host-virus arms races. Annu Rev Genet 46: 677–700. doi:10.1146/annurev-genet-110711-155522.
6. Duggal NK, Emerman M (2012) Evolutionary conflicts between viruses and restriction factors shape immunity. Nat Rev Immunol 12: 687–695. doi:10.1038/nri3295.
7. Bannert N, Kurth R (2006) The evolutionary dynamics of human endogenous retroviral families. Annu Rev Genomics Hum Genet 7: 149–173. doi:10.1146/annurev.genom.7.080505.115700.
8. Blomberg J, Benachenhou F, Blikstad V, Sperber G, Mayer J (2009) Classification and nomenclature of endogenous retroviral sequences (ERVs): problems and recommendations. Gene 448: 115–123. doi:10.1016/j.gene.2009.06.007.
9. Katzourakis A, Tristem M (2005) Phylogeny of human endogenous and exogenous retroviruses. In: Sverdlov ED, editor. Retroviruses and Primate Genome Evolution. Georgetown, Texas: LANDES BIOSCIENCE. pp. 186–200.
10. Stoye JP (2001) Endogenous retroviruses: still active after all these years? Curr Biol 11: R914–R916.
11. Belshaw R, Pereira V, Katzourakis A, Talbot G, Paces J, et al. (2004) Long-term reinfection of the human genome by endogenous retroviruses. PNAS 101: 4894–4899. doi:10.1073/pnas.0307800101.
12. Katzourakis A, Rambaut A, Pybus OG (2005) The evolutionary dynamics of endogenous retroviruses. Trends Microbiol 13: 463–468. doi:10.1016/j.tim.2005.08.004.
13. 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. doi:10.1128/JVI.74.8.3715-3730.2000.
14. Belshaw R, Katzourakis A, Paces J, Burt A, Tristem M (2005) High copy number in human endogenous retrovirus families is associated with copying mechanisms in addition to reinfection. Mol Biol Evol 22: 814–817. doi:10.1093/molbev/msi088.
15. Magiorkinis G, Gifford RJ, Katzourakis A, De Ranter J, Belshaw R (2012) Env-less endogenous retroviruses are genomic superspreaders. PNAS 109: 7385–7390. doi:10.1073/pnas.1200913109.
16. Nellåker C, Keane TM, Yalcin B, Wong K, Agam A, et al. (2012) The genomic landscape shaped by selection on transposable elements across 18 mouse strains. Genome Biol 13: R45. doi:10.1186/gb-2012-13-6-r45.
17. Hughes JF, Coffin JM (2001) Evidence for genomic rearrangements mediated by human endogenous retroviruses during primate evolution. Nat Genet 29: 487–489.
18. Campbell IM, Gambin T, Dittwald P, Beck CR, Shuvarikov A, et al. (2014) Human endogenous retroviral elements promote genome instability via nonallelic homologous recombination. BMC Biol 12: 74. doi:10.1186/s12915-014-0074-4.
19. Sanchez-Valle A, Wang X, Potocki L, Xia Z, Kang S-HL, et al. (2010) HERV-mediated genomic rearrangement of EYA1 in an individual with branchio-oto-renal syndrome. Am J Med Genet A 152A: 2854–2860. doi:10.1002/ajmg.a.33686.
20. Hermetz KE, Surti U, Cody JD, Rudd MK (2012) A recurrent translocation is mediated by homologous recombination between HERV-H elements. Mol Cytogenet 5: 6. doi:10.1186/1755-8166-5-6.
21. Kamp C, Hirschmann P, Voss H, Huellen K, Vogt PH (2000) Two long homologous retroviral sequence blocks in proximal Yq11 cause AZFa microdeletions as a result of intrachromosomal recombination events. Hum Mol Genet 9: 2563–2572.
22. Sun C, Skaletsky H, Rozen S, Gromoll J, Nieschlag E, et al. (2000) Deletion of azoospermia factor a (AZFa) region of human Y chromosome caused by recombination between HERV15 proviruses. Hum Mol Genet 9: 2291–2296.
23. Ting CN, Rosenberg MP, Snow CM, Samuelson LC, Meisler MH (1992) Endogenous retroviral sequences are required for tissue-specific expression of a human salivary amylase gene. Genes Dev 6: 1457–1465. doi:10.1101/gad.6.8.1457.
24. Romanish MT, Lock WM, van de Lagemaat LN, Dunn C a, Mager DL (2007) Repeated recruitment of LTR retrotransposons as promoters by the anti-apoptotic locus NAIP during mammalian evolution. PLoS Genet 3: e10. doi:10.1371/journal.pgen.0030010.
25. Li J, Akagi K, Hu Y, Trivett AL, Hlynialuk CJW, et al. (2012) Mouse endogenous retroviruses can trigger premature transcriptional termination at a distance. Genome Res 22: 870–884. doi:10.1101/gr.130740.111.
26. Sheets RL, Pandey R, Jen W, Roy-Burman P (1993) Recombinant Feline Leukemia Virus Genes Detected in Naturally Occurring Feline Lymphosarcomas. J Virol 67.
27. Overbaugh J, Riedel N, Hoover EA, Mullins JI (1988) Transduction of endogenous envelope genes by feline leukemia virus in vitro. Nature 332.
28. Stewart MA, Warnock M, Wheeler A, Wilkie N, Mullins JI, et al. (1986) Nucleotide Sequences of a Feline Leukemia Virus Subgroup A Envelope Gene and Long Terminal Repeat and Evidence for the Recombinational Origin of Subgroup B Viruses. J Virol 58: 825–834.
29. van der Kuyl a C, Dekker JT, Goudsmit J (1999) Discovery of a new endogenous type C retrovirus (FcEV) in cats: evidence for RD-114 being an FcEV(Gag-Pol)/baboon endogenous virus BaEV(Env) recombinant. J Virol 73: 7994–8002.
30. Manel N, Battini J-L, Taylor N, Sitbon M (2005) HTLV-1 tropism and envelope receptor. Oncogene 24: 6016–6025. doi:10.1038/sj.onc.1208972.
31. Bartosch B, Stefanidis D, Myers R, Patience C, Takeuchi Y, et al. (2004) Evidence and Consequence of Porcine Endogenous Retrovirus Recombination. J Virol 78. doi:10.1128/JVI.78.24.13880.
32. Young GR, Eksmond U, Salcedo R, Alexopoulou L, Stoye JP, et al. (2012) Resurrection of endogenous retroviruses in antibody-deficient mice. Nature 491: 774–778. doi:10.1038/nature11599.
33. Sanville B, Dolan MA, Wollenberg K, Yan Y, Martin C, et al. (2010) Adaptive evolution of Mus Apobec3 includes retroviral insertion and positive selection at two clusters of residues flanking the substrate groove. PLoS Pathog 6: e1000974. doi:10.1371/journal.ppat.1000974.
34. Mi S, Lee X, Li X, Veldman GM, Finnerty H, et al. (2000) Syncytin is a captive retroviral envelope protein involved in human placental morphogenesis. Nature 403: 785–789. doi:10.1038/35001608.
35. Lavialle C, Cornelis G, Dupressoir A, Esnault C, Heidmann O, et al. (2013) Paleovirology of “syncytins”, retroviral env genes exapted for a role in placentation. Philos Trans R Soc Lond B Biol Sci. doi:10.1098/rstb.2012.0507.
36. Cornelis G, Vernochet C, Malicorne S, Souquere S, Tzika AC, et al. (2014) Retroviral envelope syncytin capture in an ancestrally diverged mammalian clade for placentation in the primitive Afrotherian tenrecs. PNAS 111: E4332–E4341. doi:10.1073/pnas.1412268111.
37. Aswad A, Katzourakis A (2012) Paleovirology and virally derived immunity. Trends Ecol Evol 27: 627–636. doi:10.1016/j.tree.2012.07.007.
38. Best S, Le Tissier P, Towers G, Stoye JP (1996) Positional cloning of the mouse retrovirus restriction

gene Fv1. Nature 382: 826–829. doi:10.1038/382826a0.
39. Taylor GM, Gao Y, Sanders DA (2001) Fv-4 : Identification of the Defect in Env and the Mechanism of Resistance to Ecotropic Murine Leukemia Virus. J Virol 75: 1–6. doi:10.1128/JVI.75.22.11244.
40. Armitage AE, Katzourakis A, de Oliveira T, Welch JJ, Belshaw R, et al. (2008) Conserved footprints of APOBEC3G on Hypermutated human immunodeficiency virus type 1 and human endogenous retrovirus HERV-K(HML2) sequences. J Virol 82: 8743–8761. doi:10.1128/JVI.00584-08.
41. Stremlau M, Perron M, Lee M, Li Y, Song B, et al. (2006) Specific recognition and accelerated uncoating of retroviral capsids by the TRIM5alpha restriction factor. PNAS 103: 5514–5519. doi:10.1073/pnas.0509996103.
42. Pertel T, Hausmann S, Morger D, Züger S, Guerra J, et al. (2011) TRIM5 is an innate immune sensor for the retrovirus capsid lattice. Nature 472: 361–365. doi:10.1038/nature09976.
43. Sawyer SL, Emerman M, Malik HS (2007) Discordant evolution of the adjacent antiretroviral genes TRIM22 and TRIM5 in mammals. PLoS Pathog 3: e197. doi:10.1371/journal.ppat.0030197.
44. McEwan W a, Schaller T, Ylinen LM, Hosie MJ, Towers GJ, et al. (2009) Truncation of TRIM5 in the Feliformia explains the absence of retroviral restriction in cells of the domestic cat. J Virol 83: 8270–8275. doi:10.1128/JVI.00670-09.
45. Martínez Barrio Á, Ekerljung M, Jern P, Benachenhou F, Sperber GO, et al. (2011) The First Sequenced Carnivore Genome Shows Complex Host-Endogenous Retrovirus Relationships. PLoS One 6: 10.
46. Patel MR, Emerman M, Malik HS (2011) Paleovirology—ghosts and gifts of viruses past. Curr Opin Virol 1: 304–309. doi:10.1016/j.coviro.2011.06.007.
47. Lee YN, Malim MH, Bieniasz PD (2008) Hypermutation of an ancient human retrovirus by APOBEC3G. J Virol 82: 8762–8770. doi:10.1128/JVI.00751-08.
48. Soll SJ, Neil SJD, Bieniasz PD (2010) Identification of a receptor for an extinct virus. PNAS 107: 19496–19501. doi:10.1073/pnas.1012344107.
49. Perez-Caballero D, Soll SJ, Bieniasz PD (2008) Evidence for restriction of ancient primate gammaretroviruses by APOBEC3 but not TRIM5alpha proteins. PLoS Pathog 4: e1000181. doi:10.1371/journal.ppat.1000181.
50. Neil SJD, Zang T, Bieniasz PD (2008) Tetherin inhibits retrovirus release and is antagonized by HIV-1 Vpu. Nature 451: 425–430. doi:10.1038/nature06553.
51. Neil SJD (2013) The antiviral activities of tetherin. Curr Top Microbiol Immunol 371: 67–104. doi:10.1007/978-3-642-37765-5_3.
52. Lemaître C, Harper F, Pierron G, Heidmann T, Dewannieux M (2014) The HERV-K human endogenous retrovirus envelope protein antagonises Tetherin antiviral activity. J Virol: JVI.02234–14 – . doi:10.1128/JVI.02234-14.
53. Katzourakis A, Gifford RJ, Tristem M, Gilbert MTP, Pybus OG (2009) Macroevolution of complex retroviruses. Science 325: 1512. doi:10.1126/science.1174149.
54. Han G-Z, Worobey M (2012) An endogenous foamy virus in the aye-aye (Daubentonia madagascariensis). J Virol 86: 7696–7698. doi:10.1128/JVI.00650-12.
55. Katzourakis A, Aiewsakun P, Jia H, Wolfe ND, LeBreton M, et al. (2014) Discovery of prosimian and afrotherian foamy viruses and potential cross species transmissions amidst stable and ancient mammalian co-evolution. Retrovirology 11: 61. doi:10.1186/1742-4690-11-61.
56. Holmes EC (2003) Molecular clocks and the puzzle of RNA virus origins. J Virol 77: 3893–3897.
57. Katzourakis A, Tristem M, Pybus OG, Gifford RJ (2007) Discovery and analysis of the first endogenous lentivirus. PNAS 104: 6261–6265. doi:10.1073/pnas.0700471104.
58. Han G-Z, Worobey M (2012) Endogenous lentiviral elements in the weasel family (Mustelidae). Mol Biol Evol: mss126 – . doi:10.1093/molbev/mss126.
59. 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. doi:10.1016/j.virol.2008.10.045.
60. Larue RS, Lengyel J, Jónsson SR, Andrésdóttir V, Harris RS (2010) Lentiviral Vif degrades the APOBEC3Z3/APOBEC3H protein of its mammalian host and is capable of cross-species activity. J Virol 84: 8193–8201. doi:10.1128/JVI.00685-10.
61. Martin J, Herniou E, Cook J, Neill WO, Tristem M, et al. (1999) Interclass Transmission and Phyletic Host Tracking in Murine Leukemia Virus-Related Retroviruses. J Virol.
62. Hayward JA, Tachedjian M, Cui J, Field H, Holmes EC, et al. (2013) Identification of diverse full-length endogenous betaretroviruses in megabats and microbats. Retrovirology 10: 35. doi:10.1186/1742-4690-10-35.
63. Katzourakis A, Magiorkinis G, Lim AG, Gupta S, Belshaw R, et al. (2014) Larger mammalian body size leads to lower retroviral activity. PLoS Pathog 10: e1004214. doi:10.1371/journal.ppat.1004214.
64. Katzourakis A (2013) Paleovirology: inferring viral evolution from host genome sequence data. Phil Trans R Soc B 368: 20120493 – . doi:10.1098/rstb.2012.0493.

Chapter One Supplementary Items

For clarity and simplicity, the legends for table S1-S3, as well as supplementary data 1 are grouped together on page three of this document.

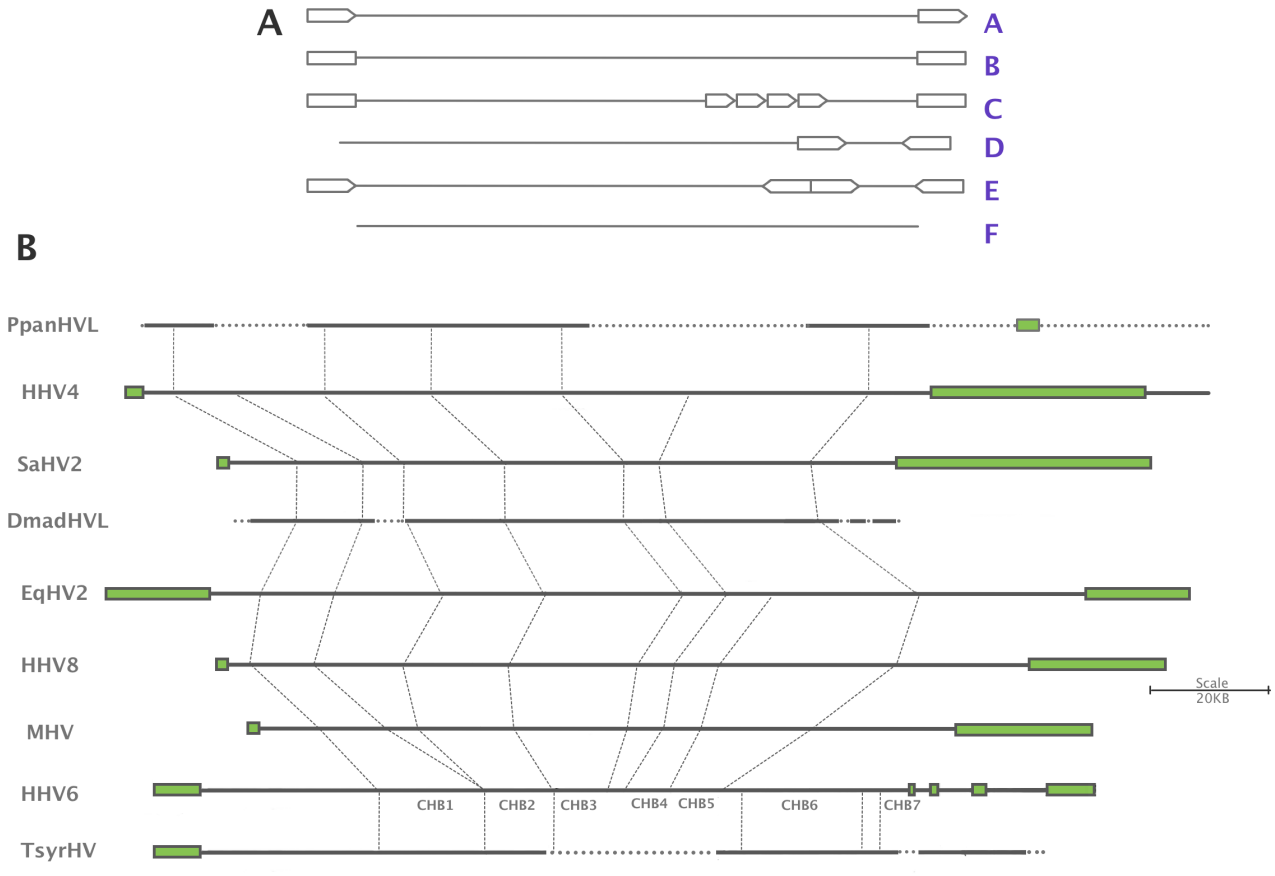


Figure S1.
Schematic diagram showing general herpesvirus genome layouts. Panel A depicts the genome maps of the 6 general herpesvirus layouts, A–F as described (and adapted from) in [4]. Boxes/Arrow boxes represent major repetitive regions. Types A, B, C and E depict genome layouts that are flanked repetitive regions, while layout D represents genomes where the major repeats are not at the termini. Type C and E also contain internal repeats - in C the number of internal reiterations is variable, as are the terminal reiterations in B. Type A represents genomes with direct terminal repeat regions. Type F represents genomes without terminal repeats. Panel B shows the layouts of various herpesvirus genomes, including the HVLs and those used as mapping references. Green boxes indicate repetitive regions and dotted lines indicate the missing regions of the HVLs, where only significant gaps are represented. HHV4, SaHV2 and HHV6 are considered types C, B and A, respectively, as are the PpanHVLs, DmadHVLs and TsyrHVLs.
doi:10.1371/journal.pgen.1004332.s001

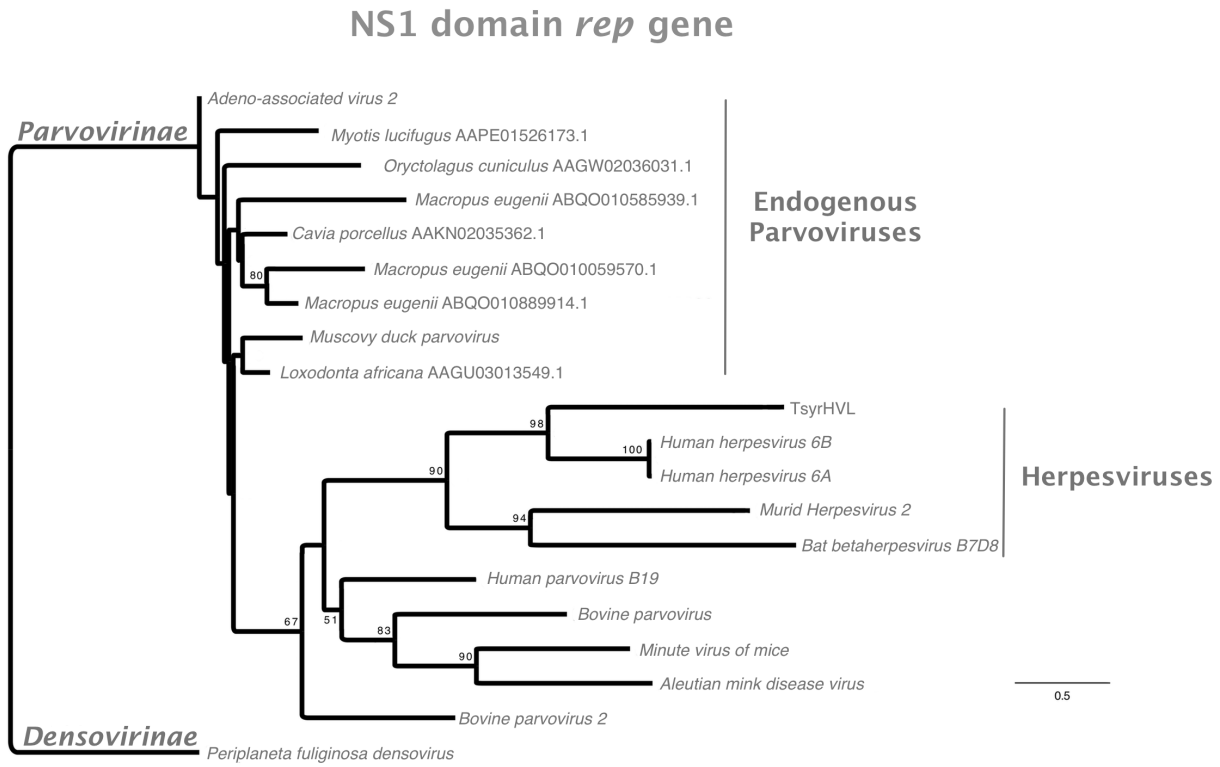


Figure S2.

Phylogeny of captured Parvovirus rep gene in herpesviruses. ML tree of parvovirus NS1 domain including endogenous parvoviruses in mammals, captured genes in herpesviruses and type species from each genus of Parvovirinae. The NCBI accession number for the contig on which each parvovirus EVEs is located is show along side each species name. Only bootstrap values above 50 are shown at nodes. Given the positional orthology of rep in the TsyrHVLs and HHV6a, this tree is consistent with a common origin for the gene capture event. This suggests that HHV7 subsequently lost the gene as our phylogenetic analysis indicates that TsyrHVLs from diverged from other roseoloviruses before the HHV6/HHV7 split.

doi:10.1371/journal.pgen.1004332.s002

Table S1.
Accession numbers for the protein cluster database hosted by NCBI of related sequences used for initial BLAST search for herpesviruses in wgs database as of September 2012.
doi:10.1371/journal.pgen.1004332.s003

Table S2.
Summary of BLAST report detailing initial hits to bonobo, aye-aye and tarsier using consensus sequences of 33 NCBI protein clusters.
doi:10.1371/journal.pgen.1004332.s004

Table S3.
The spreadsheet contains a catalogue of reciprocal BLASTx results. Shortlisted viral contigs were checked against protein databases in order to eliminate false positive results from earlier rounds of BLAST searching. Information for each of the viruses investigated in this study is found on separate sheets. Sheet 1: DmadHVLs, sheet 2: PpanHVLs, sheet 3: TsyrHVLs. 'Top Alternative Hit' documents the best non-viral BLAST hit if the best hit is viral, or else shows the top viral hit if the best hit is host. The latter scenario occurs for a false positive hit.
doi:10.1371/journal.pgen.1004332.s005

Text S1.
Text file containing scaffolded DmadHVL contigs and sequences generated by PCR experiments.
doi:10.1371/journal.pgen.1004332.s006

Table S1

Accession	Description	Consensus sequence
PHA3302	Envelope glycoprotein L	MCRRXXCXXXXXPXXVXLXWELLFXIVLIALAFXXXXXPXXXXXXXXXXXXXXIPALDXNSXYPELTRRCNIAGXXXXXXXXXXXXXXXXXXXXXXXXXXVXXLXXXAXXXXXXXVXXCXEAEXYXXLLTEPJYQPXHRLDDGPXXIRPDRWXLSSLJRYXPXRHQQQLLQGXTGXLRSPKLIXVDDEFLDXLXLLXNNPXQLRTLLTJRSDXAPDNWMKFLKGYSZCGDSXXIYTCVDDVCRKYDLMFKLFGSYLZCDGSXXIYTCVDDVCRKYDLMFKLFGSYLVLALRNEATKTRXVRIPVSTAAXPNXJTLFDALYNLVKXFLLRHGLXDXDLDIRNXLTKRT

Table S2

Query	Subject taxon	Subject accession	Query start	Query end	Subject Start	Subject End	E-value	Identities
ATPase subunit terminase	Daubentonia madagascariensis	gi|368812353|gb|AGTM011948974.1|	331	540	1039	1641	0.000408279	50
hypothetical protein	Daubentonia madagascariensis	gi|370376862|gb|AGTM010395004.1|	77	256	690	1217	4.87E-24	65
hypothetical protein	Daubentonia madagascariensis	gi|368406985|gb|AGTM012354265.1|	250	461	4	606	8.92E-24	68
hypothetical protein	Daubentonia madagascariensis	gi|370222609|gb|AGTM010547017.1|	515	701	67	621	1.14E-86	139
hypothetical protein	Daubentonia madagascariensis	gi|370370516|gb|AGTM010401350.1|	694	813	1	354	1.48E-42	82
hypothetical protein	Daubentonia madagascariensis	gi|370745063|gb|AGTM010130173.1|	234	468	13	567	1.16E-14	71
hypothetical protein	Daubentonia madagascariensis	gi|370377335|gb|AGTM010394531.1|	40	213	1560	1985	3.97E-12	61
uracil DNA glycosylase	Daubentonia madagascariensis	gi|369057494|gb|AGTM011703833.1|	7	261	557	1309	1.73E-92	163
capsid triplex subutnit 1	Daubentonia madagascariensis	gi|370047245|gb|AGTM010722312.1|	15	231	102	671	5.57E-49	106
capsid triplex subutnit 1	Daubentonia madagascariensis	gi|370376421|gb|AGTM010395445.1|	268	383	41	385	3.43E-17	50
capsid triplex subunit 2	Daubentonia madagascariensis	gi|367791523|gb|AGTM012969727.1|	117	303	2	556	7.30E-47	114
capsid triplex subunit 2	Daubentonia madagascariensis	gi|370377301|gb|AGTM010394565.1|	1	120	2	361	4.85E-35	72
FGAM-synthase	Daubentonia madagascariensis	gi|368941814|gb|AGTM011819513.1|	481	1006	3	1406	1.14E-50	173
FGAM-synthase	Daubentonia madagascariensis	gi|368447111|gb|AGTM012314139.1|	1070	1529	630	1901	5.61E-49	154
FGAM-synthase	Daubentonia madagascariensis	gi|368989409|gb|AGTM011771918.1|	1093	1438	1	945	4.47E-38	124
DNA packaging terminase subutnit 1	Daubentonia madagascariensis	gi|368812353|gb|AGTM011948974.1|	448	791	1033	2025	2.94E-67	133
DNA packaging terminase subutnit 1	Daubentonia madagascariensis	gi|370377158|gb|AGTM010394708.1|	304	447	110	541	1.95E-15	48
DNA packaging prot UL32	Daubentonia madagascariensis	gi|370377307|gb|AGTM010394559.1|	484	716	367	942	2.81E-08	61
DNA packaging prot UL32	Daubentonia madagascariensis	gi|367815637|gb|AGTM012945613.1|	228	317	1060	1335	0.0871759	32
DNA packaging terminase subutnit 1	Daubentonia madagascariensis	gi|368812353|gb|AGTM011948974.1|	301	679	1033	2163	2.73E-89	166
DNA packaging terminase subutnit 1	Daubentonia madagascariensis	gi|370377158|gb|AGTM010394708.1|	126	300	35	541	1.80E-23	62
ATPase subunit terminase	Daubentonia madagascariensis	gi|368812353|gb|AGTM011948974.1|	331	540	1039	1641	0.000408279	50
hypothetical protein	Daubentonia madagascariensis	gi|370376862|gb|AGTM010395004.1|	77	256	690	1217	4.87E-24	65
hypothetical protein	Daubentonia madagascariensis	gi|368406985|gb|AGTM012354265.1|	250	461	4	606	8.92E-24	68
hypothetical protein	Daubentonia madagascariensis	gi|370222609|gb|AGTM010547017.1|	515	701	67	621	1.14E-86	139
hypothetical protein	Daubentonia madagascariensis	gi|370370516|gb|AGTM010401350.1|	694	813	1	354	1.48E-42	82
hypothetical protein	Daubentonia madagascariensis	gi|370745063|gb|AGTM010130173.1|	234	468	13	567	1.16E-14	71
hypothetical protein	Daubentonia madagascariensis	gi|370377335|gb|AGTM010394531.1|	40	213	1560	1985	3.97E-12	61
uracil DNA glycosylase	Daubentonia madagascariensis	gi|369057494|gb|AGTM011703833.1|	7	261	557	1309	1.73E-92	163
capsid triplex subutnit 1	Daubentonia madagascariensis	gi|370047245|gb|AGTM010722312.1|	15	231	102	671	5.57E-49	106
capsid triplex subutnit 1	Daubentonia madagascariensis	gi|370376421|gb|AGTM010395445.1|	268	383	41	385	3.43E-17	50
capsid triplex subunit 2	Daubentonia madagascariensis	gi|367791523|gb|AGTM012969727.1|	117	303	2	556	7.30E-47	114
capsid triplex subunit 2	Daubentonia madagascariensis	gi|370377301|gb|AGTM010394565.1|	1	120	2	361	4.85E-35	72
FGAM-synthase	Daubentonia madagascariensis	gi|368941814|gb|AGTM011819513.1|	481	1006	3	1406	1.14E-50	173
FGAM-synthase	Daubentonia madagascariensis	gi|368447111|gb|AGTM012314139.1|	1070	1529	630	1901	5.61E-49	154
FGAM-synthase	Daubentonia madagascariensis	gi|368989409|gb|AGTM011771918.1|	1093	1438	1	945	4.47E-38	124
DNA packaging terminase subutnit 1	Daubentonia madagascariensis	gi|368812353|gb|AGTM011948974.1|	448	791	1033	2025	2.94E-67	133
DNA packaging terminase subutnit 1	Daubentonia madagascariensis	gi|370377158|gb|AGTM010394708.1|	304	447	110	541	1.95E-15	48
DNA packaging prot UL32	Daubentonia madagascariensis	gi|370377307|gb|AGTM010394559.1|	484	716	367	942	2.81E-08	61
DNA packaging prot UL32	Daubentonia madagascariensis	gi|367815637|gb|AGTM012945613.1|	228	317	1060	1335	0.0871759	32
DNA packaging terminase subutnit 1	Daubentonia madagascariensis	gi|368812353|gb|AGTM011948974.1|	301	679	1033	2163	2.73E-89	166
DNA packaging terminase subutnit 1	Daubentonia madagascariensis	gi|370377158|gb|AGTM010394708.1|	126	300	35	541	1.80E-23	62
ATPase subunit terminase	Pan paniscus	gi|383075026|gb|AJFE01003225.1|	331	467	20383	20784	9.05E-05	43
hypothetical protein	Pan paniscus	gi|383075026|gb|AJFE01003225.1|	40	893	28569	30845	2.60E-157	372
uracil DNA glycosylase	Pan paniscus	gi|383075026|gb|AJFE01003225.1|	2	261	1655	2419	7.21E-111	194
capsid triplex subunit 2	Pan paniscus	gi|383075026|gb|AJFE01003225.1|	1	304	23468	24373	4.02E-102	210
DNA packaging terminase subutnit 1	Pan paniscus	gi|383075026|gb|AJFE01003225.1|	447	822	20374	21441	1.97E-67	158
DNA packaging terminase subutnit 1	Pan paniscus	gi|383075026|gb|AJFE01003225.1|	304	447	16619	17050	6.26E-14	50
DNA packaging prot UL32	Pan paniscus	gi|383069699|gb|AJFE01006050.1|	228	346	3387	3752	3.84E-06	35
DNA packaging prot UL32	Pan paniscus	gi|383069699|gb|AJFE01006050.1|	620	716	2440	2721	0.162857	30
DNA packaging terminase subutnit 1	Pan paniscus	gi|383075026|gb|AJFE01003225.1|	301	668	20377	21480	6.76E-88	169
DNA packaging terminase subutnit 1	Pan paniscus	gi|383075026|gb|AJFE01003225.1|	48	300	16256	17050	1.45E-23	77
ATPase subunit terminase	Pan paniscus	gi|383075026|gb|AJFE01003225.1|	331	467	20383	20784	9.05E-05	43
hypothetical protein	Pan paniscus	gi|383075026|gb|AJFE01003225.1|	40	893	28569	30845	2.60E-157	372
uracil DNA glycosylase	Pan paniscus	gi|383075026|gb|AJFE01003225.1|	2	261	1655	2419	7.21E-111	194
capsid triplex subunit 2	Pan paniscus	gi|383075026|gb|AJFE01003225.1|	1	304	23468	24373	4.02E-102	210
DNA packaging terminase subutnit 1	Pan paniscus	gi|383075026|gb|AJFE01003225.1|	447	822	20374	21441	1.97E-67	158
DNA packaging terminase subutnit 1	Pan paniscus	gi|383075026|gb|AJFE01003225.1|	304	447	16619	17050	6.26E-14	50
DNA packaging prot UL32	Pan paniscus	gi|383069699|gb|AJFE01006050.1|	228	346	3387	3752	3.84E-06	35
DNA packaging prot UL32	Pan paniscus	gi|383069699|gb|AJFE01006050.1|	620	716	2440	2721	0.162857	30
DNA packaging terminase subutnit 1	Pan paniscus	gi|383075026|gb|AJFE01003225.1|	301	668	20377	21480	6.76E-88	169
DNA packaging terminase subutnit 1	Pan paniscus	gi|383075026|gb|AJFE01003225.1|	48	300	16256	17050	1.45E-23	77
DNA packaging terminase subutnit 1	Tarsius syrichta	gi|202855770|gb|ABRT010424292.1|	315	405	4733	5002	1.74E-07	34
tegument protein UL16	Tarsius syrichta	gi|202855770|gb|ABRT010424292.1|	107	270	5379	5843	4.48E-18	60
tegument protein UL16	Tarsius syrichta	gi|202855770|gb|ABRT010424292.1|	308	356	5138	5281	4.48E-18	22
tegument protein UL16	Tarsius syrichta	gi|202593353|gb|ABRT010686709.1|	11	131	3	347	3.75E-06	52
DNA packaging terminase subutnit 1	Tarsius syrichta	gi|202855770|gb|ABRT010424292.1|	171	259	4736	5002	5.02E-67	62
DNA packaging terminase subutnit 1	Tarsius syrichta	gi|202855770|gb|ABRT010424292.1|	1	172	4230	4742	5.02E-67	73
DNA packaging terminase subutnit 1	Tarsius syrichta	gi|202855770|gb|ABRT010424292.1|	257	300	4993	5121	5.02E-67	27
DNA packaging terminase subutnit 1	Tarsius syrichta	gi|202855770|gb|ABRT010424292.1|	315	405	4733	5002	1.74E-07	34
tegument protein UL16	Tarsius syrichta	gi|202855770|gb|ABRT010424292.1|	107	270	5379	5843	4.48E-18	60
tegument protein UL16	Tarsius syrichta	gi|202855770|gb|ABRT010424292.1|	308	356	5138	5281	4.48E-18	22
tegument protein UL16	Tarsius syrichta	gi|202593353|gb|ABRT010686709.1|	11	131	3	347	3.75E-06	52
DNA packaging terminase subutnit 1	Tarsius syrichta	gi|202855770|gb|ABRT010424292.1|	171	259	4736	5002	5.02E-67	62
DNA packaging terminase subutnit 1	Tarsius syrichta	gi|202855770|gb|ABRT010424292.1|	1	172	4230	4742	5.02E-67	73
DNA packaging terminase subutnit 1	Tarsius syrichta	gi|202855770|gb|ABRT010424292.1|	257	300	4993	5121	5.02E-67	27

Table S3

Primary Result							Top alternative hit				
Query	length (bp)	Species	Description	Hit ID	E-value	Coverage	Species	Description	Hit ID	E-value	Coverage
AGTM010394531.1	2006	Saimiriine herpesvirus 2	major capsid protein	NP_040227.1	0	96%	Vibrio anguillarum	Topoisomerase IV subunit A	YP_004565394.1	0.48	40%
			hypothetical protein	NP_040226.1	1.34E-28	96%	None				
AGTM010049177.1	475	Equid herpesvirus 2	ORF 29a	NP_042630.1	7.00E-23	82%	Thermus sp.	Methyltransferase	YP_001966068.1	4.2	31%
AGTM010086645.1	329	Bovine herpesvirus 4	hypothetical protein	NP_076558.1	4.00E-36	99%	Branchiostoma floridae	hypothetical protein	XP_002601625.1	36.6	28%
AGTM010130173.1	613	Saimiriine herpesvirus 2	hypothetical protein	NP_040226.1	1.00E-37	95%	Saprospira grandis	hypothetical protein	YP_005320966.1	4.3	30%
AGTM010365164.1	4430	Ovis aries	thymidylate synthase	XP_004020543.1	2.00E-26	5%	European catfish virus	thymidylate synthase	YP_006347622.1	3.00E-20	5%
AGTM010394582.1	1925	Mus musculus	thymidylate synthase	NP_067263.1	2.00E-175	46%	European catfish virus	thymidylate synthase	YP_006347622.1	2.00E-161	45%
AGTM010394559.1	1236	Equid herpesvirus 2	ORF 68	NP_042667.1	3.00E-85	90%	Methylibium petroleiphilum	ferredoxin	YP_001020565.1	8.7	9%
			ORF 69	NP_042668.1	3.00E-14						
AGTM010394565.1	1059	Bovine herpesvirus 4	major capsid protein	NP_076517.1	6.00E-120	96%	Bacteroides stercoris	hypothetical protein	ZP_02435406.1		
		Saimiriine herpesvirus 2	hypothetical protein	NP_040228.1	1.00E-48						
AGTM010394574.1	1107	Equid herpesvirus 2	Helicase-primase complex	NP_042653.1	1.00E-127	99%	None				
AGTM010394781.1	1125	Equid herpesvirus 2	ribonucleotide reductase	NP_042657.1	1.65E-154	81%	Naegleria gruberi	predicted protein	XP_002680453.1	1.00E-32	72%
AGTM010394600.1	1794	Human herpesvirus 8	latency antigen	YP_001129431.1	4.00E-13	22%	None				
AGTM010394632.1	2640	Bovine herpesvirus 4	DNA replication protein	NP_076551.1	1.00E-57	76%	Ornithorhynchus anatinus	fibrocystin-like	XP_003428622.1	4.6	12%
		Ateline herpesvirus 3	ORF 58	NP_048029.1	5.00E-47						
AGTM010394693.1	3229	Human herpesvirus 8	ORF 49	YP_001129402.1	2.00E-62	53%	Burkholderia multivorans	PF06906 family protein	ZP_15925465.1	1.3	3%
		Saimiriine herpesvirus 2	R transactivator	NP_040252.1	2.00E-55		Bos taurus	zinc finger protein	XP_003587411.1	3.7	6%
AGTM010394708.1	1234	Ateline herpesvirus 3	ORF 29a	NP_048006.1	7.00E-100	102%	Synechococcus phage	terminase	YP_007001615.1	2.3	23%
		Bovine herpesvirus 4	hypothetical protein	NP_076525.1	7.00E-75	(overlap)					
AGTM010394888.1	1775	Ateline herpesvirus 3	ORF 48	NP_048021.1	5.00E-29	48%	None				
AGTM010394834.1	1133	Equid herpesvirus 2	glycoprotein H	NP_042618.1	1.00E-27	71%	Trichomonas vaginalis	polymerase zeta subunit	XP_001326973.1	4.3	20%
		Saimiriine herpesvirus 2	DNA polymerase	NP_040211.1	8.00E-21						
AGTM010394847.1	701	Macacine herpesvirus 5	tegument protein	NP_570811.1	1.00E-60	98%	Nematostella vectensis	predicted protein	XP_001634054.1	0.78	28%
AGTM010394854.1	4653	Saimiriine herpesvirus 2	unnamed protein	NP_040222.1	2.00E-53	90%	Dickeya dadantii	ribosome-associated factor	YP_003881078.1	4.5	3%
			unnamed protein	NP_040221.1	6.00E-163						
			unnamed protein	NP_040220.1	5.00E-80						
			unnamed protein	NP_040219.1	5.00E-79						
AGTM010394873.1	2266	Bovine herpesvirus 4	ssDNA-binding protein	NP_040208.1	0	98%	Metaseiulus occidentalis	NCOAT-like protien	XP_003743615.1	2.1	26%
		Saimiriine herpesvirus 2	transport protein	NP_040209.1	4.29E-20						
AGTM010394913.1	782	Bovine herpesvirus 4	FGAM-synthetase	NP_076568.1	1.00E-13	49%	Callithrix jacchus	dehydrogenase	XP_002760484.2	9.9	62%
AGTM010395291.1	1725	Ateline herpesvirus 3	tegument protein	NP_048035.1	5.00E-13	35%	Megavirus lba	collagen-like protein 2	YP_007418171.1	3.3	15%
AGTM010394923.1	2334	Saimiriine herpesvirus 2	alkaline exonuclease	NP_040239.1	6.33E-119	98%	Tribolium castaneum	CG11489-like	XP_968798.1	0.018	27%
		Equid herpesvirus 2	virion protein kinase	NP_042633.1	7.00E-87						
AGTM010395004.1	1221	Saimiriine herpesvirus 2	hypothetical protein	NP_040212.1	1.00E-39	92%	Rhizobium sp	nidazolesuccinocarboxamide (SA	ZP_10732526.1	2.8	19%
			hypothetical protein	NP_040213.1	2.00E-14						
AGTM010395025.1	1038	Equid herpesvirus 2	ORF 55	NP_042652.1	1.00E-76	103%	None				
			Helicase-primase complex	NP_042653.1	6.00E-26	(overlap)					
AGTM010395096.1	2107	Bovine herpesvirus 4	hypothetical protein	NP_076544.1	4.01E-25	19%	None				
		Equid herpesvirus 2	ORF 53	NP_042650.1	0.89						
AGTM010401350.1	1336	Bovine herpesvirus 4	hypothetical protein	NP_076516.1	1.11E-55	96%	Naumovozyma castellii	hypothetical protein	XP_003675768.1	2.1	13%
		Macacine herpesvirus 5	hypothetical protein	NP_570764.1	1.00E-53						
AGTM010395359.1	533	Equid herpesvirus 2	glycoprotein M	NP_042636.1	1.00E-62	87%	None				
AGTM010395382.1	867	Bovine herpesvirus 4	tegument protein	NP_076556.1	5.00E-37	98%	Bifidobacterium animalis	hypothetical protein	YP_006279945.1	1.4	22%
AGTM010395445.1	1504	Saimiriine herpesvirus 2	ribonucleotide reductase	NP_040263.1	2.00E-127	95%	None				
			hypothetical protein	NP_040264.1	3.00E-31						
AGTM010395446.1	1115	Bovine herpesvirus 4	se-primase complex comp	NP_076532.1	3.66E-46	81%	None				
AGTM010456527.1	722	Equid herpesvirus 2	capsid protein	NP_042621.1	2.00E-107	71%	None				
AGTM010446062.1	394	Ateline herpesvirus 3	large tegument protein	NP_048035.1	3.00E-16	95%	Bacteroides fragilis	hypothetical protein	ZP_17249539.1	8.1	57%
AGTM010447026.1	1203	Equid herpesvirus 2	glycoprotein M	NP_042636.1	1.00E-67	88%	None				
			deoxyribonuclease	NP_042634.1	5.65E-18						
		Saimiriine herpesvirus 2	unnamed protein	NP_040240.1	1.00E-03						
AGTM010448489.1	710	Equid herpesvirus 2	tegument protein	NP_042661.1	2.00E-45	90%					
AGTM010448782.1	714	Macacine herpesvirus 5	tegument protein	NP_570811.1	1.00E-46	98%	Lactobacillus gigeriorum	1-phosphofructokinase	ZP_10417669.1	2.2	27%
AGTM010448810.1	323	felis catus	transferase	XP_003985863.1	8.00E-60	99%	None				
AGTM010455736.1	717	Saimiriine herpesvirus 2	hypothetical protein	NP_040266.1	3.00E-44	97%	Chlorobium limicola	maf protein	YP_001943015.1	2.1	37%
AGTM010461791.1	440	Ateline herpesvirus 3	ORF 34	NP_048007.1	9.00E-39	98%	Brevibacillus laterosporus	SNF2 family domain protein	ZP_16303839.1	2.9	36%
AGTM010479572.1	815	Ateline herpesvirus 3	DNA polymerase	NP_047983	6.42E-132	98%	Anolis carolinensis	DNA polymerase subunit-like	XP_003222764.1	2.00E-52	96%
AGTM010462116.1	644	Ateline herpesvirus 3	ORF 34	NP_048007.1	3.00E-18	90%	Glaciecola pallidula	hypothetical protein	ZP_11333206.1	0.053	41%
			ORF 35	NP_048008.1	3.00E-19						
AGTM010463876.1	1011	Saimiriine herpesvirus 2	ssDNA-binding protein	NP_040208.1	4.00E-117	100%	Desulfovibrio hydrothermalis	DNA gyrase	YP_007325172.1	0.057	21%
		Bovine herpesvirus 4	transport protein	NP_076499.1	1.12E-53	98%	None				
AGTM010473543.1	503	Bovine herpesvirus 4	v-FGAM-synthase	NP_076495.1	6.00E-12	96%	Actinobacillus pleuropneumoniae	hypothetical protein	ZP_07538257.1	1.7	88%
AGTM010635235.1	618	Callitrichine herpesvirus 3	ORF23	NP_733876.1	1.00E-20	99%	Branchiostoma floridae	hypothetical protein	XP_002590671.1	3.2	24%
		Human herpesvirus 8	ORF32	YP_001129385.1	4.00E-14						
AGTM010480454.1	515	Ovine herpesvirus 2	ORF 22	YP_438146.1	1.70E-01	97%					
			ORF 23	YP_438147.1	5.00E-28		Leptospira biflexa	permease	YP_001838069.1	0.36	43%
AGTM010486910.1	379	Human herpesvirus 8	ORF61	YP_001129418.1	3.00E-37	96%	Enterocytozoon bieneusi	pnucleoside-diphosphate reduct	XP_002650066.1	3.00E-10	91%
AGTM010493322.1	678	Equid herpesvirus 2	Helicase-primase	NP_042653.1	9.00E-67	84%	Azospirillum brasilense	phosphohydrolase	YP_005032122.1	1.1	20%
AGTM010496494.1	500	Saimiriine herpesvirus 3	envelope glycoprotein B	YP_004940228.1	2.00E-29	95%	Cucumis sativus	nucleosidase 1-like	XP_004144832.1	2.7	57%
AGTM010515626.1	596	Equid herpesvirus 2	ORF 54	NP_042651.1	2.00E-46	79%	Hydrogenobaculum sp.	triphosphate deaminase	YP_007499923.1	0.015	32%
AGTM010547017.1	622	Bovine herpesvirus 4	hypothetical protein	NP_076516.1	8.00E-96	98%	Neisseria elongata glycolytica	hypothetical protein	ZP_06735599.1	3.00E-23	28%
AGTM010633131.1	315	Saimiri boliviensis boliviensis	interleukin-17A	XP_003923160.1	8.00E-43	74%	None				
AGTM010731808.1	672	Bovine herpesvirus 4	randed DNA-binding prote	NP_076498.1	1.00E-18	65%	Alteromonas macleodii	SecY interacting protein	YP_006797974.1	0.52	24%
		Equid herpesvirus 2	ORF E4	NP_042601.1	1.00E-07						
AGTM010653738.1	269	Ateline herpesvirus 3	ORF 63	NP_048034.1	3.00E-19	90%		None			
AGTM010659043.1	946	Equid herpesvirus 2	tegument protein	NP_042672.1	5.70E-14	23%	Desulfotalea psychrophila	horibosylformylglycinamidine syr	YP_064440.1	3.00E-14	21%
AGTM010668737.1	422	Ateline herpesvirus 3	major capsid protein	NP_047996.1	5.00E-51	96%	Haloarcula japonica	conjugation protein	ZP_21693618.1	0.39	69%
AGTM010722312.1	672	Ateline herpesvirus 3	ORF 62	NP_048033.1	4.16E-70	86%	Providencia rettgeri	lycoside hydrolase family protei	ZP_16353551.1	0.53	40%
AGTM010722683.1	464	Bovine herpesvirus 4	ribonucleotide reductase	NP_076553.1	2.68E-59	99%	Encephalitozoon intestinalis	pnucleoside diphosphate reduct	XP_003073772.1	2.00E-16	73%
AGTM010723649.1	323	Saimiriine herpesvirus 2	glycoprotein H	NP_040224.1	6.00E-23	99%	Meleagris gallopavo	contactin-3-like	XP_003210246.1	1.8	81%
AGTM011564346.1	10759	Homo sapiens	thymidylate synthase	NP_001062.1	3.00E-38	7%	European catfish virus	thymidylate synthase	YP_006347622.1	8.00E-24	5%
AGTM010808082.1	592	Saimiriine herpesvirus 2	unnamed protein	NP_040265.1	4.60E-13	69%	Physcomitrella patens	predicted protein	XP_001763433.1	0.21	46%
AGTM010904390.1	243	Macacine herpesvirus 5	DNA replication protein	NP_570796.1	1.12E-20	98%	Photobacterium profundu	bacteriophage integrase	YP_133335.1	5.7	48%
AGTM011204162.1	260	Saimiriine herpesvirus 2	major capsid protein	NP_040227.1	7.00E-40	96%	Paenibacillus dendritiformis	group 1 glycosyl transferase	ZP_09677647.1	1.4	31%
AGTM011244974.1	212	Otolemur garnettii	glycoprotein beta-1,6-N-ac	XP_003787766.1	8.00E-38	99%	None				
AGTM011531363.1	4716	Otolemur garnettii	glycoprotein beta-1,6-N-ac	XP_003784439.1	0	27%	None				
AGTM011536785.1	19290	Macaca mulatta	membrane glycoprotein pre	NP_001244473.1	5.00E-56	3%	Lumpy skin disease virus	DV138 Ig domain OX-2-like prote	NP_150572.1	4.00E-04	1%
AGTM011537755.1	8267	Equus caballus	sidylate synthase-like isofo	XP_001492275.1	3.07E-29	6%	European catfish virus	thymidylate synthase	YP_006347622.1	2.00E-24	6%
AGTM011544182.1	8417	Nomascus leucogenys	type A, EGF and pentraxin	XP_003260519.1	0	90%	None				
AGTM011552975.1	8104	Otolemur garnettii	omplement receptor type	XP_003792361.1	9.00E-68	24%	Cowpox virus	CPXV034	NP_619823.1	1.00E-13	12%
AGTM011563296.1	6290	Otolemur garnettii	interleukin-17A	XP_003789688.1	1.00E-40	6%	None				
AGTM011636650.1	4630	Otolemur garnettii	glycoprotein	XP_003781590.1	2.00E-25	5%	None				
AGTM011564795.1	5624	Pan troglodytes	oprotein beta-1,6-N-acetyl	XP_001145936.1	1.00E-131	11%	None				
AGTM011579171.1	12618	Sus scrofa	membrane glycoprotein	XP_003483369.1	2.99E-58	5%	Goatpox virus Pellor	hypothetical protein	YP_001293329.1	0.002	2%
AGTM011602539.1	18693	Callithrix jacchus	glycoprotein beta-1,6-N-ac	XP_002747748.1	0	6%	None				
AGTM011606314.1	17433	Callithrix jacchus	l-glycoprotein beta-1,6-N-a	XP_002742885.1	2.00E-128	6%	Bovine herpesvirus 4	ta-1,6-N-acetylglucosaminyltran	NP_076572.1	1.00E-80	5%
AGTM011611918.1	7152	Otolemur garnettii	thymidylate synthase	XP_003784845.1	5.00E-128	11%	Cercopithecine herpesvirus 9	thymidylate synthase	NP_077428.1	2.00E-94	11%
AGTM011618091.1	2656	Papio anubis	interleukin-17F	XP_003897778.1	2.07E-13	8%	Saimiriine herpesvirus 2	unnamed protein product	NP_040215.1	3.00E-20	8%
AGTM011629352.1	5409	Ailuropoda melanoleuca	)-1,6-N-acetylglucosaminyl	XP_002915964.1	2.00E-99	10%	Bovine herpesvirus 4	ta-1,6-N-acetylglucosaminyltran	NP_076572.1	3.00E-40	10%
AGTM011769130.1	1954	Ovis aries	membrane cofactor	XP_004013973.1	2.00E-36	29%	Saimiriine herpesvirus 2	dihydrofolate reductase	NP_040203.1	6.00E-08	1%
AGTM011643537.1	5770	Loxodonta africanus	dihydrofolate reductase	XP_003408165.1	6.00E-11	1%	Saimiriine herpesvirus 2	dihydrofolate reductase	NP_040203.1	6.00E-08	1%
AGTM011679335.1	3436	Nomascus leucogenys	oprotein beta-1,6-N-acetyl	XP_003267479.1	1.00E-56	11%	Bovine herpesvirus 4	ta-1,6-N-acetylglucosaminyltran	NP_076572.1	8.00E-12	7%
AGTM011703833.1	3762	Saimiriine herpesvirus 2	helicase	NP_040246.1	0	67%	Hydra magnipapillata	uracil-DNA glycosylase-like	XP_002166406.1	2.00E-59	18%
		Ateline herpesvirus 3	hypothetical protein	NP_048019.1	3.00E-108						
AGTM011704110.1	1165	Bovine herpesvirus 4	v-FGAM-synthase	NP_076495.1	3.00E-60	95%	atus Kinetoplastibacterium desouzaii TC	horibosylformylglycinamidine syr	YP_007459298.1	3.00E-16	69%
AGTM011717059.1	1750	Human herpesvirus 8	Glycoprotein B	YP_001129354.1	8.00E-104	89%	Escherichia coli KTE163	hypothetical protein WKG_03335		7.8	9%
		Saimiriine herpesvirus 2	transport protein	NP_040209.1	7.80E-83						
AGTM011718482.1	5084	felis catus	transposable element der	XP_003990972.1	3.00E-40	10%	Saimiriine herpesvirus 2	dihydrofolate reductase	NP_040203.1	2.00E-09	2%
		European catfish virus	dihydrofolate reductase	YP_006347707.1	2.00E-12						
AGTM011750305.1	2680	Otolemur garnettii	uncharacterized protein	XP_003788410.1	0	34%	Bovine herpesvirus 4	ta-1,6-N-acetylglucosaminyltran	NP_076572.1	4.00E-71	26%
AGTM011754680.1	5122	Otolemur garnettii	glycoprotein beta-1,6-N-ac	XP_003785965.1	0	26%	ephant endotheliotropic herpesvirus 1	a-Acetyl-Glucosamine Transfera	YP_007517197.1	1.00E-99	22%
AGTM011876226.1	3319	Otolemur garnettii	chemokine receptor	XP_003785015.1	0	32%	None	None			
AGTM011771171.1	7616	Otolemur garnettii	omplement receptor type	XP_003792361.1	2.00E-64	26%	Macacine herpesvirus 5	complement binding protein	NP_570746.1	2.00E-09	10%
AGTM011771918.1	968	Bovine herpesvirus 4	FGAM-synthetase	NP_076568.1	9.00E-78	99%	Drosophila grimshawi	GH10940	XP_001988021.1	8.00E-16	87%
AGTM011792607.1	4739	Otolemur garnettii	dihydrofolate reductase	XP_003786000.1	7.00E-31	4%	Saimiriine herpesvirus 2	dihydrofolate reductase	NP_040203.1	1.00E-22	4%
AGTM011819513.1	1408	Bovine herpesvirus 4	hypothetical protein	NP_076568.1	7.00E-77	36%	Anolis carolinensis	ribosylformylglycinamidine synt	XP_003227285.1	3.00E-08	72%
AGTM011819567.1	1009	Saimiriine herpesvirus 2	unnamed protein	NP_040244.1	5.00E-61	100%	Nasonia vitripennis	-dependent transcription factor ATF-2-like			
		Macaca fuscata rhadinovirus	Helicase-primase complex	NP_042638.1	1.00E-06				XP_001603069.1	0.88	29%
		Saimiri boliviensis boliviensis	glycoprotein beta-1,6-N-ac	XP_003927479.1	5.00E-157		Bovine herpesvirus 4	ta-1,6-N-acetylglucosaminyltran	NP_076572.1	7.00E-55	31%
AGTM011835354.1	3893	Otolemur garnettii	type A, EGF and pentraxin	XP_003785430.1	4.00E-132	25%	None	None			
AGTM011914563.1	5415	Ailuropoda melanoleuca	thetical protein PANDA_00	XP_002915951.1	1.00E-176	17%	Bovine herpesvirus 4	ta-1,6-N-acetylglucosaminyltran	NP_076572.1	2.00E-65	13%
AGTM012005779.1	284	Saimiri boliviensis boliviensis	glycoprotein beta-1,6-N-ac	XP_003927479.1	3.25E-38	99%	Bovine herpesvirus 4	ta-1,6-N-acetylglucosaminyltran	NP_076572.1	1.00E-12	58%
AGTM011921304.1	364	Oryctolagus cuniculus	thymidylate synthase	XP_002713484.1	1.17E-57	80%	Cercopithecine herpesvirus 9	thymidylate synthase	NP_077428.1	5.00E-47	80%
AGTM011948974.1	2508	Ateline herpesvirus 3	pothetical protein AtHV3gp	NP_048000.1	0	89%	Branchiostoma floridae	thetical protein BRAFLDRAFT_10	XP_002591195.1	0.081	14%
			pothetical protein AtHV3gp	NP_048003.1	1.01E-60						
		Bovine herpesvirus 4	unnamed protein	NP_076522.1	1.00E-07		Physcomitrella patens subsp. patens	predicted protein	XP_001766999.1	1.8	5%
			DNA packaging	NP_076524.1	1.24E-05						
AGTM011952769.1	683	Equus caballus	brand factor type A, EGF an	XP_001916184.1	2.37E-142	99%	Macacine herpesvirus 5	complement binding protein	NP_570746.1	2.00E-14	99%
AGTM011994144.1	440	Nomascus leucogenys	type A, EGF and pentraxin	XP_003260519.1	1.00E-84	99%	Deerpox virus W-848-83	EEV host range protein	YP_227531.1	1.00E-12	90%
AGTM012314371.1	663	Equid herpesvirus 2	ORF 54	NP_042651.1	1.79E-49	64%	acterium catenulatum DSM 16992 = JC	ypothetical protein BIFCAT_0055	ZP_03323783.1	2.00E-04	62%
AGTM012015256.1	301	Ateline herpesvirus 3	transport protein	NP_047981.1	4.30E-25	99%	Pan troglodytes	reonine synthase-like 1 isoform	XP_521671.1	1.8	36%
AGTM012090034.1	217	Ateline herpesvirus 3	mitogen	NP_047986.1	3.02E-12	66%	Synechococcus sp. WH 8016	dehalogenase domain protein h	ZP_08957352.1	9.3	73%
AGTM012095796.1	210	Ateline herpesvirus 3	mitogen	NP_047986.1	1.52E-12	78%	Ralstonia solanacearum K60-1	naldehyde dehydrogenase (NAD	ZP_16319481.1	2.5	48%
AGTM012163629.1	216	Ateline herpesvirus 3	mitogen	NP_047986.1	5.72E-11	66%	Microcystis aeruginosa PCC 9808	Modification methylase HindV	ZP_18841800.1	2.2	97%
AGTM012291937.1	3322	Ailuropoda melanoleuca	)-1,6-N-acetylglucosaminyl	XP_002915964.1	4.61E-175	28%	Bovine herpesvirus 4	ta-1,6-N-acetylglucosaminyltran	NP_076572.1	3.00E-67	21%
AGTM012314139.1	1906	Saimiriine herpesvirus 2	unnamed protein	NP_040204.1	2.95E-99	67%	Trichinella spiralis	horibosylformylglycinamidine syr	XP_003372964.1	1.00E-21	64%
AGTM012580127.1	234	Mus musculus	thymidylate synthase	NP_067263.1	6.00E-24	60%	Human herpesvirus 8	ORF70	YP_001129361.1	9.00E-17	58%
AGTM012318916.1	602	Human herpesvirus 8	ORF 22	YP_001129375.1	2.24E-27	91%	Cellulophaga algicola DSM 14237	hypothetical protein	YP_004163066.1	3.1	17%
AGTM012354265.1	606	Ovine herpesvirus 2	ORF 11	YP_438139.1	3.00E-26	100%	None	None			
AGTM012375206.1	744	Saimiriine herpesvirus 2	unnamed protein	NP_040229.1	3.00E-15	58%	Pseudomonas putida DOT-T1E	809 family phage-associated pro	YP_006532001.1	0.35	16%
AGTM012418162.1	341	Ateline herpesvirus 3	mitogen	NP_047986.1	5.43E-10	48%	Mouse mammary tumor virus	Pr4B	NP_056884.1	0.41	24%
AGTM012969727.1	800	Ateline herpesvirus 3	hypothetical protein	NP_047997.1	5.70E-80	84%	Tetrahymena thermophila	othetical protein TTHERM_0006	XP_001007710.1	0.61	45%
		Human herpesvirus 8	ORF 27	YP_001129380.1	8.00E-82		Leadbetterella byssophila DSM 17132	hypothetical protein Lbys_1126	YP_003997201.1	2.1	18%
AGTM012703695.1	215	Saimiri boliviensis boliviensis	)-1,6-N-acetylglucosaminyl	XP_003927404.1	4.26E-36	99%	Bovine herpesvirus 4	ta-1,6-N-acetylglucosaminyltran	NP_076572.1	1.00E-19	94%
AGTM012815981.1	202	Macaca mulatta	dylate synthetase, isoform	NP_001182436.1	2.00E-29	92%	Human herpesvirus 3	thymidylate synthase	NP_040136.1	2.00E-13	65%
AGTM012840725.1	252	Papio anubis	dihydrofolate reductase-like	NP_001089311.1	1.27E-47	98%	Saimiriine herpesvirus 2	dihydrofolate reductase	NP_040203.1	1.00E-38	97%
AGTM012927422.1	2254	Anolis carolinensis	thymidylate synthase-like	XP_003219704.1	1.69E-18	51%	Human herpesvirus 3	thymidylate synthase	NP_040136.1	3.00E-16	19%
		Plasmodium yoelii yoelii	hypothetical protein	XP_726997.1	2.00E-11						
AGTM012945613.1	1348	Bovine herpesvirus 4	unnamed protein	NP_076560.1	5.00E-13	83%	None				
			tegument protein	NP_076559.1	4.00E-75						
			probable glycoprotein	NP_076561.1	4.00E-42						
AGTM012948373.1	818	Saimiriine herpesvirus 2	unnamed protein	NP_040268.1	9.00E-06	66%	None				
AGTM012954910.1	593	Saimiri boliviensis boliviensis	N-acetylglucosaminyl-tran	XP_003927402.1	1.68E-68	66%	Bovine herpesvirus 4	ta-1,6-N-acetylglucosaminyltran	NP_076572.1	5.00E-14	30%
AGTM013180837.1	312	Macaca mulatta	dihydrofolate reductase	XP_001110551.2	8.90E-42	74%	Macacine herpesvirus 5	dihydrofolate reductase	NP_570745.1	6.00E-16	49%
AGTM012985849.1	510	Macacine herpesvirus 5	large tegument protein	NP_570812.1	5.00E-21	88%	Bacteroidetes oral taxon 274 str. F0058	peptidyl-prolyl cis-trans isomeras	ZP_06982726.1	0.19	35%
AGTM012997089.1	1272	Saimiriine herpesvirus 2	iE-G	NP_040216.1	5.84E-24		None				
		Gorilla Gorilla	tigger-TE	XP_004033412.1	8.00E-14						
AGTM013034944.1	517	Macacine herpesvirus 5	hypothetical protein	NP_570817.1	6.78E-59	99%	Cavia porcellus	myosin-XV-like	XP_003466009.1	1.8	23%
AGTM013041174.1	475	Bovine herpesvirus 4	DNA polymerase	NP_076501.1	1.75E-63	99%	hanomethylovorans hollandica DSM 15	DNA polymerase	YP_007301837.1	8.00E-10	96%
AGTM013121095.1	244	Bovine herpesvirus 4	hypothetical protein	NP_076558.1	8.57E-13	71%	None				
AGTM013080733.1	277	Papio anubis	6-N-acetylglucosaminyl-tra	XP_003897091.1	2.00E-57	99%	Bovine herpesvirus 4	ta-1,6-N-acetylglucosaminyltran	NP_076572.1	7.00E-31	97%
AGTM013133622.1	220	Pan troglodytes	eta-1,6-N-acetylglucosamir	XP_003311128.1	2.72E-39	98%	Bovine herpesvirus 4	ta-1,6-N-acetylglucosaminyltran	NP_076572.1	3.00E-24	98%

Table S3 continued

Primary Hit							Top Alternative Hit				
Query ID	Length	Species	Description	Hit ID	E-value	coverage	Species	Description	Hit ID	E-value	coverage
AJFE01001036.1	4833	Pan paniscus	viral interleukin-10	XP_003804206.1	4.00E-94	10	Human Herpesvirus 4	BCRF1	YP_001129439.1	9.00E-78	10
AJFE01001973.1	1331	Human herpesvirus 4 type 2	BOLF1	YP_001129450.1	4.00E-178	78	Halorubrum tebenquichense	Bacterio-opsin activator	ZP_21586426.1	1.8	14
AJFE01002280.1	1014	Human herpesvirus 4 type 2	BOLF1	YP_001129450.1	1.19E-102	75	NONE				
AJFE01001185.1	7204	Human herpesvirus 4	BALF2	YP_401717.1	0.00E+00	80	Pseudovibrio sp.	utative membrane associated protei	YP_005079425.1	3.2	41
		Human herpesvirus 4	BARF1	YP_401719.1	4.00E-107						
		Human herpesvirus 4	BALF1	YP_401718.1	2.00E-104						
		Human herpesvirus 4	BALF3	YP_401715.1	1.00E-95						
		Human herpesvirus 4	BNLF2b	YP_401720.1	1.00E-21						
AJFE01005652.1	4766	Macacine herpesvirus 4	BALF4	YP_068009.1	0.00E+00	98	Acyrthosiphon pisum	DNA polymerase	XP_001948892.2	4.00E-10	10
		Human herpesvirus 4 type 2	BALF5	YP_001129507.1	0.00E+00						
		Macacine herpesvirus 4	BALF3	YP_001129508.1	0.00E+00						
AJFE01003225.1	44035	Human herpesvirus 4	BKRF3	YP_401679.1	2.00E-160	100	None				
		Human herpesvirus 4	BBLF4	YP_401681.1	0.00E+00						
		Human herpesvirus 4 type 2	BBRF1	YP_001129476.1	0.00E+00						
		Human herpesvirus 4	BBRF2	YP_401683.1	5.00E-169						
		Human herpesvirus 4	BBLF2/BBLF3	YP_401684.1	0.00E+00						
		Human herpesvirus 4 type 2	BBRF3	YP_001129479.1	0.00E+00						
		Human herpesvirus 4 type 2	BGLF5	YP_001129481.1	0.00E+00						
		Human herpesvirus 4	BGLF4	YP_401688.1	0.00E+00						
		Human herpesvirus 4	BGLF3	YP_401689.1	0.00E+00						
		Human herpesvirus 4	BGRF1/BDRF1	YP_401690.1	2.90E-179						
		Human herpesvirus 4	BGLF2	YP_401691.1	0.00E+00						
		Human herpesvirus 4 type 2	BGLF1	YP_001129487.1	0.00E+00						
		Human herpesvirus 4	BDLF4	YP_401693.1	2.00E-133						
		Human herpesvirus 4	BGRF1/BDRF1	YP_401690.1	0.00E+00						
		Human herpesvirus 4	BDLF2	YP_401695.1	2.00E-173						
		Macacine herpesvirus 4	BDLF1	YP_067993.1	5.30E-179						
		Human herpesvirus 4	BcLF4	YP_401697.1	0.00E+00						
		Human herpesvirus 4	BcRF1	YP_401698.1	0.00E+00						
		Human herpesvirus 4	BTRF1	YP_401699.1	0.00E+00						
		Human herpesvirus 4	BXLF2	YP_401700.1	0.00E+00						
		Human herpesvirus 4	BXLF1	YP_401701.1	0.00E+00						
		Human herpesvirus 4 type 2	BXRF1	YP_001129498.1	3.00E-144						
		Human herpesvirus 4	BVRF1	YP_401703.1	0.00E+00						
		Human herpesvirus 4 type 2	BVLF1	YP_001129500.1	4.00E-136						
		Human herpesvirus 4	BVRF2	YP_401704.1	0.00E+00						
		Human herpesvirus 4	membrane glycoprotein	YP_401706.1	8.30E-98						
AJFE01004538.1	12095	Human herpesvirus 4	BPLF1	YP_401652.1	0.00E+00	71	None				
		Human herpesvirus 4 type 2	BFRF2	YP_001129447.1	0.00E+00	14					
		Human herpesvirus 4	BFRF1	YP_401649.1	1.00E-91	5					
AJFE01121077.1	1720	Human herpesvirus 4	BHRF1	YP_401646.1	8.00E-94	33	Microlunatus phosphovorus	hypothetical protein	YP_004571555.1	0.62	22
AJFE01006050.1	5011	Human herpesvirus 4	BHRF1	YP_401646.1	6.10E-55	7	None				
		Human herpesvirus 4	BFLF1	YP_401648.1	0.00E+00	31					
		Human herpesvirus 4	BFLF2	YP_401647.1	0.00E+00	18					
		Human herpesvirus 4	BFRF1A	YP_401728.1	2.20E-68	8					
		Human herpesvirus 4	BFRF1	YP_401649.1		12					
AJFE01025133.1	55674	Homo sapiens	uncharacterized protein	XP_002831604.2	6.00E-31	0.08	None				
		Papio anubis	Uronyl 2-sulfotransferase	XP_003919522.1	1.00E-30	8					
AJFE01035256.1	22558	Saimirine boliviensis	Uracil-DNA glycosylase	XP_003932325.1	7.38E-148	3	None				
		Pan paniscus	Uracil-DNA glycosylase	XP_003832558.1	2.00E-75	2					
AJFE01061695.1	31769	Gorilla gorilla	phoribosylformylglycinamidine synt	XP_004058598.1	8.00E-52	6	None				
		Homo sapiens	uncharacterized protein	XP_003846754.1	1.00E-39	5					
AJFE01068078.1	30507	Pan paniscus	uncharacterized protein	XP_003808117.1	2.00E-28	0.01	None				
		Otolemur garnettii	uracil-DNA glycosylase	XP_003789937.1	2.00E-62	0.02					
AJFE01002611.1	1196	Rhizobium sp	cyl-(acyl carrier protein) synthase II	WP_008535308.1	8.50E-01	0.2	None				
AJFE01002662.1	1003	Human herpesvirus 4	BWRF1	YP_401635.1	6.00E-11	0.2	None				
AJFE01080043.1	24258	Pan paniscus	uncharacterized protein	XP_003832743.1	2.00E-26	0.02	ORF virus	ORF127	NP_957904.1	1.00E-12	1
		cterized protein encoded by LINC0026	uncharacterized protein	XP_004092751.1	5.00E-33	0.03					
AJFE01100746.1	32067	Streptococcus agalactiae	hypothetical	WP_000321103.1	7.00E-115	0.06	[Lymphocystis disease virus	RNA-dependent DNA polymerase	YP_073558.1	8.00E-42	1
		Streptococcus agalactiae	Endonuclease	WP_000148030.1	0.00E+00	5					

Table S3 continued

Primary Hit							Alternative Hit				
Query ID	Length	Species	Description	Hit ID	E-value	coverage	Species	Description	Hit ID	E-value	Coverage
ABRT02259801.1	53666	Human herpesvirus 7	ssDNA-binding protein	YP_073781.1	3.00E-126	2%					
		Human herpesvirus 6B	transport/capsid assembly	NP_050221.1	0	4%					
		Human herpesvirus 6B	Glycoprotein B	NP_050220.1	0	4%					
		Human herpesvirus 6B	DNA polymerase	NP_050219.1	0	5%	Callithrix jacchus	DNA polymerase delta catalytic subunit	XP_002762441.1	6.00E-86	28%
		Human herpesvirus 7	DNA packaging protein UL32	YP_073776.1	3.00E-123	4%					
		Human herpesvirus 7	DNA packaging protein UL33	YP_073775.1	2.00E-26	4%	Sinorhizobium medicae	LPS:glycosyltransferase	WP_018209614.1	6.5	24%
		Human herpesvirus 6A	nuclear egress membrane protein	NP_042927.1	1.00E-59	1%					
		Human herpesvirus 7	protein UL49	YP_073773.1	2.00E-112	2%					
		Human herpesvirus 6B	large tegument protein	NP_050212.1	0	11%					
		Human herpesvirus 7	tegument protein UL37	YP_073770.1	3.00E-103	4%					
		Human herpesvirus 6A	capsid triplex subunit 1	NP_042922.1	1.00E-83	1%					
		Human herpesvirus 6B	large ribonuclease reductase	NP_050209.1	1.00E-151	1%					
		Human herpesvirus 7	DNA polymerase processivity subunit	YP_073767.1	2.00E-109	1%					
		Human herpesvirus 6B	putative membrane glycoprotein	NP_050201.1	5.00E-14	1%					
		Human herpesvirus 6B	immediate-early protein 4	NP_050199.1	2.00E-24	1%					
		Human herpesvirus 7	envelope glycoprotein UL37	YP_073758.1	1.00E-41	1%					
		Human herpesvirus 7	tegument protein vICA	YP_073757.1	6.00E-76	4%					
		Human herpesvirus 7	protein U15	YP_073756.1	6.00E-45	1%					
		Human herpesvirus 7	tegument protein UL35	YP_073755.1	3.00E-79	2%					
		Human herpesvirus 7	envelope glycoprotein UL33	YP_073753.1	3.00E-28	1%					
		Human herpesvirus 6B	antigenic virion protein	NP_050192.1	2.00E-47	1%					
		Human herpesvirus 6B	hypothetical protein HhV6Bgp016	NP_050191.1	2.00E-100	2%					
		Human herpesvirus 6A	protein UL28	NP_042898.2	0	7%					
		Human herpesvirus 6B	hypothetical protein HhV6Bgp009	NP_050186.1	2.00E-107	2%					
		Human herpesvirus 6A	tegument protein UL24	NP_042894.1	3.00E-47	2%					
		Human herpesvirus 6A	tegument protein UL23	NP_042893.1	6.00E-41	4%					
ABRT02337366.1	15583	Human herpesvirus 6B	IE-A transactivator	NP_050266.1	4.00E-11	9%	Clostridium sp. CAG:470	phosphoglycerate kinase	WP_022412440.1	2.1	2%
		Human herpesvirus 6B	Parvovirus rep homolog	NP_050269.1	2.00E-78	8%	Alistipes sp. CAG:53	uncharacterized protein	WP_022061947.1	2.2	3%
		Human herpesvirus 6A	protein U94/rep	NP_042987.1	6.00E-78	8%					
ABRT02391417.1	48621	Human herpesvirus 6A	protein DR6	NP_042888.2	1.00E-34	2%					
		Gibbon ape leukemia virus	Hypothetical mature RT protein	P_002000606	7.00E-25	2%	Chrysemys picta bellii	uncharacterized protein LOC101930828	XP_005294278.1	1.00E-25	2%
		Oreochromis niloticus	uncharacterized protein	P_005461302	2.00E-22	1%					
		Pongo abelii	UPF0764 protein C16orf89-like	P_003777030	4.00E-09	9%					
		Pongo abelii	histone demethylase UTY-like	P_003776766	6.00E-09	9%					
		Gorilla gorilla gorilla	histone demethylase UTY-like	P_004063026	1.00E-08	11%					
		Pongo abelii	UPF0764 protein C16orf89-like	P_003778043	3.00E-07	10%					
ABRT02275240.1	34290	Human herpesvirus 6B	Helicase/primase complex	NP_050256.1	9.00E-164	7%	Pseudoalteromonas agarivorans	hypothetical protein	WP_004589690.1	1.00E-41	2%
		Human herpesvirus 6B	major immediate-early protein	NP_050265.1	6.00E-77	6%					
		Human herpesvirus 6B	Glycoprotein L	NP_050261.1	2.00E-27	2%					
		Human herpesvirus 7	protein UL117	YP_073821.1	2.00E-38	1%					
		Human herpesvirus 7	uracil-DNA glycosylase	YP_073819.1	9.00E-61	2%	Trichomonas vaginalis G3	uracil-DNA glycosylase family protein	XP_001296939.1	8.00E-27	2%
		Cyanothece sp. CCY0110	hypothetical protein, partial	P_008278995	6.00E-94						
		Human herpesvirus 6B	putative viron protein	NP_050255.1	3.00E-152	5%					
		Human herpesvirus 6A	helicase-primase helicase subunit	NP_042970.1	3.00E-137	7%					
		Human herpesvirus 7	DNA replication origin-binding helicase	YP_073813.1	2.00E-111	6%					
		Human herpesvirus 7	envelope glycoprotein M	YP_073812.1	3.00E-119	2%					
		Human herpesvirus 6A	deoxyribonuclease	NP_042963.1	2.00E-109	4%					
		Human herpesvirus 6B	tegument protein	NP_050245.1	1.00E-99	2%					
		Human herpesvirus 6A	protein UL95	NP_042960.1	1.00E-57	2%					
		Human herpesvirus 6B	Putative terminase	NP_050241.1	6.00E-79	2%					
		Human herpesvirus 6B	tegument protein	NP_050245.1	1.00E-99	2%	Glycine max	lectin S-receptor-like serine/threonine-protein kinase At2g19130	XP_003547437.1	0.1	1%
ABRT02394679.1	5129	Ceratotherium simum simum	uracil-DNA glycosylase	P_004429930	1.00E-134	16%					
ABRT02359611.1	17917	Pongo abelii	UPF0764 protein C16orf89-like	P_003777030	6.00E-14	30%					
		Homo sapiens	histone demethylase UTY isoform X3	P_005262577	9.00E-06	26%					
		Pan paniscus	lecule in BECN1-regulated autophagy protei	P_003815253	0.13	27%					
		Macaca fascicularis	chemokine (C-C motif) receptor 7 precursor	P_001272026	0	6%					
ABRT02400840.1	8590	Pan troglodytes	NA polymerase delta catalytic subunit, partia	P_003316601	8.00E-35	17%					
ABRT02455554.1	25188	Nomascus leucogenys	uncharacterized protein encoded by LINC00	P_004092751	2.00E-12	24%					
		Nomascus leucogenys	uracil-DNA glycosylase-like	P_003278500	3.00E-27	6%					

Chapter two Supplementary items

Table S1

NCBI accession numbers of viral-like DNA polymerase- containing contigs in each of the species investigated in this study. These sequences were retrieved using BLASTn with each of the sequences obtained in the initial shortlist against a database of each respective species.

Table S2

The table shows the BLAST results from the two rounds of search used to retrieve the viral-like sequences from fish genome data. In the first round of results, the highlighted hits were considered since they aligned to at least 60% of the query, retrieving one representative from each species as indicted by an 'x' in the Initial shortlist column. A second round of BLAST was performed using the hits indicated to identify additional fish species not in the initial shortlist (highlighted).

Table S3

Predicted ORFs of the *Salmo salar* contig, including the best hit to known proteins in the NCBI nr database.

Table S4

Predicted ORFs of the *Amphilophus citrinellus* contig, including the best hit to known proteins in the NCBI nr database.

Table S5

Predicted ORFs of the *Nothobranchius furzeri* contig, including the best hit to known proteins in the NCBI nr database.

Table S6

Predicted ORFs of the *Larimichthys crocea* contig, including the best hit to known proteins in the NCBI nr database.

Species	Accession
Melanochromis auratus	ABPL01007469.1
	ABPL01018272.1
Pampus argenteus	JHEK01266819.1
Neolamprologus brichardi	AFNY01056855.1
Oreochromis niloticus	AERX01063490.2
	AERX01067013.1
	AERX01066994.2
Amphilophus citrinellus	CCOE01000729.1
Boleophthalmus pectinirostris	JACK01087215.1
Nothobranchius furzeri	JNBZ01173257.1
	JNBZ01073941.1
Periophthalmus magnuspinnatus	JACL01023553.1
Pundamilia nyererei	AFNX01063967.1
Cynoglossus semilaevis	AGRG01030024.1
	AGRG01009146.1
	AGRG01017733.1
	AGRG01009869.1
Larimichthys crocea	JRPU01020623.1
	JRPU01005872.1
	JRPU01018873.1
	JRPU01000937.1
Salmo salar	AGKD01000001.1
Oncorhynchus mykiss	CCAF010051366.1
	CCAF010058454.1
Oryzias latipes	BAAE01103386.1
	BAAF04126113.1
	BAAE01009841.1
	BAAF04111357.1
	BAAF04126497.1
	BAAE01297766.1
	BAAE01052622.1
	BAAE01092067.1
Esox lucius	AZJR02000460.1
	AZJR02001005.1

First round tBLASTn									Second Round BLASTn						
Used as query for second blast	Initial shortlist	Description	Max score	Total score	Query coverage	E value	Identitiy	Accession	Description	Max score	Total score	Query coverage	E value	Identitiy	Accession
	x	Oryzias latipes DNA, contig103270 in scaffold4278, strain: HNI, whole genome shotgun sequence	397	397	92%	3.00E-109	27%	BAAE01103386.1	Amphilophus citrinellus genome assembly Midas_v5, contig scaffold-728, whole genome shotgun sequence	2683	2683	100%	0	99%	CCOE01000729.1
	x	Oryzias latipes DNA, contig126113 in scaffold3897, strain: Hd-rR, whole genome shotgun sequence	385	385	86%	5.00E-107	28%	BAAF04126113.1	Periophthalmus magnuspinnatus scaffold1146_10, whole genome shotgun sequence	1904	1904	100%	0	69%	JACL01023553.1
	x	Periophthalmus magnuspinnatus scaffold1146_10, whole genome shotgun sequence	395	395	85%	5.00E-109	27%	JACL01023553.1	Salmo salar breed double haploid jcf1001231046_0_0, whole genome shotgun sequence	1859	1859	100%	0	66%	AGKD04031988.1
		Larimichthys crocea contig18883, whole genome shotgun sequence	318	361	80%	1.00E-84	29%	JRPU01018873.1	Larimichthys crocea contig20633, whole genome shotgun sequence	1735	1735	100%	0	63%	JRPU01020623.1
		Larimichthys crocea contig5874, whole genome shotgun sequence	355	355	80%	3.00E-96	28%	JRPU01005872.1	Larimichthys crocea contig5874, whole genome shotgun sequence	1.73E+03	173400%	100%	0	62%	JRPU01005872.1
	x	Oreochromis niloticus contig063490, whole genome shotgun sequence	299	405	80%	1.00E-102	28%	AERX01063490.2	Nothobranchius furzeri breed GRZ contig_37132, whole genome shotgun sequence	1694	169400%	100%	0%	62%	JNBZ01173257.1
x	x	Oryzias latipes DNA, contig9841 in scaffold204, strain: HNI, whole genome shotgun sequence	363	363	80%	7.00E-99	27%	BAAE01009841.1	Pundamilia nyererei contig063967, whole genome shotgun sequence	1669	1669	93%	0	65%	AFNX01063967.1
	x	Oryzias latipes DNA, contig111357 in scaffold1115, strain: Hd-rR, whole genome shotgun sequence	358	358	80%	4.00E-97	27%	BAAF04111357.1	Oryzias latipes DNA, contig111357 in scaffold1115, strain: Hd-rR, whole genome shotgun sequence	1661	1661	100%	0	61%	BAAF04111357.1
	x	Larimichthys crocea contig20633, whole genome shotgun sequence	358	358	80%	6.00E-97	27%	JRPU01020623.1	Oryzias latipes DNA, contig103270 in scaffold4278, strain: HNI, whole genome shotgun sequence	1.66E+03	166100%	99%	0	61%	BAAE01103386.1
	x	Nothobranchius furzeri breed GRZ contig_37132, whole genome shotgun sequence	368	368	80%	4.00E-100	26%	JNBZ01173257.1	Oryzias latipes DNA, contig126113 in scaffold3897, strain: Hd-rR, whole genome shotgun sequence	1.66E+03	166000%	99%	0	61%	BAAF04126113.1
		Larimichthys crocea contig937, whole genome shotgun sequence	193	357	79%	4.00E-88	30%	JRPU01000937.1	Oryzias latipes DNA, contig9841 in scaffold204, strain: HNI, whole genome shotgun sequence	1.66E+03	165800%	100%	0	61%	BAAE01009841.1
x	x	Amphilophus citrinellus genome assembly Midas_v5, contig scaffold-728, whole genome shotgun sequence	356	356	77%	2.00E-96	26%	CCOE01000729.1	Salmo salar breed double haploid tcf1000000046_0-0_ssa08_2_0, whole genome shotgun sequence	1.53E+03	183000%	99%	0	66%	AGKD04000303.1
	x	Salmo salar breed double haploid jcf1001231046_0_0, whole genome shotgun sequence	377	377	76%	7.00E-104	27%	AGKD04031988.1	Salmo salar breed double haploid jcf1000675607_0-0_ssa10_0_0, whole genome shotgun sequence	1.45E+03	194200%	98%	0	66%	AGKD04000114.1
	x	Pundamilia nyererei contig063967, whole genome shotgun sequence	356	356	68%	1.00E-98	28%	AFNX01063967.1	Larimichthys crocea contig18883, whole genome shotgun sequence	1422	1725	99%	0%	64%	JRPU01018873.1
x	x	Boleophthalmus pectinirostris scaffold1197_21, whole genome shotgun sequence	151	705	68%	3.00E-33	24%	JACK01087215.1	Salmo salar breed double haploid jcf1000081466_0_0, whole genome shotgun sequence	1321	1321	69%	0	68%	AGKD04026083.1
		Salmo salar breed double haploid jcf1004662566_0-0_ssa27_0_0, whole genome shotgun sequence	263	432	67%	3.00E-67	29%	AGKD04000107.1	Salmo salar breed double haploid jcf1000035191_0-0_ssa16-ssa17_0_0, whole genome shotgun sequence	1.24E+03	124100%	67%	0	67%	AGKD04000855.1
		Salmo salar breed double haploid tcf2000000048_0-0_ssa06_0_0, whole genome shotgun sequence	119	381	67%	3.00E-44	26%	AGKD04000476.1	Oreochromis niloticus contig063490, whole genome shotgun sequence	1.18E+03	184700%	99%	0	70%	AERX01063490.2
		Salmo salar breed double haploid tcf1000000046_0-0_ssa08_2_0, whole genome shotgun sequence	317	317	66%	5.00E-84	28%	AGKD04000303.1	Salmo salar breed double haploid jcf1004662566_0-0_ssa27_0_0, whole genome shotgun sequence	1159	1823	84%	0	70%	AGKD04000107.1
		Salmo salar breed double haploid jcf1000675607_0-0_ssa10_0_0, whole genome shotgun sequence	328	328	65%	1.00E-87	28%	AGKD04000114.1	Salmo salar breed double haploid tcf2000000012_0-0_ssa09_1560_0, whole genome shotgun sequence	1053	1053	53%	0	70%	AGKD04000080.1
	x	Esox lucius Scaffold649, whole genome shotgun sequence	92.8	318	63%	2.00E-15	26%	AZJR02000460.1	Larimichthys crocea contig937, whole genome shotgun sequence	978	1723	100%	0	64%	JRPU01000937.1
	x	Oncorhynchus mykiss WGS project data, contig 51366, whole genome shotgun sequence	192	338	60%	1.00E-45	30%	CCAF010051366.1	Oryzias latipes DNA, contig126497 in scaffold4023, strain: Hd-rR, whole genome shotgun sequence	804	804	48%	0	63%	BAAF04126497.1
		Salmo salar breed double haploid jcf1000081466_0_0, whole genome shotgun sequence	277	277	54%	6.00E-72	28%	AGKD04026083.1	Oncorhynchus mykiss WGS project data, contig 51366, whole genome shotgun sequence	7.26E+02	152800%	83%	0	70%	CCAF010051366.1
		Salmo salar breed double haploid jcf1000035191_0-0_ssa16-ssa17_0_0, whole genome shotgun sequence	266	266	52%	3.00E-68	28%	AGKD04000855.1	Boleophthalmus pectinirostris scaffold1197_21, whole genome shotgun sequence	702	2444	98%	0%	50%	JACK01087215.1
		Salmo salar breed double haploid tcf2000000012_0-0_ssa09_1560_0, whole genome shotgun sequence	243	243	44%	2.00E-61	29%	AGKD04000080.1	Pundamilia nyererei contig052002, whole genome shotgun sequence	682	682	40%	0	60%	AFNX01062002.1
		Oreochromis niloticus contig067013, whole genome shotgun sequence	171	265	42%	8.00E-41	32%	AERX01067013.1	Neolamprologus brichardi contig056855, whole genome shotgun sequence	679	679	33%	0	71%	AFNY01056855.1
		Oryzias latipes DNA, contig126497 in scaffold4023, strain: Hd-rR, whole genome shotgun sequence	206	206	37%	3.00E-52	30%	BAAF04126497.1	Esox lucius Scaffold649, whole genome shotgun sequence	674	1502	99%	0	47%	AZJR02000460.1
		Oncorhynchus mykiss WGS project data, contig 58454, whole genome shotgun sequence	97.8	265	35%	2.00E-54	35%	CCAF010058454.1	Esox lucius Scaffold1210, whole genome shotgun sequence	6.72E+02	80900%	55%	0	57%	AZJR02001005.1
		Nothobranchius furzeri breed GRZ contig_167764, whole genome shotgun sequence	169	383	33%	1.00E-38	33%	JNBZ01073941.1	Oreochromis niloticus contig066994, whole genome shotgun sequence	632	632	34%	0	67%	AERX01066994.2
		Cynoglossus semilaevis scaffold1781_2, whole genome shotgun sequence	77.4	261	33%	4.00E-19	29%	AGRG01030024.1	Melanochromis auratus strain Domwe Island MAURA1011390_1, whole genome shotgun sequence	626	626	31%	0%	72%	ABPL01018272.1
		Esox lucius Scaffold1210, whole genome shotgun sequence	152	152	29%	2.00E-33	31%	AZJR02001005.1	Oreochromis niloticus contig067013, whole genome shotgun sequence	613	103300%	53%	0%	73%	AERX01067013.1
		Neolamprologus brichardi contig056855, whole genome shotgun sequence	190	190	28%	4.00E-48	30%	AFNY01056855.1	Oreochromis niloticus contig071449, whole genome shotgun sequence	6.00E+02	60000%	36%	0	57%	AERX01071449.2
		Oreochromis niloticus contig066994, whole genome shotgun sequence	134	134	28%	2.00E-29	27%	AERX01066994.2	Salmo salar breed double haploid tcf2000000048_0-0_ssa06_0_0, whole genome shotgun sequence	5.60E+02	181900%	99%	0	65%	AGKD04000476.1
		Melanochromis auratus strain Domwe Island MAURA1011390_1, whole genome shotgun sequence	189	189	25%	3.00E-48	32%	ABPL01018272.1	Oncorhynchus mykiss WGS project data, contig 58454, whole genome shotgun sequence	3.12E+02	98700%	51%	0	64%	CCAF010058454.1
		Pundamilia nyererei contig062002, whole genome shotgun sequence	96.3	96.3	25%	5.00E-17	23%	AFNX01062002.1	Nothobranchius furzeri breed GRZ contig_167764, whole genome shotgun sequence	5.12E+02	194800%	53%	3.00E-180	66%	JNBZ01073941.1
		Salmo salar breed double haploid jcf1003466616_0_0, whole genome shotgun sequence	55.5	106	24%	2.00E-13	30%	AGKD04055112.1	Salmo salar breed double haploid jcf1003466616_0_0, whole genome shotgun sequence	4.41E+02	62400%	35%	1.00E-168	61%	AGKD04055112.1
		Salmo salar breed double haploid jcf1000499427_0_0, whole genome shotgun sequence	135	199	23%	4.00E-41	31%	AGKD04197423.1	Salmo salar breed double haploid jcf1000499427_0_0, whole genome shotgun sequence	480	638	30%	1.00E-153	67%	AGKD04197423.1
		Oreochromis niloticus contig071449, whole genome shotgun sequence	93.2	93.2	23%	4.00E-16	25%	AERX01071449.2	Salmo salar breed double haploid jcf1000986304_0_0, whole genome shotgun sequence	475	475	23%	4.00E-153	72%	AGKD04281921.1
		Oryzias latipes DNA, contig126496 in scaffold4023, strain: Hd-rR, whole genome shotgun sequence	63.9	115	21%	4.00E-07	39%	BAAF04126496.1	Salmo salar breed double haploid ccf1000000011_0-0_ssa14_0_0, whole genome shotgun sequence	483	483	28%	2.00E-138	60%	AGKD04000035.1
		Cynoglossus semilaevis scaffold606_2, whole genome shotgun sequence	122	122	20%	1.00E-24	31%	AGRG01017733.1	Oreochromis niloticus contig066995, whole genome shotgun sequence	4.49E+02	44900%	27%	7.00E-138	58%	AERX01066995.2
		Cynoglossus semilaevis scaffold291_6, whole genome shotgun sequence	69.7	298	19%	2.00E-08	33%	AGRG01009146.1	Salmo salar breed double haploid jcf1000070863_0_0, whole genome shotgun sequence	381	439	25%	1.00E-117	60%	AGKD04239310.1
		Salmo salar breed double haploid jcf1000986304_0_0, whole genome shotgun sequence	133	133	19%	6.00E-30	31%	AGKD04281921.1	Cynoglossus semilaevis scaffold606_2, whole genome shotgun sequence	3.80E+02	507000%	35%	4.00E-112	56%	AGRG01017733.1
		Salmo salar breed double haploid ccf1000000011_0-0_ssa14_0_0, whole genome shotgun sequence	91.3	91.3	19%	5.00E-15	26%	AGKD04000035.1	Cynoglossus semilaevis scaffold1781_2, whole genome shotgun sequence	292	1090	66%	1.00E-111	57%	AGRG01030024.1
		Salmo salar breed double haploid tcf1000000134_0-0_ssa11-ssa26_0_0, whole genome shotgun sequence	79.7	79.7	17%	2.00E-11	25%	AGKD04000842.1	Salmo salar breed double haploid tcf1000000134_0-0_ssa11-ssa26_0_0, whole genome shotgun sequence	3.97E+02	39700%	23%	2.00E-110	62%	AGKD04000842.1
		Cynoglossus semilaevis scaffold315_3, whole genome shotgun sequence	56.6	107	15%	2.00E-04	27%	AGRG01009869.1	Oryzias latipes DNA, contig11641 in scaffold251, strain: HNI, whole genome shotgun sequence	364	364	22%	8.00E-110	55%	BAAE01011641.1
		Oreochromis niloticus contig066995, whole genome shotgun sequence	54.7	54.7	15%	4.00E-04	23%	AERX01066995.2	Salmo salar breed double haploid jcf1000223190_0_0, whole genome shotgun sequence	348	348	20%	1.00E-106	61%	AGKD04048741.1
		Salmo salar breed double haploid jcf1000094186_0-870771_ssa23_0_0, whole genome shotgun sequence	85.1	85.1	13%	4.00E-13	30%	AGKD04000549.1	Salmo salar breed double haploid jcf1000894696_0_0, whole genome shotgun sequence	2.95E+02	40600%	21%	5.00E-103	66%	AGKD04126327.1
		Salmo salar breed double haploid jcf1003038296_0_0, whole genome shotgun sequence	82	82	13%	1.00E-13	29%	AGKD04759675.1	Salmo salar breed double haploid jcf1000971150_0_0, whole genome shotgun sequence	3.35E+02	37800%	21%	4.00E-101	59%	AGKD04300645.1
		Salmo salar breed double haploid jcf1000070863_0_0, whole genome shotgun sequence	80.1	80.1	13%	1.00E-12	28%	AGKD04239310.1	Melanochromis auratus strain Domwe Island MAURA1007280_1, whole genome shotgun sequence	339	465	27%	4.00E-100	60%	ABPL01012874.1
		Pampus argenteus breed Wild-type scaffold266821, whole genome shotgun sequence	72.4	115	12%	5.00E-16	32%	JHEK01266819.1	Cynoglossus semilaevis scaffold291_6, whole genome shotgun sequence	218	1179	52%	1.00E-98	66%	AGRG01009146.1
		Salmo salar breed double haploid jcf1000863852_0_0, whole genome shotgun sequence	82.8	82.8	12%	2.00E-12	31%	AGKD04014003.1	Melanochromis auratus strain Domwe Island MAURA1054855_1, whole genome shotgun sequence	311	311	17%	2.00E-93	62%	ABPL01073471.1
		Salmo salar breed double haploid jcf1000971150_0_0, whole genome shotgun sequence	70.1	70.1	12%	2.00E-09	30%	AGKD04300645.1	Salmo salar breed double haploid jcf1000094186_0-870771_ssa23_0_0, whole genome shotgun sequence	337	337	18%	5.00E-91	66%	AGKD04000549.1
		Salmo salar breed double haploid jcf1000223190_0_0, whole genome shotgun sequence	72	72	12%	3.00E-10	29%	AGKD04048741.1	Salmo salar breed double haploid jcf1000729702_0_0, whole genome shotgun sequence	299	299	17%	3.00E-89	64%	AGKD04644989.1
		Salmo salar breed double haploid jcf1000677822_0_0, whole genome shotgun sequence	75.1	75.1	12%	2.00E-11	28%	AGKD04786333.1	Salmo salar breed double haploid jcf1003038296_0_0, whole genome shotgun sequence	291	29100%	15%	5.00E-87	68%	AGKD04759675.1
		Oryzias latipes DNA, contig11641 in scaffold251, strain: HNI, whole genome shotgun sequence	54.3	54.3	12%	4.00E-04	27%	BAAE01011641.1	Salmo salar breed double haploid jcf1001205191_0_0, whole genome shotgun sequence	277	351	18%	2.00E-86	60%	AGKD04520858.1
		Salmo salar breed double haploid jcf1000052362_0-0_ssa05_0_0, whole genome shotgun sequence	73.6	73.6	11%	1.00E-09	27%	AGKD04001468.1	Oryzias latipes DNA, contig126496 in scaffold4023, strain: Hd-rR, whole genome shotgun sequence	294	481	29%	2.00E-85	61%	BAAF04126496.1
		Salmo salar breed double haploid jcf1000894696_0_0, whole genome shotgun sequence	42	42	11%	350%	23%	AGKD04126327.1	Salmo salar breed double haploid jcf1000052362_0-0_ssa05_0_0, whole genome shotgun sequence	311	779	31%	1.00E-82	56%	AGKD04001468.1
		Salmo salar breed double haploid jcf1000672973_0_0, whole genome shotgun sequence	100	100	10%	3.00E-18	36%	AGKD04070803.1	Salmo salar breed double haploid jcf1000337496_0_0, whole genome shotgun sequence	267	267	15%	6.00E-77	61%	AGKD04268163.1
		Oryzias latipes DNA, contig92110 in scaffold3621, strain: HNI, whole genome shotgun sequence	73.6	73.6	10%	2.00E-10	31%	BAAE01092067.1	Salmo salar breed double haploid jcf1000672973_0_0, whole genome shotgun sequence	278	278	12%	9.00E-77	80%	AGKD04070803.1
		Oryzias latipes DNA, contig297681 in scaffold36800, strain: HNI, whole genome shotgun sequence	52	52	10%	0%	28%	BAAE01297766.1	Salmo salar breed double haploid jcf1000677822_0_0, whole genome shotgun sequence	260	260	14%	4.00E-76	66%	AGKD04786333.1
		Salmo salar breed double haploid jcf1000584432_0_0, whole genome shotgun sequence	69.3	69.3	10%	1.00E-08	27%	AGKD04082652.1	Oryzias latipes DNA, contig297681 in scaffold36800, strain: HNI, whole genome shotgun sequence	260	417	22%	2.00E-74	59%	BAAE01297766.1
		Salmo salar breed double haploid jcf1000729702_0_0, whole genome shotgun sequence	76.3	76.3	9%	1.00E-11	32%	AGKD04644989.1	Pampus argenteus breed Wild-type scaffold296198, whole genome shotgun sequence	199	393	34%	2.00E-72	40%	JHEK01296196.1
		Pampus argenteus breed Wild-type scaffold296198, whole genome shotgun sequence	48.1	71.2	8%	5.00E-03	42%	JHEK01296196.1	Pampus argenteus breed Wild-type scaffold266821, whole genome shotgun sequence	179	296	16%	3.00E-70	59%	JHEK01266819.1
		Pampus argenteus breed Wild-type scaffold219622, whole genome shotgun sequence	60.5	60.5	7%	9.00E-07	34%	JHEK01219620.1	Salmo salar breed double haploid jcf1000584432_0_0, whole genome shotgun sequence	242	242	13%	5.00E-65	67%	AGKD04082652.1
		Salmo salar breed double haploid jcf1000981944_0_0, whole genome shotgun sequence	61.2	61.2	5%	2.00E-06	41%	AGKD04241185.1	Salmo salar breed double haploid jcf1000863852_0_0, whole genome shotgun sequence	253	25300%	15%	7.00E-65	63%	AGKD04014003.1
		Salmo salar breed double haploid jcf1000267125_0_0, whole genome shotgun sequence	58.9	58.9	5%	6.00E-06	35%	AGKD04414339.1	Cynoglossus semilaevis scaffold315_3, whole genome shotgun sequence	225	52200%	32%	7.00E-64	61%	AGRG01009869.1
		Pimephales promelas scaffold-189977, whole genome shotgun sequence	41.6	41.6	3%	6.3	35%	JNCE01189963.1	Oryzias latipes DNA, contig126500 in scaffold4023, strain: Hd-rR, whole genome shotgun sequence	211	211	13%	7.00E-56	53%	BAAF04126500.1
		Pimephales promelas scaffold514, whole genome shotgun sequence	41.6	41.6	3%	6.3	35%	JNCD01000512.1	Salmo salar breed double haploid jcf1000948982_0_0, whole genome shotgun sequence	203	203	10%	3.00E-52	64%	AGKD04099775.1
									Salmo salar breed double haploid jcf1000901130_0_0, whole genome shotgun sequence	195	19500%	10%	5.00E-51	60%	AGKD04228511.1
									Oryzias latipes DNA, contig92110 in scaffold3621, strain: HNI, whole genome shotgun sequence	192	19200%	13%	2.00E-50	57%	BAAE01092067.1
									Oreochromis niloticus contig067011, whole genome shotgun sequence	196	196	11%	2.00E-49	56%	AERX01067011.1
									Cynoglossus semilaevis scaffold81_12, whole genome shotgun sequence	118	226	12%	5.00E-49	65%	AGRG01002782.1
									Salmo salar breed double haploid jcf1000267125_0_0, whole genome shotgun sequence	184	184	9%	2.00E-48	70%	AGKD04414339.1
									Pampus argenteus breed Wild-type scaffold219622, whole genome shotgun sequence	174	277	13%	1.00E-45	66%	JHEK01219620.1
									Neolamprologus brichardi contig056854, whole genome shotgun sequence	137	137	8%	3.00E-33	52%	AFNY01056854.1
									Nothobranchius furzeri breed GRZ contig_167765, whole genome shotgun sequence	139	204	10%	5.00E-30	60%	JNBZ01073942.1
									Salmo salar breed double haploid jcf1001055536_0_0, whole genome shotgun sequence	139	229	12%	5.00E-30	55%	AGKD04032731.1
									Esox lucius Scaffold445, whole genome shotgun sequence	136	281	17%	1.00E-28	69%	AZJR02000277.1
									Salmo salar breed double haploid jcf1000871019_0_0, whole genome shotgun sequence	117	154	9%	1.00E-27	53%	AGKD04788890.1
									Salmo salar breed double haploid jcf1000981944_0_0, whole genome shotgun sequence	126	126	5%	1.00E-27	82%	AGKD04241185.1
									Esox lucius Scaffold650, whole genome shotgun sequence	126	126	8%	3.00E-27	52%	AZJR02000461.1
									Salmo salar breed double haploid jcf1000526599_0_0, whole genome shotgun sequence	130	130	8%	4.00E-27	52%	AGKD04046694.1
									Salmo salar breed double haploid jcf1000240329_0_0, whole genome shotgun sequence	121	121	6%	1.00E-26	55%	AGKD04582342.1
									Oryzias latipes DNA, contig257531 in scaffold22450, strain: HNI, whole genome shotgun sequence	107	107	9%	5.00E-22	43%	BAAE01257592.1
									Pampus argenteus breed Wild-type scaffold6215, whole genome shotgun sequence	76.3	207	13%	6.00E-22	73%	JHEK01006214.1
									Pundamilia nyererei contig065325, whole genome shotgun sequence	107	107	6%	1.00E-20	52%	AFNX01065325.1
									Oryzias latipes DNA, contig334693 in scaffold70980, strain: HNI, whole genome shotgun sequence	97.1	97.1	5%	1.00E-19	60%	BAAE01334625.1
									Pampus argenteus breed Wild-type scaffold283868, whole genome shotgun sequence	79	79	5%	2.00E-11	47%	JHEK01283866.1
									Melanochromis auratus strain Domwe Island MAURA1003503_1, whole genome shotgun sequence	73.9	73.9	3%	5.00E-10	68%	ABPL01007469.1
									Oryzias latipes DNA, contig113468 in scaffold1340, strain: Hd-rR, whole genome shotgun sequence	70.1	70.1	4%	4.00E-09	48%	BAAF04113468.1
									Pampus argenteus breed Wild-type scaffold244281, whole genome shotgun sequence	59.7	59.7	6%	4.00E-06	42%	JHEK01244279.1
									Oryzias latipes DNA, contig307647 in scaffold44110, strain: HNI, whole genome shotgun sequence	59.3	59.3	4%	8.00E-06	50%	BAAE01307755.1
									Scartelaos histophorus scaffold95395_1, whole genome shotgun sequence	45.1	45.1	4%	0.31	44%	JACN01118715.1
									Boleophthalmus pectinirostris scaffold7332_1, whole genome shotgun sequence	42.7	42.7	5%	0.86	32%	JACK01099270.1
									Salmo salar breed double haploid jcf1000109986_0_0, whole genome shotgun sequence	42.7	42.7	1%	1.8	67%	AGKD04080171.1
									Oryzias latipes DNA, contig52599 in scaffold1692, strain: HNI, whole genome shotgun sequence	42	42	2%	1.9	74%	BAAE01052622.1
									Salmo salar breed double haploid jcf1000732633_0-0_ssa23_0_0, whole genome shotgun sequence	42.7	42.7	1%	2.5	67%	AGKD04000443.1
									Scleropages formosus scaffold443, whole genome shotgun sequence	42	42	8%	4.6	29%	JARO01000443.1

ORF	Start	Stop	strand	length	Description	species	viral group	Coverage	E value	Identity	Accession
1	2191	2538	+	348							
2	12536	13051	+	516	PREDICTED: piggyBac transposable element-derived protein 4-like	Takifugu rubripes	TE	71%	1.00E-64	81%	XP_011605214.1
3	13292	13678	-	387	transcriptional regulator, LuxR family	Klebsiella pneumoniae		16%	1.00E-05	100%	EJK89612.1
4	13752	14381	+	630	PREDICTED: piggyBac transposable element-derived protein 4-like	Poecilia formosa	TE	100%	3.00E-97	70%	XP_007576285.1
5	15078	15869	+	792							
6	16415	18445	+	2031	ORF77L	Turbot reddish body iridovirus	Iridoviridae; Megalocytivirus;	30%	2.00E-08	28%	ADE34421.1
					VP80	Buzura suppressaria nucleopolyhedrovirus	Baculoviridae; Alphabaculovirus	14%	2.00E-06	32%	YP_009001857.1
7	19742	20869	-	1128							
8	20868	21314	+	447							
9	21259	22167	+	909	ORF60	Siberian sturgeon herpesvirus	Alloherpesviridae; Ictalurivirus;	81%	4.00E-25	27%	ADC79712.1
10	22491	24467	+	1977	ORF54	Acipenserid herpesvirus 2	Alloherpesviridae; Ictalurivirus	55%	3.00E-10	24%	AEF97714.1
11	26037	26354	-	318							
12	26341	26781	-	441							
13	27011	27364	-	354							
14	27742	28824	-	1083	membrane protein ORF147B	Cyprinid herpesvirus 1	Alloherpesviridae; Cyprinivirus	62%	4.00E-17	29%	YP_007003802.1
15	29245	31062	+	1818	membrane protein ORF147A	Cyprinid herpesvirus 1	Alloherpesviridae; Cyprinivirus	59%	5.00E-73	39%	YP_007003801.1
16	31624	32331	+	708							
17	32626	34053	+	1428	contains TOPRIM domain, potential nuclease COG4026						
18	34134	35927	+	1794	ORF25	Acipenserid herpesvirus 2	Alloherpesviridae; Ictalurivirus	93%	5.00E-47	27%	AEF97684.1
19	36048	36479	+	432							
20	36556	37425	+	870	ORF27	Ictalurid herpesvirus 1	Alloherpesviridae; Ictalurivirus	39%	8.00E-06	29%	NP_041118.1
21	37679	38146	-	468							
22	38426	38953	-	528							
23	38958	39557	-	600							
24	39648	40250	-	603	Baculoviral inhibition of apoptosis protein						
25	40338	40706	-	369	unnamed protein product Oncorhynchus mykiss	173	173	69%	CDQ74451.1		
26	41024	43027	-	2004	ORF88	Ranid herpesvirus 2	Alloherpesviridae; Batrachovirus	26%	1.00E-13	30%	YP_656596.1
27	43070	43654	-	585							
28	43846	44535	+	690							
29	44680	48699	+	4020	ORF46	Acipenserid herpesvirus 2	Alloherpesviridae; Ictalurivirus	55%	3.00E-97	31%	AEF97704.1
30	49111	50007	-	897							
31	51851	52762	-	912	ORF62	Ictalurid herpesvirus 1	Alloherpesviridae; Ictalurivirus	44%	3.00E-27	44%	NP_041153.2
32	52949	53581	+	633							
33	53777	54754	+	978							
34	55096	55857	-	762							
35	56235	56756	+	522							
36	56917	57531	+	615							
37	57904	58443	+	540							
38	58515	59039	+	525	GM22539	Drosophila	sechellia	98%	3.00E-14	33%	XP_002044161.1
39	59053	59850	-	798	ORF68	Ranid herpesvirus 2	Alloherpesviridae; Batrachovirus	96%	1.00E-13	25%	YP_656576.1
40	59960	60748	+	789	Cbp/p300-interacting transactivator 1	Larimichthys crocea	345	6.00E-112	66%	KKF33343.1	
41	60671	61291	-	621	terminase	Acipenserid herpesvirus 2	Alloherpesviridae; Ictalurivirus	80%	2.00E-37	42%	ABQ10592.1
42	61547	62764	+	1218							
43	63410	64909	-	1500	ORF64	Ictalurid herpesvirus 1	Alloherpesviridae; Ictalurivirus	70%	3.00E-13	23%	NP_041156.1
44	64906	66438	-	1533							
45	66534	70034	-	3501							
46	70101	71138	-	1038							
47	71190	72710	+	1521							
48	72797	76642	+	3846							
49	76883	81688	-	4806	putative WD repeat-containing protein	Acanthamoeba polyphaga mimivirus	Mimiviridae; Mimivirus	5%	2.00E-08	44%	AKI79997.1
50	82090	83391	-	1302							
51	83390	84811	+	1422	PREDICTED: uncharacterized protein LOC101156739	Oryzias latipes		98%	3.00E-149	50%	XP_004071716.2
52	85368	85817	+	450							
53	87491	89581	+	2091							
54	89656	90690	+	1035							
55	90754	91461	+	708							
56	93004	93567	-	564							
57	93651	94307	-	657							
58	94340	97051	-	2712	ORF39	Ictalurid herpesvirus 1	Alloherpesviridae; Ictalurivirus	89%	3.00E-34	24%	NP_041130.1
59	97203	97610	-	408							
60	97710	100127	+	2418	ORF37	Acipenserid herpesvirus 2	Alloherpesviridae; Ictalurivirus	18%	9.00E-12	26%	AEF97695.1
61	101295	102137	-	843	PREDICTED: uncharacterized protein LOC105358139	Oryzias latipes		24%	5.00E-15	57%	XP_011491977.1
62	102194	103447	-	1254	ORF34	Acipenserid herpesvirus 2	Alloherpesviridae; Ictalurivirus	76%	9.00E-45	37%	AEF97692.1
63	103476	104327	-	852							
64	104359	105141	+	783							
65	105391	108120	-	2730	hypothetical protein EH28_02830	Larimichthys crocea		53%	3.00E-165	55%	KKF18232.1
66	108175	109683	+	1509	ORF44	Acipenserid herpesvirus 2	Alloherpesviridae; Ictalurivirus	42%	6.00E-20	31%	AEF97702.1
67	109702	110472	+	771							
68	110679	112529	+	1851							
69	112465	114960	+	2496	PREDICTED: GRIP and coiled-coil domain-containing protein 2-like	Neolamprologus brichardi		38%	5.00E-15	26%	XP_006800602.1
70	115269	115748	-	480	PREDICTED: uncharacterized protein LOC102076604	Oreochromis niloticus		67%	4.00E-08	32%	XP_005466037.1
71	115726	116259	-	534	hypothetical protein EH28_02830	Larimichthys crocea		99%	7.00E-36	42%	KKF18232.1
72	116382	116840	-	459							
73	118811	119173	-	363							
74	119483	120127	-	645							
75	120541	121056	+	516							
76	121457	122122	-	666	N1R/p28 family protein	Pigeonpox virus	Poxviridae; Avipoxvirus	42%	1.00E-16	43%	YP_009046385.1
77	122426	123268	-	843							
78	123341	127252	+	3912	ORF73	Ictalurid herpesvirus 1	Alloherpesviridae; Ictalurivirus	17%	3.00E-11	29%	NP_041163.1
79	127359	128291	+	933	C19 subfamily protein	Moumouvirus goulette	Mimiviridae; Moumouvirus	50%	2.00E-06	31%	AGF85227.1
80	128817	132776	-	3960	PREDICTED: uncharacterized protein LOC106024916	Esox lucius		31%	3.00E-37	51%	XP_012995787.1
81	133060	133650	+	591							
82	133672	134193	+	522							
83	135033	135674	-	642							
84	135720	136187	-	468							
85	136354	137799	-	1446							
86	137735	139384	-	1650	ORF57	Ictalurid herpesvirus 1	Alloherpesviridae; Ictalurivirus	90%	8.00E-52	30%	NP_041148.2
87	139507	141327	-	1821	ORF57	Ictalurid herpesvirus 1	Alloherpesviridae; Ictalurivirus	89%	2.00E-32	26%	NP_041148.2
88	142174	142734	+	561							
89	143018	146236	+	3219	ORF56	Ictalurid herpesvirus 1	Alloherpesviridae; Ictalurivirus	48%	2.00E-46	27%	NP_041147.1
90	146764	147327	+	564							
91	147757	148452	-	696	PREDICTED: type-2 angiotensin II receptor-like	Oryzias latipes		90%	5.00E-35	35%	XP_004084729.1
92	148493	148897	-	405	PREDICTED: uncharacterized protein LOC105358166	Oryzias latipes		85%	3.00E-14	39%	XP_011492012.1
93	149960	150628	+	669	hypothetical protein	European catfish virus	Iridoviridae; Ranavirus	80%	3.00E-14	30%	YP_006347673.1
94	150832	151452	+	621							
95	151772	152560	+	789							
96	153056	153838	+	783	PREDICTED: uncharacterized protein LOC102080423	Oreochromis niloticus		76%	2.00E-34	38%	XP_005466421.1
97	154332	155282	-	951	putative serine/threonine-protein kinase	Megavirus courdo11	Mimiviridae; Megavirus chiliensis	68%	2.00E-17	28%	AFX92917.1
98	156112	156462	-	351							
99	156677	157438	-	762							
100	157862	158488	+	627							
101	159447	160637	+	1191							
102	161454	161774	-	321	hypothetical protein LDVICp188	Lymphocystis disease virus - isolate China	Iridoviridae; Lymphocystivirus	81%	3.00E-14	36%	YP_073693.1
103	162613	167430	-	4818							
104	168222	169493	-	1272	PREDICTED: uncharacterized protein LOC105358165	Oryzias latipes		56%	2.00E-25	32%	XP_011492011.1
105	169982	171052	-	1071	protein ORF35	Anguillid herpesvirus 1	Alloherpesviridae; Cyprinivirus	63%	3.00E-11	25%	YP_003358174.1
106	171151	171711	-	561							
107	172198	173919	-	1722							
108	174414	178016	-	3603							
109	178102	178962	+	861	DNA methyltransferase	Lymphocystis disease virus 1	Iridoviridae; Lymphocystivirus	90%	3.00E-10	25%	NP_078617.1
110	179170	180348	-	1179	ORF63	Acipenserid herpesvirus 2	Alloherpesviridae; Ictalurivirus	99%	8.00E-19	25%	AEF97717.1
111	180452	181432	-	981							
112	182030	182761	-	732							
113	183961	184263	+	303							
114	185640	186749	+	1110	PREDICTED: piggyBac transposable element-derived protein 4-like	Takifugu rubripes		50%	2.00E-83	71%	XP_011605214.1
115	187403	188329	+	927	PREDICTED: piggyBac transposable element-derived protein 4-like	Esox lucius		100%	2.00E-148	71%	XP_012987620.1

ORF	Start	Stop	strand	length	Description	species	Coverage	E value	Identity	Accession
1	1065	1454	+	390						
2	4701	5372	-	672						
3	5461	6138	-	678						
4	7870	8460	-	591						
5	16224	17585	-	1362	PREDICTED: uncharacterized protein LOC101156739	Oryzias latipes	96%	5.00E-120	42%	XP_004071716.2
6	17521	18945	+	1425						
7	19436	19789	+	354						
8	23588	24145	-	558	PREDICTED: LOW QUALITY PROTEIN: cadherin-23	Astyanax mexicanus	61%	1.00E-32	58%	XP_007238651.1
9	24232	28383	+	4152						
10	28540	28884	-	345						
11	29245	29580	-	336						
12	30801	31169	+	369						
13	32303	32962	-	660						
14	32961	34238	+	1278	ORF60	Ictalurid herpesvirus 1	44%	5.00E-23	31%	NP_041151.1
15	34314	36155	+	1842	ORF54	Acipenserid herpesvirus 2	64%	1.00E-07	22%	AEF97714.1
16	36756	37085	-	330						
17	37655	38926	+	1272						
18	38945	40690	+	1746	ORF25	Acipenserid herpesvirus 2	94%	4.00E-49	27%	AEF97684.1
19	40799	41167	+	369						
20	41181	42047	+	867	ORF27	Acipenserid herpesvirus 2	38%	1.00E-08	32%	AEF97686.1
21	42144	42587	-	444						
22	42779	43303	-	525						
23	43307	43900	-	594						
24	43964	44962	-	999	unnamed protein product	Oncorhynchus mykiss	54%	5.00E-39	47%	CDQ74451.1
25	45094	46950	-	1857	ORF88	Ranid herpesvirus 2	28%	4.00E-15	31%	YP_656596.1
26	46971	47501	-	531						
27	47669	48271	+	603						
28	48382	52365	+	3984	ORF46	Acipenserid herpesvirus 2	56%	3.00E-89	30%	AEF97704.1
29	52584	53915	+	1332						
30	54475	55308	-	834	hypothetical protein	Acanthamoeba polyphaga mimivirus	28%	3.00E-06	28%	AKI79331.1
31	55538	55969	+	432						
32	56952	57899	+	948						
33	58080	59192	-	1113						
34	59245	64089	-	4845	ORF57+58	Acipenserid herpesvirus 2	95%	1.00E-104	26%	ACZ55868.2
35	64477	68280	+	3804	ORF56	Acipenserid herpesvirus 2	64%	3.00E-51	24%	AEF97716.1
36	70033	71148	+	1116						
37	71192	72106	+	915	putative cyclin-dependent kinase 2	Chelonid herpesvirus 5	61%	1.00E-16	30%	AHA93304.1
38	72862	73884	-	1023	protein ORF35	Anguillid herpesvirus 1	56%	4.00E-12	27%	YP_003358174.1
39	73932	74405	-	474						
40	74521	76035	-	1515						
41	76032	79727	-	3696						
42	79791	80657	+	867	DNA methyltransferase	Lymphocystis disease virus 1	69%	1.00E-09	27%	NP_078617.1
43	81086	83332	-	2247	ORF63	Acipenserid herpesvirus 2	53%	2.00E-15	24%	AEF97717.1
44	83412	84455	-	1044	ORF62	Ictalurid herpesvirus 1	57%	4.00E-44	45%	NP_041153.2
45	84655	85278	+	624						
46	85348	86556	+	1209						
47	86667	87011	-	345						
48	87156	87470	+	315						
49	87666	88007	-	342						
50	88110	88646	-	537						
51	88646	89095	-	450						
52	89064	89429	-	366						
53	89732	93064	+	3333	PREDICTED: cGMP-specific 3'	5'-cyclic phosphodiesterase-like	11%	5.00E-05	28%	XP_012560851.1
54	93448	94365	-	918	putative ubiquitin specific peptidase C19	Organic Lake phycodnavirus 1	22%	9.00E-05	34%	ADX06043.1
55	94379	96778	-	2400	ORF73	Ictalurid herpesvirus 1	12%	3.00E-10	35%	NP_041163.1
56	96783	97730	-	948						
57	97769	98590	+	822						
58	98639	99046	+	408	N1R/p28 family protein	Pigeonpox virus	63%	4.00E-15	43%	YP_009046385.1
59	99333	99914	-	582						
60	100111	100866	+	756	unknown protein	Grouper iridovirus	60%	6.00E-11	28%	AAV91038.1
61	100926	101423	+	498						
62	102042	102608	-	567						
63	102628	105984	-	3357	ORF39	Ictalurid herpesvirus 1	92%	2.00E-51	24%	NP_041130.1
64	106015	106386	-	372						
65	106511	108733	+	2223	ORF37	Acipenserid herpesvirus 2	52%	5.00E-11	21%	AEF97695.1
66	108561	109583	-	1023	PREDICTED: uncharacterized protein LOC105358139	Oryzias latipes	7.00E-20	46%	XP_011491977.1	
67	109585	110355	-	771	PREDICTED: uncharacterized protein LOC105358139	Oryzias latipes	9.00E-14	49%	XP_011491977.1	
68	110367	111815	-	1449	ORF34	Acipenserid herpesvirus 2	70%	1.00E-47	34%	AEF97692.1
69	112463	113200	+	738						
70	113649	116354	-	2706	hypothetical protein EH28_02830	Larimichthys crocea	8.00E-126	44%	KKF18232.1	
71	116409	117575	+	1167	ORF44	Acipenserid herpesvirus 2	52%	2.00E-22	32%	AEF97702.1
72	117575	118150	+	576						
73	118229	119887	+	1659						
74	119838	121904	+	2067	PREDICTED: GRIP and coiled-coil domain-containing prot	Neolamprologus brichardi	4.00E-19	28%	XP_006800602.1	
75	122015	122941	-	927	PREDICTED: uncharacterized protein LOC106097631	Oreochromis niloticus	9.00E-24	30%	XP_013123506.1	
76	123042	123671	-	630						
77	123745	124824	+	1080	PREDICTED: uncharacterized protein LOC105358165	Oryzias latipes	7.00E-05	24%	XP_011492011.1	
78	124969	126285	+	1317	PREDICTED: uncharacterized protein LOC102211585	Pundamilia nyererei	1.00E-06	30%	XP_005755678.1	
79	126718	127035	+	318	hypothetical protein LDVICp188	Lymphocystis disease virus - isolate China	75%	3.00E-09	32%	YP_073693.1
80	127216	128247	-	1032						
81	129040	129480	+	441						
82	129589	130080	+	492						
83	130231	130953	-	723	DNA packaging terminase subunit 1	Anguillid herpesvirus 1	3.00E-08	29%	YP_003358149.1	
84	131107	131958	+	852	Cbp/p300-interacting transactivator 1	Larimichthys crocea	1.00E-78	48%	KKF33343.1	
85	132578	133765	+	1188						
86	133963	134271	-	309						
87	134276	135700	-	1425	ORF64	Ictalurid herpesvirus 1	72%	1.00E-10	21%	NP_041156.1
88	135721	141636	-	5916						
89	141619	143004	+	1386						
90	143042	146761	+	3720						

ORF	Start	Stop	strand	length	Description	species	Coverage	E value	Identity	Accession
1	3	146	+	144						
2	288	1511	-	1224						
3	1590	6512	-	4923	ORF57+58	Acipenserid herpesvirus 2]	91%	3.00E-104	27%	ACZ55868.2
4	7530	11501	+	3972	ORF56	Acipenserid herpesvirus 2]	34%	8.00E-42	25%	AEF97716.1
5	12141	12548	-	408	PREDICTED: type-2 angiotensin II receptor-like	Oryzias latipes]	94%	3.00E-10	28%	XP_004084729.1
6	13730	15526	-	1797	ORF54	Ictalurid herpesvirus 1]	28%	5.00E-08	25%	NP_041145.1
7	15627	16751	-	1125	ORF60	Siberian sturgeon herpesvirus]	53%	6.00E-25	29%	ADC79712.1
8	16881	17618	+	738						
9	20008	21156	+	1149						
10	21195	22976	+	1782	ORF25	Acipenserid herpesvirus 2]	75%	9.00E-35	28%	AEF97684.1
11	23479	24360	+	882	ORF27	Acipenserid herpesvirus 2]	38%	3.00E-10	32%	AEF97686.1
12	24549	24998	-	450						
13	25219	25734	-	516						
14	26408	27358	-	951	unnamed protein product	Oncorhynchus mykiss]	40%	1.00E-20	42%	CDQ74451.1
15	27641	29503	-	1863	ORF28	Ictalurid herpesvirus 1]	26%	3.00E-12	34%	NP_041119.1
16	30056	30805	+	750						
17	30912	34868	+	3957	ORF46	Acipenserid herpesvirus 2]	66%	9.00E-99	28%	AEF97704.1
18	35684	35950	+	267						
19	36634	37443	-	810	PREDICTED: uncharacterized protein LOC105920738	Fundulus heteroclitus]	88%	5.00E-48	40%	XP_012711817.1
20	37876	38445	+	570	ORF077R	Rock bream iridovirus]	80%	8.00E-29	40%	AAT71892.1
21	39738	41069	+	1332						
22	41606	42544	+	939						
23	43132	43881	-	750	PREDICTED: NEDD4 family-interacting protein 1-like isofo	Astyanax mexicanus]	74%	6.00E-09	28%	XP_007258638.1
24	43953	44810	-	858						

ORF	Start	Stop	strand	length	Description	species	Coverage	E value	Identity	Accession
1	2	646	+	645	ORF25	Ictalurid herpesvirus 1	87%	1.00E-16	31%	NP_041116.1
2	727	1557	+	831	DNA helicase	Cyprinid herpesvirus 2	40%	3.00E-17	41%	AGG54974.1
3	1982	2368	+	387						
4	3028	5442	-	2415	ORF46	Acipenserid herpesvirus 2	76%	1.00E-92	29%	AEF97704.1
5	5420	6328	-	909						
6	6440	6880	-	441						
7	7019	7936	-	918						
8	8491	10467	+	1977	ORF88	Ranid herpesvirus 2	24%	6.00E-14	32%	YP_656596.1
9	10776	11843	+	1068	unnamed protein product	Oncorhynchus mykiss	52%	1.00E-56	56%	CDQ74451.1
10	11903	12382	+	480						
11	12487	13026	+	540						
12	13302	13766	+	465						
13	14257	15126	-	870	ORF27	Acipenserid herpesvirus 2	45%	9.00E-08	30%	AEF97686.1
14	15906	16775	-	870	ORF27	Acipenserid herpesvirus 2	45%	9.00E-08	30%	AEF97686.1
15	16871	17305	-	435						
16	18739	19515	-	777	hypothetical protein EH28_00657	Larimichthys crocea	72%	1.00E-131	100%	KKF32811.1
17	19711	20094	-	384						
18	20318	21988	-	1671						
19	22005	25037	-	3033						
20	25713	26393	+	681	hypothetical protein	European catfish virus	69%	2.00E-13	34%	YP_006347673.1
21	26490	27392	+	903	ORF072R	Infectious spleen and kidney necrosis virus	53%	6.00E-18	35%	NP_612294.1
22	28521	29483	-	963	PREDICTED: uncharacterized protein LOC105358165	Oryzias latipes	67%	1.00E-18	28%	XP_011492011.1
23	29912	30925	-	1014	protein ORF35	Anguillid herpesvirus 1	64%	3.00E-15	28%	YP_003358174.1
24	31045	31359	-	315						
25	32442	32855	-	414						
26	32776	34959	-	2184						
27	35026	35904	+	879	cytosine DNA methyltransferase	Turbot reddish body iridovirus	87%	3.00E-10	26%	ADE34390.1
28	36124	38415	-	2292	ORF63	Acipenserid herpesvirus 2	54%	5.00E-21	25%	AEF97717.1
29	38929	39984	-	1056	ORF62	Ictalurid herpesvirus 1	55%	2.00E-42	43%	NP_041153.2
30	39986	40819	+	834						
31	41698	42045	+	348						
32	42176	42619	+	444						
33	43689	44207	-	519						
34	44715	45632	-	918	ORF68	Ranid herpesvirus 2	82%	8.00E-14	26%	YP_656576.1
35	45631	46434	+	804	Cbp/p300-interacting transactivator 1	Larimichthys crocea	100%	0	98%	KKF33343.1
36	46396	46695	-	300	ORF62	Ictalurid herpesvirus 1	90%	3.00E-08	32%	NP_041153.2
37	47426	48541	+	1116						
38	48979	51192	-	2214	ORF56	Ictalurid herpesvirus 1	39%	1.00E-35	30%	NP_041147.1
39	51201	51515	-	315						
40	51822	52634	+	813						
41	53619	55748	+	2130	putative reverse transcriptase	Cotesia sesamiae	46%	7.00E-20	27%	AGO14420.1
42	56499	58067	-	1569						
43	58617	58916	-	300						
44	59129	61504	+	2376	ORF57+58	Acipenserid herpesvirus 2	78%	5.00E-54	29%	ACZ55868.2
45	61737	63833	+	2097	DNA polymerase	Salmonid herpesvirus 1	65%	8.00E-32	25%	AAC59316.1
46	63809	64174	+	366						
47	64228	64851	+	624						
48	64911	65375	+	465						
49	66842	67558	-	717						
50	67592	68005	-	414						
51	68336	68650	-	315						
52	68907	69446	+	540						
53	70189	71118	-	930						
54	72132	73718	-	1587						
55	73723	74061	-	339						

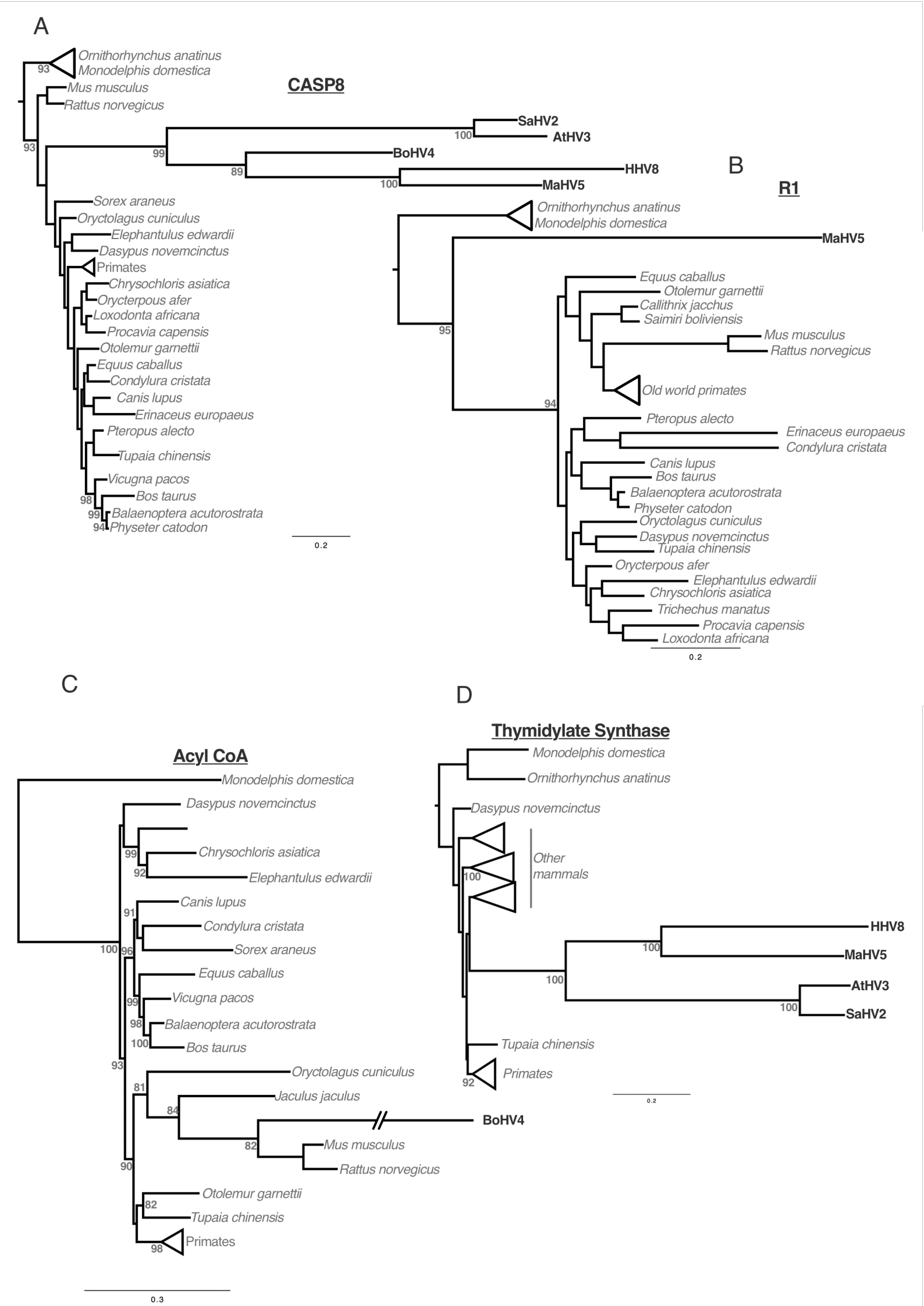
Chapter Four Supplementary Items

Figure S1
Additional phylogenetic trees that did not offer a clear host origin, but were nonetheless shown to be most closely related to mammalian genes with a high posterior probability.

Table S1
A table of coding sequences in the rhadinoviruses, where the columns include the NCBI GI number, reference ID and description for each viral genome. An arbitrary rhadinovirus ID (RVID) is assigned to each row, representing a homologous group of sequences. They are organised relative to HHV8.

Table S2
A table of containing the loadings for the first three components of the principal components analysis for the eight main genes investigated (see figure 7). The 22 variables tested were gene length, CpG bias, all four mononucleotide frequencies and 16 dinucleotide frequencies (denoted as 'nucfreq_x').

Table S3
A table of results of the selection analyses that is identical to that in table one, except that the analyses were conducted using an alignment containing a representative subset of the primate species included in the main analysis.


A
CASP8
Ornithorhynchus anatinus
Monodelphis domestica
Mus musculus
Rattus norvegicus
SaHV2
AtHV3
BoHV4
HHV8
MaHV5
Sorex araneus
Oryctolagus cuniculus
Elephantulus edwardii
Dasypus novemcinctus
Primates
Chrysochloris asiatica
Orycterpous afer
Loxodonta africana
Procavia capensis
Otolemur garnettii
Equus caballus
Condylura cristata
Canis lupus
Erinaceus europaeus
Pteropus alecto
Tupaia chinensis
Vicugna pacos
Bos taurus
Balaenoptera acutorostrata
Physeter catodon
0.2
B
R1
Ornithorhynchus anatinus
Monodelphis domestica
MaHV5
Equus caballus
Otolemur garnettii
Callithrix jacchus
Saimiri boliviensis
Mus musculus
Rattus norvegicus
Old world primates
Pteropus alecto
Erinaceus europaeus
Condylura cristata
Canis lupus
Bos taurus
Balaenoptera acutorostrata
Physeter catodon
Oryctolagus cuniculus
Dasypus novemcinctus
Tupaia chinensis
Orycterpous afer
Elephantulus edwardii
Chrysochloris asiatica
Trichechus manatus
Procavia capensis
Loxodonta africana
0.2
C
Acyl CoA
Monodelphis domestica
Dasypus novemcinctus
Chrysochloris asiatica
Elephantulus edwardii
Canis lupus
Condylura cristata
Sorex araneus
Equus caballus
Vicugna pacos
Balaenoptera acutorostrata
Bos taurus
Oryctolagus cuniculus
Jaculus jaculus
BoHV4
Mus musculus
Rattus norvegicus
Otolemur garnettii
Tupaia chinensis
Primates
0.3
D
Thymidylate Synthase
Monodelphis domestica
Ornithorhynchus anatinus
Dasypus novemcinctus
Other
mammals
HHV8
MaHV5
AtHV3
SaHV2
Tupaia chinensis
Primates
0.2

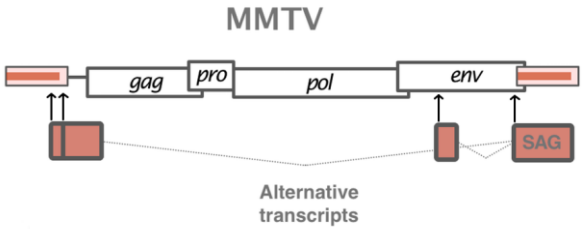
	Human herpesvirus 8				Murine Herpesvirus 4			Macacine Herpesvirus 5			Bovine herpesvirus 4			Saimiriine herpesvirus 2			Ateline herpesvirus 3			Rodent Herpesvirus peru		
RVID	CoreBlock/variable region	Description	GI number	Reference	Description	GI number	Reference	Description	GI number	Reference	Description	GI number	Reference	Description	GI number	Reference	Description	GI number	Reference	Description	GI number	Reference
RV1	A	K1	139472859	YP_001129350.1				R1	18653810	NP_570744.1												
RV2	A	KCP	139472860	YP_001129351.1	complement control protein	9629556	NP_044845.1	complement binding protein	18653812	NP_570746.1				complement control protein homologue complement control protein homologue	9625960 9625959	NP_040205.1 NP_040206.1	complement control protein homolog ccph	9631196	NP_047979.1	ccph complement regulatory protein	GI:321496560 GI:321496554	YP_004207845.1 YP_004207839.1
RV3	core block 1	ORF6	139472807	YP_001129352.1	single-stranded DNA-binding protein	9629557	NP_044846.1	ssDNA binding protein	18653813	NP_570747.1	single-stranded DNA-binding protein MDBP	13095583	NP_076498.1	major ssDNA-binding protein	9625962	NP_040208.1	major ssDNA binding protein	9631197	NP_047980.1	single-stranded DNA binding protein; Provisional	GI:321496561	YP_004207846.1
RV4	core block 1	ORF7	139472861	YP_001129353.1	DNA packaging terminase subunit 2	9629558	NP_044847.1	transport protein	18653814	NP_570748.1	transport protein	13095584	NP_076499.1	transport protein	9625963	NP_040209.1	transport protein	9631205	NP_047981.1	transport protein	GI:321496562	YP_004207847.1
RV5	core block 1	ORF8	139472808	YP_001129354.1	envelope glycoprotein B	209976837	NP_044848.2	glycoprotein B	18653815	NP_570749.1	glycoprotein B	13095585	NP_076500.1	glycoprotein B	9625964	NP_042010.1	glycoprotein B	9631198	NP_047982.1	glycoprotein B	GI:321496563	YP_004207848.1
RV6	core block 1	ORF9	139472809	YP_001129355.1	DNA polymerase catalytic subunit	9629560	NP_044849.1	DNA polymerase	18653816	NP_570750.1	DNA polymerase	13095586	NP_076501.1	DNA polymerase	9625965	NP_042011.1	DNA polymerase	9631199	NP_047983.1	DNA polymerase;	GI:321496564	YP_004207849.1
RV7	B	ORF10	139472862	YP_001129356.1	protein G10	9629561	NP_044850.1	hypothetical protein MmVgp08	18653817	NP_570751.1	unknown	13095587	NP_076502.1	unnamed protein product	9625966	NP_042012.1	orf 10; similar to Raji LF1	9631200	NP_047984.1	RHVP.10; dUTPase protein	GI:321496565	YP_004207850.1
RV8	B	ORF11	139472810	YP_001129357.1	virion protein G11	9629562	NP_044851.1	hypothetical protein MmVgp09	18653818	NP_570752.1				unnamed protein product	9625967	NP_042013.1	orf 11; similar to Raji LF2	9631201	NP_047985.1	RHVP.11; hypothetical protein; Provisional	GI:321496566	YP_004207851.1
RV9	B	K2	139472811	YP_001129358.1				R2	18653819	NP_570753.1												
RV10	B,A	ORF2	139472863	YP_001129359.1				dihydrofolate reductase	18653811	NP_570745.1				dihydrofolate reductase	9625957	NP_040203.1						
RV11	B	K3	139472864	YP_001129360.1	E3 ubiquitin ligase MIR1	9629601	NP_044852.1				hypothetical protein pBo5 hypothetical protein pBo4	13095589 13095588	NP_076504.1 NP_076503.1				RHVP.R12; RING-variant domain	GI:321496567	YP_004207852.1			
RV12	B/C	ORF70	139472812	YP_001129361.1				thymidylate synthase	18653820	NP_570754.1				thymidylate synthase	9626026	NP_040272.1	thymidylate synthase	9631261	NP_048042.1			
RV13	B	K4	139472814	YP_001129362.1				R3	18653821	NP_570755.1												
RV14	B	K4.1	139472813	YP_001129363.1																		
RV15	B	K4.2	139472865	YP_001129364.1																		
RV16	B	K5	139472815	YP_001129365.1																		
RV17	B	K6	139472866	YP_001129366.1																		
RV18	B	K7	139472816	YP_001129367.1																		
RV19	B/C	ORF16	139472817	YP_001129368.1	apoptosis regulator M11	9629595	NP_044912.1	Bcl-2 homolog	18653822	NP_570756.1	v-Bcl-2 protein	13095593	NP_076508.1	unnamed protein product	9625972	NP_040218.1	orf 16; bcl-2 family member	9631206	NP_047987.1	v-Bcl-2-like protein; inhibitor of apoptosis	GI:321496623	YP_004207908.1
RV20	core block 2	ORF17	139472867	YP_001129369.1	capsid maturation protease	9629602	NP_044855.1	capsid protein	18653823	NP_570757.1	capsid protein	13095594	NP_076509.1	unnamed protein product	9625974	NP_040219.1	orf 17; protease and capsid protein (N-term); minor capsid scaffold protein (C-term)	9631207	NP_047988.1	capsid protein;	GI:321496568	YP_004207853.1
	core block 2	ORF17.5	139472868	YP_001129370.1																		
RV21	core block 2	ORF18	139472818	YP_001129371.1	protein UL79	9629565	NP_044856.1	hypothetical protein MmVgp17	18653824	NP_570758.1	unknown	13095595	NP_076510.1	unnamed protein product	9625973	NP_040220.1	orf 18	9631208	NP_047989.1	RHVP.18;	GI:321496569	YP_004207854.1
RV22	core block 2	ORF19	139472819	YP_001129372.1	DNA packaging tegument protein UL25	9629603	NP_044857.1	tegument protein	18653825	NP_570759.1	tegument protein	13095596	NP_076511.1	unnamed protein product	9625975	NP_040221.1	virion tegument protein	9631209	NP_047990.1	tegument UL25 family	GI:321496570	YP_004207855.1
RV23	core block 2	ORF20	139472820	YP_001129373.1	nuclear protein UL24	9629604	NP_044858.1	hypothetical protein MmVgp19	18653826	NP_570760.1	unknown	13095597	NP_076512.1	unnamed protein product	9625977	NP_040222.1	fusion protein	9631210	NP_047991.1	RHVP.20; protein UL24	GI:321496571	YP_004207856.1
RV24	core block 2	ORF21	139472869	YP_001129374.1	thymidine kinase	9629566	NP_044859.1	thymidine kinase	18653827	NP_570761.1	thymidine kinase	13095598	NP_076513.1	thymidine kinase	9625976	NP_040223.1	thymidine kinase	9631211	NP_047992.1	thymidine kinase;	GI:321496572	YP_004207857.1
RV25	core block 2	ORF22	139472821	YP_001129375.1	envelope glycoprotein H	9629567	NP_044860.1	glycoprotein H	18653828	NP_570762.1	glycoprotein H	13095599	NP_076514.1	glycoprotein H	9625978	NP_040224.1	glycoprotein H	9631212	NP_047993.1	glycoprotein H;	GI:321496573	YP_004207858.1
RV26	core block 2	ORF23	139472870	YP_001129376.1	protein UL88	9629605	NP_044861.1	hypothetical protein MmVgp22	18653829	NP_570763.1	unknown	13095600	NP_076515.1	unnamed protein product	9625979	NP_040225.1	orf 23; similar to BTRF1	9631213	NP_047994.1	RHVP.23;	GI:321496574	YP_004207859.1
RV27	core block 2	ORF24	139472822	YP_001129377.1	virion protein UL87	9629606	NP_044862.1	hypothetical protein MmVgp23	18653830	NP_570764.1	unknown	13095601	NP_076516.1	unnamed protein product	9625980	NP_040226.1	orf 24; similar to BCRF1	9631214	NP_047995.1	RHVP.24;	GI:321496575	YP_004207860.1
RV28	core block 2	ORF25	139472823	YP_001129378.1	major capsid protein	9629568	NP_044863.1	major capsid protein	18653831	NP_570765.1	major capsid protein	13095602	NP_076517.1	major capsid protein	9625981	NP_040227.1	major capsid protein	9631215	NP_047996.1	major capsid protein;	GI:321496576	YP_004207861.1
RV29	core block 2	ORF26	139472871	YP_001129379.1	capsid protein subunit 2	9629569	NP_044864.1	capsid protein	18653832	NP_570766.1	capsid protein	13095603	NP_076518.1	unnamed protein product	9625982	NP_040228.1	orf 26	9631216	NP_047997.1	Capsid triplex subunit 2	GI:321496577	YP_004207862.1
RV30	core block 3	ORF27	139472872	YP_001129380.1	envelope glycoprotein 48	9629570	NP_044865.1	hypothetical protein MmVgp26	18653833	NP_570767.1				unnamed protein product	9625983	NP_040229.1	orf 27	9631217	NP_047998.1	; hypothetical RHVP.27	GI:321496578	YP_004207863.1
RV31	core block 3	ORF28	139472873	YP_001129381.1	envelope glycoprotein 150	19343406	NP_597855.1										orf 28	9631218	NP_047999.1	; hypothetical RHVP.R13	GI:321496579	YP_004207864.1
RV32	core block 3	ORF29	139472824	YP_001129382.1													orf 29B	9631221	NP_048001.1	cleavage/packaging protein;	GI:321496580	YP_004207865.1
RV33	core block 3	ORF30	139472826	YP_001129383.1	protein UL91	9629571	NP_044867.1	hypothetical protein MmVgp29	18653836	NP_570770.1	unknown	13095606	NP_076522.1	unnamed protein product	9625985	NP_040232.1	orf 30	9631222	NP_048002.1	RHVP.30;	GI:321496581	YP_004207866.1
RV34	core block 3	ORF31	139472825	YP_001129384.1	protein UL92	9629572	NP_044868.1	hypothetical protein MmVgp30	18653837	NP_570771.1	unknown	13095607	NP_076523.1	unnamed protein product	9625986	NP_040233.1	orf 31	9631223	NP_048003.1	RHVP.31; UL92 family	GI:321496582	YP_004207867.1
RV35	core block 3	ORF32	139472827	YP_001129385.1	DNA packaging tegument protein UL17	9629573	NP_044869.1	hypothetical protein MmVgp31	18653838	NP_570772.1	viral DNA cleavage/packaging protein	13095608	NP_076524.1	unnamed protein product	9625987	NP_040234.1	orf 32	9631224	NP_048004.1	viral DNA cleavage/packaging; UL17 protein	GI:321496583	YP_004207868.1
RV36	core block 3	ORF33	139472828	YP_001129386.1	tegument protein UL16	9629574	NP_044870.1	hypothetical protein MmVgp32	18653839	NP_570773.1	unknown	13095609	NP_076525.1	unnamed protein product	9625988	NP_040235.1	orf 33	9631225	NP_048005.1	RHVP.33; UL16/UL94 family	GI:321496584	YP_004207869.1
RV37	core block 3	ORF34	139472829	YP_001129387.1	protein UL95	9629575	NP_044872.1	hypothetical protein MmVgp34	18653841	NP_570775.1	unknown	13095611	NP_076526.1	unnamed protein product	9625989	NP_040236.1	orf 34	9631226	NP_048007.1	RHVP.34; UL95 family	GI:321496586	YP_004207871.1
RV38	core block 3	ORF35	139472830	YP_001129388.1	tegument protein UL14	9629576	NP_044873.1	hypothetical protein MmVgp35	18653842	NP_570776.1	unknown	13095612	NP_076527.1	unnamed protein product	9625991	NP_040237.1	orf 35	9631227	NP_048008.1	RHVP.35; unknown function (DUF848)	GI:321496587	YP_004207872.1
RV39	core block 3	ORF36	139472874	YP_001129389.1	tegument serine/threonine protein kinase	9629577	NP_044874.1	kinase	18653843	NP_570777.1	kinase	13095613	NP_076528.1	tyrosine-specific protein kinase	9625992	NP_040238.1	orf 36	9631228	NP_048009.1	Protein kinases; catalytic domain	GI:321496588	YP_004207873.1
RV40	core block 3	ORF37	139472831	YP_001129390.1	deoxyribonuclease	9629578	NP_044875.1	alkaline exonuclease	18653844	NP_570778.1	alkaline exonuclease	13095614	NP_076529.1	alkaline exonuclease	9625993	NP_040239.1	orf 37	9631229	NP_048010.1	alkaline exonuclease;	GI:321496589	YP_004207874.1
RV41	core block 3	ORF38	139472875	YP_001129391.1	myristylated tegument protein	9629579	NP_044876.1	hypothetical protein MmVgp38	18653845	NP_570779.1	unknown	13095615	NP_076530.1	unnamed protein product	9625994	NP_042040.1	orf 38	9631230	NP_048011.1	RHVP.38; unknown function (DUF2733)	GI:321496590	YP_004207875.1
RV42	core block 3	ORF39	139472832	YP_001129392.1	envelope glycoprotein M	9629609	NP_044877.1	glycoprotein M	18653846	NP_570780.1	glycoprotein M	13095616	NP_076531.1	integral membrane protein	9625995	NP_042041.1	orf 39	9631231	NP_048012.1	glycoprotein M;	GI:321496591	YP_004207876.1
RV43	core block 3	ORF40	139472833	YP_001129393.1	helicase-primase subunit	9629580	NP_044878.1	helicase/primase	18653847	NP_570781.1	helicase-primase complex component	13095617	NP_076532.1	unnamed protein product	9625996	NP_042042.1	orf 40	9631232	NP_048013.1	helicase-primase;	GI:321496592	YP_004207877.1
RV44	core block 3	ORF42	139472834	YP_001129394.1	tegument protein UL7	9629610	NP_044879.1	hypothetical protein MmVgp42	18653849	NP_570783.1	unknown	13095619	NP_076534.1	unnamed protein product	9625998	NP_042044.1	orf 42	9631234	NP_048015.1	RHVP.42; UL7 like	GI:321496593	YP_004207878.1
RV45	core block 3	ORF43	139472835	YP_001129395.1	capsid portal protein	9629611	NP_044880.1	capsid protein	18653850	NP_570784.1	capsid protein	13095620	NP_076535.1	unnamed protein product	9625999	NP_042045.1	orf 43	9631241	NP_048016.1	capsid protein; UL6 like	GI:321496594	YP_004207879.1
RV46	core block 3	ORF44	139472836	YP_001129396.1	helicase-primase helicase subunit	9629581	NP_044881.1	helicase/primase	18653851	NP_570785.1	helicase	13095621	NP_076536.1	helicase	9625999	NP_040246.1	helicase	9631235	NP_048017.1	helicase;	GI:321496595	YP_004207880.1
RV47	core block 3	ORF45	139472837	YP_001129397.1	tegument protein G45	9629612	NP_044882.1	hypothetical protein MmVgp45	18653852	NP_570786.1	unknown	13095622	NP_076537.1	unnamed protein product	9626001	NP_042047.1	orf 45	9631236	NP_048018.1	RHVP.45; Cobalamin biosynthesis protein CobT	GI:321496596	YP_004207881.1
RV48	core block 3	ORF46	139472838	YP_001129398.1	uracil-DNA glycosylase	9629613	NP_044883.1	uracil DNA glycosidase	18653853	NP_570787.1	uracil-DNA-glycosidase	13095623	NP_076538.1	uracil DNA glucosidase	9626002	NP_042048.1	orf 46	9631237	NP_048019.1	uracil DNA glycosidase;	GI:321496597	YP_004207882.1
RV49	core block 3	ORF47	139472876	YP_001129399.1	envelope glycoprotein L	209976841	NP_044884.2	glycoprotein L	18653854	NP_570788.1	glycoprotein L	13095624	NP_076539.1	unnamed protein product	9626003	NP_040249.1	orf 47	9631238	NP_048020.1	glycoprotein L;	GI:321496598	YP_004207883.1
RV50	core block 4	ORF48	139472839	YP_001129400.1	tegument protein G48	9629615	NP_044885.1	hypothetical protein MmVgp48	18653855	NP_570789.1	unknown	13095625	NP_076540.1	unnamed protein product	9626004	NP_040250.1	orf 48	9631239	NP_048021.1	RHVP.48; unknown function (DUF832)	GI:321496599	YP_004207884.1
RV51	core block 4	ORF50	139472840	YP_001129401.1	protein R1a	209976842	NP_044887.2	transactivator	18653857	NP_570791.1	R transactivator protein	13095626	NP_076541.1	HVS R transactivator, partial	9626005	NP_040252.1	orf 50	9631240	NP_048022.1	transcriptional activator; unknown function (DUF4551)	GI:321496601	YP_004207886.1
RV52	core block 4	ORF49	139472877	YP_001129402.1	protein G49	9629616	NP_044886.1	hypothetical protein MmVgp49	18653856	NP_570790.1	unknown	13095627	NP_076542.1	unnamed protein product	9626006	NP_040251.1	orf 49	9631240	NP_048022.1	RHVP.49; BRRF1-like protein	GI:321496600	YP_004207885.1
RV53	core block 4	K8	139472841	YP_001129403.1																		
RV54	core block 4	K8.1	139472842	YP_001129404.1																		
RV55	core block 4	ORF52	139472843	YP_001129405.1	virion protein G52	9629617	NP_044889.1	hypothetical protein MmVgp53	18653858	NP_570792.1	unknown	13095629	NP_076544.1	unnamed protein product	9626008	NP_040254.1	orf 52	9631243	NP_048024.1	RHVP.52; BLRF2 protein	GI:321496603	YP_004207888.1
RV56	core block 4	ORF53	139472844	YP_001129406.1	envelope glycoprotein N	9629618	NP_044890.1	hypothetical protein MmVgp54	18653859	NP_570793.1	unknown	13095630	NP_076545.1	unnamed protein product	9626009	NP_040255.1	orf 53	9631244	NP_048025.1	RHVP.53; UL73 viral envelope glycoprotein	GI:321496604	YP_004207889.1
RV57	core block 4	ORF54	139472845	YP_001129407.1	deoxyuridine triphosphatase	9629584	NP_044891.1	dUTPase	18653860	NP_570794.1	dUTPase	13095631	NP_076546.1	unnamed protein product	9626010	NP_040256.1	orf 54	9631245	NP_048026.1	dUTPase;	GI:321496605	YP_004207890.1
RV58	core block 4	ORF55	139472846	YP_001129408.1	tegument protein UL51	9629619	NP_044892.1	hypothetical protein MmVgp56	18653861	NP_570795.1	unknown	13095632	NP_076547.1	unnamed protein product	9626012	NP_040257.1	orf 55	9631246	NP_048027.1	RHVP.55; virus U44 protein	GI:321496606	YP_004207891.1
RV59	core block 4	ORF56	139472847	YP_001129409.1	helicase-primase primase subunit	9629585	NP_044893.1	DNA replication protein	18653862	NP_570796.1	DNA replication protein	13095633	NP_076548.1	unnamed protein product	9626011	NP_040258.1	primase	9631247	NP_048028.1	; helicase-primase primase subunit BSLF1; Provisional	GI:321496607	YP_004207892.1
RV60	core block 4	ORF57	139472848	YP_001129410.1	multifunctional expression regulator	209976843	NP_044895.2	immediate-early gene product	18653863	NP_570797.1	possible post-transcriptional transactivator	13095634	NP_076549.1	unnamed protein product	9626013	NP_040259.1				RHVP.57; transcriptional regulator family	GI:321496608	YP_004207893.1
RV61	HHV8 & MaHV5 specific region	K9	139472878	YP_001129411.1				R10	18653868	NP_570802.1												
RV62	HHV8 & MaHV5 specific region	vIRF-4	139472849	YP_001129412.1																		
RV63	HHV8 & MaHV5 specific region	vIRF-3	139472850	YP_001129413.1																		
RV64	HHV8 & MaHV5 specific region	vIRF-2	139472851	YP_001129414.1																		
RV65	core block 4	ORF58	139472879	YP_001129415.1	envelope protein UL43	9629620	NP_044896.1	hypothetical protein MmVgp68	18653872	NP_570806.1	unknown	13095635	NP_076550.1	unnamed protein product	9626014	NP_040260.1	orf 58	9631248	NP_048029.1	RHVP.58; BMRF2 protein	GI:321496609	YP_004207894.1
RV66	core block 4	ORF59	139472852	YP_001129416.1	DNA polymerase processivity subunit	9629621	NP_044897.1	DNA replication protein	18653873	NP_570807.1	DNA replication protein	13095636	NP_076551.1	unnamed protein product	9626015	NP_040261.1	orf 59	9631249	NP_048030.1	; DNA replication accessory factor	GI:321496610	YP_004207895.1
RV67	core block 4	ORF60	139472853	YP_001129417.1	ribonucleotide reductase subunit 2	9629622	NP_044898.1	small ribonucleotide reductase	18653874	NP_570808.1	ribonucleotide reductase small subunit	13095637	NP_076552.1	ribonucleotide reductase, small subunit	9626016	NP_040262.1	small subunit of ribonucleotide reductase	9631250	NP_048031.1	ribonucleotide reductase small subunit;	GI:321496611	YP_004207896.1
RV68	core block 4	ORF61	139472880	YP_001129418.1	ribonucleotide reductase large	9629623	NP_044899.1	large ribonucleotide reductase	18653875	NP_570809.1	ribonucleotide reductase large subunit	13095638	NP_076553.1	ribonucleotide reductase, large subunit	9626017	NP_040263.1	large subunit of ribonucleotide reductase	9631251	NP_048032.1	ribonucleotide reductase large subunit;	GI:321496612	YP_004207897.1
RV69	core block 4	ORF62	139472854	YP_001129419.1	capsid triplex subunit 1	209976844	NP_044900.2	assembly/DNA maturation protein	18653876	NP_570810.1	assembly/DNA maturation protein	13095639	NP_076554.1	unnamed protein product	9626018	NP_040264.1	orf 62	9631252	NP_048033.1	; Capsid triplex subunit 1; Provisional	GI:321496613	YP_004207898.1
RV70	core block 4	ORF63	139472855	YP_001129420.1	tegument protein UL37	209976845	NP_044901.2	tegument protein	18653877	NP_570811.1	tegument protein	13095640	NP_076555.1	unnamed protein product	9626019	NP_040265.1	orf 63	9631253	NP_048034.1	; virus tegument protein U30	GI:321496614	YP_004207899.1
RV71	core block 4	ORF64	139472881	YP_001129421.1	large tegument protein	9629589	NP_044902.1	tegument protein	18653878	NP_570812.1	tegument protein	13095641	NP_076556.1	unnamed protein product	9626020	NP_040266.1	large tegument protein	9631254	NP_048035.1	; tegument protein, N-terminal conserved region	GI:321496615	YP_004207900.1
RV72	core block 4	ORF65	139472858	YP_001129422.1	small capsid protein	9629625	NP_044903.1	capsid protein	18653879	NP_570813.1	capsid protein	13095642	NP_076557.1	unnamed protein product	9626021	NP_040267.1	orf 65	9631255	NP_048036.1	RHVP.R15;	GI:321496616	YP_004207901.1
RV73	core block 4	ORF66	139472857	YP_001129423.1	protein UL49	9629626	NP_044904.1	hypothetical protein MmVgp76	18653880	NP_570814.1	unknown	13095643	NP_076558.1	unnamed protein product	9626022	NP_040268.1	orf 66	9631256	NP_048037.1	RHVP.66; UL49 family	GI:321496617	YP_004207902.1
RV74	core block 4	ORF67	139472856	YP_001129424.1	nuclear egress type 2 membrane protein	9629627	NP_044905.1	tegument protein	18653881	NP_570815.1	tegument protein	13095644	NP_076559.1	unnamed protein product	9626023	NP_040269.1	orf 67	9631258	NP_048038.1	tegument protein; virion protein U34	GI:321496618	YP_004207903.1
RV75	core block 4	ORF67A	139472882	YP_001129425.1	DNA packaging protein UL33	209976846	YP_002300518.1	Similar to ORF67.5, Macaca mulatta rhadinovirus 26-95	19424754	NP_598363.1	unknown	13095645	NP_076560.1	unnamed protein product	9626024	NP_040270.1	orf 67A	9631257	NP_048039.1	RHVP.67a; UL33-like protein	GI:321496619	YP_004207904.1
RV76	C	ORF68	139472883	YP_001129426.1	DNA packaging protein UL32	209976847	NP_044906.2	glycoprotein	18653882	NP_570816.1	probable glycoprotein	13095646	NP_076561.1	unnamed protein product	9626024	NP_040270.1	orf 68	9631259	NP_048040.1	RHVP.68; putative major envelope glycoprotein	GI:321496620	YP_004207905.1
RV77	C	ORF69	139472884	YP_001129427.1	nuclear egress lamina protein	9629591	NP_044907.1	hypothetical protein MmVgp80	18653883	NP_570817.1	unknown	13095647	NP_076562.1	unnamed protein product	9626025	NP_040271.1	orf 69	9631260	NP_048041.1	RHVP.69; UL31-like protein	GI:321496621	YP_004207906.1
RV78	C	K12	139472803	YP_001129428.1																		
RV79	C	K13	139472802	YP_001129429.1				FLIP	18653884	NP_570818.1	v-FLIP	13095651	NP_076565.1	unnamed protein product	9626027	NP_040273.1	orf 71; similar to family of FLIP proteins	9631262	NP_048043.1			
RV80	C	ORF72	139472885	YP_001129430.1	cyclin	9629628	NP_044911.1	cyclin D homolog	18653885	NP_570819.1	cyclin homologue	13095652	NP_076566.1	unnamed protein product	9626028	NP_040274.1	v-cyclin	9631263	NP_048044.1	cyclin D-like protein; Cyclin box fold.long note online	GI:321496622	YP_004207907.1
RV81	C	LANA	139472804	YP_001129431.1	nuclear antigen LANA	9629629	NP_044913.1	latent nuclear antigen	18653886	NP_570820.1	unknown	13095653	NP_076567.1	unnamed protein product	9626029	NP_040275.1	orf 73	9631264	NP_048045.1	LANA-like protein; hypothetical protein; Provisional	GI:321496624	YP_004207909.1
RV82	C	K14	139472886	YP_001129432.1				R15	18653887	NP_570821.1										G-protein coupled receptor	GI:321496625	YP_004207910.1
RV83	C	ORF74	139472805	YP_001129433.1	membrane protein G74	9629596	NP_044914.1	G protein coupled receptor	18653888	NP_570822.1				G protein-coupled receptor homologue	9626030	NP_040276.1	orf 74; IL-8 receptor family	9631265	NP_048046.1	tegument protein/FGARAT; FGAM-synthase; Provisional	GI:321496626	YP_004207911.1
RV84	C	ORF75	139472887	YP_001129434.1	tegument protein G75C	9629630	NP_044915.1	tegument protein	18653889	NP_570823.1	tegument protein/v-FGAM-synthetase	13095653	NP_076568.1	unnamed protein product	9626031	NP_040277.1	orf 75	9631266	NP_048047.1	tegument protein/FGARAT; FGAM-synthase; Provisional	GI:321496627	YP_004207912.1
RV85	C	K15	139472806	YP_001129435.1																		
RV86	core block 3				DNA packaging terminase subunit 1	209976840	NP_044866.2	hypothetical protein MmVgp28	18653835	NP_570769.1	cleavage/packaging protein	13095610	NP_076521.1	unnamed protein product	9625990	NP_040231.1	orf29a/b	9631219	NP_048000.1	cleavage/packaging protein; Probable DNA packing protein, N-terminus	GI:321496585	YP_004207870.1
RV87	core block 3																orf 29A	9631220	NP_048006.1			
RV88	core block 4							helicase/primase	18653848	NP_570782.1	helicase-primase complex component	13095618	NP_076533.1	unnamed protein product	9625997	NP_040243.1	orf 41	9631233	NP_048014.1			
RV89	core block 4				envelope glycoprotein 350	9629583	NP_044888.1							glycoprotein	9626007	NP_040253.1	glycoprotein	9631242	NP_048023.1	RHVP.R14; large tegument protein UL36; Provisional	GI:321496602	YP_004207887.1
RV90	C				protein G75B	9629631	NP_044916.1															
RV91	C				protein G75A	9629632	NP_044917.1															
RV92	B				protein 13M	209976838	YP_002300516.1															
RV93	core block 2				capsid scaffold protein	209976839	YP_002300517.1															
RV94	A				protein M1	9629554	NP_044841.1															
RV95	A				protein M2	209976835	NP_044842.2															
RV96	A				chemokine-binding protein M3	9629600	NP_044843.1															
RV97	A				chemokine-binding protein M4	209976836	NP_044844.2															
RV98	core block 3							hypothetical protein MmVgp27	18653834	NP_570768.1												
RV99	HHV8 & MaHV5 specific region							R6	18653864	NP_570798.1												
RV100	HHV8 & MaHV5 specific region							R7	18653865	NP_570799.1												
RV101	HHV8 & MaHV5 specific region							R8	18653866	NP_570800.1												
RV102	HHV8 & MaHV5 specific region							R9	18653867	NP_570801.1												
RV103	HHV8 & MaHV5 specific region							R11	18653869	NP_570803.1												
RV104	HHV8 & MaHV5 specific region							R12	18653870	NP_570804.1												
RV105	HHV8 & MaHV5 specific region							R13	18653871	NP_570805.1												
RV106	B							Similar to ORFRU8-L, Macaca mulatta rhadinovirus 26-95	19424749	NP_598358.1												
RV107	B							Similar to ORFRU1-R, Macaca mulatta rhadinovirus 26-95	19424750	NP_598359.1												
RV108	HHV8 & MaHV5 specific region							bZIP transcription factor fragment	19424751	NP_598360.1												
RV109	HHV8 & MaHV5 specific region							glycoprotein R8.1	19424752	NP_598361.1												
RV110	HHV8 & MaHV5 specific region							Similar to vIRF, Macaca mulatta rhadinovirus 26-95	19424753	NP_598362.1												
RV111	HHV8 & MaHV5 specific region							Similar to ORFRU9-L, Macaca mulatta rhadinovirus 26-95	19424755	NP_598364.1												
RV112	HHV8 & MaHV5 specific region							Similar to ORFRU12-L, Macaca mulatta rhadinovirus 26-95	19424756	NP_598365.1												
RV113	HHV8 & MaHV5 specific region							Similar to ORFRU14-L, Macaca mulatta rhadinovirus 26-95	19424757	NP_598366.1												
RV114	A										hypothetical protein pBo1	13095579	NP_076494.1									
RV115	A										v-FGAM-synthase	13095580	NP_076495.1	unnamed protein product	9625958	NP_040204.1	orf03; similar to FGARATs	9631203	NP_047978.1			
RV116	A										hypothetical protein pBo2	13095581	NP_076496.1									
RV117	A										hypothetical protein pBo3	13095582	NP_076497.1									
RV118	B										hypothetical protein	13095590	NP_076505.1									
RV119	B										hypothetical protein	13095591	NP_076506.1									
RV120	B										hypothetical prot

		Interleukin 6			Major histocompatibility complex			CD59			Interleukin 17		
		Component 1	Component 2	Component 3	Component 1	Component 2	Component 3	Component 1	Component 2	Component 3	Component 1	Component 2	Component 3
variables	length			0.813		0.417	0.197		0.227	0.969		0.378	0.923
	cpgbias		0.189	0.564			0.978	-0.194	0.104		-0.212	0.101	
	nuqfreq_g		0.435		0.17	-0.431		0.156	-0.472	0.126	0.176	-0.501	0.222
	nuqfreq_c	-0.243	-0.134		0.229	0.146		0.226	0.157		0.222	0.146	
	nuqfreq_t	-0.243	-0.134		0.229	0.146		0.226	0.157		0.222	0.146	
	nuqfreq_a	-0.243	-0.134		0.229	0.146		0.226	0.157		0.222	0.146	
	nuqfreq_gc	-0.213	0.248		0.226	-0.178		0.223	-0.181		0.223	-0.122	
	nuqfreq_gt	-0.213	0.248		0.226	-0.178		0.223	-0.181		0.223	-0.122	
	nuqfreq_gg		0.435		0.17	-0.431		0.156	-0.472	0.126	0.176	-0.501	0.222
	nuqfreq_ga	-0.213	0.248		0.226	-0.178		0.223	-0.181		0.223	-0.122	
	nuqfreq_ag	-0.213	0.248		0.226	-0.178		0.223	-0.181		0.223	-0.122	
	nuqfreq_at	-0.243	-0.134		0.229	0.146		0.226	0.157		0.222	0.146	
	nuqfreq_ac	-0.243	-0.134		0.229	0.146		0.226	0.157		0.222	0.146	
	nuqfreq_aa	-0.243	-0.134		0.229	0.146		0.226	0.157		0.222	0.146	
	nuqfreq_cc	-0.243	-0.134		0.229	0.146		0.226	0.157		0.222	0.146	
	nuqfreq_ct	-0.243	-0.134		0.229	0.146		0.226	0.157		0.222	0.146	
	nuqfreq_ca	-0.243	-0.134		0.229	0.146		0.226	0.157		0.222	0.146	
	nuqfreq_cg	-0.213	0.248		0.226	-0.178		0.223	-0.181		0.223	-0.122	
	nuqfreq_tt	-0.243	-0.134		0.229	0.146		0.226	0.157		0.222	0.146	
	nuqfreq_ta	-0.243	-0.134		0.229	0.146		0.226	0.157		0.222	0.146	
	nuqfreq_tg	-0.213	0.248		0.226	-0.178		0.223	-0.181		0.223	-0.122	
	nuqfreq_tc	-0.243	-0.134		0.229	0.146		0.226	0.157		0.222	0.146	

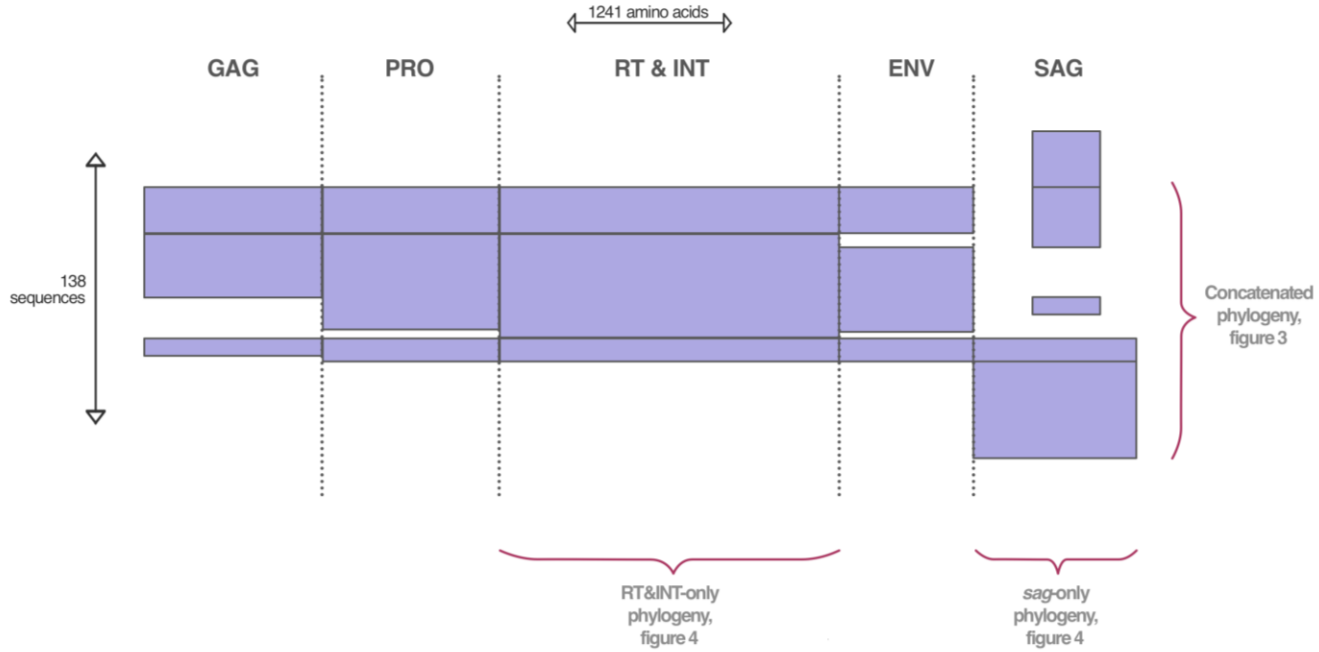
		C2GnT-M			CD200			CCL3			Dihydrofolate reductase		
		Component 1	Component 2	Component 3	Component 1	Component 2	Component 3	Component 1	Component 2	Component 3	Component 1	Component 2	Component 3
variables	length			0.99		0.392	-0.176		0.238	0.878		0.408	0.893
	cpgbias	0.153	0.194	-0.118		-0.208	-0.97		-0.167	-0.381	-0.121		0.228
	nuqfreq_g	-0.102	-0.479		0.181	-0.435		-0.134	-0.463	0.144	-0.176	-0.46	0.182
	nuqfreq_c	-0.234	0.143		0.228	0.143		-0.233	0.145		-0.227	0.14	
	nuqfreq_t	-0.234	0.143		0.228	0.143		-0.233	0.145		-0.227	0.14	
	nuqfreq_a	-0.234	0.143		0.228	0.143		-0.233	0.145		-0.227	0.14	
	nuqfreq_gc	-0.224	-0.204		0.226	-0.173		-0.226	-0.197		-0.225	-0.167	
	nuqfreq_gt	-0.224	-0.204		0.226	-0.173		-0.226	-0.197		-0.225	-0.167	
	nuqfreq_gg	-0.102	-0.479		0.181	-0.435		-0.134	-0.463	0.144	-0.176	-0.46	0.182
	nuqfreq_ga	-0.224	-0.204		0.226	-0.173		-0.226	-0.197		-0.225	-0.167	
	nuqfreq_ag	-0.224	-0.204		0.226	-0.173		-0.226	-0.197		-0.225	-0.167	
	nuqfreq_at	-0.234	0.143		0.228	0.143		-0.233	0.145		-0.227	0.14	
	nuqfreq_ac	-0.234	0.143		0.228	0.143		-0.233	0.145		-0.227	0.14	
	nuqfreq_aa	-0.234	0.143		0.228	0.143		-0.233	0.145		-0.227	0.14	
	nuqfreq_cc	-0.234	0.143		0.228	0.143		-0.233	0.145		-0.227	0.14	
	nuqfreq_ct	-0.234	0.143		0.228	0.143		-0.233	0.145		-0.227	0.14	
	nuqfreq_ca	-0.234	0.143		0.228	0.143		-0.233	0.145		-0.227	0.14	
	nuqfreq_cg	-0.224	-0.204		0.226	-0.173		-0.226	-0.197		-0.225	-0.167	
	nuqfreq_tt	-0.234	0.143		0.228	0.143		-0.233	0.145		-0.227	0.14	
	nuqfreq_ta	-0.234	0.143		0.228	0.143		-0.233	0.145		-0.227	0.14	
	nuqfreq_tg	-0.224	-0.204		0.226	-0.173		-0.226	-0.197		-0.225	-0.167	
	nuqfreq_tc	-0.234	0.143		0.228	0.143		-0.233	0.145		-0.227	0.14	

Gene	Site model M7 & M8	Branch-site model	Model	Free parameters	lnL	2ΔlnL	p-value (better than 1 ratio)	omega virus	omega host	omega virus tip	Δω for models significantly better than ω0	fold change
IL-17	no result	no result	ω0	1	-4904.815274							
			ω0,ωSaHV2	2	-4900.098440	9.433668	0.002130378	0.1715	0.3759	n/a	-0.2044	2.191836735
DHFR	0	no result	ω0	1	-5070.829093							
			ω0,ωiv	2	-5070.824936	0.008314	0.927348662					
			ω0,ωiv=ωHHV8=ωMaHV5=ωSaHV2	2	-5067.744211	6.169764	0.012995194	0.088	0.163	n/a	-0.075	1.852272727
			ω0,ωiv,ωHHV8=ωMaHV5=ωSaHV2	3	-5066.497408	8.66337	0.013145379	infinity	0.1615	0.0737	n/a	n/a
glucosaminyl transferase 3	5	no result	ω0	1	-12261.5084							
			ω0,ωBoHV4	2	-12261.50839	0.000016	0.99680847					
MHC class II	19	no result	ω0	1	-9994.611265							
			ω0,ωRHVP	2	-9958.226164	72.770202	0	0.0504	0.5212	n/a	-0.4708	10.34126984
Interleukin 6	2	no result	ω0	1	-5551.633195							
			ω0,ωiv	2	-5544.803438	13.659514	0.000219128	0.0101	0.7113	n/a	-0.7012	70.42574257
CD59 glycoprotein	5	no result	ω0	1	-3254.722179							
			ω0,ωSaHV2	2	-3237.269203	34.905952	0.000000003	0.0675	0.6972	n/a	-0.6297	10.32888889
CD200 immunoglobulin	1	no result	ω0	1	-5214.998836							
			ω0,ωiv	2	-5210.020032	9.957608	0.00160186	0.0035	0.3256	n/a	-0.3221	93.02857143
			ω0,ωiv=ωHHV8=ωMaHV5	2	-5211.006682	7.984308	0.004718453	0.0163	0.3261	n/a	-0.3098	20.00613497
			ω0,ωiv,ωHHV8=ωMaHV5	3	-5209.461637	11.074398	0.00393754	0.0034	0.3272	0.033	n/a	n/a
CCL3 chemokine	1	no result	ω0	1	-1933.422135							
			ω0,ωiv	2	-1931.16556	4.513150	0.033635243	0.002	0.3831	n/a	-0.3811	191.55
			ω0,ωiv=ωHHV8=ωMaHV5	2	-1913.505912	39.832446	0	0.008	0.4931	n/a	-0.4851	61.6375
			ω0,ωiv,ωHHV8=ωMaHV5	3	-1913.318422	40.207426	0.000000002	0.0022	0.4935	0.0079	n/a	n/a

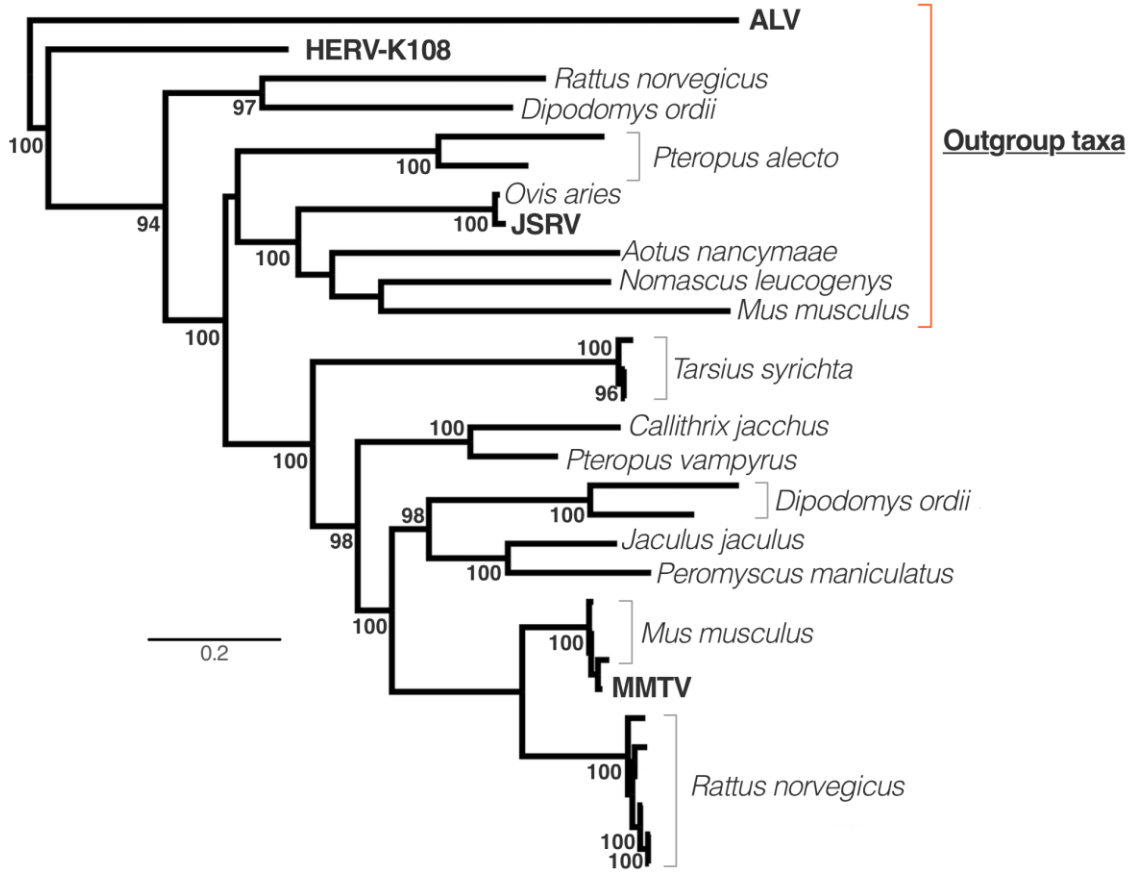
Chapter Four Supplementary Items



Supplementary Figure 1
Genome map of Mouse Mammary Tumour Virus (MMTV)
An approximately scaled diagram of the MMTV genome, representing the encoded genes as labelled boxes with LTRs at either end. The arrows indicate the approximate positions of alternative promoters, and alternative transcripts that also contain the *sag* gene in different conditions are represented below the genome map. The position of alternative transcripts were drawn according to[1].



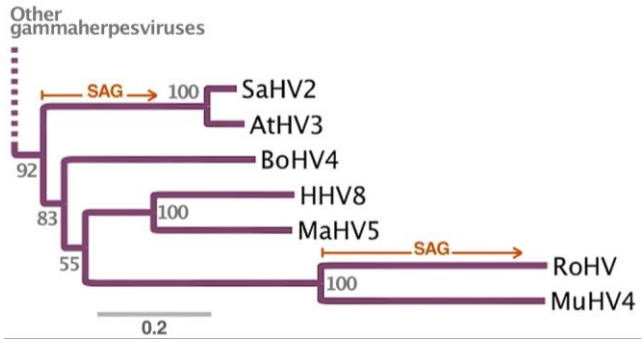
Supplementary Figure 2
Diagram of concatenated alignment of sequences used for phylogenetic analyses
The diagram depicts a schematic of the concatenated alignment created using mammalian ERVs and herpesviral *sag* sequences. Each of the three trees reported in the study were reconstructed using the subset of sequences indicated.



Supplementary Figure 3

Phylogenetic tree used to determine root of Figure 3 tree

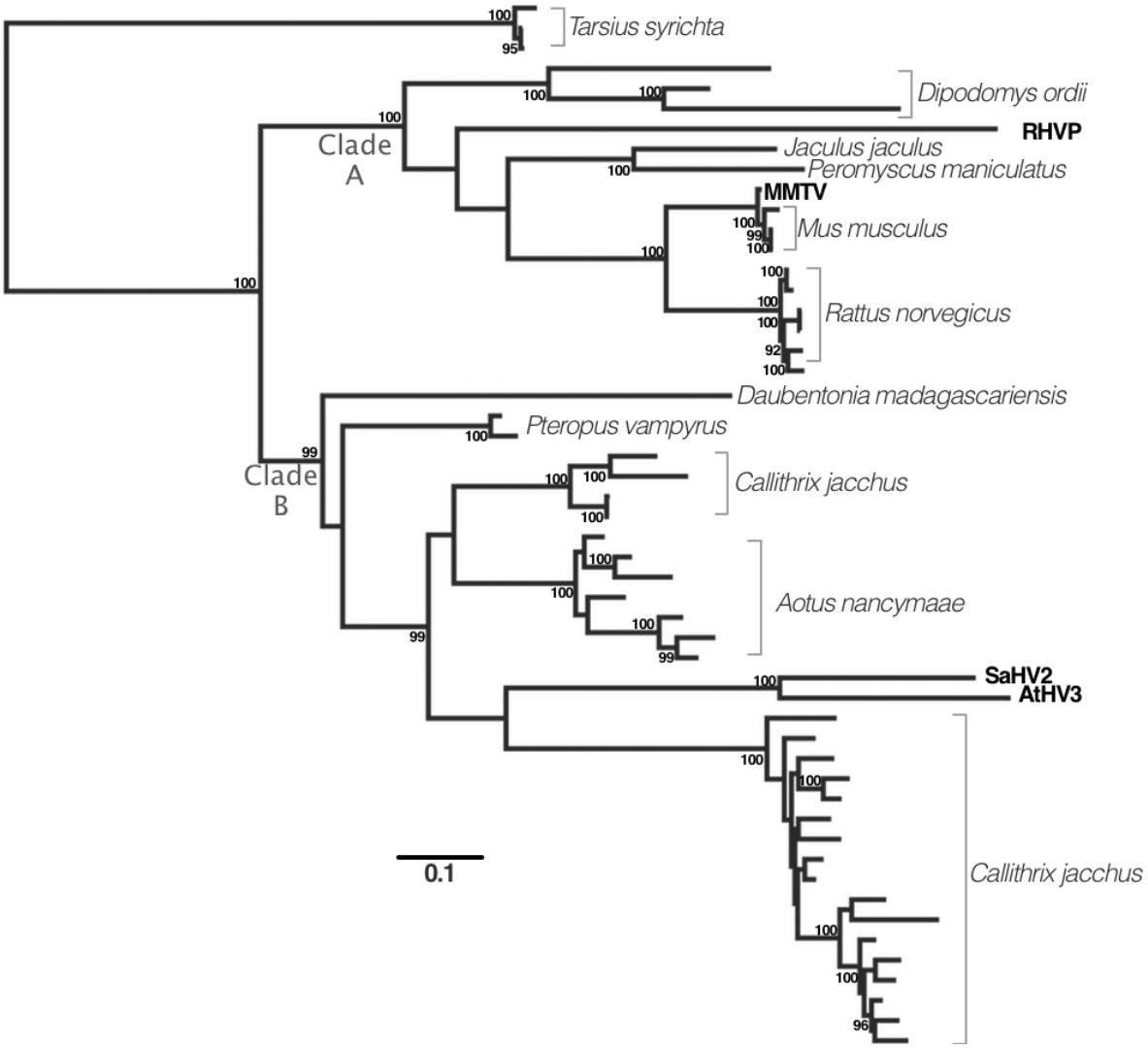
This figure shows the phylogeny used to determine the placement of the root for the tree in Figure 3. It depicts a midpoint-rooted Bayesian tree of a polymerase alignment that included the outgroup sequence set with the pol-containing sequences in the concatenated tree. The numbers at each node represent posterior probability values (only those above 90% are shown). This tree shows that the in-group sequences are rooted at the branch leading to the *Tarsius syrichta* sequences. The scale bar indicated represents amino acid substitutions per site.



Supplementary Figure 4

Phylogenetic tree of rhadinoviruses using six conserved genes

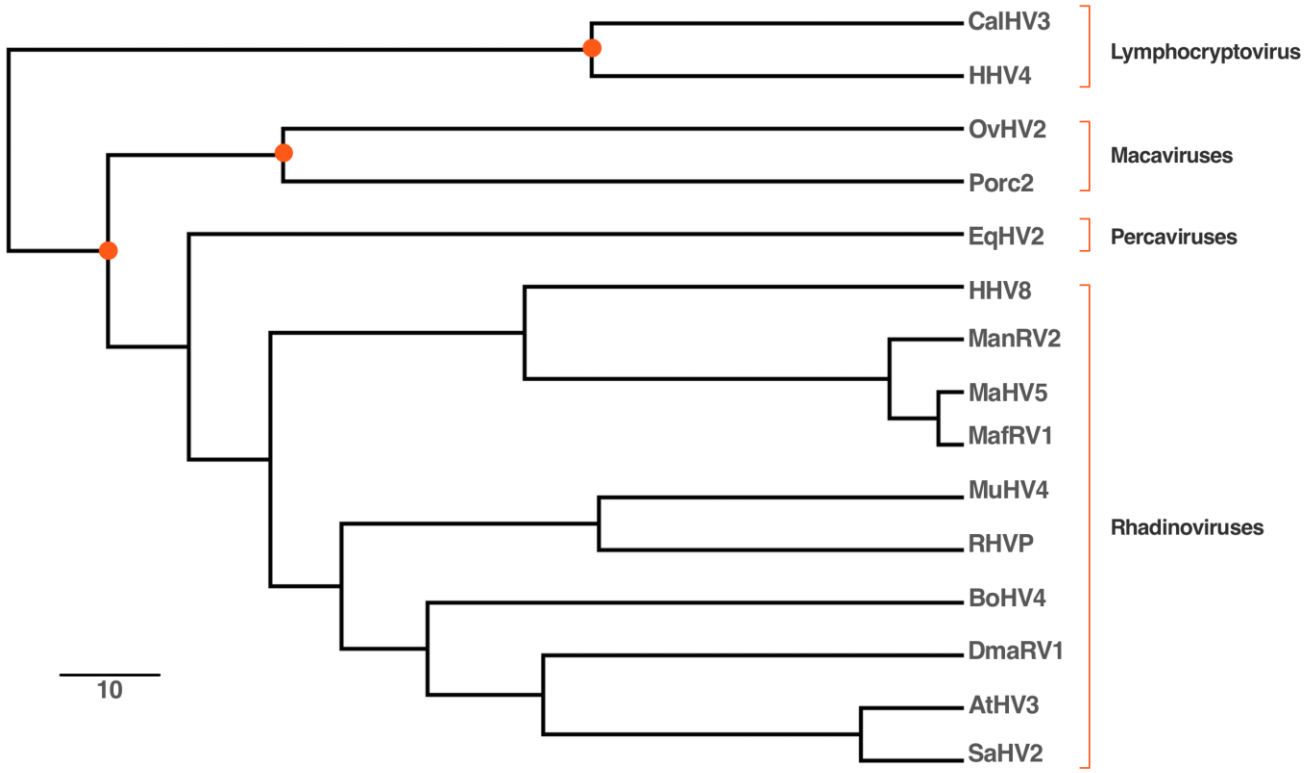
The phylogeny shown is a subtree of a *gamaherpesvirinae* maximum likelihood phylogeny estimated using a concatenated alignment of 6 core genes (terminase, large tegument, uracil-DNA glycosylase, kinase, capsid protein and helicase). The scale bar indicated represents nucleotide substitutions per site. Values at each node represent bootstrap percentages as estimated by 1,000 non-parametric pseudoreplicates. Taxon abbreviation are RHVP: Rodent herpesvirus peru, MuHV4: Murid herpesvirus 4, AtHV3: Ateline herpesvirus 3, SaHV2: Saimirine herpesvirus 2, BoHV4: Bovine herpesvirus 4, HHV8: human herpesvirus 8, MaHV5: Macacine herpesvirus 5. The hypothesized model of *sag* acquisition is represented by a red arrow.



Supplementary Figure 5

Phylogenetic test of the monophyly of SaHV2 and AtHV3

A phylogeny of *gag*, *pro*, *pol*, *env* and *sag* genes with sites detected under positive selection removed from the alignment. This did not have any effect on the topology, although a slight change in branch length is observed as well as a slight increase of the posterior probability for some nodes (indicated at each node if above 90%). The scale bar indicated represents the number of amino acid substitutions per site.



Supplementary Figure 6

Time-calibrated herpesvirus tree

The figure shown represents the time-calibrated phylogeny of herpesviruses estimated in BEAST, which was pruned down to the relevant tips for the tree in figure 5. The scale bar represents indicated is in million of years. Taxon abbreviation are RHVP: Rodent herpesvirus peru, MuHV4: Murid herpesvirus 4, AtHV3: Ateline herpesvirus 3, SaHV2: Saimirine herpesvirus 2, BoHV4: Bovine herpesvirus 4, HHV8: human herpesvirus 8, CalHV3: Callitrichine herpesvirus 3 , EqHV2: Equine herpesvirus 2, HHV4:Human herpesvirus 4, ManRV2: Macaca nemestrina rhadinovirus 2, MafRV1:Macaca fuscata rhadinovirus 1, OvHV2: Ovine herpesvirus 2, Porc2: Porcine lymphotropic herpesvirus 2, MaHV5: Macacine herpesvirus 5 , DmaRV1: Daubentonia madagascariensis radinovirus 1. (DmaRV1 sequences were identified in our previous publication[2]).

Genus	Aotus	Aotus	Aotus	Aotus	Aotus	Aotus	Aotus	Rhadinovirus	Rhadinovirus	Rhadinovirus	Alpharetrovirus	Class III ERV	Betarerovirus	Betarerovirus	Daubentonia	Callithrix	Callithrix	Callithrix	Pteropus
Species	nancymaae	nancymaae	nancymaae	nancymaae	nancymaae	nancymaae	nancymaae	saHV2	atHv3	roHV	ALV	HERV-K	JSRV	MMTV	madagascariensis	jacchus	jacchus	jacchus	vampyrus
Accession	JYKP01018661.1	JYKP01073282.1	JYKP01181974.1	JYKP01239066.1	JYKP01018659.1	JYKP01112511.1	JYKP01112426.1	NC_001350.1	NC_001987.1	NC_015049.1	NC_015116.1	NC_022518.1	NC_001494.1	NC_001503.1	AGTM012997089.1	ACFV01171011.1_2	ACFV01044201.1	ACFV01126740.1	ABRP01228684.1
gag																			
pro																			
pol																			
env																			
sag																			
Genus	Pteropus	Pteropus	Pteropus	Pteropus	Pteropus	Pteropus	Callithrix	Callithrix	Callithrix	Callithrix	Callithrix	Callithrix	Callithrix	Callithrix	Callithrix	Callithrix	Callithrix	Callithrix	Callithrix
Species	vampyrus	vampyrus	alecto	alecto	alecto	alecto	jacchus	jacchus	jacchus	jacchus	jacchus	jacchus	jacchus	jacchus	jacchus	jacchus	jacchus	jacchus	jacchus
Accession	ABRP01191652.1	ABRP01087430.1	ALWS01056471.1	ALWS01044416.1	ALWS01016527.1	ALWS01000833.1	ACFV01183383.1	ACFV01156521.1	ACFV01154330.1	ACFV01114849.1	ACFV01097500.1	ACFV01093596.1	ACFV01041432.1	ACFV01013848.1	ACFV01088484.1	ACFV01145903.1	ACFV01138800.1	ACFV01115463.1	ACFV01157657.1
gag																			
pro																			
pol																			
env																			
sag																			
Genus	Callithrix	Callithrix	Callithrix	Callithrix	Pteropus	Pteropus	Pteropus	Callithrix	erinaceus	erinaceus	erinaceus	Orycteropus	Orycteropus	erinaceus	erinaceus	erinaceus	erinaceus	erinaceus	erinaceus
Species	jacchus	jacchus	jacchus	jacchus	alecto	alecto	vampyrus	jacchus	europaeus	europaeus	europaeus	afer	afer	europaeus	europaeus	europaeus	europaeus	europaeus	europaeus
Accession	ACFV01083091.1	ACFV01070902.1	ACFV01067173.1	ACFV01173401.1	ALWS01040448.1	ALWS01070041.1	ABRP01272947.1	ACFV01171011.1	AMDU01213627.1	AMDU01124824.1	AMDU01056098.1	ALYB01353094.1	ALYB01313857.1	AMDU01014656.1	AANN01453706.1	AMDU01096953.1	AMDU01040360.1	AMDU01023674.1	AMDU01022785.1
gag																			
pro																			
pol																			
env																			
sag																			
Genus	erinaceus	erinaceus	erinaceus	Mus	Mus	Mus	Mus	Mus	Mus	Cricetulus	Cricetulus	Cricetulus	Cricetulus	Mus	Mus	Mus	Rattus	Rattus	Balaenoptera
Species	europaeus	europaeus	europaeus	musculus	musculus	musculus	musculus	musculus	musculus	griseus	griseus	griseus	griseus	musculus	musculus	musculus	norvegicus	norvegicus	acutorostrata
Accession	AMDU01015984.1	AANN01268976.1	AANN01232404.1	AHBB01142544.1	AEKQ02152907.1	AEKR01205738.1	CAAA01025638.1	AEKQ02014752.1	AEKQ02050583.1	AFTD01100812.1	AMDS01087150.1	AFTD01011754.1	AMDS01100063.1	AAHY01150974.1	AAHY01024305.1	CAAA01148393.1	AAHX01013382.1	AABR06028593.1	ATDI0100021
gag																			
pro																			
pol																			
env																			
sag																			
Genus	Bos	Bos	Chrysochloris	Chrysochloris	Cricetulus	Dipodomys	Daubentonia	Elephantulus	Equus	Erinaceus	Erinaceus	Ictidomys	Leptonychotes	Lipotes	Lipotes	Loxodonta	Microtus	Mus	Mus
Species	indicus	taurus	asiatica	asiatica	griseus	ordii	madagascariensis	edwardii	caballus	europaeus	europaeus	tridecemlineatus	weddellii	vexillifer	vexillifer	africana	ochrogaster	musculus	musculus
Accession	AGFL01181997.1	AAFC03079735.1	AMDV01321751.1	AMDV01361218.1	APMK01250014.1	ABRO01007532.1	AGTM011639756.1	AMGZ01235975.1	AAWR02042168.1	AANN01090537.1	AMDU01141785.1	AAQQ0101014	APMU01073177.1	AUPI01040582.1	AUPI01145360.1	AAGU03002582.1	AHZW01004478.1	AAHY01071927.1	AEKQ02053596.1
gag																			
pro																			
pol																			
env																			
sag																			
Genus	Mus	Mus	Mus	Mus	Peromyscus	Peromyscus	Myotis	Myotis	Ochotona	Odobenus	Orycteropus	Ovis	Peromyscus	Peromyscus	Procavia	Procavia	Rattus	Rattus	Rattus
Species	musculus	musculus	musculus	musculus	maniculatus	maniculatis	brandtii	lucifugus	princeps	rosmarus	afer	aries	maniculatus	maniculatus	capensis	capensis	norvegicus	norvegicus	norvegicus
Accession	AEKQ02185596.1	AEKQ02186966.1	CAAA01150610.1	CAAA01198845.1	AYHN01003103.1	EU204642.1	ANKR01221297.1	AAPE02008149.1	AAYZ01229114.1	ANOP01040725.1	ALYB01231012.1	AMGL01129945.1	AYHN01129821.1	AYHN01152242.1	ABRQ01011474.1	ABRQ01381071.1	AABR06024640.1	AABR06032972.1	AABR06085381.1
gag																			
pro																			
pol																			
env																			
sag																			
Genus	Rattus	Rattus	Rattus	Rattus	Sus	Tarsius	Tarsius	Tarsius	Tarsius	Tupaia	Tursiops	Aotus	Cricetulus	Dipodomys	Dipodomys	Dipodomys	Dipodomys	Echinops	Erinaceus
Species	norvegicus	norvegicus	norvegicus	norvegicus	scrofa	syrichta	syrichta	syrichta	syrichta	chinensis	truncatus	nancymaae	griseus	ordii	ordii	ordii	ordii	telfairi	europaeus
Accession	AABR06098027.1	AAHX01003470.1	AAHX01032432.1	AAHX01097670.1	AOCR01001337.1	ABRT010264555.1	ABRT010313192.1	ABRT02366711.1	ABRT02472330.1	ALAR01194664.1	ABRN01205561.1	AC201708	AMDS01143269.1	ABRO01000005.1	ABRO01000060.1	ABRO01005184.1	ABRO01009415.1	AAIY02121914.1	AMDU01169997.1
gag																			
pro																			
pol																			
env																			
sag																			
Genus	Jaculus	Mus	Peromyscus	Nomascus	Ochotona	Tarsius													
Species	jaculus	musculus	maniculatus	leucogenys	princeps	syrichta													
Accession	AKZC01311602.1	CAAA01158679.1	AYHN01193045.1	ADFV01049394.1	AAYZ01351999.1	ABRT010387126.1													
gag																			
pro																			
pol																			
env																			
sag																			

Supplementary Table 1

Details of sequences included in the concatenated alignment

For each sequence, the presence/absence of each gene in the alignment is denoted as a filled or blank cell, respectively. Species names and NCBI accession codes for the source contigs are indicated.

	Model	Free parameters	lnL	AIC
	One-ratio	1	-5884.97839	11771.95679
	Free-ratio	50	-5846.46793	11792.93586
	ωt,ωi	2	-5884.49031	11772.98062
	ωt,ωi=ω0	2	-5884.49031	11772.98062
	ωimh=ωth=ωt=ω0,ωia=ωim	2	-5884.49031	11772.98062
	ωimh=ωth=ωt=ω0,ωia=1,ωim	2	-5903.75545	11811.51089
Two-ratio	ωimh=ωth=ωt=ω0,ωia,ωib=1	2	-5910.25074	11824.50149
	ωth=ωt=ω0,ωia=ωib=1,ωimh	2	-5928.62362	11861.24724
	ωth=ωt=ω0,ωia=ωib,ωimh=1	2	-5885.83891	11775.67781
	ωt=ω0,ωia=ωib=ωimh,ωth=1	2	-5909.30869	11822.61738
	ω0,ωimh	2	-5884.97754	11773.95508
	ωimh=ωth=ωt=ω0,ωia,ωim	3	-5883.49986	11772.99972
	ωth=ωt=ω0,ωia,ωib,ωimh=1	3	-5884.92127	11775.84254
	ωth=ωt=ω0,ωia=ωib,ωimh	3	-5884.47495	11774.9499
	ωth=ωt=ω0,ωia=1,ωib,ωimh	3	-5900.66225	11807.32449
	ωth=ωt=ω0,ωia,ωib=1,ωimh	3	-5910.22637	11826.45273
Three-ratio	ωt=ω0,ωia=ωib,ωimh,ωth=1	3	-5909.30822	11824.61643
	ωt=ω0,ωia=ωib=ωimh,ωth	3	-5883.29022	11772.58043
	ωt=ω0,ωia=ωib,ωimh=ωth	3	-5883.28052	11772.56104
	ωt=ω0,ωia=ωib=ωimh=1,ωth	3	-5931.70421	11869.40842
	ωt=ω0,ωia,ωib=ωimh,ωth=1	3	-5907.98418	11821.96837
	ωt=ω0,ωia,ωib=ωimh=1,ωth	3	-5910.95674	11827.91348
	ωt=ω0,ωia=1,ωib=ωimh,ωth	3	-5900.45884	11806.91768
	ωth=ωt=ω0,ωia,ωib,ωimh	4	-5883.48601	11774.97202
	ωt=ω0,ωia=ωib,ωimh,ωth	4	-5883.26153	11774.52306
Multi-ratio	ωt=ω0,ωia,ωib=ωimh,ωth	4	-5882.59582	11773.19164
	ωt=ω0,ωia,ωib,ωimh,ωth	5	-5882.4447	11774.88939
	ωt=ω0,ωia,ωib,ωimh,ωtph	5	-5883.14786	11776.29572
	ωt=ω0,ωia,ωib,ωimh,ωtmh,ωtrhp	6	-5881.73242	11775.46484

Omega (ω) types

ω0	background
ωt	tips
ωia	internal clade A
ωib	internal clade B
ωimh	internal monkey herpesviruses
ωth	tips herpesviruses
ωtmh	tips monkey herpesviruses
ωtrhp	tip rodent herpesvirus peru

Supplementary Table 2

Maximum likelihood test of selection for main phylogeny in figure 3

Because there are a high number of ways that the 50 branches could be grouped, we only tested a series of biologically realistic hypotheses. The key labeled 'Omega (ω) types' indicates the symbols used to denote branch types. Although none of the models tested offered a significantly improved likelihood according to the likelihood ratio test, 16 of the hypotheses tested ranked higher than the free-ratio model according to the Akaike Information Criterion (AIC). The free-ratio model has a significantly better likelihood than the one-ratio, and underlined models are those that performed better than the free ratio test according to the AIC.

Pteropus alecto	*Pteropus vampyrus*	ω	estimated	fixed to one	P-value
ALWS01167045.1	ABRP01226597.1	1.2887	-1468.284955	-1468.362177	0.69432418
ALWS01038626.1	ABRP01001304.1	1.2344	-1705.567551	-1705.635601	0.712189342
ALWS01093920.1	ABRP01162448.1	1.0573	-1642.736517	-1642.739836	0.935065063
ALWS01169798.1	ABRP01020384.1	1.3079	-1805.989282	-1806.245837	0.473795651
ALWS01105028.1	ABRP01171138.1	1.1196	-1736.627038	-1736.654892	0.813412943
ALWS01104918.1	ABRP01114503.1	1.25	-1767.382375	-1767.515113	0.606382783
ALWS01103775.1	ABRP01158624.1	0.8619	-1734.607552	-1734.651535	0.766779068
ALWS01097461.1	ABRP01033558.1	1.1834	-1714.53515	-1714.610998	0.69691984
ALWS01078826.1	ABRP01079374.1	1.0683	-1795.940301	-1795.954459	0.866368052
ALWS01054040.1	ABRP01167243.1	0.7182	-1756.449987	-1756.721645	0.461061708
ALWS01009353.1	ABRP01087430.1	0.7691	-1690.028939	-1690.115408	0.677513806
ALWS01048320.1	ABRP01157105.1	1.0241	-1653.640581	-1653.641404	0.967637963
ALWS01055306.1	ABRP01174128.1	0.7097	-1588.763236	-1589.245527	0.32603514
ALWS01095756.1	ABRP01176227.1	2.8441	-1698.397185	-1699.612614	0.118967769
ALWS01010934.1	ABRP01014798.1	1.0723	-1667.086515	-1667.0953	0.894547793
ALWS01014665.1	ABRP01173071.1	0.7673	-1786.965452	-1787.173963	0.518426761
ALWS01020099.1	ABRP01184288.1	1.5341	-1665.587038	-1665.742479	0.577139459
ALWS01025422.1	ABRP01046600.1	1.2849	-1745.829143	-1745.974319	0.589995049
ALWS01032946.1	ABRP01053790.1	0.8868	-1666.080683	-1666.12071	0.777224245
ALWS01033198.1	ABRP01142907.1	1.1876	-1879.600579	-1879.726197	0.616206113
ALWS01037244.1	ABRP01034590.1	1.6693	-1716.619141	-1717.322582	0.235574689
ALWS01040448.1	ABRP01159955.1	0.4293	-1143.51778	-1144.625893	0.136566538
ALWS01040757.1	ABRP01170161.1	1.2433	-1698.251283	-1698.306593	0.739439376
ALWS01043129.1	ABRP01096911.1	1.0729	-1493.436316	-1493.458623	0.832715399
ALWS01046598.1	ABRP01032009.1	0.2822	-1662.960313	-1664.65104	0.065933865

Supplementary Table 3

Maximum likelihood test of selection for orthologous sequences in *Pteropus* species

The Table shows the accession numbers of the two species for each orthologous pair of LTRs and their respective Pairwise dN/dS (ω) values obtained through CODEML. The likelihood score for the estimated value was compared to the likelihood of fixing ω to one, and P-values obtained from a chi-squared test with one degree of freedom.

Supplementary References

1. Mustafa, F., Lozano, M. & Dudley, J. P. C3H mouse mammary tumor virus superantigen function requires a splice donor site in the envelope gene. *J. Virol*. **74,** 9431–40 (2000).

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