Evolutionary biology of members of the viral kingdom *Bamfordvirae*

A thesis submitted for the degree of Doctor of Philosophy



José Gabriel Niño Barreat

Pembroke College

Trinity Term 2022

“I am thankful that I can see much to admire in all religions. [...] But whether there be a God and whatever be his nature. Whether we have an immortal soul or not, and whatever may be our state after death, I can have no fear of having to suffer for the study of nature and the search after truth".

ALFRED RUSSEL WALLACE

Dili, Timor Island, 1861

I wish to dedicate this thesis to the memory of my grandmother, Herminia Margarita Montero de Barreat (1942—2006), and to the legacy of Dr. José Gregorio Hernández Cisneros (1864—1919).

Statement of contribution and associated publications

This thesis presents chapters which have been published in peer-reviewed journals and others which are in the process of being submitted for publication. The introduction and discussion have been written specifically for this thesis.

Chapter I

Barreat JGN, Katzourakis A. 2022. Paleovirology of the DNA viruses of eukaryotes. *Trends in Microbiology*. doi: 10.1016/j.tim.2021.07.004.

I conceived the review with supervision and guidance from Aris Katzourakis, and wrote the review with supervision from Aris Katzourakis who provided advice, feedback and editorial assistance.

Chapter II

Barreat JGN, Katzourakis A. 2022. Phylogenomics of the *Maverick* virus-like mobile genetic elements of vertebrates. *Molecular Biology and Evolution*. doi: 10.1093/molbev/msaa291

I designed the study together with Aris Katzourakis, collected the data, wrote the code, conducted the analyses, prepared the figures and wrote the manuscript. The work was supervised by Aris Katzourakis who provided advice, feedback and editorial assistance.

Chapter III

I designed the study together with Aris Katzourakis, collected the data, wrote the code, conducted the analyses, prepared the figures and wrote the manuscript. The work was supervised by Aris Katzourakis who provided advice, feedback and editorial assistance.

Chapter IV

I designed the study together with Aris Katzourakis, collected the data, wrote the code, conducted the analyses, prepared the figures and wrote the manuscript. The work was supervised by Aris Katzourakis who provided advice, feedback and editorial assistance.

Chapter V

Barreat, JGN, Katzourakis A. (in press). Evolutionary analysis of placental orthologues reveals two ancient DNA virus integrations. *Journal of Virology*.

I designed the study together with Aris Katzourakis, collected the data, wrote the code, conducted the analyses, prepared the figures and wrote the manuscript. The work was supervised by Aris Katzourakis who provided advice, feedback and editorial assistance.

Table of contents

1. Abstract **17**

2. Acknowledgements **18**

3. Introduction **20**

3.1 The origin of viruses 23

3.2 Major patterns in virus evolution 29

3.3 Virus species and viral taxonomy 32

3.4 The kingdom *Bamfordvirae* 35

3.4.1 Prokaryotic relatives 37

3.4.2 Eukaryotic elements 38

3.4.2a Adenoviruses 40

3.4.2b Nucleocytoplasmic Large DNA viruses 41

3.4.2c Virophages 43

3.4.2d *Mavericks*/*Polintons* 44

3.4.2e Polinton-like viruses 45

3.4.2f Capsidless elements 46

3.5 Open questions 47

3.6 Aims and scope of this thesis 49

3.7 References 51

4. Chapter I: *Paleovirology of the DNA viruses of eukaryotes* **67**

4.1 Abstract 67

4.2 Paleovirology: probing virus origins and evolution 69

4.3 The diversity of DNA viruses found in eukaryotic genomes 72

4.4 Coevolution of DNA viruses with eukaryotes 76

4.5 Transitions between viruses and transposons 80

4.6 Functional significance and potential uses of DNA virus EVEs 82

4.7 Concluding remarks and future perspectives 82

4.8 Box 1: Evidence of ancient viruses from fossils and aDNA 83

4.9 Box 2: Mechanisms of DNA virus integration in eukaryotes 85

4.10 Box 3: Endogenous DNA viruses in immunity 87

4.11 Box 4: Virus-host horizontal gene transfers and evolutionary novelty 88

4.12 References 91

5. Chapter II: *Phylogenomics of the Maverick virus-like mobile genetic elements of vertebrates* 102

5.1 Abstract 102

5.2 Introduction 103

5.3 Materials and methods 106

5.4 Results 115

5.5 Discussion 126

5.6 References 135

6. Chapter III: *A billion years arms-race between viruses, virophages and eukaryotes* 144

6.1 Abstract 144

6.2 Introduction 145

6.3 Materials and methods 149

6.4 Results 156

6.5 Discussion 163

6.6 Concluding remarks 169

6.7 References 170

7. Chapter IV: *Ecological and evolutionary dynamics of cell-virus-virophage systems* 179

7.1 Abstract 179

7.2 Introduction 180

7.3 Materials and methods 182

7.4 Results 191

7.5 Discussion 206

7.6 References 210

8. Chapter V: *An ancient endogenous DNA virus in the human* genome 215

8.1 Abstract 215

8.2 Introduction 216

8.3 Materials and methods 220

8.4 Results 223

8.5 Discussion 230

8.6 References 236

9. Discussion 243

9.1 Main findings and implications 243

9.2 Limitations 252

9.3 Future perspectives 253

9.4 Conclusion 256

9.5 References 257

10. Appendix to Chapter I **262**

10.1 Glossary 262

10.2 Highlights 264

10.3 Outstanding questions 265

11. Appendix to Chapter II **267**

11.1 List S1. Taxonomy of host species whose genomes harbour *Mavericks* 267

12. Appendix to Chapter III **285**

13. Appendix to Chapter IV **301**

13.1 Text S1. Derivation of the displacements of cells and viruses relative to virophages 301

14. Appendix to Chapter V **312**

15. Appendix to the Discussion 319

List of figures

1. **Chapter I**

1.1 Figure 1. Endogenous viral element (EVE) formation and persistence in eukaryotic genomes 70

1.2 Figure 2. Evolutionary trees of the major lineages of DNA viruses found in eukaryotic genomes 77

1.3 Figure 3. Evidence of viruses in the geological fossil record 85

2. **Chapter II**

2.1 Figure 1. Genome organisation and diagram of the virion of Human adenovirus type 5 (HAdV-C5) 104

2.2 Figure 2. The genomic architectures of vertebrate *Mavericks* 117

2.3 Figure 3. Copy numbers of *Mavericks* are low in most taxa 119

2.4 Figure 4. Maximum credibility tree of vertebrate *Mavericks* based on the protein-primed DNA polymerase B 121

2.5 Figure 5. *Mavericks* have switched hosts frequently in cyprinid fish 124

3. **Chapter III**

3.1 Figure 1. The two main hypotheses for the origin of eukaryotic bamfordviruses including virophages and NCLDVs 148

3.2 Figure 2. Rooted Bayesian maximum clade credibility tree of the major lineages of eukaryotic *Bamfordvirae* 158

3.3 Figure 3. Outgroup-rooted maximum likelihood tree of the major lineages of eukaryotic *Bamfordvirae* 159

3.4 Figure 4. Evolutionary model for the origin of the major lineages of eukaryotic *Bamfordvirae* 167

4. Chapter IV

4.1 Figure 1. Model of virophage integration into the cell genome following the Mavirus mechanism 184

4.2 Figure 2. Model of virophage integration into the cell genome following the Sputnik mechanism (1a) 186

4.3 Figure 3. Model of virophage integration into the host virus genome following the Sputnik mechanism (1b) 188

4.4 Figure 4. Dynamics of the Mavirus and Sputnik 1a models in the absence of virophage integration 193

4.5 Figure 5. Dynamics of the Mavirus and Sputnik 1a models in the presence of virophage integration 195

4.6 Figure 6. Outcomes depending on the virophage infection mechanism 197

4.7 Figure 7. Dynamics of the Sputnik 1b model with integration into the host virus genome 200

4.8 Figure 8. Maximum amplitude of the virus wave 203

4.9 Figure 9. Time to virus extinction 204

4.10 Figure 10. Number of surviving cells 205

5. Chapter V

5.1 Figure 1. Comparison of the regions orthologous to the chromosome 7 *Maverick* integration in humans (*Homo sapiens*) 227

5.2 Figure 2. Comparison of the regions orthologous to the chromosome 8 *Maverick* integration in humans (*Homo sapiens*) 228

5.3 Figure 3. Phylogenetic tree of the mammalian polymerase B-like sequences orthologous to the viral fossils in human chromosomes 7 and 8 229

6. Discussion

6.1 Figure 1. Monomeric structures of the major capsid proteins (hexons) of Mavirus (*Lavidaviridae*), Faustovirus (NCLDV) and Chimpanzee adenovirus type 68 (*Adenoviridae*) 245

6.2 Figure 2. Monomeric structures of the minor capsid proteins (pentons) from Mavirus (*Lavidaviridae*), *Paramecium bursaria Chlorella virus 1* (PBCV1, NCLDV) and Human adenovirus 5 (*Adenoviridae*) 245

6.3 Figure 3. *De novo* predicted structures of the major capsid (hexon) and minor capsid (penton) monomers for two intact *Mavericks* 247

6.4 Figure 4. *De novo* predicted structure for the trimer of hexon proteins in two intact *Mavericks* 248

6.5 Figure 5. *De novo* predicted structure for the minor capsid pentamer in two intact *Mavericks* 249

7. Appendix to Chapter II

7.1 Figure S1. Genome and TIR size distributions of intact *Mavericks* found in the genomes of 17 species of vertebrates 279

7.2 Figure S2. Sequence logos of the conserved motifs at the termini of vertebrate *Mavericks* 280

7.3 Figure S3. Time tree showing the integration events that gave rise to different orthologue groups as inferred from the host phylogeny 281

7.4 Figure S4. Majority-rule consensus tree of the intact vertebrate *Mavericks* based on the protein-primed DNA polymerase B 282

8. Appendix to Chapter III

8.1 Figure S1. Bayesian maximum credibility tree of the major capsid protein inferred with a relaxed molecular clock 293

8.2 Figure S2. Bayesian maximum credibility tree of the minor capsid protein inferred with a relaxed molecular clock 294

8.3 Figure S3. Bayesian maximum credibility tree of the ATPase inferred with a relaxed molecular clock 295

8.4 Figure S4. Bayesian maximum credibility tree of the protease inferred with a relaxed molecular clock 296

8.5 Figure S5. Bayesian maximum credibility tree of the ATPase inferred with a relaxed molecular clock (monophyletic constraint on virophages) 297

8.6 Figure S6. Bayesian maximum credibility tree of the protease inferred with a relaxed molecular clock (monophyletic constraint on virophages) 298

8.7 Figure S7. Maximum-likelihood unrooted phylogenetic trees of the transcriptional homologues encoded in cytoplasmic linear plasmids 299

9. Appendix to Chapter IV

9.1 Figure S1. Dynamics of the Mavirus and Sputnik mechanisms in the presence of programmed cell-death (virophages assumed to be neutral) 304

9.2 Figure S2. Time-course of the population of viruses during an ABM simulation 305

9.3 Figure S3. Time-course of the population of virophages during an ABM simulation 306

9.4 Figure S4. Time-course of the number of surviving cells during an ABM simulation 307

10. Appendix to Chapter V

10.1 Figure S1. Distribution of observed genetic distances in the chromosome 7 *Maverick* integration and in neighbouring noncoding regions 312

10.2 Figure S2. Distribution of observed genetic distances in the chromosome 8 *Maverick* integration and in the neighbouring noncoding regions 313

List of tables

1. **Chapter II**

   1.1 Table 1. Maximum-likelihood estimates for the non-synonymous to synonymous ratios (ω) of the 8 conserved genes in cyprinid *Mavericks* 125

   1.2 Table 2. Proposed species-level classification of vertebrate *Mavericks*, with designated type sequences 132

2. **Chapter III**

   2.1 Table 1. Comparison between the nuclear-escape and alternative maximum-likelihood models based on the Akaike information criterion (AIC) 161

   2.2 Table 2. Posterior model odds of the nuclear-escape and alternative hypotheses using concatenated and single-protein data sets 162

   2.3 Table 3. Bayesian stepping-stone analysis of the nuclear-escape and alternative hypotheses 163

3. **Chapter V**

   3.1 Table 1. Posterior estimates for the age of the root, and the minimum ages of the chromosome 7/8 integrations 229

4. **Appendix to Chapter II**

   4.1 Table S1. Copy numbers of intact *Mavericks* in the genomes of 17 vertebrate species 283

   4.2 Table S2. Root ages and evolutionary rates estimated from the Bayesian evolutionary analyses 284

5. **Appendix to Chapter III**

   5.1 Table S1. Viral species included in the final multiple sequence alignments (54 taxa) with their accession numbers and source reference 285

5.2 Table S2. Distribution of root positions calculated from the MCMC posterior tree sample 291

5.3 Table S3. Cytoplasmic linear plasmids used for querying the databases in search for protein homologues 300

6. Appendix to Chapter V

6.1 Table S1. Calibration points from TimeTree used for the divergence-time analysis in BEAST 314

6.2 Table S2. Significant matches to the *Maverick*-PolB protein of *Terrapene carolina* in marsupial and monotreme genomes 315

Abstract

Members of the viral kingdom *Bamfordvirae* arguably represent the most diverse lineage of viruses infecting eukaryotes. They comprise the Nucleocytoplasmic Large DNA viruses (NCLDVs), virophages, adenoviruses, *Maverick*/*Polintons* and Polinton-like viruses. Most elements encode a conserved module of genes involved in capsid formation: two jelly-roll fold capsid proteins, an FtsK-HerA DNA-packaging ATPase and an adenoviral-like protease. However, over the course of evolution, these viruses diversified to an unprecedented extent by numerous gene gains and losses, replacement of the ancestral replicase and acquisition of integrases. I investigated aspects of the evolutionary biology of this viral kingdom by combining results from paleovirology, phylogenetics, mathematical modelling and computational simulations. First, I show that *Mavericks* have undergone frequent cross-species transmissions and the genes involved in capsid formation are under strong purifying selection. These observations imply that *Mavericks* are endogenous viruses and not merely transposons. *Maverick* viruses were estimated to have infected vertebrates for at least 419 million years. Next, I found support for an alternative evolutionary scenario for the origin of the bamfordviruses of eukaryotes, which is at odds with the nuclear-escape hypothesis (origin from a *Maverick*-like element that escaped from the nucleus). The topology of the phylogenetic tree suggests that this viral lineage has infected eukaryotes for > 1 billion years. By modelling the interactions between cells, viruses and virophages, I argue that the virophage entry mechanism is potentially the main driver of the different integration patterns observed for Mavirus/Spuntik-like virophages. Theoretical analyses also supported the role of virophage inhibition, programmed cell-death and multicellularity as effective antiviral mechanisms in microbial eukaryotes. Finally, I explored the evolutionary biology of two *Maverick* integrations found in the human genome, which integrated at least ~102 million years ago. These results shed light on the ancient interactions of viruses with eukaryotes, and present new questions and hypotheses open for investigation.

Acknowledgements

My deepest gratitude belongs to my family, and in particular to my mum, brother, uncles, aunts and cousins, whose love, support and encouragement has been essential throughout my life and in completing this DPhil.

I am thankful to my supervisor, Professor Aris Katzourakis, for his guidance throughout these years, his constant support and insights during the various stages of this DPhil project. I am also grateful to the support, feedback and lessons from Professors Peter Simmonds (Medicine), Stuart West (Biology), Nick Lloyd Trefethen (Mathematics) and Alan Garfinkel (Visiting Professor, Computer Science), which served to enrich and improve the chapters presented in this thesis. I thank Professor Peter Simmonds and Dr. Alexander Suh for being the examiners of my thesis, and for their valuable and constructive feedback.

This thesis was possible thanks to the generous support of the National Academy of Medicine of Venezuela and Pembroke College, through the *"Dr. José Gregorio Hernández"* Award. In particular, I owe much gratitude to Dr. Alfredo Díaz Bruzual, Dr. Enrique López Loyo and Dr. Guillermo Colmenares, for supporting my application to Oxford and for kindly inviting me to present a talk at the Solemn Session of the Academy in Caracas. I am also immensely grateful to the staff at Pembroke College; in particular to Nancy Braithwaite, Alison Franklin (now at Linacre), Caroline Barnes, and to the former and current Masters, Dame Lynn Brindley and Sir Ernest Ryder.

I would also like to give a heartfelt acknowledgement to my friends at Pembroke College, the Department of Biology and at the Peter Medawar Building, who were a constant source of joy and support during my studies. I would like to mention Simonetta Spavieri, Susana Rostro García, Pak Hei Yeung, Flaminia Pisscheda, Carolina Toso, Gonzalo Gonzalez de Diego, Marc Rigter, Siddarth Kumar, Flora Charbonnier, Fanny Clemente, Scott Moynihan, Kathryn White, Hannah McIntyre, Cilian Brophy, Bénjamin Bréant, Rohan Watt, Bestin Samuel, Joseph Rosing, Patricia Cipollitti, Patricia Castillo Silva, Filippo Ugolini, Corina Smith, Emilia Skirmuntt, Alex Zarebski, Mahan Ghafari, Lucy Matkin, Darlan Candido, Bernando Guitiérrez Granja, Marina Escalera Zamudio, Sarah François, Charles R. de Souza and Yuyang Chen. I would also like to thank Juliana Choi for her love and company during the final stages of my DPhil.

Introduction

Viruses are highly diverse and abundant parasitic genetic replicators that use protein capsids for transmission between cells. The word *virus* – meaning "poison" in Latin – came to replace the earlier term *contagium vivum fluidum*, which was used to denote an infectious substance that could pass through bacterial filters and cause a variety of diseases in plant and animal hosts (Enquist and Racaniello, 2013). However, it was soon realised that the infectious substance could not be grown in pure culture (i.e. in isolation from its host), thus defying Koch's 2nd postulate (Artenstein, 2012; Enquist and Racaniello, 2013). These observations made it clear that viruses were a distinct class of infectious agents fundamentally different from the bacterial pathogens known at the turn of the 19th century (Artenstein, 2012).

A virus is built from two essential components: an RNA or DNA genome, which carries the genetic information, and a protein shell that coats the genome, and serves the dual purpose of protection in the environment and as a vehicle for transmission between hosts (Harrison, 2013). Some viral particles have an extra membrane derived from the host which may also aid in protection and facilitate fusion and release of the virus genome into the host cell (Lenard, 2008). In addition to these components, some viral particles may also package an array of host-derived proteins, RNAs and even organelles such as ribosomes (Bechtel et al., 2005; Compans, 1993; Onafuwa-Nuga et al., 2005). The mature and infectious viral particles are called virions.

Unlike a bacterium that grows exponentially at the right conditions in pure culture, when viruses are grown together with host cells they follow a characteristic one-step growth curve with “eclipse” and “latent” phases; the periods between infection and formation of the first intracellular viral particles, and until they are relasead extracellularly (Condit, 2013; Ellis and Delbrück, 1939). At first, the viral particles enter cells by binding to specific cell surface receptors and so the number of particles in solution decreases which constitutes the start of the eclipse and latent phases (Condit, 2013). Once the virus is inside the cell, the virions dis-assemble and the genome can gain access to host cell resources that are needed for expression of viral genes, protein synthesis and replication (Whelan, 2013). Although some large DNA viruses encode components used for translation, such as their own transfer RNAs (tRNAs) and amino-acyl RNA synthetases (aaRS), they all depend on ribosomes of cellular organisms which seem to be absent from the virosphere (Abrahão et al., 2018). Self-assembly of new viral particles then takes place and the viral progeny is released, either by cell lysis or budding, leading to an increase in the number of infectious viral particles in the extracellular medium (Hunter, 2013).

The unique parasitic lifestyles and constitution of viruses have led to heated debates over whether they should be considered living organisms. Viruses seem to lie at the boundary between cellular organisms, which are universally agreed to be living, and biological macromolecules such as DNA/RNA which are considered to be non-living. As a system, viruses can reproduce and evolve with the help of cells, not unlike obligate intracellular bacteria which depend on their eukaryotic hosts for survival and reproduction (Loterio et al., 2021; Zientz et al., 2004). However, virions do not have an active metabolism and so do not exhibit the properties of system self-renewal and maintenance (*autopoiesis*) which is characteristic of cells and present in parasitic cellular organisms (Luisi, 2003). To bridge

this divide, the concept of the "virocell" has been proposed which emphasises the distinction between virions and viruses, and recognises the dual nature of viruses. The virocell concept argues that the living state of the virus is the metabolically-active infected cell, reprogrammed to produce virions, while virions are the non-living state of the virus used for dispersion (Forterre, 2013, 2011). According to this concept, the "real" viral organism is not the virion but the infected cell (Forterre, 2013, 2011). However, it may be argued that viruses are simply non-living molecular components that become part of a living system when they infect a cell; the living state being an emergent property determined by the dynamic interactions of components encoded by the cell, not the virus.

In addition to contributing to an important burden of disease in humans, animals and plants, illustrated most recently by the SARS-CoV-2 pandemic, viruses play major roles regulating ecosystem function and have had a major impact on the evolutionary history of life on Earth. For example, until its eradication through a global vaccination programme, smallpox, a disease caused by *Variola virus*, killed an estimated 300 million people from 1900 to 1980 (Henderson, 2011). Viruses are also emerging as key regulators of microbial species in aquatic and soil ecosystems thus affecting nutrient availability and impacting biogeochemical cycles, e.g. carbon, nitrogen, sulphur and phosphorous (Bi et al., 2021; Coutinho et al., 2020; Jover et al., 2014; Roux et al., 2017; Shelford et al., 2012). Perhaps most intriguingly, viruses have contributed to driving key evolutionary innovations in host species. For instance, viruses that may integrate into the host genome can be later used as an antiviral restriction factor against exogenous viruses, resulting in the evolution of an immune defence mechanism (Aswad and Katzourakis, 2012). In placental mammals, repurposed retrovirus *env* genes are expressed during formation of the placenta and contribute fusing cells to each other, thus forming the syncytiotrophoblast (Cornelis et al.,

2015; Dupressoir et al., 2009; Roberts et al., 2021). These proteins also function in protecting the developing embryo from attack of the mother's immune system (Dupressoir et al., 2012). Interestingly, the genus *Mabuya* of viviparous lizards has independently repurposed one of these retroviral genes for formation of its own placenta (Cornelis et al., 2017).

Viruses are therefore fascinating subjects of study, which can illuminate central aspects about the biology of host organisms and the history of life on Earth. Furthermore, they can help us elucidate general biological principles which may be true for extra-terrestrial life forms and their parasites. In this chapter, I will discuss our current knowledge on the origin of viruses and the main patterns of virus evolution, describe the concept of a viral species and the scheme of viral classification, and finally introduce the viral lineage which will be the focus of this thesis: the kingdom *Bamfordvirae*. I will conclude by discussing the open questions in the field and by defining the aims and scope of the work presented in this thesis.

The origin of viruses

The history of cellular life forms on Earth can be traced back by studying fossil remains in the geological record, the presence of chemical biosignatures on rocks, and by comparative genomic analyses among extant organisms. The oldest uncontroversial microfossils, whose size and morphology resemble bacteria and which occur together with chemical signs of sulphur metabolism (pyrite crystals), occur in the Strelley Pool formation of Western Australia and have been dated back to 3.4 billion years (Wacey et al., 2011). Older fossil evidence of life has been described more recently from Greenland and Canada, ranging from 3.7 to 4.28 billion years ago, although their biogenic origin is debated (Allwood et al., 2018;

Dodd et al., 2017; Dominic et al., 2022; Nutman et al., 2016). Enriched isotopic ratios of $^{12}C/^{13}C$ in graphite inclusions of zircon crystals in Jack Hills (Western Australia), have been interpreted as evidence for the activity of carbon-fixating enzymes 4.1 billion years ago (Bell et al., 2015). These observations are consistent with integrated analyses of phylogenetic and palaeontological data which suggest an early origin of life on Earth, placing the time of the Last Universal Common Ancestor (LUCA) of cellular organisms, 3.9 – 4.1 billion years ago during the Hadean or early Archaean Eons (Battistuzzi et al., 2004; Betts et al., 2018).

The RNA world hypothesis is a widely accepted view that RNA replicators predated the origin of DNA and proteins. Support for this idea comes from the ability of RNA to store genetic information (e.g. some viruses have RNA genomes), the enzymatic capacity of some RNA molecules (ribozymes), and the observation that RNA components can form abiotically (Higgs and Lehman, 2015; Cafferty and Hudd, 2014). It is assumed that DNA could not have evolved before the origin of a ribonucleotide reductase (Teichert et al., 2019), and so the easiest path to biological complexity is from these early RNA replicators. However, an RNA ribozyme that can catalyse its own replication has yet to be described, and recent works have shown that both DNA nucleosides can arise from abiotic reactions (Teichert et al., 2019) and that RNAs with non-canonical bases (present in tRNAs and rRNAs) are capable of synthesis of small peptides (Müller et al., 2022). Therefore, it seems that early RNA-peptide or DNA-RNA-peptide worlds are plausible alternatives to a pure RNA world, and that these molecular species may have coevolved from early times.

Since viruses are obligate parasites of cells and use a capsid for crossing cell membranes, any of the first viruses could not have existed before the origin of the first cells; however, individual viral components may be even older. The first viruses could have evolved at a

time earlier to LUCA, but only after membranes were recruited to demarcate "individuals" from a pre-cellular gene pool. As such, it seems that simpler selfish genetic elements, which do not encode for capsids, would have predated viruses and evolved to be the first genetic parasites. These genetic parasites would have taken advantage of the first ensembles of cooperating genes and used their ribosomes to increase in number at the expense of the co-operators (perhaps the evolution of cell membranes was an early counter-measure to fight off these parasites and stabilise communities of co-operators?). It is possible that some replication enzymes that survive to the present day in viruses and transposable elements, such as the reverse-transcriptase (RT), RNA-dependent RNA polymerases, protein-primed and primer-independent DNA polymerases, may be relics of this pre-cellular world (Krupovič et al., 2019a).

By the time LUCA evolved, cells would have been considerably complex systems. Phenotypic and genomic reconstructions of LUCA indicate it was likely a motile, anaerobic, carbon- and nitrogen-fixating organism that used hydrogen as an electron donor (el Baidouri et al., 2020; Weiss et al., 2016). It had a single cell membrane with a mix of bacterial/archaeal lipids, living in aquatic environments with moderate to high temperatures (Catchpole and Forterre, 2019; el Baidouri et al., 2020; Krupovič et al., 2020a; Weiss et al., 2016). LUCA probably had a DNA genome replicated by an archaeal-like DNA polymerase (PolD), an inference based on the structural homologies between the polymerases used for replication and transcription, and the universal presence of key proteins involved in DNA-replication (e.g. sliding clamp, clamp loader ATPase, ssDNA binding protein) across the domains of life (Koonin et al., 2020b). The genome of LUCA was also complex, with an estimated size of 2.32–2.48 million base pairs and 355–2624 different genes (el Baidouri et al., 2020; Weiss et al., 2016).

In addition to this complexity, it is likely that LUCA had its own parasitic genetic elements, an inference based on theoretical arguments, comparative studies of viruses that infect bacteria and archaea, and phylogenetic evidence for the coevolution of these viral lineages with their hosts over geological timescales (Krupovič et al., 2020a). The appearance of parasitic genetic elements seems to be an inevitable evolutionary outcome of Darwinian replicator systems (Iranzo et al., 2016a; Koonin et al., 2017), and so it is likely that LUCA was already infected by a number of these genetic parasites. An attempt to reconstruct the virome of LUCA has been made by examining the host ranges and evolutionary relationships of viruses that infect bacteria and archaea (Krupovič et al., 2020a). It was suggested by the authors of this study, that double-stranded DNA viruses belonging to the realms *Duplodnaviria* and *Varidnaviria* would have been part of LUCA's virome (Krupovič et al., 2020a). Viruses in the realm *Duplodnaviria* include the tailed bacteriophages that infect bacteria and archaea (*Caudovirales*), as well as the herpesviruses that infect animals (*Herpesvirales*). They share a common major capsid protein with an HK97 fold in addition to homologous terminases (DNA-packaging proteins), portal proteins and maturation proteases (Cheng et al., 2004; Duda and Teschke, 2019; Iranzo et al., 2016b; Koonin et al., 2020a). The realm *Varidnaviria* also includes viruses that infect bacteria, archaea and eukaryotes, which have capsid proteins with either single or double vertical double-jelly roll folds (Koonin et al., 2020a). Phylogenies of the *Duplodnaviria* and *Varidnaviria* infecting bacterial and archaeal hosts, do suggest that these viruses have maintained stable associations with their host lineages and coevolved with them for considerable periods of time (Krupovič et al., 2019b; Low et al., 2019). However, it is still not entirely clear which aspects influence the distribution of viral lineages across different organisms, and so deep

inferences about the viral evolutionary past based on their current hosts should be taken with caution.

Unlike cellular organisms, which have evolved from a single common ancestor, the main viral lineages seem to have distinct origins, given that viruses lack universal genetic markers (Koonin et al., 2021a). The emerging scenario in virus evolution points to a chimeric origin of viruses, with replication modules potentially derived from various sources (primordial pool of genetic replicators, other mobile elements, captured host replicases), while the capsid-formation modules have been recruited on multiple occasions independently from repurposed host proteins (Krupovič et al., 2009; Krupovič and Koonin, 2017). Some viruses and mobile genetic elements use replication strategies which have no counterpart in the replication of cellular genomes, suggesting an origin before the emergence of LUCA. In the view of the RNA-world hypothesis, the RNA-dependent RNA polymerases (RdRps) of RNA viruses may be descended from some of the first protein enzymes responsible for replicating pre-cellular RNA genomes after the origin of the genetic code (Koonin et al., 2020a). Structural analyses of the core catalytic domains suggest that reverse-transcriptase (RT), the DNA-dependent RNA and DNA polymerases, may have evolved from these primitive RdRps with RNA-recognition motif (RRM, palm) and double-psi beta-barrel (DPBB) domains at a pre-cellular stage (Koonin et al., 2020b), during the transition from RNA to DNA-based genomes.

Acquisition of a capsid is the critical event underlying the origin of a viral lineage from a lineage of non-viral mobile genetic elements. Sequence and structure comparisons of the capsid, nucleocapsid and matrix proteins of diverse viruses have revealed homologies to cellular proteins, suggesting these structural elements of viruses were acquired

independently on multiple occasions by non-viral replicators (Krupovič et al., 2019a; Krupovič and Koonin, 2017). For example, diverse cellular homologues of the most common protein fold found in viral capsid proteins, the horizontal single jelly-roll fold, were found in the genomes of archaea, bacteria and eukaryotes (Krupovič and Koonin, 2017). These cellular homologues are related to carbohydrate-binding proteins, chaperones involved in nucleosome formation (nucleoplasmins/nucleophosmins) and members of the tumour-necrosis factor superfamily (Krupovič et al., 2019a; Krupovič and Koonin, 2017), which can form oligomeric assemblies inside cells (Liu et al., 2002; Namboodiri et al., 2003). Clustering of these proteins with their viral relatives, suggests viruses acquired them independently on two to three occasions (Krupovič et al., 2019a). Another example are the alphaviruses (family *Togaviridae*), whose capsid proteins seem to be derived from chymotrypsin-like serine proteases (Krupovič and Koonin, 2017). The matrix and nucleocapsid domains of the structural polyprotein of retroviruses (Gag), also appear to have evolved from a cellular source since they resemble diverse helix-turn-helix (HTH) domain and Zn-knuckle domain host proteins (Krupovič et al., 2019a; Krupovič and Koonin, 2017). However, the origin of the capsid proteins in some viral lineages remains uncertain, for example the HK97 fold in *Duplodnaviria* (Krupovič and Koonin, 2017), which suggests that cellular homologues remain to be characterised or that some viral lineages may have acquired their capsid proteins via alternative evolutionary pathways.

The worlds of viruses and transposons are closely intertwined. Transposons are mobile genetic elements that are integrated into host genomes and lack genes for capsid formation. They are classified into two main classes, the retroelements or “copy-paste” transposons (Class I), which use an RNA and reverse transcriptase for replication, and the DNA or “cut-and-paste” transposons (Class II), which use a transposase to excise themselves and insert

in a new location (Wells and Feschotte, 2020). In addition, transposons can be classified as autonomous or non-autonomous, depending on whether they either encode all the genes and regulatory elements required for their own transposition (e.g. LINE elements), or if they require complementation by another element (an example being the *Alu* elements which are mobilised by LINEs) (Wells and Feschotte, 2020). It has been noted that mobile elements traditionally referred to as LTR retrotransposons, i.e. *Bel/Pao*, *Ty3/Gypsy*, and *Ty1/Copia*, are actually endogenous viruses now formally classified in the families *Belpaoviridae*, *Metaviridae* and *Pseudoviridae* (Koonin et al., 2021b). Recently, it has been proposed that capsid-encoding "transposons" be incorporated into the "orthovirsophere" (virosphere proper), while transposons and other mobile elements with virus-like properties join the "perivirosphere" (Koonin et al., 2021b).

Major patterns in virus evolution

At the ecological level, the evolution of viruses occurs by two main processes: stable association and codivergence with their hosts or cross-species transmissions and adaptation to new hosts. The degree to which viruses have codiverged with their hosts or undergone cross-species transmissions can be estimated by comparisons of the host and virus phylogenies. Analyses of this kind have revealed that although many viruses show a strong signal for stable host associations and co-divergence over tens of thousands to millions of years, cross-species transmissions are also frequent events during viral evolution (Geoghegan et al., 2017; Sharp and Simmonds, 2011). In RNA viruses, for example, cross-species transmissions may even be the dominant evolutionary pattern (Leopardi et al., 2018; Shi et al., 2018). These observations can be understood in the context of viral adaptation to the ecological niche provided by their hosts. While in the short-term viruses can show high

genetic variation (they are effectively exploring the sequence space tolerated by their hosts), over longer timescales, the virus will have optimised its phenotype to match its host niche and so accumulate relatively few genetic changes (Simmonds et al., 2019). This gives rise to a time-dependent phenomenon in the evolutionary rates of viruses, which are high during short timescales but which approximate the host rate over longer periods of time (Aiewsakun and Katzourakis, 2016; Ghafari et al., 2021). Viruses that jump between species are under high selective pressure to adapt to their new host niche, but once adapted, they can remain in stable association with their new host for prolonged periods of time (Simmonds et al., 2019). This time-dependent rate phenomenon is not unique to viruses, and has also been observed in the rates of molecular evolution for vertebrates, insects and bacteria (Ho et al., 2015).

At the genetic level, recombination and horizontal gene transfer are also major drivers of virus evolutionary change. Recombination is a mechanism of reciprocal exchange of genetic material that occurs when multiple viral strains coinfect a cell, and which can result in the origin of new combinations of genetic variants from different parental strains (Pérez-Losada et al., 2015). This mechanism can increase the genetic variability between homologous genes shared by different viral strains/species, which can increase the fitness of the virus when it is adapting to its host (Pérez-Losada et al., 2015). The process of recombination has been observed for both RNA and DNA viruses, and has been implicated as an important factor enabling the change in virus virulence, tropism and in cross-species transmissions (Graham and Baric, 2010; Martin et al., 2011; Xiao et al., 2016). Horizontal gene transfer (HGT) refers to the acquisition of new genetic material, usually carrying non-homologous genes, from the genome of a different species. In viruses, this can involve acquisition of new genes from other viruses, non-viral mobile genetic elements and from cellular organisms.

Phycodnaviruses and mimiviruses for example, which are large dsDNA viruses that infect algae and amoeba, encode tens to hundreds of proteins derived from bacterial or eukaryotic sources, and which seem to have been captured relatively recently by these lineages (Filée et al., 2008; Iyer et al., 2006). The entomopoxviruses and baculoviruses, which are unrelated dsDNA viruses that infect insect hosts, show signs of virus-to-virus gene exchanges given that multiple shared genes seem exclusively viral (with no cellular homologues), or are each other's closest relatives in phylogenetic trees (Iyer et al., 2006; Thézé et al., 2015). Non-autonomous transposable elements (MITEs) have been discovered in the genome of the giant virus *Pandoravirus salinus*; they were probably acquired from the host amoeba and have potentially contributed to the genetic novelty in this lineage given that they are found in the proximity to genes or within coding sequences (Sun et al., 2015).

Another important aspect that influences the evolutionary patterns of viruses is their strategy for persistence in host populations. All viruses must infect a new host organism in order to persist in time, however this can be achieved by two main strategies: 1) exclusive horizontal transmission between susceptible hosts (exogenous viruses), or 2) a mixed strategy with vertical transmission in the host germline coupled with cycles of reinfection and/or infection of new individuals (endogenous viruses). Exogenous viruses can also happen to integrate into the host germline and become endogenous "by chance" (Katzourakis and Gifford, 2010); however, since they have not evolved dedicated mechanisms for integration into the host genome, I would argue these viruses persist mainly by exclusive horizontal transmission and so are not mixed strategists. Viruses that use the exogenous strategy rely on a number of active infections in the host population, or in a different species (host reservoir), which enable them to remain in circulation by producing infectious viral particles. In contrast, viruses that adopt the mixed endogenous/exogenous strategy have

evolved mechanisms to infect cells of the germline and integrate the viral genome into the host chromosome. The viral genetic information is stored and replicated with the host genome, and can be passed to the next generation of hosts without the formation of viral particles (Aiewsakun and Katzourakis, 2015). An endogenous virus can still reactivate formation of viral particles, which allows them to increase their copy number in the same genome by reinfection or to colonise the genome of a new individual (Belshaw et al., 2004; Tarlinton et al., 2006). The mixed strategy has the advantage that the virus can be passively replicated by the host genome and survive prolonged periods of time hiding from the hosts' immune system; however, it must invest in an integration mechanism, be capable of reactivating and infecting cells of the host germline. Although most viral lineages are exogenous, and this seems to be the most successful strategy, the endogenous strategy has evolved independently in several lineages of RNA and DNA viruses. Examples of RNA and DNA viruses that have endogenous forms are the retroviruses, metaviruses, pseudoviruses, belpaoviruses, caulimoviruses, *Mavericks*, *Teratorns*, and some endogenous virophages.

Virus species and viral taxonomy

The International Committee on the Taxonomy of Viruses (ICTV) defines viruses as *"a type of MGE [mobile genetic element] that encode at least one protein that is a major component of the virion encasing the nucleic acid of the respective MGE and therefore the gene encoding the major virion protein itself"* (Lefkowitz et al., 2018). A virus species is the lowest taxonomic rank used in viral classification and is defined by the ICTV as *"a monophyletic group of MGEs whose properties can be distinguished from those of other species by multiple criteria"* (Lefkowitz et al., 2018). However, this definition does not encompass capsidless fungal and plant RNA viruses currently classified in the families

Narnaviridae, *Hypoviridae* and *Endornaviridae* (Dolja and Koonin, 2012). Members in these three families contain the RNA-dependent RNA polymerase (RdRp) gene, which is a viral hallmark gene with a monophyletic origin in (-)ssRNA, (+)ssRNA and dsRNA viruses, and have clearly evolved from capsid-forming RNA viruses (Dolja and Koonin, 2012; Koonin et al., 2020a). The criteria by which species of viruses may be distinguished from each other include the host range of the virus, its virulence, tissue and cell tropism, and its degree of genetic relatedness to other viruses (Lefkowitz et al., 2018). The classification of viruses has 8 hierarchal and nested primary ranks which aim to reflect the evolutionary affinities of viruses within common lines of descent (Gorbalenya et al., 2020). These ranks are from lowest to highest: species (ending in *virus*), genus (-*virus*), family (-*viridae*), order (-*virales*), class (-*viricetes*), phylum (-*viricota*), kingdom (-*virae*) and realm (-*viria*) (Gorbalenya et al., 2020).

A virus realm encompasses a group of viruses with the deepest evolutionary connections pointing to a common origin. A number of viral taxa await formal classification into one of the existing realms or new ones, which will depend on investigations of the deep evolutionary history of these viruses to the rest of the virosphere. Currently the ICTV recognises 6 virus realms: *Adnaviria* (archaeal viruses with filamentous virions and a dsDNA genome), *Duplodnaviria* (tailed bacterial/archaeal dsDNA viruses and herpesviruses), *Monodnaviria* (ssDNA viruses encoding a HUH-superfamily endonuclease), *Riboviria* (viruses encoding an RdRp or RT, with both RNA or DNA genomes), *Ribozyviria* (RNA virus satellites related to hepatitis delta virus) and *Varidnaviria* (dsDNA viruses with vertical jelly-roll fold capsid proteins) (Lefkowitz et al., 2018; Walker et al., 2021, 2020).

An alternative functional classification of viruses is the Baltimore scheme which groups viruses by the structure of their genomes and the strategy used for expression of their genetic information. The Baltimore scheme classifies viruses into 7 "classes": dsDNA viruses (group I), ssDNA viruses (group II), dsRNA viruses (group III), (+)-ssRNA viruses (group IV), (-)-ssRNA viruses (group V), ssRNA-RT viruses (group VI) and dsDNA-RT viruses (group VII) (Condit, 2013). An advantage of the Baltimore classification is that it is all-encompassing so every virus can be assigned to a Baltimore group, and it provides information about the virus lifecycle; however, the major limitation is that the classification is not based on evolutionary criteria. Viruses from Baltimore groups III-VII correspond to the realm *Riboviria*, most ssDNA viruses in Baltimore group II belong to the realm *Monodnaviria* (with the exception of the families *Finnlakeviridae*, *Anelloviridae* and *Spiraviridae*), while dsDNA viruses in Baltimore group I have either been classified in the realms *Adnaviria*, *Duplodnaviria* or *Varidnaviria* (Koonin et al., 2020a; Lefkowitz et al., 2018), or remain unclassified at the realm level. However, it is worth noting that there is no direct mapping between the ICTV and Baltimore classification, and that the Baltimore groups were not designed to represent monophyletic groups nor correspond to viral realms.

The primary aim in viral taxonomy is to be able to recognise and name monophyletic groups of viruses, that is, to devise a categorical system that groups viruses which are related by a succession of common virus ancestors. A detailed understanding of the evolutionary history of viruses is required, especially since the replication and capsid-formation modules may have distinct origins. In the case of RNA viruses classified in the realm *Riboviria*, all members share a viral hallmark gene which encodes the RNA-dependent RNA polymerase (RdRp). Some authors have suggested that given that viruses in this realm have distinct, non-homologous capsid proteins, they may have evolved independently from a capsid-less

ancestor (Krupovič and Koonin, 2017; Wolf et al., 2018). However, it is not necessary that the ancestor was a capsidless element, as the RdRp may have been acquired through a continuous line of descent from capsid-encoding ancestors, while the capsid gene was exchanged many times over diversification, leaving no trace of the ancestral capsid gene in modern viruses (Peter Simmonds, personal communication). In the case of the kingdom *Bamfordvirae*, the genes of the structural module have been conserved but the replicases have been replaced independently on multiple occasions. However, given that the capsid formation module is conserved we can be confident that these viruses are descended from virus ancestors. Chimaeric viruses, with capsid proteins and replicases derived from different viral lineages, are difficult to place in the current taxonomic framework. For example, bidnaviruses are ssDNA viruses with major capsid proteins derived from parvoviruses (*Monodnaviria*), whilst the replicases are protein-primed DNA polymerases derived from *Mavericks*/*Polintons* (*Varidnaviria*), with additional genes derived from reoviruses (*Riboviria*) and baculoviruses (Krupovič and Koonin, 2014). Should bidnaviruses be classified in *Monodnaviria*, *Varidnaviria*, *Riboviria* or something else? Currently bidnaviruses are classified in the realm *Monodnaviria*, kingdom *Shotokuvirae* (Lefkowitz et al., 2018).

The kingdom *Bamfordvirae*

Viruses in the realm *Varidnaviria* (meaning Various DNA viruses) are divided into two kingdoms: *Bamfordvirae* and *Helvetiavirae*. They are mostly dsDNA viruses (with the exception of the family *Finnlakeviridae* with ssDNA genomes), which assemble pseudo-hexameric or hexameric capsomers using vertical jelly-roll major capsid proteins (Laanto et al., 2017; Walker et al., 2020). The kingdom *Helvetiavirae* (*Helvetia* L. = Swiss, in reference

to the swiss-roll or jelly-roll), contains viruses which encode one or two single jelly-roll fold capsid proteins and comprises the families *Sphaerolipoviridae*, *Simuloviridae*, *Matshushitaviridae* and *Portogloboviridae* which infect bacteria and archaea (Krupovič et al., 2022; Lefkowitz et al., 2018). The kingdom *Bamfordvirae* (named after Dennis Bamford, who first proposed the common origins of this group), encompasses viruses with a double jelly-roll major capsid protein and which infect bacterial, archaeal and eukaryotic hosts (Koonin et al., 2020a; Lefkowitz et al., 2018). It includes viruses in the families *Tectiviridae*, *Turriviridae*, *Corticoviridae*, *Autolykiviridae*, *Finnlakeviridae*, *Adenoviridae*, *Lavidaviridae*, *Mavericks*/*Polintons*, Polinton-like viruses and viruses in the phylum *Nucleocytoviricota* or NCLDVs (Nucleo-cytoplasmic Large DNA viruses) (Lefkowitz et al., 2018).

Previously, it was considered that viruses in the kingdoms *Bamfordvirae* and *Helvetiavirae* had a common origin and were thus monophyletic. This was based on the shared presence of vertical jelly-roll fold capsid proteins as well as a DNA-packaging ATPase of the FtsK-HerA superfamily (Koonin et al., 2020a; Krupovič and Koonin, 2017). The leading evolutionary hypothesis was that the double jelly-roll fold in bamfordviruses had evolved by gene duplication and fusion of the single jelly-roll major capsid gene of helvetiaviruses (Koonin et al., 2020a; Krupovič and Koonin, 2017). However, evolutionary analyses guided by protein structure have identified a closest relative of the double jelly-roll capsid protein in a family of bacterial proteins which have the double jelly-roll fold and are involved in carbohydrate metabolism, DUF2961 (Krupovič et al., 2022). These findings suggest that the jelly-roll fold capsid proteins of bamfordviruses and helvetiaviruses have distinct origins and thus, do not trace back to a common viral ancestor (Krupovič et al., 2022). The current evolutionary scenario proposes that the two single jelly-roll capsid proteins of

helvetiaviruses evolved by gene duplication of a single-jelly roll gene in a portoglobovirus-like ancestor, while the double jelly-roll capsid proteins of bamfordviruses were captured independently from cellular organisms (Krupovič et al., 2022). These discoveries call for a revision of the realm *Varidnaviria*, which will probably require splitting between helvetiaviruses and bamfordviruses.

Prokaryotic relatives

Members of five viral families classified in the kingdom *Bamfordvirae* infect prokaryotes: *Tectiviridae* (e.g. phages PRD1, Bam35, AP50), *Turriviridae* (archaeoviruses STIV1/2), *Corticoviridae* (phages PM2, Cr39582), *Autolykiviridae* (*Paulavirus*/*Livvievirus*) and *Finnlakeviridae* (finnlakevirus FLiP) (Lefkowitz et al., 2018). Viruses in these families form icosahedral virions with an internal lipid membrane, built from trimers of the double jelly-roll major capsid protein (hexon) and pentamers of the single jelly-roll minor capsid protein (pentons) at the vertices; they also share a DNA packaging ATPase of the FtsK-HerA superfamily which forms one of the vertices of the virion, with the exception of finnlakeviruses that do not seem to require this protein for genome encapsidation (Kauffman et al., 2018; King, 2011; Laanto et al., 2017; Oksanen, 2017; Rice et al., 2004). Tectiviruses, turriviruses and autolykiviruses have linear, dsDNA genomes flanked by inverted repeats with a terminal protein covalently bound at the 5' ends; this protein acts as a primer during replication by the protein-primed family B DNA polymerase (pPOLB) (Kauffman et al., 2018; Oksanen and Bamford, 2012; Rice et al., 2004). By contrast, the genomes of corticoviruses (dsDNA) and finnlakeviruses (ssDNA) are circular, and are replicated by non-homologous endonucleases using the rolling-circle mechanism (Laanto et al., 2017; Mäntynen et al., 2020; Oksanen, 2017).

Tectiviruses are the most diverse family, with 12 species in 5 genera; they infect mesophilic and thermophilic bacteria in the genera *Pseudomonas*, *Bacillus* and *Thermus* (Oksanen and Bamford, 2012). Autolykiviruses and corticoviruses are both viruses of marine bacteria: the family *Autolykiviridae* contains 5 species (2 genera) and infect marine *Vibrio*, while the family *Corticoviridae* comprises 2 species (1 genus) which have been isolated from marine *Pseudoalteromonas* (Kauffman et al., 2018; Oksanen, 2017). The family *Finnlakeviridae* has a single species, *Flavobacterium virus FLiP*, which was isolated from bacteria in a freshwater ecosystem in Finland (Laanto et al., 2017). Turriviruses are the only bamfordviruses known to infect archaea; the family contains 2 species in 1 genus, both of which infect the thermophilic archaeon *Saccharolobus solfataricus* (Rice et al., 2004; Sakai and Kurosawa, 2018).

Eukaryotic elements

In eukaryotes, bamfordviruses underwent a major diversification giving rise to some of the most conspicuous viral lineages. They are broadly extended in eukaryotes as a whole, having been found infecting members of every major host lineage. Some of these, such as the poxviruses and adenoviruses, are known to be human pathogens causing mild to severe disease. The group of eukaryotic bamfordviruses includes 242 recognised species in 11 viral families (Krupovič and Bamford, 2008; Lefkowitz et al., 2018); however, many representatives have yet to be recognised by the official viral taxonomy, and their inclusion will surely increase the number of species and viral families in this lineage. The bamfordviruses of eukaryotes share with their prokaryotic relatives an ancestral gene module comprising the double and single jelly-roll capsid proteins, the DNA packaging

ATPase of the FtsK-HerA superfamily, and the protein-primed DNA polymerase (Koonin et al., 2020a; Koonin and Krupovič, 2017; Krupovič and Koonin, 2015). In addition, the eukaryotic bamfordviruses uniquely acquired a cysteine protease (adenoviral-like protease or adenain), perhaps captured from a DNA transposon, which is involved in capsid maturation (Krupovič and Koonin, 2015).

The ancestral condition of the virion in this lineage seems to be an icosahedral capsid with an inner lipid membrane, as in the bamfordviruses of prokaryotes, which is still observed in most groups today (Krupovič and Bamford, 2008). However, several lineages assemble more derived virions: for example, adenoviruses have lost the internal membrane together with the FtsK-HerA DNA packaging ATPase (which is commonly found in viruses with internal membranes) (Krupovič and Bamford, 2008), while pandoraviruses seem to have exchanged the double jelly roll major capsid with an unrelated bacterial glycoside hydrolase, which allows them to build amphora-like virions (Krupovič et al., 2020b). In three lineages, transpovirons, mitochondrial and cytoplasmic linear plasmids, the capsid-formation module has been lost altogether; these are capsidless elements which evolved from a virus ancestor (Desnues et al., 2012; Meinhardt et al., 1997). The ancestral protein-primed DNA polymerase has also been replaced, most notably in the NCLDVs by a nucleotide-primed DNA polymerase B, but also in many Polinton-like viruses and virophages which encode derived superfamily 1/3 helicases potentially involved in replication (Fischer, 2020; Krupovič et al., 2016; Krupovič and Koonin, 2015). In the sections below, I will describe the main groups of eukaryotic bamfordviruses: adenoviruses, NCLDVs, virophages, *Mavericks*/*Polintons*, Polinton-like viruses and the three types of capsidless elements.

Adenoviruses

Viruses in the family *Adenoviridae* are currently classified into 87 species belonging to 6 genera (Lefkowitz et al., 2018). They infect a wide range of vertebrates, including birds, mammals, non-avian reptiles, amphibians and ray-finned fish (Benko et al., 2022). The first adenovirus was discovered in 1953 when a new cytopathogenic agent was isolated from human adenoid tissue (Rowe et al., 1953). The viral capsid is icosahedral with 240 non-vertex capsomers (hexons) and 12 vertex capsomers (pentons) which bind to a protein fibre. Their genome is a linear, dsDNA molecule 26—48 kb in size, which is flanked by inverted repeats and covalently bound by terminal proteins (Benko et al., 2022). The genome of adenoviruses is replicated in the nucleus by the virally encoded protein-primed DNA polymerase using a "jumping-back" mechanism, where a trinucleotide synthesised at position 4—6 slides back to positions 1—3 before proceeding with synthesis of the new DNA strand (Hoeben and Uil, 2013). Unlike other bamfordviruses in eukaryotes, adenoviruses encode an ATPase from the ABC superfamily for packaging their genome (Krupovič and Bamford, 2008). Although adenoviruses normally cause mild or subclinical infections in humans, they are commonly isolated from the faeces of children with acute gastroenteritis (only second to rotaviruses) and can also cause severe conditions in immunocompromised people, with high mortality rates (Echavarría, 2008; Hierholzer, 1992).

Nucleocytoplasmic Large DNA viruses

NCLDVs are a diverse group of dsDNA viruses with genomes ranging from 106 kb to 2.47 Mb, which are currently classified into 7 families: *Phycodnaviridae*, *Mimiviridae*, *Ascoviridae*, *Iridoviridae*, *Marseilleviridae*, *Asfarviridae* and *Poxviridae* (Lefkowitz et al., 2018). In addition, many members that have been characterised experimentally and from metagenomic isolates await family-level classification (e.g. pandoraviruses, pithoviruses, faustoviruses, molliviruses, kaumoebaviruses, medusaviruses, klosneuviruses). NCLDVs have a broad host range in eukaryotes, infecting animals, amoebae, dinoflagellates, chlorophytes, cryptophytes, stramenopiles, rhizarians, alveolates, haptophytes and excavates (Mönttinen et al., 2021; Schulz et al., 2020). Several lineages of NCLDVs have genome sizes exceeding 500 kb, and are considered to be "giant viruses". Pandoraviruses, for example, are the largest viruses characterised so far in terms of physical size and genome length, with particles around 1 μm in size and genomes > 2 Mb (Legendre et al., 2013). A group of even larger viral pathogens, the klothoviruses, with particles exceeding 4 μm in length have been isolated from chaetognaths and could possibly also belong to the group of NCLDVs, but genomic analyses have not been yet conducted (Barthélémy et al., 2019). The viral gigantism seen in NCLDVs has probably evolved independently at least twice, driven by extensive horizontal gene captures from varied sources (eukaryotes, bacteria and other viruses) (Koonin and Yutin, 2018; Mönttinen et al., 2021).

The virions in NCLDVs can be icosahedral with an inner membrane (*Phycodnaviridae*, *Mimiviridae*, *Marseilleviridae*, *Iridoviridae*, *Asfarviridae*, medusavirus, faustovirus), or have more complex shapes such as bacilliform, brick-shaped or ovoid (*Ascoviridae*, *Poxviridae*, pandoravirus, pithovirus) (Lefkowitz et al., 2018). Replication of NCLDVs

occurs either in the cytoplasm (*Poxviridae*, *Mimiviridae*, *Marseilleviridae*) or it may start in the nucleus and transition to the cytoplasm (*Ascoviridae*, *Phycodnaviridae*, *Iridoviridae*, *Asfarviridae*) (Lefkowitz et al., 2018). NCLDVs share a core genome composed of the viral late transcription factor 3 (VLTF-like), major capsid protein (double jelly-roll), helicase-primase (D5-like; displaced by a bacteriophage homologue in phycodnaviruses), DNA-packaging ATPase (FtsK-HerA), and the DNA polymerase elongation subunit family B (Mönttinen et al., 2021). Although the single jelly-roll penton protein is not universally present, it can be found in some lineages such as the chloroviruses (Fang et al., 2019). The ancestral protein-primed DNA polymerase was replaced in NCLDVs by a nucleotide-primed family B DNA polymerase, which resembles archaeo-eukaryotic polymerases ε, α, δ and ζ (with homologues in herpesviruses and baculoviruses), and was probably involved in several gene transfers with the early eukaryotes (Mönttinen et al., 2021; Takemura et al., 2015).

In addition, NCLDVs share a nearly universal set of 47 genes that were inferred to be present in their most recent common ancestor (Yutin et al., 2009). These include genes involved in replication, gene expression, virion morphogenesis and virus-host interactions, which indicate that their ancestor was already a complex virus (Yutin et al., 2009). NCLDVs can also encode tRNAs, amino-acyl tRNA-synthetases, translation factors, proteins involved in tRNAs/mRNA maturation, 18S rRNA introns (which can be highly expressed), as well as histones that assemble into nucleosomes (Abrahão et al., 2018; Liu et al., 2021). This complex gene set allows most NCLDVs to induce the formation of viral factories in the cytoplasm, which are perinuclear foci that can exclude host proteins and recruit organelles such as mitochondria, cytoskeletal filaments (e.g. vimentin) and membranes, to promote viral replication and assembly (Netherton and Wileman, 2011; Novoa et al., 2005).

Viruses in the families *Poxviridae*, *Asfarviridae* and *Iridoviridae* are important pathogens of vertebrates. Two human poxviruses, *Variola virus* and *Molluscum contagiosum virus*, cause the lethal smallpox (now eradicated) and molluscum contagiosum, a mild skin condition, respectively (Chen et al., 2013; Henderson, 2011). Other poxviruses are known to be zoonotically transmitted to humans, for example, the monkeypox, cowpox and tanapox viruses (Essbauer et al., 2010; Lewis-Jones, 2004). *African swine fever virus* (family *Asfarviridae*) causes a haemorrhagic fever with 50—100% mortality, which is transmitted by ticks and affects wild and domestic swine, warthogs and bush hogs (Alonso et al., 2018). The iridoviruses are pathogens of fish, amphibians and non-avian reptiles (lizards, snakes, turtles and crocodiles) (Chinchar et al., 2017). While the genera *Megalocytivirus* and *Ranavirus*, can be highly virulent reaching mortalities of 100%, the genus *Lymphocystivirus* produces warts on the epidermis that resolve spontaneously, with low mortality rates (Chinchar et al., 2017).

Virophages

Virophages are specialised viral parasites of NCLDVs (Fischer, 2020). They are classified in the family *Lavidaviridae*, which currently has two genera *Sputnikvirus* and *Mavirus* (Fischer, 2020; Lefkowitz et al., 2018). However, additional virophages have been isolated and many others have been detected in diverse metagenomes (Fischer, 2020; Gaia et al., 2014; Mougari et al., 2019; Paez-Espino et al., 2019; Roux et al., 2017). The genome of virophages is a linear or circular dsDNA molecule which ranges from 17 to 36 kb in size (Fischer, 2020). The most conserved genes in virophages are those of the structural module, encoding the major and minor jelly-roll capsid proteins, the FtsK-HerA DNA packaging

ATPase, and the maturation cysteine protease (Fischer, 2020). Most virophage genomes also encode a protein with a superfamily 3 helicase (SH3) domain, however, it seems to have polyphyletic origins (Fischer, 2020). The Mavirus-type virophages encode a protein-primed DNA polymerase which is probably the main replicase, while Sputnik-type virophages may use the SH3 domain protein which is fused to a bacterial polymerase domain (TVPol), however the exact replication mechanisms are still unknown (Fischer, 2020; Yutin et al., 2013). Virophages share promoter and polyadenylation sequences with their giant virus host, and use their transcriptional machinery for gene expression (Fischer and Suttle, 2011; la Scola et al., 2008). Therefore, virophages depend on coinfection of the cell by the giant virus in order to replicate, which occurs in the host's viral factory (Fischer, 2020). Virophages such as Sputnik, Mavirus and Guarani have been shown to negatively impact replication of the giant virus, which is beneficial to host eukaryote population, but it seems that virophages such as Zamilon can have a neutral (commensal) effect (Fischer and Suttle, 2011; Gaia et al., 2014; la Scola et al., 2008; Mougari et al., 2019). Interestingly, some virophages have been found integrated in the genomes of their giant virus or eukaryotic host, and so seem to adopt a dual strategy of exogenous and endogenous persistence (Blanc et al., 2015; Desnues et al., 2012; Fischer and Hackl, 2016). Although all virophages that have been characterised so far infect members of the family *Mimiviridae*, a possible host has been identified in the Asfar/faustovirus clade (Paez-Espino et al., 2019), and it seems plausible that they may infect a wider number of NCLDV families.

Mavericks/Polintons

Mavericks, or *Polintons*, are elements found integrated in the genomes of a wide range of eukaryotes including animals, fungi, amoebae, stramenopiles and parabasalids (Kapitonov

and Jurka, 2006; Pritham et al., 2007). The genomes of *Mavericks* are 15—30 kb in length, flanked by inverted repeats and in addition to the protein-primed DNA polymerase, the FtsK-HerA packaging ATPase, and maturation protease, they encode a retroviral-like integrase (Kapitonov and Jurka, 2006; Pritham et al., 2007). Although *Mavericks* have been traditionally thought of as transposons, structural analyses of the encoded PY and PX proteins suggest they are homologues of the jelly-roll fold major and minor capsid proteins of other bamfordviruses (Krupovič et al., 2014). Therefore, they seem to encode the complete set of genes required for replication and the formation of virions. However, since the elements have not been characterised experimentally, many questions remain surrounding their lifecycle and natural history. From the presence of the pPOLB and a pair of trinucleotide repeats at the ends of the genome, it can be inferred that the protein-primed polymerase uses a "jumping-back" mechanism as observed in adenoviruses (Kapitonov and Jurka, 2006). The integration mechanism seems to follow a "staggered-cut" mechanism given that target site duplications of 5 to 6 bp occur next to the inverted repeats (Kapitonov and Jurka, 2006; Pritham et al., 2007). It does not seem that *Mavericks* cause any pathologies in hosts, given that they have not been isolated in association with any disease.

Polinton-like viruses

The Polinton-like viruses (PLVs) are a diverse and abundant assemblage of viral sequences identified in marine and freshwater metagenomes. They have small genomes (< 41 kb in size) and resemble *Mavericks* in their genetic composition (Bellas and Sommaruga, 2021). However, they lack the maturation protease seen in *Mavericks* and virophages, and only a few elements encode the *Maverick* pPOLB and integrase proteins (Bellas and Sommaruga, 2021). PLVs seem to infect a wide range of eukaryotes, including chlorophytes,

haptophytes, cryptophytes, stramenopiles, amoeba and fungi (Bellas and Sommaruga, 2021). A single PLV has been characterised experimentally, TsV-N1, which infects the marine chlorophyte *Tetraselmis striata* (Pagarete et al., 2015). TsV-N1 has a dsDNA genome of 31 kb, it replicates in the nucleus and encodes a eukaryotic-like DNA polymerase δ (Pagarete et al., 2015). Other PLVs encode a fusion protein composed of a superfamily 1 helicase and an inactivated pPOLB domain (Krupovič et al., 2016). The *Tlr* elements found integrated in the genome of the ciliate *Tetrahymena thermophila* also share this fusion protein, which in addition to the lack of a protease, suggest *Tlr* elements are endogenous PLVs (Krupovič et al., 2016; Wuitschick et al., 2002). Additional PLVs — called PgV and CpV virophages, but which form a distinct group from lavidaviruses — have been sequenced in association with or integrated in the genomes of phycodnaviruses, raising the possibility that some PLVs may depend on coinfection with a host NCLDV for replication (Bellas and Sommaruga, 2021; Santini et al., 2013; Stough et al., 2019).

Capsidless elements

In addition to the capsid-encoding elements, there are three lineages of capsid-less elements in the kingdom *Bamfordvirae* which infect eukaryotes. All evolved separately from an ancestral virus via reductive evolution by losing the genes involved in capsid formation. Based on phylogenies of the protein-primed DNA polymerase, the oldest group appears to be the mitochondrial linear plasmids. These plasmids are found in the mitochondria of fungi, plants, amoeba and ciliates, and their genomes are flanked by inverted repeats covalently linked to terminal proteins at their 5' ends (6.8—9 kb in size) (Meinhardt et al., 1997; Nishimura et al., 2019). The plasmids encode the protein-primed DNA polymerase and an RNA polymerase, and may include additional genes of unknown function (Handa, 2008;

Meinhardt et al., 1997). Mitochondrial linear plasmids may integrate into the host mitochondrial genome, which can have important phenotypic effects; for example, induce senescence in *Neurospora*, or increase longevity in *Podospora* (Bertrand et al., 1985; Hermanns et al., 1994). Cytoplasmic linear plasmids have a similar genetic organisation to mitochondrial linear plasmids and genomes bound by terminal proteins (8—18 kb), but are found in the cytoplasm of fungi in the Order Saccharomycetales (yeasts) (Meinhardt et al., 1997). Some cytoplasmic linear plasmids encode toxin-antitoxin systems and confer killer phenotypes on their hosts (Frank and Wolfe, 2009; Meinhardt et al., 1997). Multiple elements have been found integrated in their host's nuclear genome (Frank and Wolfe, 2009; Satwika et al., 2012). The third group, transpovirons, are plasmid-like elements associated with giant viruses of the family *Mimiviridae*, and also commonly found in association with virophages (de Azevedo et al., 2022; Desnues et al., 2012; Jeudy et al., 2020). Their genomes are also flanked by inverted repeats and have a size of approximately 7 kb (Desnues et al., 2012). Transpovirons encode the superfamily 1 helicase fused to an inactivated pPOLB domain that is also observed in PLVs; phylogenetic analysis of this protein suggests that transpovirons are a highly derived type of PLV (Krupovič et al., 2016).

Open questions

The sheer diversity and ancient origin of the eukaryotic bamfordviruses present many questions and unresolved evolutionary conundrums. One major question in the field relates to the nature of *Mavericks*, which are still being portrayed as transposons in the recent scientific literature despite the evidence that *Mavericks* encode capsid-formation protein homologues (e.g. Nguyen et al., 2022; Chang et al., 2022). In addition to the structural homology of predicted *Maverick* proteins to those of other viruses (Krupovič et al., 2014),

study of the patterns of transmission and selection over millions of years can provide evidence for their existence as virions. Another intriguing question is the disappearance of *Mavericks* from the genomes of mammals. It is widely believed that *Mavericks* went extinct, leaving no trace of their existence in mammals by the time of their most recent common ancestor (Kapitonov and Jurka, 2006; Pritham et al., 2007). However, a *Maverick*-like sequence was recently reported on human chromosome 7 which constitutes a surprising discovery (Starrett et al., 2020). The authors of this study did not validate the locus or perform an evolutionary investigation into its origins, leaving open questions about the authenticity of the sequence (i.e. it could result from contamination in the genome assembly) and its evolutionary history in mammals. A systematic study of the locus and its present state across other species can shed light on the past of *Mavericks* in mammalian genomes.

Another question involves the origin of the main lineages of eukaryotic bamfordviruses. Evidence of early gene exchanges between NCLDVs and eukaryotes, suggest that the most recent common ancestor of NCLDVs existed more than 1 billion years ago (Guglielmini et al., 2019). However, the relations of NCLDVs to other lineages, and in particular virophages, remain unclear. Several hypotheses have been proposed to explain the origin of these viral lineages, based on single-protein phylogenies of the protein-primed DNA polymerase (nuclear-escape hypothesis) or the close connection between *Mavericks* and the Mavirus virophage (virophage-first hypothesis) (Krupovič and Koonin, 2015; Fischer and Suttle, 2011). However, although the protein-primed DNA polymerase seems to be the ancestral replicase for viruses in the kingdom *Bamfordvirae*, it has been replaced for non-homologous replicases on multiple occasions (*Corticoviridae*, *Finnlakeviridae*, NCLDVs, some virophages and some PLVs). Therefore, the proteins of the capsid-formation module (major and minor capsid proteins, FtsK-HerA ATPase and maturation protease), which are

more conserved throughout the lineage, could be potentially used to differentiate between competing evolutionary scenarios of virus origins. The integration dynamics of virophages into eukaryotic and virus genomes also present questions with important ecological and evolutionary implications. Although mathematical models of the interaction between cells, viruses and virophages exist (Taylor et al., 2014; Wodarz, 2013), they have not taken this crucial aspect of virophage integration into account. It remains puzzling how such systems can persist in time and especially in face of very strong inhibition of some virophages on their virus hosts. The evolutionary consequences of these integration dynamics also remain unexplored, but seem particularly relevant since multiple examples of integrated virophages are being discovered (Blanc et al., 2015; Hackl et al., 2021).

Aims and scope of this thesis

The aim of this thesis was to investigate aspects of the evolutionary biology of viruses in the kingdom *Bamfordvirae*, focusing on the lineage that infects eukaryotes (eukaryoviruses). A special emphasis has been made on understanding the long-term evolutionary process of these viruses with their eukaryotic hosts, over timescales of hundreds of millions to billions of years in the past. To achieve this goal, a variety of techniques have been used drawing from paleovirology, comparative genomics, phylogenetics, mathematical modelling and computational simulations. The first chapter, *"Paleovirology of the DNA viruses of eukaryotes"*, provides an in-depth review of the scientific literature which introduces the field of paleovirology, and gives context for the evolution of *Bamfordvirae* within the larger diversity of ancient eukaryotic DNA virus families. The second chapter, "*Phylogenomics of the Maverick virus-like mobile genetic elements of vertebrates*", explores the evolutionary history of *Mavericks* in vertebrates which led to discovery of multiple new species and an

extensive genomic fossil record in these animals, also finding strong evidence for the existence of *Mavericks* as a lineage of integrating, mostly aquatic DNA viruses. The third chapter, "*A billion-years arms race between viruses, virophages and eukaryotes*", examines alternative evolutionary scenarios for the origin of the major lineages of eukaryotic bamfordviruses based on analyses of the capsid-formation proteins; this work presents a new evolutionary model that supports the early origins of virophages and NCLDVs from an exogenous virus ancestor. The fourth chapter, "*Ecological and evolutionary dynamics of cell-virus-virophage systems*", presents mathematical models and computational simulations used to understand the effect of virophage genome integration on the interactions of this tripartite system. The possible roles of virus-induced cell death, virus inhibition by virophages, and multicellularity as antiviral defence strategies used by the early eukaryotes, have also been investigated. The fifth chapter, "*An ancient endogenous DNA virus in the human genome*", analyses two *Maverick* integrations in the human genome which give insights into the DNA viruses that infected our direct ancestors more than 100 million years ago; these findings are placed in the context of the possible reasons for the demise of *Mavericks* in mammalian genomes. The final chapter discusses the wider implications of these discoveries as they relate to the ancient origins of bamfordviruses and their coevolution with eukaryotes, limitations of these studies, and prospects for future research.

References

Abrahão J, Silva L, Silva LS, Khalil JYB, Rodrigues R, Arantes T, Assis F, Boratto P, Andrade M, Kroon EG, Ribeiro B, Bergier I, Seligmann H, Ghigo E, Colson P, Levasseur A, Kroemer G, Raoult D, la Scola B. 2018. Tailed giant Tupanvirus possesses the most complete translational apparatus of the known virosphere. *Nature Communications* 9. doi:10.1038/s41467-018-03168-1

Aiewsakun P, Katzourakis A. 2016. Time-dependent rate phenomenon in viruses. *Journal of Virology* 90:7184–7195. doi:10.1128/JVI.00593-16

Aiewsakun P, Katzourakis A. 2015. Endogenous viruses: Connecting recent and ancient viral evolution. *Virology*. doi:10.1016/j.virol.2015.02.011

Allwood AC, Rosing MT, Flannery DT, Hurowitz JA, Heirwegh CM. 2018. Reassessing evidence of life in 3,700-million-year-old rocks of Greenland. *Nature* 563. doi:10.1038/s41586-018-0610-4

Alonso C, Borca M, Dixon L, Revilla Y, Rodriguez F, Escribano JM. 2018. ICTV virus taxonomy profile: *Asfarviridae*. *Journal of General Virology* 99. doi:10.1099/jgv.0.001049

Artenstein AW. 2012. The discovery of viruses: Advancing science and medicine by challenging dogma. *International Journal of Infectious Diseases*. doi:10.1016/j.ijid.2012.03.005

Aswad A, Katzourakis A. 2012. Paleovirology and virally derived immunity. *Trends in Ecology and Evolution*. doi:10.1016/j.tree.2012.07.007

Barthélémy R, Faure E, Goto T. 2019. Serendipitous discovery in a marine invertebrate (Phylum Chaetognatha) of the longest giant viruses reported till date. *Virology: Current Research* 3:1–13.

Battistuzzi FU, Feijao A, Hedges SB. 2004. A genomic timescale of prokaryote evolution: insights into the origin of methanogenesis, phototrophy, and the colonization of land. *BMC Evolutionary Biology* 4. doi:10.1186/1471-2148-4-44

Bechtel JT, Winant RC, Ganem D. 2005. host and viral proteins in the virion of Kaposi's sarcoma-associated herpesvirus. *Journal of Virology* 79. doi:10.1128/jvi.79.8.4952-4964.2005

Bell EA, Boehnke P, Harrison TM, Mao WL. 2015. Potentially biogenic carbon preserved in a 4.1 billion-year-old zircon. *Proc Natl Acad Sci U S A* 112. doi:10.1073/pnas.1517557112

Bellas CM, Sommaruga R. 2021. Polinton-like viruses are abundant in aquatic ecosystems. *Microbiome* 9. doi:10.1186/s40168-020-00956-0

Belshaw R, Pereira V, Katzourakis A, Talbot G, Pačes J, Burt A, Tristem M. 2004. Long-term reinfection of the human genome by endogenous retroviruses. *Proc Natl Acad Sci U S A* 101:4894–4899. doi:10.1073/pnas.0307800101

Benko M, Aoki K, Arnberg N, Davison AJ, Echavarria M, Hess M, Jones MS, Kajan GL, Kajon AE, Mittal SK, Podgorski II, Martin CS, Wadell G, Watanabe H, Harrach B. 2022. ICTV Virus Taxonomy Profile: *Adenoviridae*. *Journal of General Virology* 103. doi:10.1099/jgv.0.001721

Bertrand H, Chan BSS, Griffiths AJF. 1985. Insertion of a foreign nucleotide sequence into mitochrondrial DNA causes senescence in neurospora intermedia. *Cell* 41. doi:10.1016/S0092-8674(85)80068-4

Betts HC, Puttick MN, Clark JW, Williams TA, Donoghue PCJ, Pisani D. 2018. Integrated genomic and fossil evidence illuminates life's early evolution and eukaryote origin. *Nature Ecology and Evolution* 2. doi:10.1038/s41559-018-0644-x

Bi L, Yu DT, Du S, Zhang LM, Zhang LY, Wu CF, Xiong C, Han LL, He JZ. 2021. Diversity and potential biogeochemical impacts of viruses in bulk and rhizosphere soils. *Environmental Microbiology* 23. doi:10.1111/1462-2920.15010

Blanc G, Gallot-Lavallée L, Maumus F. 2015. Provirophages in the *Bigelowiella* genome bear testimony to past encounters with giant viruses. *Proc Natl Acad Sci U S A* 112:E5318–E5326. doi:10.1073/pnas.1506469112

Cafferty BJ, Hud NV. 2014. Abiotic synthesis of RNA in water: a common goal of prebiotic chemistry and bottom-up synthetic biology. *Current Opinion in Chemical Biology* 22. doi: 10.1016/j.cbpa.2014.09.015

Catchpole RJ, Forterre P. 2019. The evolution of reverse gyrase suggests a nonhyperthermophilic Last Universal Common Ancestor. *Molecular Biology and Evolution* 36. doi:10.1093/molbev/msz180

Chang N, Rovira Q, Wells J, Feschotte C, Vaquerizas JM. 2022. Zebrafish transposable elements show extensive diversification in age, genomic distribution, and developmental expression. *Genome Research*. doi: 10.1101/gr.275655.121

Chen X, Anstey A v., Bugert JJ. 2013. *Molluscum contagiosum virus* infection. *The Lancet Infectious Diseases*. doi:10.1016/S1473-3099(13)70109-9

Cheng H, Shen N, Pei J, Grishin N v. 2004. Double-stranded DNA bacteriophage prohead protease is homologous to herpesvirus protease. *Protein Science* 13. doi:10.1110/ps.04726004

Chinchar VG, Hick P, Ince IA, Jancovich JK, Marschang R, Qin Q, Subramaniam K, Waltzek TB, Whittington R, Williams T, Zhang QY. 2017. ICTV virus taxonomy profile: *Iridoviridae*. *Journal of General Virology* 98. doi:10.1099/jgv.0.000818

Compans RW. 1993. Arenavirus ultrastructure and morphogenesis: the *Arenaviridae*. doi:10.1007/978-1-4615-3028-2_1

Condit RC. 2013. Principles of virology. Fields Virology: Sixth Edition. doi:10.1136/vr.i1135

Cornelis G, Funk M, Vernochet C, Leal F, Tarazona OA, Meurice G, Heidmann O, Dupressoir A, Miralles A, Ramirez-Pinilla MP, Heidmann T, Roberts RM. 2017. An endogenous retroviral envelope syncytin and its cognate receptor identified in the viviparous placental Mabuya lizard. *Proc Natl Acad Sci U S A* 114:E10991–E11000. doi:10.1073/pnas.1714590114

Cornelis G, Vernochet C, Carradec Q, Souquere S, Mulot B, Catzeflis F, Nilsson MA, Menzies BR, Renfree MB, Pierron G, Zeller U, Heidmann O, Dupressoir A, Heidmann T. 2015. Retroviral envelope gene captures and syncytin exaptation for placentation in marsupials. *Proc Natl Acad Sci U S A* 112. doi:10.1073/pnas.1417000112

Coutinho FH, Cabello-Yeves PJ, Gonzalez-Serrano R, Rosselli R, López-Pérez M, Zemskaya TI, Zakharenko AS, Ivanov VG, Rodriguez-Valera F. 2020. New viral biogeochemical roles revealed through metagenomic analysis of Lake Baikal. *Microbiome* 8. doi:10.1186/s40168-020-00936-4

de Azevedo BL, Araújo Júnior JP, Ullmann LS, Rodrigues RAL, Abrahão JS. 2022. The discovery of a new mimivirus isolate in association with virophage-transpoviron elements in Brazil highlights the main genomic and evolutionary features of this tripartite system. *Viruses* 14. doi:10.3390/v14020206

Desnues C, la Scola B, Yutin N, Fournous G, Robert C, Azza S, Jardot P, Monteil S, Campocasso A, Koonin EV, Raoult D. 2012. Provirophages and transpovirons as the diverse mobilome of giant viruses. *Proc Natl Acad Sci U S A* 109:18078–18083. doi:10.1073/pnas.1208835109

Dodd MS, Papineau D, Grenne T, Slack JF, Rittner M, Pirajno F, O'Neil J, Little CTS. 2017. Evidence for early life in Earth's oldest hydrothermal vent precipitates. *Nature* 543. doi:10.1038/nature21377

Dolja VV, Koonin EV. 2012. Capsid-less RNA viruses. eLS. doi: 10.1002/9780470015902.a0023269

Dominic P, Zhenbing S, S DM, Francesco I, F SJ, Erik H, Paul S, S LCT. 2022. Metabolically diverse primordial microbial communities in Earth's oldest seafloor-hydrothermal jasper. *Science Advances* 8:eabm2296. doi:10.1126/sciadv.abm2296

Duda RL, Teschke CM. 2019. The amazing HK97 fold: versatile results of modest differences. *Current Opinion in Virology*. doi:10.1016/j.coviro.2019.02.001

Dupressoir A, Lavialle C, Heidmann T. 2012. From ancestral infectious retroviruses to bona fide cellular genes: role of the captured syncytins in placentation. *Placenta* 33:663–671. doi:10.1016/j.placenta.2012.05.005

Dupressoir A, Vernochet C, Bawa O, Harper F, Pierron G, Opolon P, Heidmann T. 2009. Syncytin-A knockout mice demonstrate the critical role in placentation of a fusogenic, endogenous retrovirus-derived, envelope gene. *Proc Natl Acad Sci U S A* 106. doi:10.1073/pnas.0902925106

Echavarría M. 2008. Adenoviruses in immunocompromised hosts. *Clinical Microbiology Reviews*. doi:10.1128/CMR.00052-07

el Baidouri F, Venditti C, Suzuki S, Meade A, Humphries S. 2020. Phenotypic reconstruction of the last universal common ancestor reveals a complex cell. *bioRxiv* 44. doi: 10.1101/2020.08.20.260398

Ellis EL, Delbrück M. 1939. The growth of bacteriophage. *Journal of General Physiology* 22. doi:10.1085/jgp.22.3.365

Enquist L, Racaniello V. 2013. Virology: from *contagium vivum fluidum* to virome In: Knipe D, Howley P, editors. Field's Virology. Philadelphia: Wolters Kluwer Health.

Essbauer S, Pfeffer M, Meyer H. 2010. Zoonotic poxviruses. *Veterinary Microbiology*. doi:10.1016/j.vetmic.2009.08.026

Fang Q, Zhu D, Agarkova I, Adhikari J, Klose T, Liu Y, Chen Z, Sun Y, Gross ML, van Etten JL, Zhang X, Rossmann MG. 2019. Near-atomic structure of a giant virus. *Nature Communications* 10. doi:10.1038/s41467-019-08319-6

Filée J, Pouget N, Chandler M. 2008. Phylogenetic evidence for extensive lateral acquisition of cellular genes by Nucleocytoplasmic large DNA viruses. *BMC Evolutionary Biology* 8:1–13. doi:10.1186/1471-2148-8-320

Fischer MG. 2020. The virophage family *Lavidaviridae*. *Current Issues in Molecular Biology* 40. doi:10.21775/cimb.040.001

Fischer MG, Hackl T. 2016. Host genome integration and giant virus-induced reactivation of the virophage Mavirus. *Nature* 540:288–291. doi:10.1038/nature20593

Fischer MG, Suttle CA. 2011. A virophage at the origin of large DNA transposons. *Science (1979)* 231–234. doi: 10.1126/science.1199412

Forterre P. 2013. The virocell concept and environmental microbiology. *ISME Journal* 7:233–236. doi:10.1038/ismej.2012.110

Forterre P. 2011. Manipulation of cellular syntheses and the nature of viruses: the virocell conceptComptes Rendus Chimie. doi:10.1016/j.crci.2010.06.007

Frank AC, Wolfe KH. 2009. Evolutionary capture of viral and plasmid DNA by yeast nuclear chromosomes. *Eukaryotic Cell* 8. doi:10.1128/EC.00110-09

Gaia M, Benamar S, Boughalmi M, Pagnier I, Croce O, Colson P, Raoult D, la Scola B. 2014. Zamilon, a novel virophage with *Mimiviridae* host specificity. *PLoS ONE* 9. doi:10.1371/journal.pone.0094923

Geoghegan JL, Duchêne S, Holmes EC. 2017. Comparative analysis estimates the relative frequencies of co-divergence and cross-species transmission within viral families. *PLoS Pathogens* 13. doi:10.1371/journal.ppat.1006215

Ghafari M, Simmonds P, Pybus OG, Katzourakis A. 2021. A mechanistic evolutionary model explains the time-dependent pattern of substitution rates in viruses. *Current Biology* 31. doi:10.1016/j.cub.2021.08.020

Gorbalenya AE, Krupovič M, Mushegian A, Kropinski AM, Siddell SG, Varsani A, Adams MJ, Davison AJ, Dutilh BE, Harrach B, Harrison RL, Junglen S, King AMQ, Knowles NJ, Lefkowitz EJ, Nibert ML, Rubino L, Sabanadzovic S, Sanfaçon H, Simmonds P, Walker PJ, Zerbini FM, Kuhn JH. 2020. The new scope of virus taxonomy: partitioning the virosphere into 15 hierarchical ranks. *Nature Microbiology*. doi:10.1038/s41564-020-0709-x

Graham RL, Baric RS. 2010. Recombination, Reservoirs, and the modular spike: mechanisms of coronavirus cross-species transmission. *Journal of Virology* 84. doi:10.1128/jvi.01394-09

Hackl T, Duponchel S, Barenhoff K, Weinmann A, Fischer MG. 2021. Virophages and retrotransposons colonize the genomes of a heterotrophic flagellate. *eLife* 10. doi:10.7554/eLife.72674

Handa H. 2008. Linear plasmids in plant mitochondria: Peaceful coexistences or malicious invasions? *Mitochondrion* 8. doi:10.1016/j.mito.2007.10.002

Harrison SC. 2013. Principles of virus structure. Fields Virology: Sixth Edition. doi:10.1016/b978-0-12-814515-9.00033-3

Henderson DA. 2011. The eradication of smallpox - an overview of the past, present, and future. *Vaccine* 29. doi:10.1016/j.vaccine.2011.06.080

Hermanns J, Asseburg A, Osiewacz HD. 1994. Evidence for a life span-prolonging effect of a linear plasmid in a longevity mutant of *Podospora anserina*. *MGG Molecular and General Genetics* 243. doi:10.1007/BF00301065

Hierholzer JC. 1992. Adenoviruses in the immunocompromised host. *Clinical Microbiology Reviews*. doi:10.1128/CMR.5.3.262

Higgs PG, Lehman N. 2015. The RNA World: molecular cooperation at the origins of life. *Nature Reviews Genetics* 16. doi: 10.1038/nrg3841

Ho SYW, Duchêne S, Molak M, Shapiro B. 2015. Time-dependent estimates of molecular evolutionary rates: Evidence and causes. *Molecular Ecology*. doi:10.1111/mec.13450

Hoeben RC, Uil TG. 2013. Adenovirus DNA replication. *Cold Spring Harbor Perspectives in Biology* 5. doi:10.1101/cshperspect.a013003

Hunter E. 2013. Virus assembly In: Knipe D, Howley P, editors. Fields Virology. Philadelphia: Wolters Kluwer Health.

Iranzo J, Puigbò P, Lobkovsky AE, Wolf YI, Koonin EV. 2016a. Inevitability of genetic parasites. *Genome Biology and Evolution*. doi: 10.1093/gbe/evw193

Iranzo J, Krupovič M, Koonin EV. 2016b. The double-stranded DNA virosphere as a modular hierarchical network of gene sharing. *mBio* 7. doi:10.1128/mBio.00978-16

Iyer LM, Balaji S, Koonin EV, Aravind L. 2006. Evolutionary genomics of nucleo-cytoplasmic large DNA viruses. *Virus Research* 117. doi:10.1016/j.virusres.2006.01.009

Jeudy S, Bertaux L, Alempic JM, Lartigue A, Legendre M, Belmudes L, Santini S, Philippe N, Beucher L, Biondi EG, Juul S, Turner DJ, Couté Y, Claverie JM, Abergel C. 2020. Exploration of the propagation of transpovirons within Mimiviridae reveals a unique example of commensalism in the viral world. *ISME Journal* 14. doi:10.1038/s41396-019-0565-y

Jover LF, Effler TC, Buchan A, Wilhelm SW, Weitz JS. 2014. The elemental composition of virus particles: implications for marine biogeochemical cycles. *Nature Reviews Microbiology* 12. doi:10.1038/nrmicro3289

Kapitonov VV, Jurka J. 2006. Self-synthesizing DNA transposons in eukaryotes. *Proc Natl Acad Sci U S A* 103. doi:10.1073/pnas.0600833103

Katzourakis A, Gifford RJ. 2010. Endogenous viral elements in animal genomes. *PLoS Genetics* 6. doi:10.1371/journal.pgen.1001191

Kauffman KM, Hussain FA, Yang J, Arevalo P, Brown JM, Chang WK, Vaninsberghe D, Elsherbini J, Sharma RS, Cutler MB, Kelly L, Polz MF. 2018. A major lineage of non-tailed dsDNA viruses as unrecognized killers of marine bacteria. *Nature* 554. doi:10.1038/nature25474

King A. 2011. Family Tectiviridae. *Virus Taxonomy*.

Koonin EV, Dolja VV, Krupovič M, Varsani A, Wolf YI, Yutin N, Zerbini FM, Kuhn JH. 2020a. Global organization and proposed megataxonomy of the virus world. *Microbiology and Molecular Biology Reviews* 84. doi:10.1128/mmbr.00061-19

Koonin EV, Krupovič M. 2017. *Polintons*, virophages and transpovirons: a tangled web linking viruses, transposons and immunity. *Current Opinion in Virology* 25:7–15. doi:10.1016/j.coviro.2017.06.008

Koonin EV, Wolf YI, Katsnelson MI. 2017. Inevitability of the emergence and persistence of genetic parasites caused by evolutionary instability of parasite-free states. *Biology Direct.* doi: 10.1186/s13062-017-0202-5

Koonin EV, Krupovič M, Ishino S, Ishino Y. 2020b. The replication machinery of LUCA: Common origin of DNA replication and transcription. *BMC Biology* 18. doi:10.1186/s12915-020-00800-9

Koonin EV, Krupovič M, Agol VI. 2021a. The Baltimore classification of viruses 50 years later: how does it stand in the light of virus evolution? *Microbiology and Molecular Biology Reviews.* doi: 10.1128/MMBR.00053-21

Koonin EV, Dolja VV, Krupovič M, Kuhn JH. 2021b. Viruses defined by the position of the virosphere within the replicator space. *Microbiology and Molecular Biology Reviews* 85(4). doi: 10.1128/MMBR.00193.

Koonin EV, Yutin N. 2018. Multiple evolutionary origins of giant viruses. *F1000Res*. doi:10.12688/f1000research.16248.1

Krupovič M, Bamford DH. 2008. Virus evolution: how far does the double β-barrel viral lineage extend? *Nature Reviews Microbiology* 6:941–948. doi:10.1038/nrmicro2033

Krupovič M, Bamford DH, Koonin EV. 2014. Conservation of major and minor jelly-roll capsid proteins in *Polinton* (*Maverick*) transposons suggests that they are *bona fide* viruses. *Biology Direct* 9:1–7. doi:10.1186/1745-6150-9-6

Krupovič M, Dolja VV, Koonin EV. 2020a. The LUCA and its complex virome. *Nature Reviews Microbiology* 18. doi:10.1038/s41579-020-0408-x

Krupovič M, Dolja VV, Koonin EV. 2019a. Origin of viruses: primordial replicators recruiting capsids from hosts. *Nature Reviews Microbiology* 17. doi:10.1038/s41579-019-0205-6

Krupovič M, Koonin EV. 2017. Multiple origins of viral capsid proteins from cellular ancestors. *Proc Natl Acad Sci U S A* 114. doi:10.1073/pnas.1621061114

Krupovič M, Koonin EV. 2015. *Polintons*: A hotbed of eukaryotic virus, transposon and plasmid evolution. *Nature Reviews Microbiology* 13:105–115. doi:10.1038/nrmicro3389

Krupovič M, Koonin EV. 2014. Evolution of eukaryotic single-stranded DNA viruses of the Bidnaviridae family from genes of four other groups of widely different viruses. *Scientific Reports* 4. doi:10.1038/srep05347

Krupovič M, Makarova KS, Koonin EV. 2022. Cellular homologs of the double jelly-roll major capsid proteins clarify the origins of an ancient virus kingdom. *Proc Natl Acad Sci U S A* 119. doi:10.1073/pnas.2120620119

Krupovič M, Makarova KS, Wolf YI, Medvedeva S, Prangishvili D, Forterre P, Koonin EV. 2019b. Integrated mobile genetic elements in Thaumarchaeota. *Environmental Microbiology* 21. doi:10.1111/1462-2920.14564

Krupovič M, Ravantti JJ, Bamford DH. 2009. Geminiviruses: a tale of a plasmid becoming a virus. *BMC Evolutionary Biology* 9. doi:10.1186/1471-2148-9-112

Krupovič M, Yutin N, Koonin E. 2020b. Evolution of a major virion protein of the giant pandoraviruses from an inactivated bacterial glycoside hydrolase. *Virus Evolution* 6. doi:10.1093/ve/veaa059

Krupovič M, Yutin N, Koonin EV. 2016. Fusion of a superfamily 1 helicase and an inactivated DNA polymerase is a signature of common evolutionary history of *Polintons*, Polinton-like viruses, *Tlr*1 transposons and transpovirons. *Virus Evolution* 2. doi:10.1093/ve/vew019

la Scola B, Desnues C, Pagnier I, Robert C, Barrassi L, Fournous G, Merchat M, Suzan-Monti M, Forterre P, Koonin E, Raoult D. 2008. The virophage as a unique parasite of the giant mimivirus. *Nature* 455:100–104. doi:10.1038/nature07218

Laanto E, Mäntynen S, de Colibus L, Marjakangas J, Gillum A, Stuart DI, Ravantti JJ, Huiskonen JT, Sundberg LR. 2017. Virus found in a boreal lake links ssDNA and dsDNA viruses. *Proc Natl Acad Sci U S A* 114. doi:10.1073/pnas.1703834114

Lefkowitz EJ, Dempsey DM, Hendrickson RC, Orton RJ, Siddell SG, Smith DB. 2018. Virus taxonomy: the database of the International Committee on Taxonomy of Viruses (ICTV). *Nucleic Acids Research* 46. doi:10.1093/nar/gkx932

Legendre M, Doutre G, Poirot O, Lescot M, Arslan D, Seltzer V, Bertaux L, Bruley C, Claverie J. 2013. Pandoraviruses: amoeba viruses with genomes up to 2.5 Mb reaching that of parasitic eukaryotes. *Science (1979)* 341:281–286. doi:10.1126/science.1239181

Lenard J. 2008. Viral Membranes. *Encyclopedia of Virology* 308–314. doi:10.1016/B978-012374410-4.00530-6

Leopardi S, Holmes EC, Gastaldelli M, Tassoni L, Priori P, Scaravelli D, Zamperin G, de Benedictis P. 2018. Interplay between co-divergence and cross-species transmission in the evolutionary history of bat coronaviruses. *Infection, Genetics and Evolution* 58. doi:10.1016/j.meegid.2018.01.012

Lewis-Jones S. 2004. Zoonotic poxvirus infections in humans. *Current Opinion in Infectious Diseases*. doi:10.1097/00001432-200404000-00003

Liu Y, Bisio H, Toner CM, Jeudy S, Philippe N, Zhou K, Bowerman S, White A, Edwards G, Abergel C, Luger K. 2021. Virus-encoded histone doublets are essential and form nucleosome-like structures. *Cell* 184. doi:10.1016/j.cell.2021.06.032

Liu Y, Xu L, Opalka N, Kappler J, Shu HB, Zhang G. 2002. Crystal structure of sTALL-1 reveals a virus-like assembly of TNF family ligands. *Cell* 108. doi:10.1016/S0092-8674(02)00631-1

Loterio RK, Zamboni DS, Newton HJ. 2021. Keeping the host alive - Lessons from obligate intracellular bacterial pathogens. *Pathogens and Disease*. doi:10.1093/femspd/ftab052

Low SJ, Džunková M, Chaumeil PA, Parks DH, Hugenholtz P. 2019. Evaluation of a concatenated protein phylogeny for classification of tailed double-stranded DNA viruses belonging to the order Caudovirales. *Nature Microbiology* 4. doi:10.1038/s41564-019-0448-z

Luisi PL. 2003. Autopoiesis: a review and a reappraisal. *Naturwissenschaften* 90:49–59. doi:10.1007/s00114-002-0389-9

Mäntynen S, Laanto E, Sundberg LR, Poranen MM, Oksanen HM. 2020. ICTV virus taxonomy profile: *Finnlakeviridae*. *Journal of General Virology* 101. doi:10.1099/jgv.0.001488

Martin DP, Biagini P, Lefeuvre P, Golden M, Roumagnac P, Varsani A. 2011. Recombination in eukaryotic single stranded DNA viruses. *Viruses*. doi:10.3390/v3091699

Meinhardt F, Schaffrath R, Larsen M. 1997. Microbial linear plasmids. *Applied Microbiology and Biotechnology*. doi:10.1007/s002530050936

Mönttinen HAM, Bicep C, Williams TA, Hirt RP. 2021. The genomes of nucleocytoplasmic large DNA viruses: Viral evolution writ large. *Microbial Genomics*. doi:10.1099/MGEN.0.000649

Mougari S, Bekliz M, Abrahao J, Pinto F di, Levasseur A, la Scola B. 2019. Guarani virophage, a new Sputnik-like isolate from a Brazilian lake. *Frontiers in Microbiology*. doi:10.3389/fmicb.2019.01003

Müller F, et al. 2022. A prebiotically plausible scenario of an RNA–peptide world. *Nature* 605. doi: 10.1038/s41586-022-04676-3

Namboodiri VMH, Dutta S, Akey I v., Head JF, Akey CW. 2003. The crystal structure of *Drosophila* NLP-core provides insight into pentamer formation and histone binding. *Structure* 11. doi:10.1016/S0969-2126(03)00007-8

Netherton CL, Wileman T. 2011. Virus factories, double membrane vesicles and viroplasm generated in animal cells. *Current Opinion in Virology*. doi:10.1016/j.coviro.2011.09.008

Nguyen A, Wang W, Chong E, Chatla K, Bachtrog D. 2022. Transposable element accumulation drives size differences among polymorphic Y chromosomes in *Drosophila*. *Genome Research*. doi: 10.1101/gr.275996.121

Nishimura Y, Shiratori T, Ishida K ichiro, Hashimoto T, Ohkuma M, Inagaki Y. 2019. Horizontally-acquired genetic elements in the mitochondrial genome of a centrohelid *Marophrys* sp. SRT127. *Scientific Reports* 9. doi:10.1038/s41598-019-41238-6

Novoa RR, Calderita G, Arranz R, Fontana J, Granzow H, Risco C. 2005. Virus factories: associations of cell organelles for viral replication and morphogenesis. *Biology of the Cell* 97:147–172. doi:10.1042/bc20040058

Nutman AP, Bennett VC, Friend CRL, van Kranendonk MJ, Chivas AR. 2016. Rapid emergence of life shown by discovery of 3,700-million-year-old microbial structures. *Nature* 537. doi:10.1038/nature19355

Oksanen H, Bamford D. 2012. Family *Tectiviridae* In: King AMQ, Adams MJ, Carstens EB, Lefkowitz EJ, editors. Virus Taxonomy. Oxford: Elsevier Scientific Publ. Co. pp. 317–322.

Oksanen HM. 2017. ICTV virus taxonomy profile: *Corticoviridae*. *Journal of General Virology* 98. doi:10.1099/jgv.0.000795

Onafuwa-Nuga AA, King SR, Telesnitsky A. 2005. Nonrandom packaging of host RNAs in Moloney Murine Leukemia Virus. *Journal of Virology* 79. doi:10.1128/jvi.79.21.13528-13537.2005

Paez-Espino D, Zhou J, Roux S, Nayfach S, Pavlopoulos GA, Schulz F, McMahon KD, Walsh D, Woyke T, Ivanova NN, Eloe-Fadrosh EA, Tringe SG, Kyrpides NC. 2019. Diversity, evolution, and classification of virophages uncovered through global metagenomics. *Microbiome* 7. doi:10.1186/s40168-019-0768-5

Pagarete A, Grébert T, Stepanova O, Sandaa RA, Bratbak G. 2015. Tsv-N1: A novel DNA algal virus that infects tetraselmis striata. *Viruses* 7. doi:10.3390/v7072806

Pérez-Losada M, Arenas M, Galán JC, Palero F, González-Candelas F. 2015. Recombination in viruses: mechanisms, methods of study, and evolutionary consequences. *Infection, Genetics and Evolution*. doi:10.1016/j.meegid.2014.12.022

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. doi:10.1016/j.gene.2006.08.008

Rice G, Tang L, Stedman K, Roberto F, Spuhler J, Gillitzer E, Johnson JE, Douglas T, Young M. 2004. The structure of a thermophilic archaeal virus shows a double-stranded DNA viral capsid type that spans all domains of life. *Proc Natl Acad Sci U S A* 101. doi:10.1073/pnas.0401773101

Roberts RM, Ezashi T, Schulz LC, Sugimoto J, Schust DJ, Khan T, Zhou J. 2021. Syncytins expressed in human placental trophoblast: human trophoblast syncytins. *Placenta* 113. doi:10.1016/j.placenta.2021.01.006

Roux S, Chan LK, Egan R, Malmstrom RR, McMahon KD, Sullivan MB. 2017. Ecogenomics of virophages and their giant virus hosts assessed through time series metagenomics. *Nature Communications* 8. doi:10.1038/s41467-017-01086-2

Rowe WP, Huebner RJ, Gilmore LK, Parrott RH, Ward TG. 1953. Isolation of a cytopathogenic agent from human adenoids undergoing spontaneous degeneration in tissue culture. *Proceedings of the Society for Experimental Biology and Medicine* 84. doi:10.3181/00379727-84-20714

Sakai HD, Kurosawa N. 2018. *Saccharolobus caldissimus* gen. nov., sp. nov., a facultatively anaerobic iron-reducing hyperthermophilic archaeon isolated from an acidic terrestrial hot spring, and reclassification of *Sulfolobus solfataricus* as *Saccharolobus solfataricus* comb. nov. and *Sulfolobus shibatae* as *Saccharolobus shibatae* comb. nov. *International Journal of Systematic and Evolutionary Microbiology* 68. doi:10.1099/ijsem.0.002665

Santini S, Jeudy S, Bartoli J, Poirot O, Lescot M, Abergel C, Barbe V, Wommack KE, Noordeloos AAM, Brussaard CPD, Claverie JM. 2013. Genome of *Phaeocystis globosa* virus PgV-16T highlights the common ancestry of the largest known DNA viruses infecting eukaryotes. *Proc Natl Acad Sci U S A* 110. doi:10.1073/pnas.1303251110

Satwika D, Klassen R, Meinhardt F. 2012. Repeated capture of a cytoplasmic linear plasmid by the host nucleus in *Debaryomyces hansenii*. *Yeast* 29. doi:10.1002/yea.2893

Schulz F, Roux S, Paez-Espino D, Jungbluth S, Walsh DA, Denef VJ, McMahon KD, Konstantinidis KT, Eloe-Fadrosh EA, Kyrpides NC, Woyke T. 2020. Giant virus diversity and host interactions through global metagenomics. *Nature* 578. doi:10.1038/s41586-020-1957-x

Sharp PM, Simmonds P. 2011. Evaluating the evidence for virus/host co-evolution. *Current Opinion in Virology*. doi:10.1016/j.coviro.2011.10.018

Shelford EJ, Middelboe M, Møller EF, Suttle CA. 2012. Virus-driven nitrogen cycling enhances phytoplankton growth. *Aquatic Microbial Ecology* 66. doi:10.3354/ame01553

Shi M, Lin XD, Chen X, Tian JH, Chen LJ, Li K, Wang W, Eden JS, Shen JJ, Liu L, Holmes EC, Zhang YZ. 2018. The evolutionary history of vertebrate RNA viruses. *Nature* 556. doi:10.1038/s41586-018-0012-7

Simmonds P, Aiewsakun P, Katzourakis A. 2019. Prisoners of war — host adaptation and its constraints on virus evolution. *Nature Reviews Microbiology* 17. doi:10.1038/s41579-018-0120-2

Starrett GJ, Tisza MJ, Welch NL, Belford AK, Peretti A, Pastrana DV, Buck CB. 2020. Adintoviruses: a proposed animal-tropic family of midsize eukaryotic linear dsDNA (MELD) viruses. *Virus Evolution*. doi:10.1093/ve/veaa055

Stough JMA, Yutin N, Chaban Y v., Moniruzzaman M, Gann ER, Pound HL, Steffen MM, Black JN, Koonin EV, Wilhelm SW, Short SM. 2019. Genome and environmental activity of a *Chrysochromulina parva virus* and its virophages. *Frontiers in Microbiology* 10. doi:10.3389/fmicb.2019.00703

Sun C, Feschotte C, Wu Z, Mueller RL. 2015. DNA transposons have colonized the genome of the giant virus *Pandoravirus salinus*. *BMC Biology* 13. doi:10.1186/s12915-015-0145-1

Takemura M, Yokobori SI, Ogata H. 2015. Evolution of eukaryotic DNA polymerases via interaction between cells and large DNA viruses. *Journal of Molecular Evolution* 81. doi:10.1007/s00239-015-9690-z

Tarlinton RE, Meers J, Young PR. 2006. Retroviral invasion of the koala genome. *Nature* 442. doi:10.1038/nature04841

Taylor BP, Cortez MH, Weitz JS. 2014. The virus of my virus is my friend: Ecological effects of virophage with alternative modes of coinfection. *Journal of Theoretical Biology* 354:124–136. doi:10.1016/j.jtbi.2014.03.008

Teichert JS, Kruse FM, Trapp O. 2019. Direct prebiotic pathway to DNA nucleosides. *Angewandte Chemie*. doi: 10.1002/anie.201903400

Thézé J, Takatsuka J, Nakai M, Arif B, Herniou EA. 2015. Gene acquisition convergence between entomopoxviruses and baculoviruses. *Viruses*. doi:10.3390/v7041960

Wacey D, Kilburn MR, Saunders M, Cliff J, Brasier MD. 2011. Microfossils of sulphur-metabolizing cells in 3.4-billion-year-old rocks of Western Australia. *Nature Geoscience* 4. doi:10.1038/ngeo1238

Walker PJ, Siddell SG, Lefkowitz EJ, Mushegian AR, Adriaenssens EM, Alfenas-Zerbini P, Davison AJ, Dempsey DM, Dutilh BE, García ML, Harrach B, Harrison RL, Hendrickson RC, Junglen S, Knowles NJ, Krupovič M, Kuhn JH, Lambert AJ, Łobocka M, Nibert ML, Oksanen HM, Orton RJ, Robertson DL, Rubino L, Sabanadzovic S, Simmonds P, Smith DB, Suzuki N, van Dooerslaer K, Vandamme AM, Varsani A, Zerbini FM. 2021. Changes to virus taxonomy and to the International Code of Virus Classification and Nomenclature ratified by the International Committee on Taxonomy of Viruses (2021). *Archives of Virology* 166. doi:10.1007/s00705-021-05156-1

Walker PJ, Siddell SG, Lefkowitz EJ, Mushegian AR, Adriaenssens EM, Dempsey DM, Dutilh BE, Harrach B, Harrison RL, Hendrickson RC, Junglen S, Knowles NJ, Kropinski AM, Krupovič M, Kuhn JH, Nibert M, Orton RJ, Rubino L, Sabanadzovic S, Simmonds P, Smith DB, Varsani A, Zerbini FM, Davison AJ. 2020. Changes to virus taxonomy and the Statutes ratified by the International Committee on Taxonomy of Viruses (2020). *Archives of Virology* 165. doi:10.1007/s00705-020-04752-x

Weiss MC, Sousa FL, Mrnjavac N, Neukirchen S, Roettger M, Nelson-Sathi S, Martin WF. 2016. The physiology and habitat of the last universal common ancestor. *Nature Microbiology* 1. doi:10.1038/nmicrobiol.2016.116

Wells JN, Feschotte C. 2020. A field guide to eukaryotic transposable elements. *Annual review of Genetics* 54. doi: 10.1146/annurev-genet-040620-022145

Whelan S. 2013. Viral replication strategies. Fields Virology: Sixth Edition.

Wodarz D. 2013. Evolutionary dynamics of giant viruses and their virophages. *Ecology and Evolution* 3:2103–2115. doi:10.1002/ece3.600

Wolf YI, Kazlauskas D, Iranzo J, Lucía-Sanz A, Kuhn JH, Krupovič M, Dolja V V., Koonin EV. 2018. Origins and evolution of the global RNA virome. *mBio* 9. doi:10.1128/mBio.02329-18

Wuitschick JD, Gershan JA, Lochowicz AJ, Li S, Karrer KM. 2002. A novel family of mobile genetic elements is limited to the germline genome in *Tetrahymena thermophila*. *Nucleic Acids Research* 30. doi:10.1093/nar/30.11.2524

Xiao Y, Rouzine IM, Bianco S, Acevedo A, Goldstein EF, Farkov M, Brodsky L, Andino R. 2016. RNA recombination enhances adaptability and is required for virus spread and virulence. *Cell Host and Microbe* 19. doi:10.1016/j.chom.2016.03.009

Yutin N, Raoult D, Koonin EV. 2013. Virophages, *Polintons*, and transpovirons: A complex evolutionary network of diverse selfish genetic elements with different reproduction strategies. *Virology Journal* 10. doi:10.1186/1743-422X-10-158

Yutin N, Wolf YI, Raoult D, Koonin EV. 2009. Eukaryotic large nucleo-cytoplasmic DNA viruses: Clusters of orthologous genes and reconstruction of viral genome evolution. *Virology Journal* 6. doi:10.1186/1743-422X-6-223

Zientz E, Dandekar T, Gross R. 2004. Metabolic interdependence of obligate intracellular bacteria and their insect hosts. *Microbiology and Molecular Biology Reviews* 68. doi:10.1128/mmbr.68.4.745-770.2004

Chapter I

Paleovirology of the DNA viruses of eukaryotes

Abstract

Paleovirology is the study of ancient viruses and how they have coevolved with their hosts. An increasingly detailed understanding of the diversity, origins and evolution of the DNA viruses of eukaryotes has been obtained through the lens of paleovirology in recent years. Members of multiple viral families have been found integrated in the genomes of eukaryotes, providing a rich fossil record to study. These elements have extended our knowledge of exogenous viral diversity, host ranges, the timing of viral evolution and are revealing the existence of entire new families of eukaryotic integrating dsDNA viruses and transposons. Future work in paleovirology will continue to provide insights into antiviral immunity, viral diversity, potential applications and reveal other aspects of the viral world.

Viruses (see Glossary, Appendix) are one of the dominant biological entities on Earth, containing a significant portion of the genetic diversity in the biosphere (Koonin et al., 2015) and outnumbering cells by a factor of 2.6-160 in many ecosystems (Wigington et al., 2016). They are an ancient component of the Earth's biomes and may have driven major evolutionary transitions by the interaction with their hosts (Koonin, 2016; Koonin and Krupovič, 2018). DNA viruses comprise an enormous diversity of viruses infecting eukaryotes. They range in size from the porcine circoviruses (some of the smallest viruses known), to the giant viruses that infect microbial eukaryotes (Brandes and Linial, 2019; Opriessnig et al., 2020). Their genomes are also diverse, they can be single or double-stranded, circular or linear DNA molecules, and two families undergo reverse transcription (Krupovič et al., 2020, 2018; Senkevich et al., 2015). Their close relationship as obligate intracellular parasites of eukaryotes has translated into frequent genetic exchanges throughout evolutionary time. In particular, the large genome sizes of eukaryotes, have meant that ancestral viral integrations can accumulate in the genomes of their hosts (Lynch and Marinov, 2015), forming a rich genomic fossil record and providing an opportunity to study the deep viral past. In this review, we will highlight key achievements and exciting discoveries in the paleovirology of the DNA viruses of eukaryotes, which are leading to novel insights and increasing our understanding of their diversity, origins and evolution. We will focus on DNA viruses since they have received less attention than their RNA counterparts in terms of paleovirology, even though classic work on DNA viruses laid the foundations of this science more than 70 years ago (Lwoff and Gutmann, 1950; Wollman, 1953).

Paleovirology: probing virus origins and evolution

Paleovirology is the study of ancient viruses over macroevolutionary timescales. It is concerned with how major virus lineages have originated, diversified and coevolved with their hosts. Paleovirology relies mostly on comparative genomics and phylogenetics of extant viral sequences to draw inferences about the viral past. However, there is indirect evidence of viruses in the fossil record and viral sequences that have been recovered from ancient DNA (aDNA) (**Box 1, p. 83**). One of the major challenges in paleovirology is the sequence divergence between related viruses. Over long timescales (hundreds of millions of years), viral sequences can diverge beyond recognition, but this can be overcome sometimes by performing comparisons of protein structure which is more conserved during evolution (Baker et al., 2005; Krupovič and Bamford, 2008).

Our knowledge in paleovirology comes mostly from the study of endogenous viral elements (EVEs). EVEs are viral integrations that have entered the germline of their hosts and are inherited as host alleles; their survival is determined by the balance between genetic drift and natural selection (Charlesworth, 2009) (Figure 1). On some occasions EVEs are repurposed to perform a beneficial host function and become exaptations (Koonin and Krupovič, 2018). An important property of EVEs is that they can be used to calibrate the timing of virus evolution: if an EVE is orthologous across a number of species, this gives a minimum estimate for the age of the virus that integrated into the genome (Katzourakis and Gifford, 2010). We can therefore think of EVEs as a genomic fossil record of viruses.

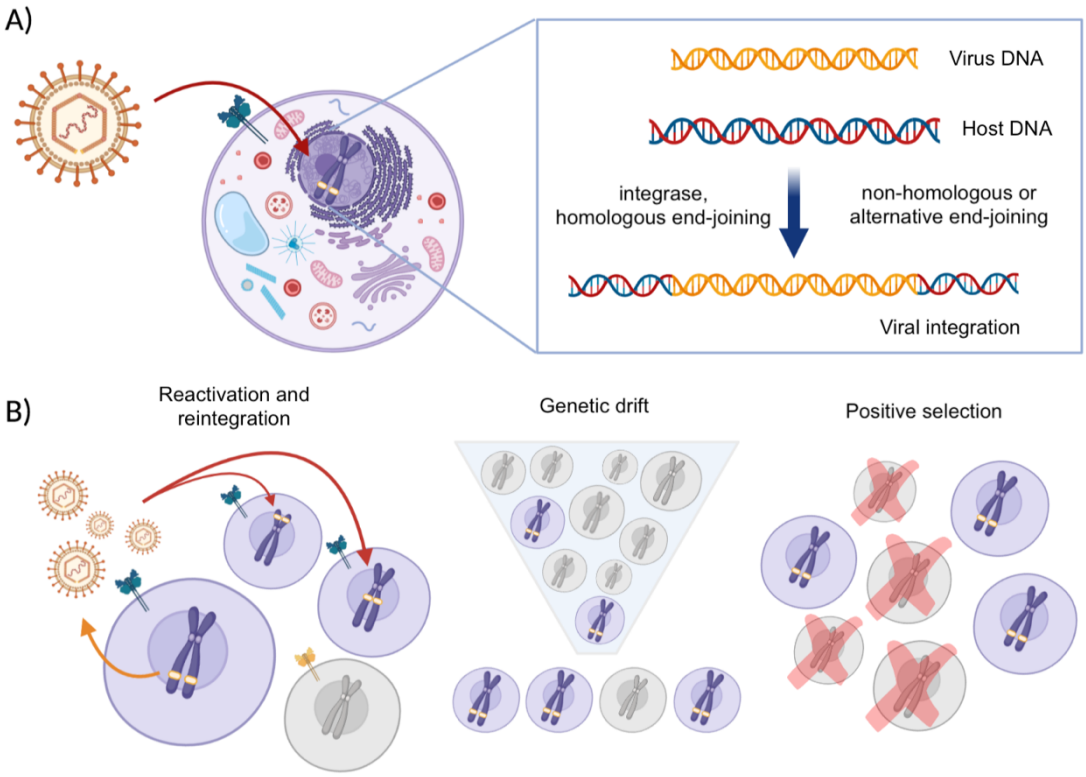


Figure 1. Endogenous viral element formation and persistence in eukaryotic genomes. A) By infecting a susceptible cell, which expresses the required receptor for entry (blue), the virus genetic material can potentially gain access to the nucleus and be integrated into the cell's genome. This viral integration is considered an EVE once it has been inherited vertically. B). The persistence of an EVE throughout evolutionary time can result from different processes. If the virus has integrated an intact viral copy it might persist by continued cycles of reactivation and reintegration, leading to target-site polymorphisms. Cells might become non-susceptible by a mutation in the receptor protein (yellow), restricting the virus' ability to colonise the genome. Alternatively, genetic drift may stochastically change the frequency of an EVE, here illustrated by a population-bottleneck. Finally, an EVE might confer a selective advantage on the host and will increase in frequency as a result of positive selection. Over evolutionary time, certain viruses such as bracoviruses have colonised all their hosts' chromosomes.

The idea of endogenous viral elements has its origins in the work with lysogenic bacteriophages by Lwoff and Gutmann (Lwoff and Gutmann, 1950), who showed that the virus had an endogenous stage which became a heritable feature of the host bacterium (termed a prophage). Wollman later suggested that the prophage is "intimately tied" to the genetic material of the bacterium, and is inherited as a host gene (Wollman, 1953). This general mechanism of integration into the host genome was later extended to some animal viruses: Lwoff proposed that polyomavirus could integrate into the chromosomes of cancerous cells (Lwoff, 1960), and Temin used this idea to construct the "provirus hypothesis" for *Rous sarcoma virus*, an oncogenic retrovirus of chicken (Temin, 1964). For many years it was believed that retroviruses, were the only viruses to enter the eukaryotic germline given that integration into the host genome is a step required to complete viral replication (forming a provirus). However, the discovery of multiple geminivirus-like sequences in the genome of tobacco plants (*Nicotiana tabacum*), demonstrated that DNA viruses could also enter the germlines of eukaryotes (Bejarano et al., 1996). Eventually, it was shown that viruses from all Baltimore groups could potentially integrate into eukaryotic genomes (Katzourakis and Gifford, 2010).

Compared to the RNA-retrotranscribing viruses, DNA virus EVEs occur less frequently (Katzourakis and Gifford, 2010). Most families of eukaryotic DNA viruses do not encode an integrase, which limits direct integration into the host genome. Overall, the integration mechanisms for DNA virus EVEs are less well understood, although progress is being made (**Box 2, p. 85)**. Many families of DNA viruses have considerably larger genomes than RNA viruses, and integration of a full viral copy may be strongly counter-selected by the host. In general, it seems that the most abundant DNA virus EVEs in eukaryotic genomes are from families with smaller genome sizes (<10 kb), those which preferentially replicate in the

nucleus and the families which do encode an integrase (Gong and Han, 2018a; Inoue et al., 2017; Katzourakis and Gifford, 2010; Huiquan Liu et al., 2011; Lytras et al., 2021; Pritham et al., 2007). Large viral integrations may also be more likely to be toxic for the host or have a negative effect on gene regulation, which may reduce their chances of fixation (Alexander Suh, personal communication). The finding of very large herpesviral EVEs (>100 kb), is probably facilitated by the tendency of these viruses to establish latent infection in the host nucleus (Aimola et al., 2020). From an evolutionary standpoint, DNA viruses evolve more slowly and probably codiverge more often with their hosts (Barreat and Katzourakis, 2021a; Gallot-Lavallée and Blanc, 2017; Geoghegan et al., 2017), suggesting that their associations with hosts are more stable as compared to RNA viruses.

The diversity of DNA viruses found in eukaryotic genomes

DNA viruses in the genomes of eukaryotes fall into four viral kingdoms (*Bamfordvirae* [realm *Varidnaviria*], *Heunggongvirae* [realm *Duplodnaviria*], *Shotokuvirae* [realm *Monodnaviria*] and *Pararnavirae* [realm *Riboviria*]) and two genera currently classified in the family "Polydnaviridae" (Figure 2). The kingdoms have been established by homology in protein structure and the molecular biology of the viruses. The kingdom *Bamfordvirae* comprises a diverse assemblage of viruses which share major and minor capsid proteins with vertical jelly-roll folds (Krupovič and Koonin, 2015). Viruses in the kingdom *Bamfordvirae* infect all major eukaryotic lineages and include the Nucleocytoplasmic Large DNA viruses (NCLDVs), virophages, *Mavericks*, Polinton-like viruses and adenoviruses. The kingdom *Heungonggvirae* includes the herpesviruses which infect animals as well as the tailed bacteriophages, which share the HK97 fold in their capsid proteins as well as similarities in the portal complex and the virus assembly pathways (Baker et al., 2005).

Another major lineage of eukaryotic viruses with endogenous elements is the kingdom *Shotokuvirae* (phyla *Cressdnaviricota* and *Cossaviricota*). The phylum *Cressdnaviricota* comprises the families *Geminiviridae*, *Genomoviridae*, *Circoviridae* and *Nanoviridae*, which are known as the "circular, rep-encoding single-stranded DNA viruses" (CRESS) (Krupovič et al., 2020). The family *Parvoviridae*, composed of viruses with linear single-stranded genomes and the families *Papillomaviridae* and *Polyomaviridae*, which have viruses with circular double-stranded DNA genomes, share replication proteins homologous to the ones in CRESS DNA viruses and thus have likely evolved from a common ancestor (Hickman et al., 2002; Kazlauskas et al., 2019). Hepadnaviruses and caulimoviruses, which infect vertebrates and plants respectively, are classified in the kingdom *Pararnavirae*. These viruses replicate by transcribing their genomes into RNA and then using a reverse transcriptase to make more dsDNA genomes (Krupovič et al., 2018). Viruses in the family "Polydnaviridae" are found integrated in the genomes of parasitoid wasps; these viruses are used as gene transfer agents to deliver virulence factors into the host where the wasp larvae develop (Drezen et al., 2017; Gauthier et al., 2018). "Polydnaviridae" is an artificial grouping since the two genera *Bracovirus* and *Ichnovirus* have distinct evolutionary origins (Gauthier et al., 2018; Thézé et al., 2011; Volkoff and Cusson, 2020).

The integrations from these viruses have revealed two common themes: 1) some virus lineages have integrated more broadly than we would expect from the exogenous host range, indicating that these groups may infect a greater number of eukaryotic species than currently thought and 2) some integrations represent new viral families/lineages with no known exogenous counterparts, effectively increasing our knowledge of eukaryotic viral diversity. Circoviruses, for example, are known to infect vertebrates and arthropods, but their EVEs

have also been identified in nematodes, gastropods and in *Giardia intestinalis* (Cui et al., 2014; Dennis et al., 2019; C. Gilbert et al., 2014; Katzourakis and Gifford, 2010). Geminiviruses and nanoviruses are viral pathogens of plants, but their EVEs have been found in fungi, *Entamoeba hystolitica* (Huiquan Liu et al., 2011), plants (Filloux et al., 2015) and insects (ter Horst et al., 2018), while nanovirus-like sequences have been detected in plants, amphibians, crustaceans, insects, placozoans, green algae, diatoms, and in the *Giardia*, *Entamoeba* and *Blastocystis* protozoan parasites (Huiquan Liu et al., 2011). It has been noted that geminiviral and nanoviral-like EVEs in insect genomes probably derive from plant viruses that use these animals as vectors, in particular by the presence of movement protein genes (ter Horst et al., 2018). Similarly, although hepadnaviruses are known to infect mammals and birds, hepadnavirus EVEs indicate that turtles and crocodilians are also potential hosts (Suh et al., 2014). This is consistent with the discovery of exogenous hepadnaviruses (and 'nakednaviruses') infecting additional vertebrate hosts (amphibians and teleost fish) (Dill et al., 2016; Lauber et al., 2017). NCLDV EVEs have also been discovered which suggest a broad association with all major eukaryotic lineages and a greater phylogenetic diversity than presently acknowledged (Gallot-Lavallée and Blanc, 2017; Gong et al., 2020; Maumus and Blanc, 2016). These insights gained from paleovirology can complement other sequence-based approaches such as metagenomics (Schulz et al., 2020), as powerful tools for virus discovery.

Multiple novel viral lineages that have not been found in exogenous form, have been discovered by analysing the endogenous elements in host genomes. Naryaviruses, nenyaviruses and vilyaviruses are new clades of CRESS DNA viruses which have been identified in the genomes of the human protozoan parasites *Giardia duodenalis* and *Entamoeba* spp. (Kinsella et al., 2020). Their distinct placement on the Rep phylogeny

suggests they potentially represent 3 new viral families (Kinsella et al., 2020). Another viral lineage distantly related to circoviruses and nanoviruses was identified in the genome of the pillbug *Armadillidium vulgare*, and a possible nonplant-infecting lineage of nanoviruses in the genomes of the molluscs *Crassostrea gigas* and *Lottia gigantea* (Thézé et al., 2014). In plants, a new lineage of endogenous caulimoviruses named 'florendoviruses' was discovered; they consist of bipartite genomes (unlike other caulimoviruses) and exist at very high copy numbers (>0.5% total genome size) (Geering et al., 2014). Importantly, new lineages of endogenous viruses (that is, functional viruses that persist integrated in the genomes of hosts), have also been discovered. *Mavericks* probably represent at least two new viral families (Barreat and Katzourakis, 2021a). They were previously described as large transposons (~20 kb), but they were later found to encode proteins homologous to capsid forming proteins of other bamfordviruses (Krupovič et al., 2014). *Mavericks* found in the genomes of vertebrates undergo frequent cross-species transmissions and their core genes are evolving under strong purifying selection, further supporting their existence as endogenous viruses (Barreat and Katzourakis, 2021a). *Teratorns* are another remarkable group of mobile genetic elements which encode proteins homologous to the proteins of alloherpesviruses (Aswad and Katzourakis, 2017). Teratorns are extremely large elements (~180 kb) which have been found in the genomes of teleost fish (Aswad and Katzourakis, 2017; Inoue et al., 2018, 2017). They likely represent a new viral family which is the sister lineage to *Alloherpesviridae* (Aswad and Katzourakis, 2017). Intriguingly, the ichnoviruses that have been domesticated by ichneumonid parasitoid wasps have an uncertain viral origin (Volkoff and Cusson, 2020). Although there is no sequence similarity to any other known lineage of viruses, their viral ancestry is supported by the absence of introns in their genes (more typical of virus rather than wasp genes) (Volkoff and Cusson, 2020). Therefore, it will be important to clarify the viral (or cellular) ancestry of ichnoviruses.

In general, oncogenic DNA virus integrations seem to be absent from the genomic fossil record of animals. Adenoviruses and polyomaviruses are known to integrate into somatic cells but have not been found in host germlines (Doerfler, 1968; Feng et al., 2008; Sutter et al., 1978), there is only one report of a potentially integrated papillomavirus in the genome of the platypus (although an integration site has not been identified) (Cui and Holmes, 2012), and herpesviral integrations have only been described in the tarsier genome or as low frequency (~1%) polymorphisms in the human population (Aimola et al., 2020; Aswad and Katzourakis, 2014). Despite the rich genomic fossil record of hepadnaviruses in vertebrates, no hepadnaviral integrations in mammals have been identified. Interestingly, the oncogenic X protein is specific to the *Orthohepadnavirus* genus that infects mammals, other vertebrate hepadnaviruses do not encode this protein and they are not known to be oncogenic to their hosts (Suh et al., 2014).

Coevolution of DNA viruses with eukaryotes

The DNA virus EVEs of eukaryotes have also revealed the long timescales of virus-host coevolution. Orthologous EVEs provide direct evidence for the existence of viral lineages tens to hundreds of millions of years in the past (Figure 2). The oldest orthologue dated is an hepadnavirus EVE integrated in the genomes of turtles 207-230.7 Mya (Suh et al., 2014). Another hepadnavirus EVE in the genomes of birds and the tuatara could be potentially 280 My old, but authors caution against this finding since orthology was difficult to assess (Lytras et al., 2021). Two orthologous *Maverick* integrations also occur in the genomes of turtles with an estimated age of 95 My (Barreat and Katzourakis, 2021a). In humans, a parvovirus capsid gene is found integrated in an intron of the *Ellis van Creveld syndrome 2*

gene, the integration occurred at least 98 Mya since orthologous EVEs are found in non-human primates, carnivores, ungulates and dolphins (H. Liu et al., 2011). Orthologous circovirus EVEs have also been dated to 72-90 My based on their presence in pythons and vipers (Dennis et al., 2019). In plants, an endogenous caulimovirus orthologue has been discovered in the genomes of *Pinus taeda* and *Pinus lambertiana*, indicating an age of at least 75 My (Gong and Han, 2018a).

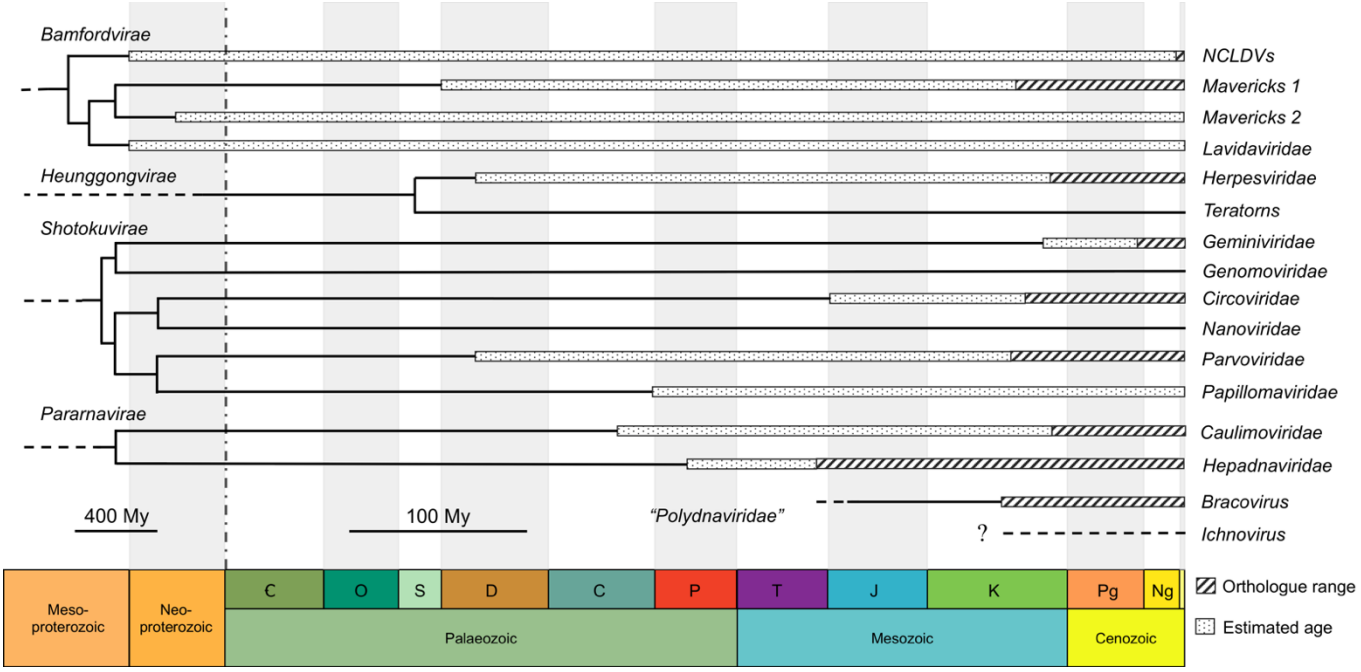


Figure 2. Evolutionary trees of the major lineages of DNA viruses found in eukaryotic genomes. Cross-hatched boxes represent the age of the oldest viral integrations identified by orthology in host species, while dotted boxes represent the age estimated for the groups based on molecular dating, genetic exchanges, codivergence or biogeographic data. The cladistic relationships shown for *Bamfordvirae* and *Shotokuvirae* are based on current evolutionary hypotheses; the tree for *Bamfordvirae* depicted as a "virophage-first" scenario (Barreat and Katzourakis, 2021b; Kazlauskas et al., 2019). Node ages reflect the divergence date for the hosts infected by the different viral lineages, but this codivergence assumption should be handled with caution. **C**: Cambrian, O: Ordovician, S: Silurian, D: Devonian, C: Carboniferous, P: Permian, T: Triassic, J: Jurassic, K: Cretaceous, Pg: Paleogene, Ng: Neogene. The timescales are in million years from the present (My). Note a shift in the measurement timescales denoted by the vertical dotted line (at 485.4 My).

Evidence for the ancient coexistence of eukaryotes and DNA viruses has also been obtained from analyses of sequence divergence or by close correspondence of the virus/host phylogenies. By using relaxed molecular clocks on a phylogeny calibrated by the dates of genomic orthologues, *Mavericks* have been estimated to have infected vertebrates since at least 419 Mya (Barreat and Katzourakis, 2021a). In a group of parvoviruses called "chapparvoviruses", endogenised elements mostly follow the host phylogeny, suggesting they have infected their insect hosts for at least 400 My (Pénzes et al., 2019). Similarly, the evolution of circovirus EVEs in ray-finned fish agrees with the host phylogeny, suggesting they potentially integrated >200 Mya (Dennis et al., 2019). From the divergence in the repeat regions (which are identical at integration) and using rates of neutral evolution reported for mammals, an endogenous herpesvirus found in the genome of the tarsier (*Tarsius syrichta*) has been dated to 56-76 My (Aswad and Katzourakis, 2014). In humans, herpesviruses 6A and 6B (HHV6A/B) can infect cells of the germline and can be inherited vertically (Wight et al., 2020). A specific HHV6B integration has been found in the genomes of three different individuals from Maasai, Pakistani or Native American descent, which suggests the integration of this virus occurred ~84,000-324,000 years ago before the modern human migrations out of Africa (Aswad et al., 2020).

Additional evidence has been obtained for the existence of NCLDV and virophages more than a billion years ago. Phylogenetic analyses indicate that the DNA-dependent RNA polymerase was exchanged on several occasions between NCLDVs and proto-eukaryotes, which occurred 1-2 billion years ago (Guglielmini et al., 2019). Comparison of different phylogenetic hypotheses using the sequences of the core viral proteins further suggest that virophages evolved to parasitise NCLDVs shortly after the origin of the NCLDV ancestor (Barreat and Katzourakis, 2021b). This is consistent with the close correspondence between

the promoter and poly-A sequences of virophages and NCLDVs, which would have been inherited from a recent common ancestor and this specificity maintained throughout evolutionary time (Campbell et al., 2017; Fischer and Suttle, 2011). This lineage of protovirophages would have further diversified into the modern virophages (lavidaviruses), *Mavericks*, Polinton-like viruses, cytoplasmic linear plasmids and adenoviruses (Barreat and Katzourakis, 2021b).

Another relevant aspect in DNA virus evolution is the role of cross-species transmissions. Although DNA viruses tend to have stable host associations and to co-speciate with their hosts, cross-species transmissions can occur frequently in some lineages. For example, it was shown that although *Mavericks* in cyprinid fish had mainly evolved by co-speciation (65%), cross-species transmissions occurred 25% of the time (Barreat and Katzourakis, 2021a). Similarly, although the evolution of hepadnaviruses is characterised by stable host associations and they mirror the host phylogeny, there is evidence suggesting multiple host-jumps between vertebrates and even a cross-class transmission between fish and mammals (Dill et al., 2016). The macroevolutionary mode in caulimoviruses, however, seems to be dominated by cross-species transmissions rather than co-speciation; these include cross-division transfers in plants (Diop et al., 2018; Gong and Han, 2018a). The reason for this difference could be due to the transmission mode of caulimoviruses which use insect vectors to spread to new plants; the finding of endogenous caulimoviruses in the genomes of these insects could support this hypothesis. Paleovirological evidence could thus be used to identify hidden vectors/intermediates during cross-species transmissions.

Transitions between viruses and transposons

The DNA viruses of eukaryotes have given rise to new lineages of endogenous viruses and replication-competent transposons. Endogenous viruses encode the proteins required to make an infectious particle, while transposons do not. The *Teratorn* lineage of endogenous viruses originated when an exogenous virus related to alloherpesviruses acquired the integrase of a *piggyBac* transposon (Aswad and Katzourakis, 2017; Inoue et al., 2018). Similarly, the Mavirus-like virophages as well as *Mavericks*, share a retroviral-like integrase that was acquired from an ancestral retroelement (Krupovič and Koonin, 2015; Pritham et al., 2007). Endogenous virophages are an example of DNA virus EVEs exapted to function in immunity (**Box 3, p. 87**). The bracoviruses, which are dispersed across the chromosomes of braconid parasitoid wasps, originate from the domestication of endogenous nudiviruses that integrated into the wasp genome ~103 Mya (Gauthier et al., 2021; Thézé et al., 2011). In fact, the nudivirus ancestors of bracoviruses encode tyrosine recombinases and have also integrated extensively across a broad range of arthropod genomes (Cheng et al., 2020).

Multiple lineages of eukaryotic transposons also descend from ancestral DNA viruses. In the black truffle, *Tuber melanosporum*, geminivirus rep-like sequences have been found in association with 2 transposase ORFs and flanked by terminal inverted repeats (TIRs); these elements form their own clade and seem to have amplified within the truffle genome (Huiquan Liu et al., 2011). Another rep-like element with TIRs appears in the planarian, *Schmidtea mediterranea*; there are multiple copies which are similar to the extrachromosomal PEVE element (Huiquan Liu et al., 2011). Circovirus rep-like elements have undergone independent intragenomic amplifications in carnivores in association with LINE-1 retroelements (Dennis et al., 2019). In birds, certain endogenous hepadnaviruses

have reached more than 300 copies in their host genomes and were hypothesised to have been mobilised by flanking TEs (Lytras et al., 2021).

The transition from transposons to viruses is illustrated by hepadnaviruses and caulimoviruses. The reverse transcriptases of viruses form a monophyletic group in the phylogeny of retroelements (Gong and Han, 2018b; Krupovič et al., 2018; Krupovič and Koonin, 2017), suggesting that they ultimately descend from a transposable element that acquired a capsid protein. However, while the capsid protein of caulimoviruses is homologous to the proteins of retroviruses, metaviruses and pseudoviruses (orterviruses), hepadnaviruses encode a non-homologous capsid protein (Krupovič and Koonin, 2017). Recently, a lineage of hepadnavirus-like retroelements (HEART1/2) was discovered in the genomes of arthropods (Gong and Han, 2018b). HEART elements are believed to be retrotransposons since they lack the core protein required for capsid formation, appear in more than 300 copies in some insect genomes and have a conserved linear genetic organisation (Gong and Han, 2018b). Discovery of the HEARTs raises two possible scenarios for the origins of retrotranscribing viruses: 1) either the orterviruses and hepadnaviruses arose independently from an ancestral retroelement by acquiring different capsids or 2) orterviruses and hepadnaviruses descend from a common viral ancestor that exchanged the ancestral capsid gene in one of the lineages (implying that HEARTs evolved from an ancestral virus by reductive evolution). The question about the nature of the ortervirus/hepadnavirus ancestor (either virus or retrotransposon), could be illuminated by searching for other closely related lineages of endogenous viruses or retrotransposons in the genomes of eukaryotes.

Functional significance and potential uses of DNA virus EVEs

In addition to the roles of endogenous virophages, bracoviruses and ichnoviruses as defence and offence mechanisms in immunity, DNA virus EVEs have been an important source of evolutionary novelty in eukaryotes. These include novel DNA binding-proteins, the eukaryotic DNA-dependent RNA polymerases as well as a role in horizontal transfer of genetic material (**Box 4, p. 88**). Importantly, the new viral diversity uncovered by paleovirology could be a source of useful biotechnological applications. For example, adeno-associated viruses (*Parvoviridae*) are well-established vectors used in genetic engineering; however, since they are small viruses, transgene sizes are limited to about 5 kb (Naso et al., 2017). As an alternative, retroviral vectors can package up to 11 kb of genetic material (Sweeney et al., 2017). Since they have the capacity to integrate and also infect vertebrates, *Mavericks* and *Teratorns* seem to be attractive candidates to develop new viral vectors that would deliver larger/multiple transgenes to the host chromosome. These endogenous viruses could also be potentially used to develop new viral-based vaccines or vectors for reprogramming T-cells against cancer (Lauer et al., 2017; Milone and O'Doherty, 2018).

Concluding Remarks and Future Perspectives

Paleovirology is advancing our understanding of the diversity and deep origins of eukaryotic DNA viruses. We have recognised endogenous elements from numerous viral families and we are constantly uncovering a hidden diversity of viral lineages integrated across the genomes of eukaryotes. These endogenous viral elements are giving us insights into the interactions between viruses and hosts over geological timescales, which span a very

detailed record in the millions to tens of millions of years and a few integrations in the hundreds of millions of years. The inferences from these findings are starting to give us insights into virus evolution in the scale of billions of years in the past. Research in paleovirology has also helped to disentangle the blurry evolutionary boundaries between eukaryotic DNA viruses and transposons. We are learning increasingly more about their impact on host biology, and especially about their role in immunity. Many questions are still open to enquiry (see Outstanding Questions, Appendix). New algorithms that bypass the need for detecting homology to known viral proteins could greatly increase our knowledge of viral diversity and evolution. In addition, it will be important to study the molecular virology of elements such as *Mavericks* and *Teratorns*, given that they could potentially be used as new vectors in genetic engineering, gene therapy or for the development of new vaccines.

Box 1. Evidence of ancient viruses from fossils and aDNA

The study of ancient viruses, i.e. viruses which have infected organisms of the past, can also be informed by the geological fossil record and molecular fossils, in particular, by the isolation and sequencing of aDNA (genetic material which has been preserved under exceptional circumstances). Two reports have been made of indirect evidence of viruses in the geological fossil record: in non-avian dinosaurs and insects. The vertebra of a Jurassic dinosaur (dated at ~150 My) was described in 2011 which showed characteristic signs of Paget's disease (Witzmann et al., 2011). Although it is not entirely clear what causes Paget's disease, it seems underlying genetic factors may interact with paramyxoviruses to produce an altered histology at the site of infection (Kravets, 2018). Structures resembling viral occlusion bodies (similar to the ones produced by baculoviruses) were reported in the gut of amber-fossilised insects from the Cretaceous Period (~100 My) (Poinar and Poinar, 2005).

Although this evidence is indirect, it hints to the possibility of finding actual fossilised virions or molecular fossils in the future. Direct sequencing of virus genomes is indeed possible from archaeological samples. A study of ancient DNA from human Bronze and Iron age samples (800—4,500 years), recovered sequences from hepatitis B viruses (family *Hepadnaviridae*) infecting these individuals (Mühlemann et al., 2018). Their analyses revealed an extended diversity of hepatitis B viruses which is not seen today, and concluded that present day diversity arose only after the split of the New and Old-World genotypes (Mühlemann et al., 2018). In exceptional circumstances, more ancient viruses have been recovered. A 30,000-year old giant virus, *Pithovirus sibericum*, was isolated from permafrost after culture in *Acanthamoeba castellanii* (Legendre et al., 2014). This new virus combined a pandoravirus amphora-like capsid morphology with a gene content more similar to icosahedral viruses (Legendre et al., 2013). A later study from the same permafrost sediment samples characterised *Mollivirus sibericum*, another nonicosahedral giant virus which appeared to be a distant relative of pandoraviruses or from another lineage which acquired pandoravirus genes from extensive horizontal gene transfers (Legendre et al., 2015). These examples illustrate a field of paleovirology which remains largely unexplored.

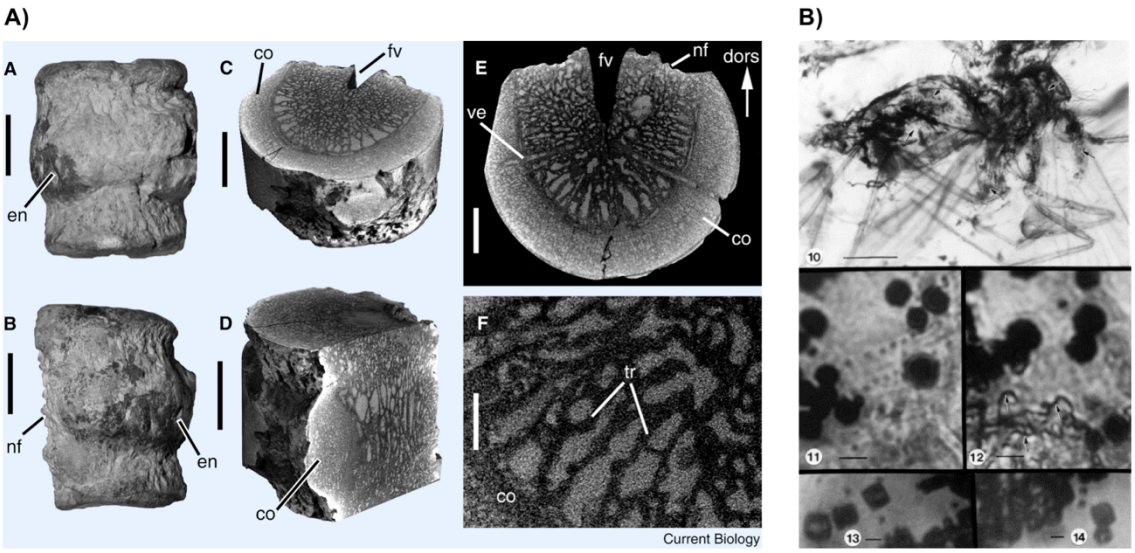


Figure 3. Evidence of viruses in the geological fossil record. A) Sections of a vertebra from *Dysalotosaurus lettowvorbecki* showing characteristic signs of Paget's disease: bone outgrowth, thickening of the bone cortex and formation of internal trabeculae. B) Specimen of an amber-fossilised sandfly (Diptera: Phlebotomidae), showing structures which resemble, in size and geometry, viral occlusion bodies present in the gut. Images reproduced with permission from Elsevier (Poinar and Poinar, 2005; Witzmann et al., 2011).

Box 2. Mechanisms of DNA virus integration in eukaryotes

DNA virus EVEs have formed by a number of mechanisms in eukaryotes. *Mavericks*, *Teratorns* and virophages, which encode their own integrase/transposase/recombinase, can mediate direct integration of the element into the host genome (Fischer and Suttle, 2011; Inoue et al., 2017; Pritham et al., 2007). The DNA circles of bracoviruses, endogenous viruses of parasitoid wasps, also integrate into host immune cells using a tyrosine recombinase and follow a unique mechanism involving sequence cleavage and loss at the host integration motifs (HIMs) (Burke et al., 2014; Chevignon et al., 2018). Human herpesviruses 6 A and B (HHV6A/B) integrate full genome copies into telomeres, probably

by using the homologous recombination pathway (either break-induced replication, requiring Rad51 or single-strand annealing, requiring MRN/RPA/Rad52), which depends on the sequence similarity of the telomeric repeats in the virus (pTMR) and the host (Aimola et al., 2020). Most other integrations are thought to occur via the non-homologous end-joining or microhomology-mediated pathways. Adeno-associated virus (AAV), a virus satellite of adenovirus in the family *Parvoviridae* (genus *Dependoparvovirus*), can integrate via site-specific nicking of the host DNA by the viral Rep, followed by host DNA synthesis, strand displacement and template switching to the AAV genome, which generates an endogenous copy of AAV (Henckaerts et al., 2009). In humans, AAV can latently infect cells by integrating into chromosome 19q13.42, causing a partial duplication of the *myosin-binding subunit 85* (*MBS85*) gene (Henckaerts et al., 2009). Proteins involved in the nonhomologous end-joining pathway, Ku70/80, PARP1 and Rad50, have been shown to participate in the site-specific integration mechanism of AAV (Daya et al., 2009). In addition, endogenous viral elements from hepadnaviruses have been proposed to integrate via non-homologous end-joining induced after double-strand breaks (Gilbert and Feschotte, 2010). Caulimoviruses have also been suggested to integrate via either the non-homologous end-joining or microhomology mediated pathways (Geering et al., 2014; Gong and Han, 2018a). It is plausible that additional DNA virus EVEs have arisen by integration through these repair mechanisms, which are mainly found as genome fragments, but experimental evidence for specific cases is still lacking. The synthesis-dependent end-joining pathway, which relies on regions of microhomology and the presence of inverted or direct repeats at virus-host junctions (Yu and McVey, 2010), could also account for the integration of some DNA viruses. Another possibility is that DNA virus EVEs may integrate by interaction of RNA transcripts with the machinery of retroelements in cells, such as has been found to occur extensively for RNA virus EVEs (Katzourakis and Gifford, 2010). The hallmarks for

integration via LINEs for example, are a poly-A tail and flanking target-site duplications (Katzourakis and Gifford, 2010), however, signs for this integration pathway have not been characterised for DNA virus EVEs.

Box 3. Endogenous DNA viruses in immunity

A number of endogenous DNA virus integrations function in immunity. That is the case for provirophages, or the integrated form of virophages. The Mavirus virophage can integrate into the genome of its host *Cafeteria burkhardae*, and upon coinfection with its giant virus *Cafeteria roenbergensis virus* (CroV), it reactivates and interferes with viral replication (Fischer and Hackl, 2016). Provirophage reactivation leads to decreased CroV titres, and although it does not protect the eukaryotic host from dying, release of newly produced virophages protects the wider host population (Berjón-Otero et al., 2019). As such, integrated provirophages may be seen as an altruistic defence system against giant virus infection in unicellular eukaryotes (Fischer and Hackl, 2016). Another remarkable case is that of polydnaviruses, a group of endogenous viral elements in the genomes of parasitoid wasps (Superfamily Ichneumonoidea) (Gauthier et al., 2018). Polydnavirus EVEs have been domesticated by parasitoid wasps on at least three separate occasions, to be expressed in the ovaries of females and package virulence genes which are delivered to the caterpillar host (Thézé et al., 2011). These are essential in suppressing the host immune system for development of the wasp larvae. Elements from the *Bracovirus* genus, which infect 'microgastroid complex' wasps, seem to be monophyletic and diverged from betanudiviruses ~190 My ago (101-340 My) (Gauthier et al., 2018; Thézé et al., 2011). In contrast, polydnaviruses in the genus *Ichnovirus*, which infect wasps in the Campopleginae and Banchinae subfamilies, do not appear to be monophyletic and suggest two independent acquisitions (Béliveau et al., 2015). In addition, some DNA virus EVEs appear to contribute

to immune defence by RNA interference mechanisms. In the *Ornithodoros* soft ticks, small-interfering RNAs (siRNA) and PIWI-interacting RNAs (piRNAs) are produced from endogenous *African Swine Fever virus*-like integrations which may confer these vectors some protection against infection by the virus (Forth et al., 2020). The piRNA pathway is based on the transcription of sequences from genomic clusters which are processed to 23-30 base-long RNAs, which can then bind to and cleave other transcripts with complementary sequences (de Araujo et al., 2022). Apart from this primary mechanism, piRNAs can also be amplified via the "ping-pong" mechanism where the cleaved templates can be used to produce more piRNAs (de Araujo et al., 2022). The piRNA pathway has been implicated in the antiviral response of mosquitoes and hemipterans (de Araujo et al., 2022). In a study of arthropod genomes (ter Horst et al., 2018), piRNA clusters were found to be enriched in EVEs and 18.49% of these integrations resembled parvoviruses. However, the small RNAs did not have a high similarity to the known viruses, which may indicate that they might interfere against uncharacterised exogenous viruses or that they had this function in the past (Forth et al., 2020). Intriguingly, ancient HHV6A/B integrations have been found closely linked to the *MOV10L1* locus (in the subtelomere of chromosome 22q) in East Asians, which is required to make piRNAs (Liu et al., 2020). Hypothetically, piRNAs from this locus could function in restricting further integrations by exogenous HHV6 (Liu et al., 2020).

Box 4. Virus-host horizontal gene transfers and evolutionary novelty

The exchanges of genetic material between hosts and viruses are the sources of significant evolutionary novelty. One striking case is that of dinoflagellates, whose chromatin is depleted in histones and which instead use DNVPs (dinoflagellate/viral nucleoproteins) as their primary DNA-binding protein (Gornik et al., 2012). The only homologue of DNVPs occurs in viruses of the family *Phycodnaviridae*, it appears these proteins were acquired by

the ancestors of dinoflagellates through a phycodnavirus infection, however the precise evolutionary source of these proteins remains unclear (Irwin et al., 2018). Interestingly, when expressed in the yeast *Saccharomyces cerevisiae*, DNVPs have a toxic effect by impairing transcription and disrupting nucleosomal chromatin, this effect is relieved by a reduction in the number of histones (Irwin et al., 2018). Thus, it might be the case that histone depletion was an adaptation to chromatin invasion by DNVPs (Irwin et al., 2018). Another remarkable case is that of the largest subunits of the DNA-dependent RNA polymerase of eukaryotes and NCLDVs (Guglielmini et al., 2019). Phylogenetic analyses of the homologous proteins, revealed that these polymerases were exchanged between protoeukaryotes and NCLDVs, and the viral polymerases gave rise to two of the three DNA-dependent RNA polymerases (polymerases I and II) used by modern eukaryotes (Guglielmini et al., 2019). NCLDVs are also prominent for their extensive capture of host gene homologues, which have been acquired independently by different lineages and seem to be a driving force for viral gigantism (Koonin and Yutin, 2018). DNA viruses have also been implicated in mediating horizontal gene transfer between eukaryotes. Nine different transposon superfamilies have been found integrated in the genome of the baculovirus *Autographa californica multiple nucleopolyhedrovirus* (AcMNPV) (Loiseau et al., 2020), and some present signs of horizontal transfer between multiple sympatric species of insects (Gilbert et al., 2014). Hitchhiking has also been reported for a *Maverick* integrated into a bracovirus in the genome of *Cotesia congregata*, raising the possibility of horizontal transfer of an endogenous virus mediated by bracoviruses between the wasp and lepidopteran host genomes (Dupuy et al., 2011). Furthermore, DNA viruses may be implicated as the main mechanism of eukaryote-to-eukaryote gene transfer, especially because there is extensive evidence for both host gene capture and cross-species transmission events occurring during viral evolution (Gilbert and Cordaux, 2017). A compelling example is that of C-type lectins

from hymenopterans which seem to have been transferred on at least two occasions to their lepidopteran hosts via infecting bracoviruses (Gasmi et al., 2015). The existence of these integrations as polymorphisms in virus populations will require the use of ultra-deep sequencing to uncover the magnitude of host gene capture by viruses.

References

Aimola G, Beythien G, Aswad A, Kaufer BB. 2020. Current understanding of human herpesvirus 6 (HHV-6) chromosomal integration. *Antiviral Research*. doi:10.1016/j.antiviral.2020.104720

Aswad A, Aimola G, Wight D, Roychoudhury P, Zimmermann C, Hill J, Lassner D, Xie H, Huang M-L, Parrish NF, Schultheiss H-P, Venturini C, Lager S, Smith GCS, Charnock-Jones DS, Breuer J, Greninger AL, Kaufer BB. 2020. Evolutionary history of endogenous Human Herpesvirus 6 reflects human migration out of Africa. *Molecular Biology and Evolution*. doi:10.1093/molbev/msaa190

Aswad A, Katzourakis A. 2017. A novel viral lineage distantly related to herpesviruses discovered within fish genome sequence data. *Virus Evolution* 3. doi:10.1093/ve/vex016

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. doi:10.1371/journal.pgen.1004332

Baker ML, Jiang W, Rixon FJ, Chiu W. 2005. Common ancestry of herpesviruses and tailed DNA bacteriophages. *Journal of Virology* 79. doi:10.1128/jvi.79.23.14967-14970.2005

Barreat JGN, Katzourakis A. 2021a. Phylogenomics of the *Maverick* virus-like mobile genetic elements of vertebrates. *Molecular Biology and Evolution*. doi:10.1093/molbev/msaa291

Barreat JGN, Katzourakis A. 2021b. Evidence for a billion-years arms race between viruses, virophages and eukaryotes. *bioRxiv*. doi:10.1101/2021.04.20.440574

Bejarano ER, Khashoggi A, Witty M, Lichtenstein C. 1996. Integration of multiple repeats of geminiviral DNA into the nuclear genome of tobacco during evolution. *Proc Natl Acad Sci U S A* 93. doi:10.1073/pnas.93.2.759

Béliveau C, Cohen A, Stewart D, Periquet G, Djoumad A, Kuhn L, Stoltz D, Boyle B, Volkoff A-N, Herniou EA, Drezen J-M, Cusson M. 2015. Genomic and proteomic analyses indicate that banchine and campoplegine polydnaviruses have similar, if not identical, viral ancestors. *Journal of Virology* 89. doi:10.1128/jvi.01001-15

Berjón-Otero M, Koslová A, Fischer MG. 2019. The dual lifestyle of genome-integrating virophages in protists. *Ann N Y Acad Sci*. doi:10.1111/nyas.14118

Brandes N, Linial M. 2019. Giant viruses-big surprises. *Viruses*. doi:10.3390/v11050404

Burke GR, Walden KKO, Whitfield JB, Robertson HM, Strand MR. 2014. Widespread genome reorganization of an obligate virus mutualist. *PLoS Genetics* 10. doi:10.1371/journal.pgen.1004660

Campbell S, Aswad A, Katzourakis A. 2017. Disentangling the origins of virophages and *Polintons*. *Current Opinion in Virology*. doi:10.1016/j.coviro.2017.07.011

Charlesworth B. 2009. Fundamental concepts in genetics: effective population size and patterns of molecular evolution and variation. *Nature Reviews Genetics* 10:195–205. doi:10.1038/nrg2526

Cheng RL, Li XF, Zhang CX. 2020. Nudivirus remnants in the genomes of arthropods. *Genome Biology and Evolution* 12. doi:10.1093/gbe/evaa074

Chevignon G, Periquet G, Gyapay G, Vega-Czarny N, Musset K, Drezen JM, Huguet E. 2018. Cotesia congregata bracovirus circles encoding PTP and ankyrin genes integrate into the DNA of parasitized *Manduca sexta* hemocytes. *Journal of Virology* 92. doi:10.1128/jvi.00438-18

Cui J, Holmes EC. 2012. Evidence for an endogenous papillomavirus-like element in the platypus genome. *Journal of General Virology* 93:1362–1366. doi:10.1099/vir.0.041483-0

Cui J, Zhao W, Huang Z, Jarvis E, Gilbert M, Walker P, Holmes E, Zhang G. 2014. Low frequency of paleoviral infiltration across the avian phylogeny. *Genome Biology* 15:539. doi: 10.1186/S13059-014-0539-3

Daya S, Cortez N, Berns KI. 2009. Adeno-associated virus site-specific integration is mediated by proteins of the nonhomologous end-joining pathway. *Journal of Virology* 83. doi:10.1128/jvi.01040-09

de Araujo AC, Huguet E, Herniou EA, Drezen JM, Josse T. (2022). Transposable element repression using piRNAs, and its relevance to Endogenous Viral Elements (EVEs) and immunity in insects. *Current Opinion in Insect Science*, 100876. doi: 10.1016/j.cois.2022.100876

Dennis TPW, de Souza WM, Marsile-Medun S, Singer JB, Wilson SJ, Gifford RJ. 2019. The evolution, distribution and diversity of endogenous circoviral elements in vertebrate genomes. *Virus Research* 262:15–23. doi:10.1016/j.virusres.2018.03.014

Dill JA, Camus AC, Leary JH, di Giallonardo F, Holmes EC, Ng TFF. 2016. Distinct viral lineages from fish and amphibians reveal the complex evolutionary history of hepadnaviruses. *Journal of Virology* 90. doi:10.1128/jvi.00832-16

Diop SI, Geering ADW, Alfama-Depauw F, Loaec M, Teycheney PY, Maumus F. 2018. Tracheophyte genomes keep track of the deep evolution of the *Caulimoviridae*. *Scientific Reports* 8. doi:10.1038/s41598-017-16399-x

Doerfler W. 1968. The fate of the DNA of adenovirus type 12 in baby hamster kidney cells. *Proc Natl Acad Sci U S A* 60:2. doi: 10.1073/pnas.60.2.636

Drezen JM, Leobold M, Bézier A, Huguet E, Volkoff AN, Herniou EA. 2017. Endogenous viruses of parasitic wasps: variations on a common theme. *Current Opinion in Virology*. doi:10.1016/j.coviro.2017.07.002

Dupuy C, Periquet G, Serbielle C, Bézier A, Louis F, Drezen JM. 2011. Transfer of a chromosomal *Maverick* to endogenous bracovirus in a parasitoid wasp. *Genetica* 139. doi:10.1007/s10709-011-9569-x

Feng H, Shuda M, Chang Y, Moore PS. 2008. Clonal integration of a polyomavirus in human Merkel cell carcinoma. *Science* (1979) 319. doi:10.1126/science.1152586

Filloux D, Murrell S, Koohapitagtam M, Golden M, Julian C, Galzi S, Uzest M, Rodier-Goud M, D'Hont A, Vernerey MS, Wilkin P, Peterschmitt M, Winter S, Murrell B, Martin DP, Roumagnac P. 2015. The genomes of many yam species contain transcriptionally active endogenous geminiviral sequences that may be functionally expressed. *Virus Evolution* 1. doi:10.1093/ve/vev002

Fischer MG, Hackl T. 2016. Host genome integration and giant virus-induced reactivation of the virophage Mavirus. *Nature* 540:288–291. doi:10.1038/nature20593

Fischer MG, Suttle CA. 2011. A virophage at the origin of large DNA transposons. *Science* (1979) 231–234. doi: 10.1126/science.1199412

Forth JH, Forth LF, Lycett S, Bell-Sakyi L, Keil GM, Blome S, Calvignac-Spencer S, Wissgott A, Krause J, Höper D, Kampen H, Beer M. 2020. Identification of African swine fever virus-like elements in the soft tick genome provides insights into the virus' evolution. *BMC Biology* 18. doi:10.1186/s12915-020-00865-6

Gallot-Lavallée L, Blanc G. 2017. A glimpse of nucleo-cytoplasmic large DNA virus biodiversity through the eukaryotic genomics window. *Viruses* 9. doi:10.3390/v9010017

Gasmi L, Boulain H, Gauthier J, Hua-Van A, Musset K, Jakubowska AK, Aury JM, Volkoff AN, Huguet E, Herrero S, Drezen JM. 2015. Recurrent domestication by Lepidoptera of genes from their parasites mediated by bracoviruses. *PLoS Genetics* 11. doi:10.1371/journal.pgen.1005470

Gauthier J, Boulain H, van Vugt JJFA, Baudry L, Persyn E, Aury JM, Noel B, Bretaudeau A, Legeai F, Warris S, Chebbi MA, Dubreuil G, Duvic B, Kremer N, Gayral P, Musset K, Josse T, Bigot D, Bressac C, Moreau S, Periquet G, Harry M, Montagné N, Boulogne I, Sabeti-Azad M, Maïbèche M, Chertemps T, Hilliou F, Siaussat D, Amselem J, Luyten I, Capdevielle-Dulac C, Labadie K, Merlin BL, Barbe V, de Boer JG, Marbouty M, Cônsoli FL, Dupas S, Hua-Van A, le Goff G, Bézier A, Jacquin-Joly E, Whitfield JB, Vet LEM, Smid HM, Kaiser L, Koszul R, Huguet E, Herniou EA, Drezen JM. 2021. Chromosomal scale assembly of parasitic wasp genome reveals symbiotic virus colonization. *Communications Biology* 4. doi:10.1038/s42003-020-01623-8

Gauthier J, Drezen JM, Herniou EA. 2018. The recurrent domestication of viruses: major evolutionary transitions in parasitic wasps. Parasitology. doi:10.1017/S0031182017000725

Geering ADW, Maumus F, Copetti D, Choisne N, Zwickl DJ, Zytnicki M, McTaggart AR, Scalabrin S, Vezzulli S, Wing RA, Quesneville H, Teycheney PY. 2014. Endogenous florendoviruses are major components of plant genomes and hallmarks of virus evolution. *Nature Communications* 5. doi:10.1038/ncomms6269

Geoghegan JL, Duchêne S, Holmes EC. 2017. Comparative analysis estimates the relative frequencies of co-divergence and cross-species transmission within viral families. *PLoS Pathogens* 13. doi:10.1371/journal.ppat.1006215

Gilbert Clément, Chateigner A, Ernenwein L, Barbe V, Bézier A, Herniou EA, Cordaux R. 2014. Population genomics supports baculoviruses as vectors of horizontal transfer of insect transposons. *Nature Communications* 5. doi:10.1038/ncomms4348

Gilbert C, Cordaux R. 2017. Viruses as vectors of horizontal transfer of genetic material in eukaryotes. *Current Opinion in Virology*. doi:10.1016/j.coviro.2017.06.005

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

Gilbert C., Meik JM, Dashevsky D, Card DC, Castoe TA, Schaack S. 2014. Endogenous hepadnaviruses, bornaviruses and circoviruses in snakes. *Proceedings of the Royal Society B: Biological Sciences* 281. doi:10.1098/rspb.2014.1122

Gong Z, Han G-Z. 2018a. Euphyllophyte paleoviruses illuminate hidden diversity and macroevolutionary mode of *Caulimoviridae*. *Journal of Virology* 92. doi:10.1128/jvi.02043-17

Gong Z, Han GZ. 2018b. Insect retroelements provide novel insights into the origin of hepatitis B viruses. *Molecular Biology and Evolution* 35. doi:10.1093/molbev/msy129

Gong Z, Zhang Y, Han GZ. 2020. Molecular fossils reveal ancient associations of dsDNA viruses with several phyla of fungi. *Virus Evolution* 6. doi:10.1093/ve/veaa008

Gornik SG, Ford KL, Mulhern TD, Bacic A, McFadden GI, Waller RF. 2012. Loss of nucleosomal DNA condensation coincides with appearance of a novel nuclear protein in dinoflagellates. *Current Biology* 22:2303–2312. doi:10.1016/j.cub.2012.10.036

Guglielmini J, Woo AC, Krupovič M, Forterre P, Gaia M. 2019. Diversification of giant and large eukaryotic dsDNA viruses predated the origin of modern eukaryotes. *Proc Natl Acad Sci U S A*. doi:10.1073/pnas.1912006116

Henckaerts E, Dutheil N, Zeltner N, Kattman S, Kohlbrenner E, Ward P, Clément N, Rebollo P, Kennedy M, Keller GM, Linden RM. 2009. Site-specific integration of adeno-associated virus involves partial duplication of the target locus. *Proc Natl Acad Sci U S A* 106. doi:10.1073/pnas.0806821106

Hickman AB, Ronning DR, Kotin RM, Dyda F. 2002. Structural unity among viral origin binding proteins: crystal structure of the nuclease domain of adeno-associated virus Rep. *Molecular Cell* 10. doi:10.1016/S1097-2765(02)00592-0

Inoue Y, Kumagai M, Zhang X, Saga T, Wang D, Koga A, Takeda H. 2018. Fusion of *piggyBac*-like transposons and herpesviruses occurs frequently in teleosts. *Zoological Letters* 4. doi:10.1186/s40851-018-0089-8

Inoue Y, Saga T, Aikawa T, Kumagai M, Shimada A, Kawaguchi Y, Naruse K, Morishita S, Koga A, Takeda H. 2017. Complete fusion of a transposon and herpesvirus created the Teratorn mobile element in medaka fish. *Nature Communications* 8. doi:10.1038/s41467-017-00527-2

Irwin NAT, Martin BJE, Young BP, Browne MJG, Flaus A, Loewen CJR, Keeling PJ, Howe LJ. 2018. Viral proteins as a potential driver of histone depletion in dinoflagellates. *Nature Communications* 9. doi:10.1038/s41467-018-03993-4

Katzourakis A, Gifford RJ. 2010. Endogenous viral elements in animal genomes. *PLoS Genetics* 6. doi:10.1371/journal.pgen.1001191

Kazlauskas D, Varsani A, Koonin EV, Krupovič M. 2019. Multiple origins of prokaryotic and eukaryotic single-stranded DNA viruses from bacterial and archaeal plasmids. *Nature Communications* 10. doi:10.1038/s41467-019-11433-0

Kinsella CM, Bart A, Deijs M, Broekhuizen P, Kaczorowska J, Jebbink MF, van Gool T, Cotten M, van der Hoek L. 2020. *Entamoeba* and *Giardia* parasites implicated as hosts of CRESS viruses. *Nature Communications* 11. doi:10.1038/s41467-020-18474-w

Koonin EV. 2016. Viruses and mobile elements as drivers of evolutionary transitions. *Philosophical Transactions of the Royal Society B: Biological Sciences*. doi:10.1098/rstb.2015.0442

Koonin EV, Dolja VV, Krupovič M. 2015. Origins and evolution of viruses of eukaryotes: The ultimate modularity. *Virology*. doi:10.1016/j.virol.2015.02.039

Koonin EV, Krupovič M. 2018. The depths of virus exaptation. *Current Opinion in Virology*. doi:10.1016/j.coviro.2018.07.011

Koonin EV, Yutin N. 2018. Multiple evolutionary origins of giant viruses. *F1000Res*. doi:10.12688/f1000research.16248.1

Kravets I. 2018. Paget's disease of bone: diagnosis and treatment. *American Journal of Medicine*. doi:10.1016/j.amjmed.2018.04.028

Krupovič M, Bamford DH. 2008. Virus evolution: how far does the double β-barrel viral lineage extend? *Nature Reviews Microbiology* 6:941–948. doi:10.1038/nrmicro2033

Krupovič M, Bamford DH, Koonin EV. 2014. Conservation of major and minor jelly-roll capsid proteins in *Polinton* (*Maverick*) transposons suggests that they are bona fide viruses. *Biology Direct* 9:1–7. doi:10.1186/1745-6150-9-6

Krupovič M, Blomberg J, Coffin JM, Dasgupta I, Fan H, Geering AD, Gifford R, Harrach B, Hull R, Johnson W, Kreuze JF, Lindemann D, Llorens C, Lockhart B, Mayer J, Muller E, Olszewski NE, Pappu HR, Pooggin MM, Richert-Pöggeler KR, Sabanadzovic S, Sanfaçon H, Schoelz JE, Seal S, Stavolone L, Stoye JP, Teycheney P-Y, Tristem M, Koonin EV, Kuhn JH. 2018. Ortervirales: new virus order unifying five families of reverse-transcribing viruses. *Journal of Virology* 92. doi:10.1128/jvi.00515-18

Krupovič M, Koonin EV. 2017. Homologous capsid proteins testify to the common ancestry of retroviruses, caulimoviruses, pseudoviruses, and metaviruses. *Journal of Virology* 91. doi:10.1128/jvi.00210-17

Krupovič M, Koonin EV. 2015. *Polintons*: A hotbed of eukaryotic virus, transposon and plasmid evolution. *Nature Reviews Microbiology* 13:105–115. doi:10.1038/nrmicro3389

Krupovič M, Varsani A, Kazlauskas D, Breitbart M, Delwart E, Rosario K, Yutin N, Wolf YI, Harrach B, Zerbini FM, Dolja V v., Kuhn JH, Koonin EV. 2020. *Cressdnaviricota*: a virus phylum unifying seven families of rep-encoding viruses with single-stranded, circular DNA genomes. *Journal of Virology* 94. doi:10.1128/jvi.00582-20

Lauber C, Seitz S, Mattei S, Suh A, Beck J, Herstein J, Börold J, Salzburger W, Kaderali L, Briggs JAG, Bartenschlager R. 2017. Deciphering the origin and evolution of Hepatitis B viruses by means of a family of non-enveloped fish viruses. *Cell Host and Microbe* 22. doi:10.1016/j.chom.2017.07.019

Lauer KB, Borrow R, Blanchard TJ. 2017. Multivalent and multipathogen viral vector vaccines. *Clinical and Vaccine Immunology*. doi:10.1128/CVI.00298-16

Legendre M, Bartoli J, Shmakova L, Jeudy S, Labadie K, Adrait A, Lescot M, Poirot O, Bertaux L, Bruley C, Couté Y, Rivkina E, Abergel C, Claverie JM. 2014. Thirty-thousand-year-old distant relative of giant icosahedral DNA viruses with a pandoravirus morphology. *Proc Natl Acad Sci U S A* 111:4274–4279. doi:10.1073/pnas.1320670111

Legendre M, Doutre G, Poirot O, Lescot M, Arslan D, Seltzer V, Bertaux L, Bruley C, Claverie J. 2013. Pandoraviruses: amoeba viruses with genomes up to 2.5 Mb reaching that of parasitic eukaryotes. *Science* (1979) 341:281–286. doi:10.1126/science.1239181

Legendre M, Lartigue A, Bertaux L, Jeudy S, Bartoli J, Lescot M, Alempic JM, Ramus C, Bruley C, Labadie K, Shmakova L, Rivkina E, Couté Y, Abergel C, Claverie JM. 2015. In-depth study of *Mollivirus sibericum*, a new 30,000-yold giant virus infecting *Acanthamoeba*. *Proc Natl Acad Sci U S A* 112. doi:10.1073/pnas.1510795112

Liu Huiquan, Fu Y, Li B, Yu X, Xie J, Cheng J, Ghabrial SA, Li G, Yi X, Jiang D. 2011. Widespread horizontal gene transfer from circular single-stranded DNA viruses to eukaryotic genomes. *BMC Evolutionary Biology* 11. doi:10.1186/1471-2148-11-276

Liu H., Fu Y, Xie J, Cheng J, Ghabrial SA, Li G, Peng Y, Yi X, Jiang D. 2011. Widespread endogenization of densoviruses and parvoviruses in animal and human genomes. *Journal of Virology* 85. doi:10.1128/jvi.00828-11

Liu X, Kosugi S, Koide R, Kawamura Y, Ito J, Miura H, Matoba N, Matsuzaki M, Fujita M, Kamada AJ, Nakagawa H, Tamiya G, Matsuda K, Murakami Y, Kubo M, Aswad A, Sato K, Momozawa Y, Ohashi J, Terao C, Yoshikawa T, Parrish NF, Kamatani Y. 2020. Endogenization and excision of human herpesvirus 6 in human genomes. *PLoS Genetics* 16. doi:10.1371/JOURNAL.PGEN.1008915

Loiseau V, Herniou EA, Moreau Y, Lévêque N, Meignin C, Daeffler L, Federici B, Cordaux R, Gilbert C. 2020. Wide spectrum and high frequency of genomic structural variation, including transposable elements, in large double-stranded DNA viruses. *Virus Evolution* 6. doi:10.1093/ve/vez060

Lwoff A. 1960. Tumor viruses and the cancer problem: a summation of the conference. *Cancer Research* 20.

Lwoff A, Gutmann A. 1950. Recherches sur un *Bacillus megatherium* lysogène. *Ann Inst Pasteur* (Paris) 78.

Lynch M, Marinov GK. 2015. The bioenergetic costs of a gene. *Proc Natl Acad Sci U S A* 112:201514974. doi:10.1073/pnas.1514974112

Lytras S, Arriagada G, Gifford RJ. 2021. Ancient evolution of hepadnaviral paleoviruses and their impact on host genomes. *Virus Evolution* 7. doi:10.1093/ve/veab012

Maumus F, Blanc G. 2016. Study of gene trafficking between *Acanthamoeba* and giant viruses suggests an undiscovered family of amoeba-infecting viruses. *Genome Biology and Evolution* 8. doi:10.1093/gbe/evw260

Milone MC, O'Doherty U. 2018. Clinical use of lentiviral vectors. *Leukemia*. doi:10.1038/s41375-018-0106-0

Mühlemann B, Jones TC, de Barros Damgaard P, Allentoft ME, Shevnina I, Logvin A, Usmanova E, Panyushkina IP, Boldgiv B, Bazartseren T, Tashbaeva K, Merz V, Lau N, Smrčka V, Voyakin D, Kitov E, Epimakhov A, Pokutta D, Vicze M, Price TD, Moiseyev V, Hansen AJ, Orlando L, Rasmussen S, Sikora M, Vinner L, Osterhaus ADME, Smith DJ, Glebe D, Fouchier RAM, Drosten C, Sjögren KG, Kristiansen K, Willerslev E. 2018. Ancient hepatitis B viruses from the Bronze Age to the Medieval period. *Nature* 557. doi:10.1038/s41586-018-0097-z

Naso MF, Tomkowicz B, Perry WL, Strohl WR. 2017. Adeno-Associated Virus (AAV) as a vector for gene therapy. *BioDrugs*. doi:10.1007/s40259-017-0234-5

Opriessnig T, Karuppannan AK, Castro AMMG, Xiao CT. 2020. Porcine circoviruses: current status, knowledge gaps and challenges. *Virus Research*. doi:10.1016/j.virusres.2020.198044

Pénzes JJ, Souza WM de, Agbandje-Mckenna M, Gifford RJ. 2019. An ancient lineage of highly divergent parvoviruses infects both vertebrate and invertebrate hosts. *Viruses* 11. doi:10.3390/v11060525

Poinar G, Poinar R. 2005. Fossil evidence of insect pathogens. *Journal of Invertebrate Pathology* 89. doi:10.1016/j.jip.2005.05.007

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. doi:10.1016/j.gene.2006.08.008

Schulz F, Roux S, Paez-Espino D, Jungbluth S, Walsh DA, Denef VJ, McMahon KD, Konstantinidis KT, Eloe-Fadrosh EA, Kyrpides NC, Woyke T. 2020. Giant virus diversity and host interactions through global metagenomics. Nature 578. doi:10.1038/s41586-020-1957-x

Senkevich TG, Bruno D, Martens C, Porcella SF, Wolf YI, Moss B. 2015. Mapping vaccinia virus DNA replication origins at nucleotide level by deep sequencing. *Proc Natl Acad Sci U S A* 112. doi:10.1073/pnas.1514809112

Suh A, Weber CC, Kehlmaier C, Braun EL, Green RE, Fritz U, Ray DA, Ellegren H. 2014. Early Mesozoic coexistence of amniotes and *Hepadnaviridae*. *PLoS Genetics* 10. doi:10.1371/journal.pgen.1004559

Sutter D, Westphal M, Doerfler W. 1978. Patterns of integration of viral DNA sequences in the genomes of adenovirus type 12-transformed hamster cells. *Cell* 14. doi:10.1016/0092-8674(78)90243-X

Sweeney NP, Meng J, Patterson H, Morgan JE, McClure M. 2017. Delivery of large transgene cassettes by foamy virus vector. *Scientific Reports* 7. doi:10.1038/s41598-017-08312-3

Temin HM. 1964. The participation of DNA in *Rous sarcoma virus* production. *Virology* 23:486–494. doi:10.1016/0042-6822(64)90232-6

ter Horst AM, Nigg JC, Dekker FM, Falk BW. 2018. Endogenous viral elements are widespread in arthropod genomes and commonly give rise to PIWI-Interacting RNAs. *Journal of Virology* 93. doi:10.1128/jvi.02124-18

Thézé J, Bézier A, Periquet G, Drezen JM, Herniou EA. 2011. Paleozoic origin of insect large dsDNA viruses. *Proc Natl Acad Sci U S A* 108. doi:10.1073/pnas.1105580108

Thézé J, Leclercq S, Moumen B, Cordaux R, Gilbert C. 2014. Remarkable diversity of endogenous viruses in a crustacean genome. *Genome Biology and Evolution* 6. doi:10.1093/gbe/evu163

Volkoff AN, Cusson M. 2020. The unconventional viruses of ichneumonid parasitoid wasps. *Viruses*. doi:10.3390/v12101170

Wight DJ, Aimola G, Aswad A, Jill Lai C-Y, Bahamon C, Hong K, Hill JA, Kaufer BB. 2020. Unbiased optical mapping of telomere-integrated endogenous human herpesvirus 6. *Proc Natl Acad Sci U S A* 202011872. doi:10.1073/pnas.2011872117

Wigington CH, Sonderegger D, Brussaard CPD, Buchan A, Finke JF, Fuhrman JA, Lennon JT, Middelboe M, Suttle CA, Stock C, Wilson WH, Wommack KE, Wilhelm SW, Weitz JS. 2016. Re-examination of the relationship between marine virus and microbial cell abundances. *Nature Microbiology* 1. doi:10.1038/nmicrobiol.2015.24

Witzmann F, Claeson KM, Hampe O, Wieder F, Hilger A, Manke I, Niederhagen M, Rothschild BM, Asbach P. 2011. Paget disease of bone in a Jurassic dinosaur. *Current Biology*. doi:10.1016/j.cub.2011.08.006

Wollman E. L. 1953. Sur le determinisme génetique de la lysogénie. *Ann Inst Pasteur* (Paris) 84.

Yu AM, McVey M. 2010. Synthesis-dependent microhomology-mediated end joining accounts for multiple types of repair junctions. *Nucleic Acids Research* 38. doi:10.1093/nar/gkq379

Chapter II

Phylogenomics of the *Maverick* virus-like mobile genetic elements of vertebrates

Abstract

Mavericks are virus-like mobile genetic elements found in the genomes of eukaryotes. Although *Mavericks* encode capsid morphogenesis homologues, their viral particles have not been observed. Here, we provide new evidence supporting the viral nature of *Mavericks* and the potential existence of virions. To this end, we conducted a phylogenomic analysis of *Mavericks* in hundreds of vertebrate genomes, discovering 134 elements with an intact coding capacity in 17 host species. We reveal an extensive genomic fossil record in 143 species and date 3 groups of elements to the Late Cretaceous. Bayesian phylogenetic analysis using genomic fossil orthologues suggests that *Mavericks* have infected osteichthyans for ~419 million years. They have undergone frequent cross-species transmissions in cyprinid fish and all core genes are subject to strong purifying selection. We conclude that vertebrate *Mavericks* form an ancient lineage of aquatic dsDNA viruses which are probably still functional in some vertebrate lineages.

Introduction

Mavericks are mobile genetic elements found integrated in the genomes of most eukaryotes, with the notable exception of mammals and land plants (Kapitonov and Jurka, 2006; Pritham et al., 2007). They were identified initially from the presence of a distinct integrase that was thought to be of cellular origin (the "c-integrases") (Gao and Voytas, 2005). However, closer inspection of the surrounding genomic architecture revealed the existence of additional open reading frames (ORFs) and flanking inverted repeats, a telltale sign of DNA transposons (Feschotte and Pritham, 2005). Nevertheless, there is mounting evidence based on sequence and structural comparisons, that *Mavericks* share striking features and a close evolutionary relationship with viruses.

Indeed, phylogenetic and comparative genomic analyses have placed *Mavericks* in the PRD1-adenovirus lineage (kingdom *Bamfordvirae*), a diverse assemblage of mobile genetic elements that infect bacteria, archaea and eukaryotes (Krupovič and Bamford, 2008). *Mavericks* encode a family B protein-primed DNA polymerase (pPOLB), an adenoviral-like protease, retroviral-like integrase and DNA packaging ATPase (Pritham et al., 2007). In addition, two conserved ORFs were found to encode proteins homologous to the double and single jelly-roll capsid proteins of diverse dsDNA viruses (Krupovič et al., 2014). The evolutionary history of the PRD1-adenovirus lineage may represent the most remarkable diversification ever seen in the virosphere, both in terms of ecology and genome complexity. Both capsid-encoding and capsidless elements are found in this group, which comprises bacteriophages (tectiviruses, corticoviruses), archaeoviruses (turriviruses), the Nucleocytoplasmatic Large DNA viruses (NCLDVs, including the giant viruses),

virophages, mitochondrial and cytoplasmic linear plasmids, adenoviruses, *Tlr*-elements, Polinton-like viruses and *Mavericks* (Koonin and Krupovič, 2017).

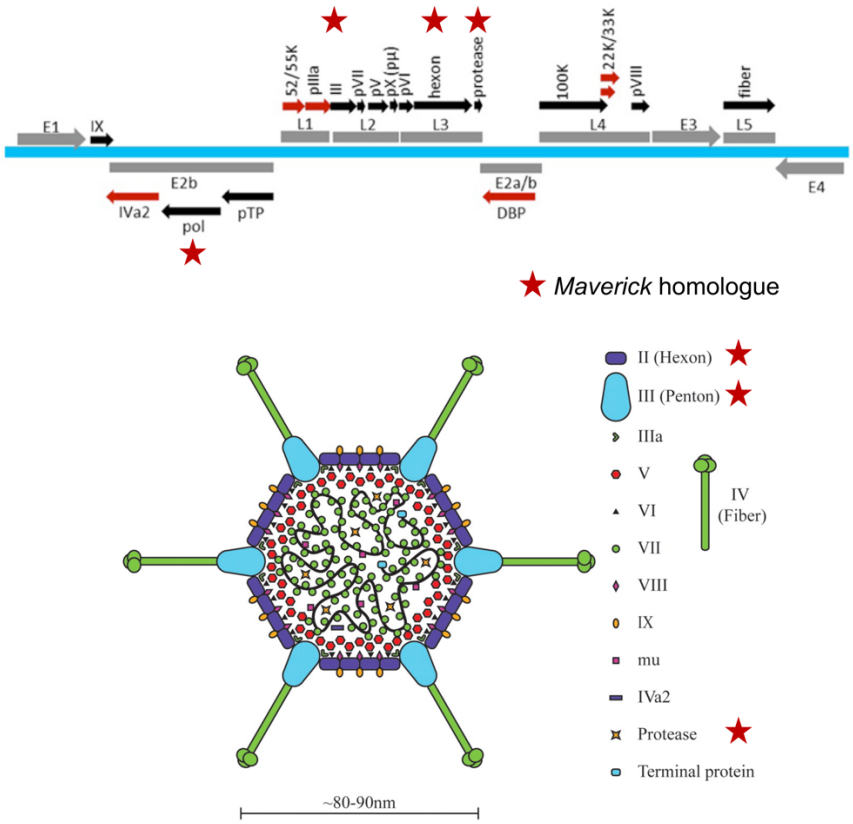


Figure 1. Genome organisation and diagram of the virion of Human adenovirus type 5 (HAdV-C5). The four gene homologues shared with *Mavericks* (hexon, penton, protease and DNA polymerase) are highlighted with stars. Modified with permission from Ahi and Mittal (2016), and Giberson et al. (2012).

Although viral particles for *Mavericks* have not been observed, the fact that they encode the full repertoire of genes required for capsid morphogenesis suggests they are viruses (Figure 1) (Krupovič et al., 2014). This is consistent with a scenario where these elements have retained the ancestral capsid-encoding capacity of the PRD1-adenovirus lineage, while the capsidless forms have originated on several occasions via reductive evolution. Precisely, the defining feature of viruses is the ability to form a protective protein shell used in the

horizontal transfer of parasitic genetic replicons. It is this feature which sets them apart from other mobile genetic elements such as viroids, transposons and plasmids. Therefore, pinpointing the precise nature of *Mavericks*, whether they are viruses or transposons, is a fundamental question with important implications in understanding the evolutionary biology of the PRD1-adenovirus lineage.

The current classification of *Mavericks* groups them into two clades, group-I and group-II, which are paraphyletic passed on the pPOLB marker (Krupovič and Koonin, 2015). There is indication that horizontal transfers have occurred during their evolutionary history, for example, both group-I and group-II elements coexist in the genomes of *Nasonia vitripennis* (Hymenoptera), *Tribolium castaneum* (Coleoptera) and *Nematostella vectensis* (Cnidaria) (Haapa-Paananen et al., 2014). Similarly, the *Mavericks* in two species of drosophilid flies, *Drosophila bipectinata* and *D. eugracilis*, were shown to descend from different lineages (Haapa-Paananen et al., 2014). In vertebrates, the elements that have been analysed seem to have been inherited vertically within their hosts (Haapa-Paananen et al., 2014). As such, it seems that *Mavericks* may undergo evolutionary dynamics similar to retroviruses, combining vertical inheritance in the host-germline with cross-species transmissions. Given that hundreds of vertebrate assemblies are now available for genome mining, vertebrates represent a suitable group of animals to test these ideas on the evolutionary biology of *Mavericks*.

Here we investigate the viral nature of vertebrate *Mavericks* by an extensive, in-depth phylogenomic analysis. This approach allows us to gain insights into the genetic diversity, abundance, taxonomic distribution and the evolutionary history of *Mavericks* in vertebrates.

Materials and Methods

Genome mining

We selected the genomes of 533 vertebrates available in the NCBI RefSeq and WGS databases (NCBI Resource Coordinators, 2016) as of December 2018, in addition to the assembly of the Iberian ribbed newt *Pleurodeles waltl* (Elewa et al., 2017), which at the time was not available in the public repositories (this genome sequence was kindly provided to us by A. Elewa). Each genome was screened individually in the NCBI-BLAST server with tblastn (Altschul et al., 1990; Johnson et al., 2008) using the integrase (INT) and pPOLB sequences of the Danio rerio *Polinton*-1 from RepBase (Bao et al., 2015) as queries; we used local BLAST version 2.8.0+ to screen *P. waltl* (Camacho et al., 2009). A set of candidate regions for downstream analysis was recovered by identifying hits to INT and pPOLB that co-localised to the same contig and were separated by 4–40 kb. The DNA sequences of these regions were downloaded with 20 kb flanks added to each end, giving 3,913 candidates in total.

To map the elements within the candidate regions, we blasted each sequence against itself (blastn 2.8.0+) and analysed the hits producing alignments of inverted repeats, which were identified as alignment pairs with hit-coordinates in reverse orientations. From 2,863 sequences with inverted repeats, 628 had the 5'–$(AG)_3$//$(CT)_3$–3' and 5'–$(AGT)_2$//$(ACT)_2$–3' end-motifs previously reported for vertebrate *Mavericks* (Kapitonov and Jurka, 2006), as well as perfect 6 bp target site duplications (TSDs). These sequences were partitioned according to species/end-motif and aligned using MAFFT version 7.407 (Katoh et al., 2002). We used the alignments to extract the first 200 bp of the 5' terminal inverted repeat and

build nucleotide Hidden Markov Models (HMMs) with hmmbuild version 3.2.1 (Eddy, 2009, 1995). Next, we used the models for each species/end-motif to find additional inverted repeats and TSDs in the remaining candidates within that species using nhmmer version 3.2.1. New elements were aligned to the initial species/end-motif partitions (MAFFT) and refined HMMs were obtained, these were used in turn as heterologous probes to annotate TIRs between species. A final set of alignments and HMMs was then used to screen all the remaining sequences. This strategy allowed us to map the locations of 1,345 *Mavericks* with recognisable TIRs (which include both elements with perfect and imperfect repeats).

Gene prediction

Genes in the candidate sequences were modelled using Genewise version 2.4.1 (Birney et al., 2004) and protein HMMs of the eight core genes of vertebrate *Mavericks* (*int*, *ppolb*, *atp*, *pro*, *major cp*, *minor cp*, *pz* and *pw*). Specifically, we compiled all the predictions from the RepBase consensus sequences (Bao et al., 2015), clustered them using Mmseqs2 (Steinegger and Söding, 2017) and assigned them to homologous protein sets using HHpred (Söding et al., 2005; Zimmermann et al., 2018). These groups were used to query the non-redundant protein database in blastp searches (restricted to "vertebrates" and evalue < 1e-100, this very conservative evalue was chosen to ensure that only homologous proteins would be included in the reference alignments used for gene modelling). Results were combined and used to search the non-redundant nucleotide database with the command line tblastn. Protein sequences were aligned with MAFFT and HMMs were built in HMMer2 format (hmmbuild version 2.3.2) for compatibility with Genewise.

We removed the host DNA from the *Maverick* candidates using the genomic coordinates of the inverted repeats found in the previous step. These sequences were masked with RepeatMasker version 4.0.8 (Smit et al., 2015) using the Dfam 2.0 (blast engine) (Hubley et al., 2016) and RepBase (release 20181026, hmmer engine) repeat libraries with *Mavericks*/*Polintons* removed. Although it was possible to mask secondary transposon insertions in several species, this was not possible for all since they are not represented in the databases. Further, comparisons in RepeatMasker are made at the nucleotide level which would produce conservative maskings.

The protein HMMs were used in Genewise (genewisedb mode) to model the core genes of the masked sequences. In addition, regions which showed positive hits to INT and pPOLB but lacked inverted repeats were also included in gene modelling, as these could represent defective copies with relevance to reconstructing the evolutionary history of *Mavericks*.

Protein Multiple Sequence Alignments

Protein predictions were extracted from the Genewise output files and if fragmented, concatenated to other sequence fragments belonging to the same prediction. We assembled multifasta files with unique predictions for each homologous protein. All sets proved difficult to align given that many sequences were not masked, some presumably had indel mutations and there were possibly false positive fragments in these predictions (especially short stretches to the N-terminal end, separated by long 'introns' to the main prediction). We used 6 different approaches to choose appropriate alignments for phylogenetic analysis.

Initial sets were aligned using MAFFT, ClustalO (2 iterations) (Sievers et al., 2011), PASTA (Mirarab et al., 2015) and FAMSA (Deorowicz et al., 2016). Columns with ≥50% gaps in the resulting alignments were removed. We noticed that trimAl (Capella-Gutiérrez et al., 2009) was not able to recover conserved positions from either the ClustalO or FAMSA alignments, so these were discarded. It was possible to trim both the MAFFT and PASTA alignments, which we did using the automated1 heuristic; this option chooses between the three modes of automatic parameter selection depending on alignment characteristics (Capella-Gutiérrez et al., 2009). The other 2 approaches consisted of "refinements" of the MAFFT alignments since these were seen to give the best summary statistics (explained below).

First, we aligned sequences with either default MAFFT or MAFFT with the –leavegappy flag (which introduces fewer gaps in gap-rich regions). We then removed columns with ≥99% gaps, unaligned the sequences and realigned them using the previous settings. Next, we removed columns with ≥50% gaps as before. At this stage, a subset of the main alignment with sequences having ≤10% gaps were trimmed in trimAl (automated1). This subset alignment was used as a scaffold to add the removed sequences with the –add and –keeplength functions of MAFFT, thus giving a merged final alignment with all sequences.

As criteria to choose the final alignments, we considered basic statistics from the SQUID package (Eddy, 2002): alignment length, mean, maximum and minimum sequence length, residue content, average identity, most related/unrelated pair and most distant sequence. Similarly, we computed the mean Shannon entropy (ignoring gaps and missing symbols) (Shannon, 1948) for each alignment as well as the mean observed, Jukes and Cantor, Poisson and Gamma distances in MATLAB (The Math Works Inc, 2019) for all pairwise

comparisons. We chose the alignments that approached the expected protein length, had identifiable conserved motifs reported in the literature (Gao and Voytas, 2005; Kapitonov and Jurka, 2006; Krupovič et al., 2014; Krupovič and Bamford, 2008; Pritham et al., 2007; Zhou et al., 2010), the highest identities and residue content as well as the lowest entropy and pairwise distances. The best alignment approach for most protein sets was MAFFT followed by trimming, whereas in the case of INT the best approach was MAFFT with realignment and trimming.

Orthology

To assess orthology of the predictions we extracted 2 kb of DNA from the flanks of each element. We then used blastn to query the database of all flanks and recovered loci which showed hits to both flanks. Candidates were downloaded in full, such that they contained a continuous sequence including both flanks and the elements, and aligned them using the progressive algorithm in Mauve version 2015-02-25 (Darling et al., 2004). In alignments involving elements without TIRs, the limits between host and *Maverick* DNA are not clear, so we walked through these loci by adding 10 kb to each flank and aligning once again in Mauve. The alignments were visualised and only those which showed consistent high similarity as well as colinear blocks were considered to be orthologues.

cDNA Multiple Sequence Alignments

We used the initial cDNA predictions recovered from Genewise to derive codon alignments for each gene. First, updated HMMs were built for each protein from the final alignments. These models were used to obtain the homologous coding subsequence from the initial

cDNA predictions. Stretches of modelled codons were seen to be either contiguous or separated by short <100 bp "introns" (possibly insertions not included in the model). Sequences were pooled and aligned using TranslatorX (Abascal et al., 2010), which aligns coding DNAs by translating them and using the aminoacid alignment as a guide for the codons, such that all the sequences are in the same reading frame. Sequences were edited manually to remove insertions/non-homologous regions.

Phylogenetic analysis

The initial alignments contained thousands of sequences and probably had some redundancy from many sequences that formed well-supported clades within the same host genomes, for these reasons, we implemented an approach to choose a phylogenetically-informative subset in order to reduce the computational expense of downstream analyses. We used Fasttree version 2.1.11 (Price et al., 2010) with the JTT+CAT model and Shimodaira-Hasegawa support to approximate a maximum likelihood phylogeny for each protein. We selected subsets of sequences ensuring that protein predictions from all genomes were represented, as well as all the orthologues and elements with apparently intact genes. In particular, we chose subsets from groups that were monophyletic with respect to the same host genome and support values that were > 0.70, also making sure that each clade was represented by at least 3 sequences. In cases such as in salmonids, chelonians and crocodilians, where the sequences did not separate into well-defined clusters by species (but were nonetheless monophyletic as a whole), we tried to obtain a representative sample from the topology although which specific sequence was selected was rather subjective. We then performed another round of sequence selection using maximum likelihood trees. Before running these analyses, we used ModelTest-NG version 0.1.5 (Darriba et al., 2020) to select the best fitting

models of protein evolution. Both the minimum AIC and BIC criteria agreed on the JTT+G model for all proteins, with estimated frequencies (+F) for the ATP, PRO and pPOLB alignments. Analyses were run on IQtree version 1.6.11 (Nguyen et al., 2015) with the selected models and 1000 ultrafast bootstrap replicates. From these alignments a final subset was chosen with the same criteria outlined above.

Finally, we used BEAST2 version 2.5.2 (Bouckaert et al., 2019) to perform Bayesian inference on the tree topology, divergence times and rates of evolution for each of the final alignments. We used age calibrations for each internal node of *Maverick* orthologues which correspond to single integration event in the genome of the most recent common ancestor of the species, and thus reflect a conservative minimum age of the integrations (Katzourakis and Gifford, 2010). For each group of *Maverick* orthologues, we calibrated the prior age distribution of their most recent common ancestor with the age of the most recent common ancestor inferred for hosts in the literature. We set the prior mean equal to the reported mean, adjusting the deviation such that the lognormal distribution contained the lower and upper bounds of the 95% confidence interval. The points and references used for all the prior calibrations are provided in the Supplementary Excel File 2 (*.zip file). The best models of protein evolution were selected as before. Analyses were started with the maximum likelihood topology estimated in RAxML-NG (Kozlov et al., 2019) and run in parallel with logs made every 5,000 generations. As the choice for the molecular clock, we used an uncorrelated relaxed clock (Drummond et al., 2006) using the lognormal distribution, which allows each branch in the tree to have its own evolutionary rate. We did this to relax the rate assumption since *Mavericks* could have complex evolutionary dynamics alternating between endogenous and exogenous modes. Runs were combined in LogCombiner using a 1-25% burn-in and inspected in Tracer (Rambaut et al., 2018) ensuring that estimates had

Effective Sample Sizes (ESS) ~200 or greater, good mixing and convergence (MCMC chain lengths were between 345 million and 4,241 million generations). Final tree files were combined by resampling at a lower frequency in LogCombiner and summarized as Maximum Credibility trees in TreeAnnotator. The resulting time trees were visualised in Figtree version 1.4.4 (Rambaut, 2018). To confirm the robustness of our results, we computed an independent Bayesian tree in MrBayes 3.2.7a (Ronquist and Huelsenbeck, 2003), using the DNA polymerase of intact *Mavericks* exclusively (MCMC length: 2,000,000 generations, model: LG+I+G4).

Cophylogenetic analysis

In order to test for possible host switches and the selective regimes affecting vertebrate *Mavericks*, we chose to focus on the elements from cyprinid fish (Cypriniformes: Cyprinidae). The reasons for this are that multiple cyprinids harbour *Mavericks*, some elements have an intact coding capacity, they have attained considerable copy numbers (e.g. in *Oxygymnocypris stewartii*) and their phylogeny was well supported but inconsistent with that of their hosts. The tree topology for each gene was analysed separately and the host tree was built by combining results from the relevant literature (Wang et al., 2012; Yang et al., 2016, 2015). We adopted a simulation-based approach using approximate Bayesian computation (ABC) as implemented in Coala 1.2.1 (Baudet et al., 2015) to quantitatively assess the importance of four classes of cophylogenetic events: cospeciation (simultaneous speciation of parasites and host), duplication (speciation of parasites without host speciation), host switch (parasite jumps from one host species to another) and loss (of a parasite from a host lineage after speciation). Analyses were run for 5 rounds to attain convergence (without overfitting) and using tolerance thresholds of 0.1 in the ABC rejection

algorithm. The mean probability of each event class was then used to calculate associated costs ($cost_i = -\log_e(p_i)$), for the computation of the Most Parsimonious Reconciliation (MPR) tree in Mowgli (Doyon et al., 2010). This reconciled tree was visualised in Sylvx (Chevenet et al., 2016).

Selection analysis

As a measure of the selective regimes influencing the evolution of these sequences (that is, natural selection acting on the level of the elements as opposed to the level of the host organisms), we estimated the non-synonymous to synonymous substitution ratios (ω) under the maximum-likelihood framework of the CODEML package from PAML 4 (Yang, 2007). We used unrooted subtrees of the protein Bayesian phylogenies for cyprinids as the *Maverick* topologies and the codon-aligned cDNA data as described above. Several models were tested: a single-ratio model with ω = 1 (the neutral expectation), a single-ratio model with ω = estimated, and a two-ratio model for internal and terminal branches ($\omega_{internal}$ = estimated, $\omega_{terminal}$ = estimated). Our alternative hypothesis is that episodes of diversifying selection acting on the viruses may have occurred coinciding with host switches, thus, we used the tree reconciliation computed in Jane (Conow et al., 2010) to label the branches involved in host switches in a two-ratio model (ω_{switch} = estimated, $\omega_{no\text{-}switch}$ = estimated). A likelihood-ratio test was conducted for each pair of nested models, and the χ^2 test statistic compared to its critical value (under the difference of the degrees of freedom) to assess the significance of these comparisons. Non-nested models were compared using the Akaike information criterion.

Data and code availability

The final datasets generated during this study are available in figshare: protein alignments (https://figshare.com/s/44ab548e4a1e855c8124), profile Hidden Markov Models of proteins (https://figshare.com/s/2c0964ae9d819fb5597d), rooted maximum credibility time trees (https://figshare.com/s/383384810dc4d086632a), codon alignments and unrooted trees for cyprinid *Mavericks* (https://figshare.com/s/e8430094ab33dfab69ec). Supplementary code files can be accessed through GitHub (https://github.com/josegabrielnb/Mavericks).

Results

Mavericks occur in all major Osteichthyian lineages except for mammals

We found evidence of *Mavericks* in the genomes of ray-finned fish (Actinopterygii), coelacanths (Coelacanthi), amphibians (Lissamphibia), lepidosaurs (Lepidosauria), turtles (Testudines), crocodiles (Crocodilia) and birds (Aves). No hits mapped to the genomes of mammals, except for the unplaced genomic scaffold NW_019367942.1 in the cat (*Felis catus*) assembly, which appears to be a contaminant from fish. This sequence shows consistent similarity to fish *Mavericks* in Censor (Jurka et al., 1996) and contains a Harbinger DNA transposon which are not known in mammals (Kapitonov and Jurka, 2006) Thus, we were able to map a total of 3,511 loci homologous to *Mavericks* in 143 species belonging to all major groups of Osteichthyes excluding mammals (List S1, Supplementary Excel File 1 in *.zip file).

The vast majority elements are mutationally degraded

The majority of these loci (96%) represent sequences that have become eroded by host mutation and have lost their coding capacity. This could be seen in the Genewise annotations where stop codons, frameshift mutations and large insertions (which Genewise detects as introns), were identified as well as in the lack of discernible TIRs in 2,166 elements. We found a total of 134 elements in 17 species which encode 8 conserved and intronless ORFs (or 7 in *Xenopus tropicalis*), terminal inverted repeats and 6 bp target site duplications. Therefore, the latter seem to be potentially active *Mavericks*, which we call 'intact', while the former are probably genomic fossils.

The genomic architecture of vertebrate *Mavericks* is conserved

The genomic architectures of the intact *Mavericks* of vertebrates are generally conserved (Figure 2). The ORFs are arranged in two modules: a first module containing the genes coding for the protein-primed DNA polymerase, ATPase and the PZ protein, and a second module coding for the integrase, minor capsid protein, the PW protein, major capsid protein and protease. We discovered an additional conserved ORF of uncertain function which lies between *atp* and *pz*: ORF *pm*. The two gene modules occupy different strands and both are oriented towards the centre of the genome. Most elements have two accessory regions flanking the core modules where more ORFs encoding diverse proteins can be found. Interestingly, the modules in the elements of the cane toad *Rhinella marina* point away from each other, while in the western clawed frog *Xenopus tropicalis* they occupy the same strand; accessory regions appear to be absent in these *Mavericks*. Sizes of the intact elements ranged from 12,723 to 24,620 bp with a mean and standard deviation of 16,862 ± 1,678 bp; the sizes

of TIRs also varied, they comprised between 46 and 1447 bp with a mean and standard deviation of 402 ± 276 bp (Figure S1).

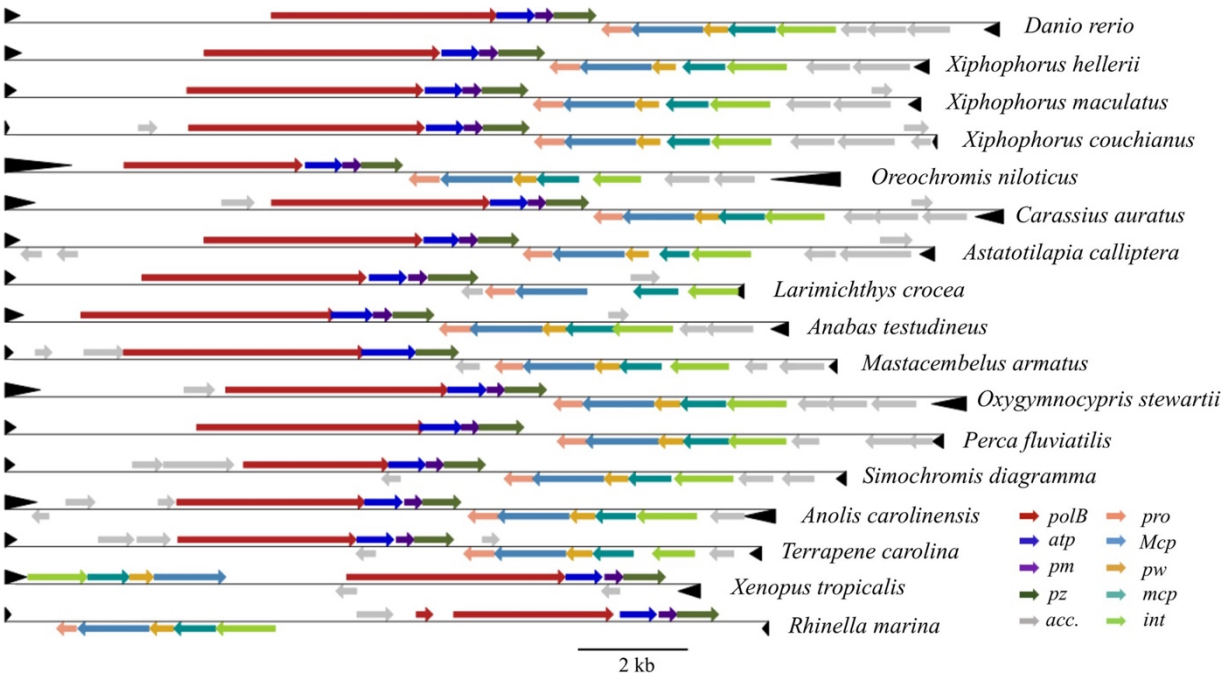


Figure 2. The genomic architectures of vertebrate *Mavericks* are conserved. Elements from teleosts are shown above and those from tetrapods below. Coloured arrows indicate open reading frames (ORFs), black arrowheads indicate TIRs. See text for details. *ppolb*: protein-primed DNA polymerase, *atp*: ATPase, *pm*: PM, *pz*: PZ, *pro*: protease, *Mcp*: major capsid protein, *pw*: PW, *mcp*: minor capsid protein, *int*: integrase, acc.: accessory genes. Scale bar = 2,000 base pairs.

The genomes of vertebrate *Mavericks* have conserved 6 bp end-motifs at the termini. The most common motif is an AGT trinucleotide repeated twice at the 5' end which occurs together with an ACT trinucleotide repeated twice at the 3' end; or 5'-$(AGT)_2$//$(ACT)_2$-3', where '//' is the *Maverick* sequence internal to the termini. Another common motif is an AG dinucleotide repeated three times with a downstream CT repeat, 5'-$(AG)_3$//$(CT)_3$-3'. A new motif type was found in *Mavericks* of the Asian swamp eel, *Monopterus albus*, which consists of dinucleotides followed by a single conserved base of the form 5'-$(AC)_3$A//T$(GT)_3$-3'. In all types, these conserved positions are characterised by an

information content > 1 bit which decreases markedly as one moves away from the motifs into the DNA of the host, and to a lesser degree inwards to the internal sequence (the bit content of a site is calculated from the formula: information content = $2 + \sum p_i \cdot \log_2(p_i)$, where p_i is the relative frequency of nucleotide i at that position; Figure S2).

Mavericks mostly attain low copy numbers, but have amplified enormously in a few genomes

In general, *Mavericks* attain low copy numbers in the genomes of their vertebrate hosts, with an average of 22 and a median of 8 elements (Figure 3). In some cases, there can be considerably more as in the genomes of the axolotl *Ambystoma mexicanum* (=287), the tuatara *Sphenodon punctatus* (=280) or the pike *Esox lucius* (=172). However, all the elements in these hosts seem to be defective since the genes for the protein-primed DNA polymerase are fragmented by stop codons/frameshift mutations. In other genomes both genomic fossils and intact elements can be found, which total 135 in the Asian swamp eel *Mastacembelus armatus* (11 intact copies), 123 in the Tibetan cyprinid *Oxygymnocypris stewartii* (38 intact copies), 105 in the Nile tilapia *Oreochromis niloticus* (5 intact copies) and 101 in the eastern box turtle *Terrapene carolina* (1 intact copy). Copy numbers of intact elements range from 1 in some genomes of cichlids, turtles and amphibians to a maximum of 38 in *Oxygymnocypris stewartii* (Table S1). Of a total of 134 elements, 129 were found in the genomes of teleosts while only 5 were identified in tetrapods. Moreover, the intact *Mavericks* of teleosts are considerably more abundant (and variable) with a mean and standard deviation of 9.92 ± 10.11 copies per genome, compared to 1.25 ± 0.50 copies per genome in tetrapods.

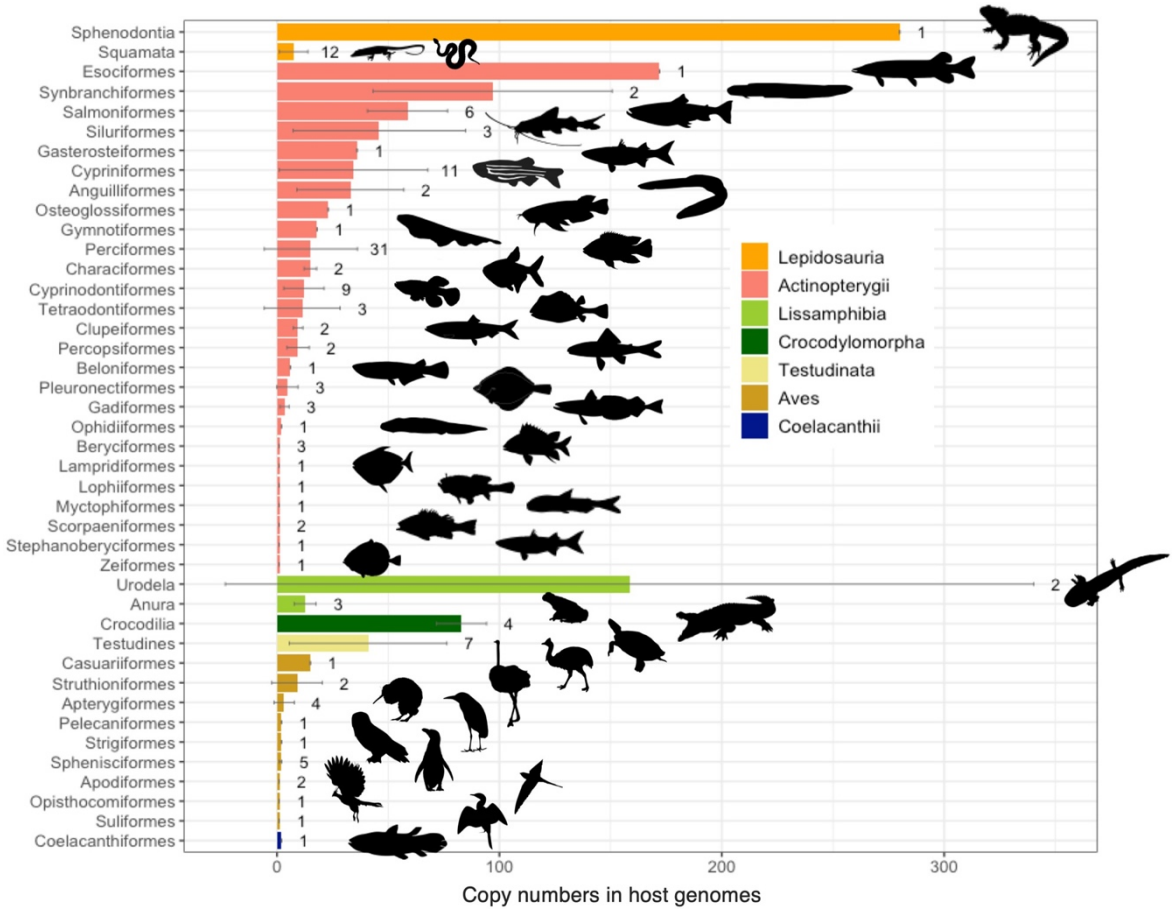


Figure 3. Copy numbers of *Mavericks* are low in most taxa. Major clades of vertebrates are colour coded in the legend. Each bar shows the mean and standard deviation, as well as the number of observations (analysed genomes) in that Order, to the right. N = 143. Silhouettes obtained from PhyloPic. For simplicity, the orders where *Mavericks* were not found are not shown in the figure. The absence of *Mavericks* may also result from lower sampling of the group or very divergent sequences, our analysis does not show definitive absence.

Orthologous loci suggest minimum insertion dates spanning 3-95 Mya

Pairwise comparisons among all loci allowed us to identify 115 sequences in 54 groups of orthologues (Supplementary Excel File 2 in *.zip file). Orthologous loci were found in the genomes of crocodiles, paleognath birds, penguins, testudinoid turtles, perciform, anguilliform, cyprinid and salmonid fish (Figure S3). Some taxa have multiple orthologue groups, which indicates that independent integration events had already occurred in the

genome of their most recent common ancestor. Three of our calibrations date back to the Late Cretaceous (~80–95 Mya) while all others are from the Cenozoic Era (~3–54 Mya). None of the orthologue groups contained intact *Mavericks*, they were all genomic fossils from defective elements. It is important to note that these ages reflect minimum conservative estimates of the age of the insertions, since they could have occurred at a point prior to the divergence of the host species (Katzourakis and Gifford, 2010), although it is unlikely they are too far away in the past.

Mavericks have infected vertebrates for hundreds of millions of years

The inferred time scale of evolution of vertebrate *Mavericks* was in the order of hundreds of millions of years. Estimated root ages vary between a mean of 454.73 and 231.86 Mya, with 7 of the 8 proteins suggesting a Palaeozoic origin for vertebrate *Mavericks*; the youngest estimate from the 95% probability densities (HPDs) was ~177 Mya for the major capsid protein (Jurassic Period of the Mesozoic Era) and the oldest, ~569 Mya for the PZ protein (Ediacaran Period of the Proterozoic Eon) (Table S2). Most evolutionary rates are in the order of $\sim10^{-9}$ amino acid substitutions per site per year, with minimum rates on the order of $\sim10^{-10}$ and the highest around $\sim10^{-7}$ (Table S2). A limitation of our approach is that the fossil orthologue calibrations implicitly assume that *Mavericks* have been evolving mostly at the host neutral rate, since the branches after the calibration points largely represent host evolution. This would translate into an overestimation of the divergence times if the rates at which the sequences had been evolving were actually higher (for example, in the case of exogenous ancestors). However, considering that *Mavericks* appear to persist mostly as endogenous copies in the genomes of their hosts, rates of this order of magnitude have been estimated using different approaches for other families of viruses (see Discussion) and that

the deep topology of the pPOLB tree is consistent with several codivergence scenarios in the order of hundreds of millions of years (Figure 4), it seems that our evolutionary rate and divergence time estimates are good approximations of reality.

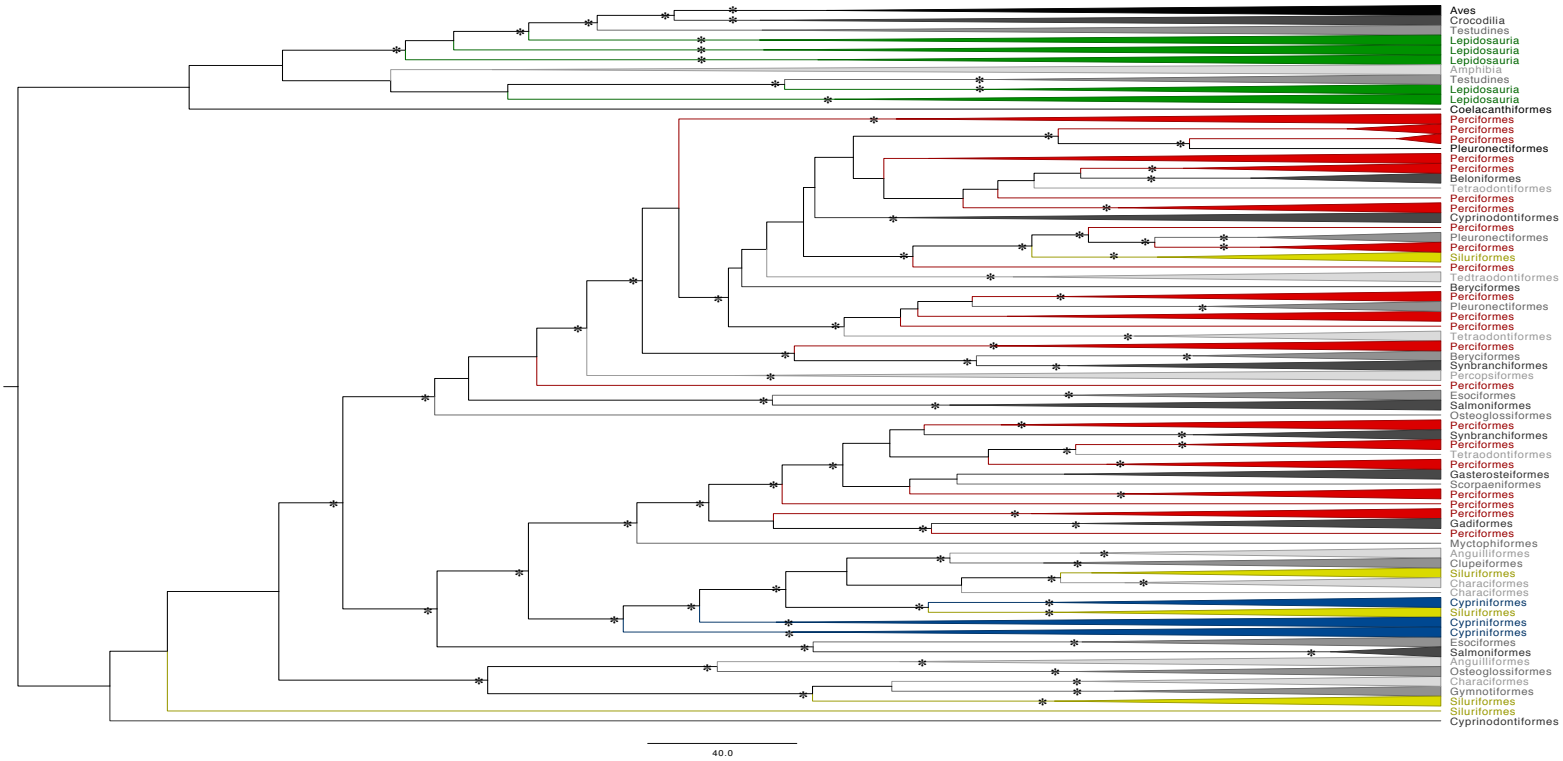


Figure 4. Maximum credibility tree of vertebrate *Mavericks* based on the protein-primed DNA polymerase B, asterisks indicate a posterior probability > 0.95. Scale bar = 40.0 My.

The phylogeny of *Mavericks* suggests vertical transmission as well as horizontal transfers

The tree from the pPOLB protein had the best overall support, with an average posterior probability of clades of 0.89. In this maximum credibility tree, the *Mavericks* from sarcopterygians and actinopterygians each form monophyletic groups (Figure 4). *Mavericks* of tetrapods are the sister group to the coelacanth's (posterior probability = 0.64), and the

recently proposed Archelosauria was also recovered, i.e. (Testudines, (Crocodilia, Aves)) with a posterior probability = 1. Similarly, we observe a sister relationship between Salmoniformes and Esociformes (pikes). These observations agree with the host topology and are consistent with the idea that some *Maverick* lineages have been inherited vertically throughout osteichthyan evolution; although we cannot rule out possible horizontal transfers with unsampled intermediates.

However, there are also apparent discordances with the host topology, most notably, the lepidosaur, perciform and siluriform *Mavericks* are polyphyletic (Figure 4), which also occurs for the elements in chelonians and the fish orders: Anguilliformes (true eels), Characiformes (piranhas and tetras), Pleuronectiformes (flatfish), Salmoniformes + Esociformes (two clades), Synbranchiformes (swamp eels) and Tetraodontiformes (ocean sunfish, pufferfish). This is also consistent with the paraphyly of *Mavericks* in the Cypriniformes (minnows and carps). In general, these patterns hold for the trees inferred independently from the other 7 proteins, especially the monophyly of tetrapod *Mavericks* (Additional Tree Files 1-8 in Figshare). The most notable difference being that, in the other trees, the *Mavericks* of teleosts are paraphyletic to the ones in tetrapods and fall in two lineages, however, given that the posterior probability for the grouping of one of the teleost lineages as the sister to tetrapods is generally low (posterior probability = 0.51 ± 0.34), we cannot reject the monophyly of the *Mavericks* of teleosts. An independent phylogeny estimated exclusively from the pPOLB sequences of intact elements is in total agreement with these findings (Figure S4), with a support >0.99 for all major clades, suggesting that our analysis is robust to the inclusion of degraded genomic fossils.

Mavericks have switched hosts frequently in cyprinid fish

We focused on cypriniform fish for the cophylogenetic and selection analyses, given that their *Mavericks* were not monophyletic, several species contained intact elements and high copy numbers were found in *Oxygymnocypris stewartii*. The coevolutionary event with the highest probability inferred from the ABC analysis was cospeciation, with a mean probability of 0.64 ± 0.05, followed by host switching with a mean probability of 0.26 ± 0.06. Reconciliation of the cyprinid *Maverick* phylogeny with the host's (using the transformed event probabilities as costs), suggests that although cospeciation and within-host amplification have been common evolutionary paths for the *Mavericks* of cyprinids, host switching has also occurred extensively (Figure 5). Moreover, species such as *Danionella dracula*, whose *Mavericks* are monophyletic, seem to all descend from an ancestral cross-species transmission, while in other host genomes multiple lineages coexist, some of which were inherited vertically and others by independent colonisation events (e.g. *Danio rerio*, *Oxygymnocypris stewartii*). We must recognise that these results rely on an accurate cyprinid host phylogeny; although we built a host tree consistent with several independent molecular works, these relationships have been problematic to resolve (Stout et al., 2016). However, we believe the presence of polyphyletic *Maverick* lineages in some genomes argues in favour of cross-species transmissions even if the true host species phylogeny is unknown.

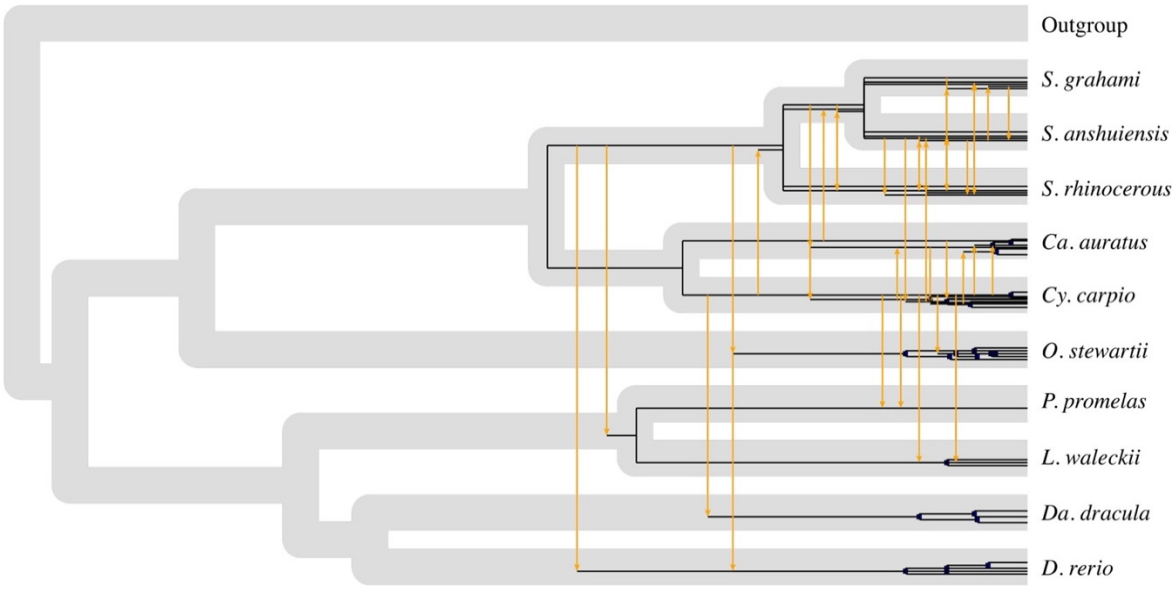


Figure 5. *Mavericks* have switched hosts frequently in cyprinid fish. The figure depicts the most-parsimonious reconciliation of the cyprinid pPOLB tree onto the host phylogeny. Black lines represent vertical inheritance within the host genome, yellow arrows indicate host switch events. Genera abbreviations: *Sinocyclocheilus* (*S.*), *Carassius* (*Ca.*), *Cyprinus* (*Cy.*), *Oxygymnocypris* (*O.*), *Pimephales* (*P.*), *Leuciscus* (*L.*), *Danionella* (*Da.*) and *Danio* (*D.*).

The core genes of cyprinid *Mavericks* are under strong purifying selection

As we have seen previously, while some *Mavericks* are intact most are degraded elements. Previous analyses had used consensus sequences and the selective forces affecting their evolution had not been explored (Haapa-Paananen et al., 2014; Kapitonov and Jurka, 2006). Evidence of selection acting on the capsid morphogenesis homologues in a phylogenetic context, is a good indication that *Mavericks* have been an active group of endogenous viruses. Here, we found that the eight core genes of *Mavericks* in cyprinid fish are under strong purifying selection with $\omega \sim 0.001$ (Table 1); meaning that for every 1000 synonymous changes that are fixed, a single non-synonymous change reaches fixation. The best models we fitted were either a single-ratio estimated ω (*int*, *Mcp*, *ppolb*, *pro*) or the 2-

ratio model with an ω for internal and terminal branches (*atp*, *mcp*, *pz*). The estimated ω for the internal branches was somewhat lower but in the same order of magnitude as the ω for terminal branches. Interestingly, *pw* was the single gene for which the 2-ratio host switch model was favoured.

Table 1. Maximum-likelihood estimates for the non-synonymous to synonymous ratios (ω) of the 8 conserved genes in cyprinid *Mavericks*. P-values were derived from a χ^2 comparison between the best and neutral models, a p-value < 0.05 indicates rejection of neutrality.

Gene	Best model	ω	p-value
atp	2-ratios, $\omega_{internal}/\omega_{terminal}$	$\omega_{internal} = 0.00044$ $\omega_{terminal} = 0.00066$	0.000
int	1-ratio, ω-estimated	ω = 0.00138	0.000
Mcp (py)	1-ratio, ω-estimated	ω = 0.00181	0.000
mcp (px)	2-ratios, $\omega_{internal}/\omega_{external}$	$\omega_{internal} = 0.00074$ $\omega_{terminal} = 0.00105$	0.000
ppolb	1-ratio, ω-estimated	ω = 0.00099	0.000
pro	1-ratio, ω-estimated	ω = 0.00101	0.000
pw	2-ratios, $\omega_{no\text{-}switch}/\omega_{switch}$	$\omega_{no\text{-}switch} = 0.001255$ $\omega_{switch} = 0.001494$	0.000
pz	2-ratios, $\omega_{internal}/\omega_{terminal}$	$\omega_{internal} = 0.00102$ $\omega_{terminal} = 0.00157$	0.000

Discussion

Mavericks are widespread across the genomes of osteichthyans but seem to have gone extinct in mammals. Their distribution is patchy and they attain low copy numbers in most genomes. Given that *Mavericks* are large elements, it is possible that high copy numbers are selected against in host genomes as they pose a risk of deleterious ectopic recombination (Petrov et al., 2003). The hundreds of defective copies in the genomes of *Ambystoma mexicanum* and *Sphenodon punctatus* are perplexing, and it is still unclear which biological processes might be responsible for these amplification events or if they have exerted any impact on these organisms. The high copy numbers found in the axolotl could be related to the low deletion rates which have been detected in salamander genomes, which contribute to their large genome sizes as well (Sun et al., 2012).

Since few chondrichthyans were available for screening, the apparent absence of *Mavericks* from this group may reflect a sampling bias rather than an actual absence. A central theme that has emerged from our analysis is that most copies (96%) of vertebrate *Mavericks* are defective and thus, constitute a rich record of highly degraded genomic fossils. The observation that most intact elements were found in ray-finned fish is relevant since diverse families of giant viruses, virophages and Polinton-like viruses (la Scola et al., 2008; Yoshikawa et al., 2019; Yutin et al., 2015), elements to which *Mavericks* are related, also infect aquatic organisms. Recently, Bellas and Sommaruga (Bellas and Sommaruga, 2021) have shown that Polinton-like viruses are amongst the most abundant viruses from an alpine lake ecosystem in Austria. Haapa-Paananen et al. (Haapa-Paananen et al., 2014) noticed that vertebrate *Mavericks* formed a clade with elements from cnidarians, echinoderms and

marine mollusks. Our comprehensive analysis across a broad diversity of vertebrates confirms the notion that water plays an important role in the transmission of *Mavericks*.

The genomic arrangement of core genes into two modules seems to be a universal characteristic of vertebrate *Mavericks*. No introns were observed in the genes of intact elements which agrees with the work of Kapitonov and Jurka (Kapitonov and Jurka, 2006). Although the *pz* and *pw* genes still have an uncharacterized function, their topological conservation indicates that they are required for the maintenance of *Mavericks* in the genomes of vertebrates. The *pm* ORF we describe, is also conserved, implying a relevant role in the lifecycle of vertebrate *Mavericks*. Other conserved features of vertebrate *Mavericks* are the genome end-motifs and the sizes of intact elements. The end-motif repetitions allow the "jumping-back" mechanism required during protein-primed DNA replication in adenoviruses and tectiviruses (Caldentey et al., 1993; King and van der Vliet, 1994), and would be expected to have the same function in *Mavericks* (Kapitonov and Jurka, 2006). In *Monopterus albus*, the existence of 14 defective copies with the unique "AC" type indicate that the ancestral element of this motif was capable of replication. From the analysis of the element sizes, the variation around the mean size of ~17 kb is about 10%, suggesting a constrained genome size. In adenoviruses, departures from the wildtype genome size have been shown to affect the stability of the virus capsid (Smith et al., 2009).

We provide extensive evidence for the presence of *Mavericks* in the genomes of birds. Pritham et al. (Pritham et al., 2007) mentioned this possibility after finding sequence homology to the *Maverick* integrase in short genomic clones of *Apteryx australis*. Subsequently, Guizard et al. (Guizard et al., 2016) described remnants of *Mavericks* in two distinct loci of *Gallus gallus* with corresponding homologues in *Meleagris gallopavo*. In our

analysis, we have found defective copies of *Mavericks* in the genomes of palaeognaths, penguins, swifts and hummingbirds, owls, cormorants, ibises, and in the hoatzin. As the *Mavericks* of birds form a well-supported monophyletic group and there are two loci in *Apteryx australis* with 7 of the 8 core genes (one is lacking *pro*, the other *atp*), it seems likely that there were active *Mavericks* in the genome of the most recent common ancestor of extant birds. We also found diverse integrations in the genomes of non-avian reptiles which expand the known range of *Mavericks*, this is the case for the elements discovered in snakes (suborder Serpentes) and in turtles from the families Cheloniidae, Testudinidae and Trionychidae, including an intact element from *Terrapene carolina*.

Identification of orthologous endogenous viral elements from the Late Cretaceous is not unprecedented. Endogenous viral elements of comparable age have been described for bornaviruses in afrotherians (Katzourakis and Gifford, 2010), an ERV-L retrovirus from placental mammals (Lee et al., 2013) and an hepadnavirus found in neoavian birds (Suh et al., 2013). From the Bayesian phylogenetic analyses, it is clear that these elements have infected vertebrates for hundreds of millions of years. In fact, we cannot rule out the presence of *Mavericks* in the genome of the most recent common ancestor of Osteichthyes, since the 95% highest probability densities for 5 proteins include the estimated age of the actinopterygian/sarcopterygian split at the Silurian/Devonian boundary, ~419 Mya (Zhu et al., 2009). Similar time scales of virus evolution have been estimated for retroviruses under the power-law decay model (~450 Mya) (Aiewsakun and Katzourakis, 2017). It would appear, therefore, that the superlineage to which *Mavericks* belong is considerably older.

Analysis of mammalian and avian herpesvirus evolution has revealed substantial codivergence with their hosts, and under this assumption, their evolutionary rates have been

estimated at $3 \cdot 10^{-9}$ amino acid substitutions per site per year (McGeoch et al., 2000). By using the power-law model of evolutionary rates that has been used to study retrovirus macroevolution (Aiewsakun and Katzourakis, 2017), and assuming that the sequences we analysed diverged ~419 Mya, we would expect to observe $2.43 \cdot 10^{-9}$ amino acid substitutions per site per year. This expected value is remarkably similar to the mean rates we have estimated independently for all core proteins of *Mavericks*. These low estimates are consistent with the time-dependent rate phenomenon of molecular evolution: estimated evolutionary rates decrease as the divergence time between sequences increases, and this has also been shown to occur generally across viruses in different Baltimore groups (Aiewsakun and Katzourakis, 2016).

We were able to demonstrate that the *Mavericks* of cyprinid fish are under strong purifying selection, with an ω of the order of 0.001–0.0001 for all genes. Purifying selection on the order of 10^{-3} has been observed before in dsDNA viruses with the captured CD200 immunoglobulin and 1 CCL3 chemokine genes of rhadinoviruses (family *Herpesviridae*) (Aswad and Katzourakis, 2018), but these values seem to be generally low. Since we used branch models for detecting selection in the viruses, our inference reflects the major evolutionary pressures acting on the whole-gene level. Instances of genes in which the internal/terminal branch model was favoured were probably due to the effect of relaxed selection acting on recent pseudogenes, as could be inferred from the higher ω for terminal branches. The low ω for internal branches for all genes suggests that *Mavericks* have continuously produced viral particles capable of germ-line reinfection throughout their history. Interestingly, the gene *pw* was the only one consistent with a host-switch model, indicating that it may be associated with a role in host adaptation since host-switch branches had higher estimated values of ω. In the case of human endogenous retroviruses, most

elements show evidence of purifying selection on the *env* gene which is consistent with an origin through formation of viral particles and reinfection (Belshaw et al., 2004). On the other hand, the *env* gene of some human endogenous retroviruses evolves neutrally, suggesting the gene is no longer functional in these cases and that these elements have been copied by complementation *in trans* or intragenomic replication *in cis* (Belshaw et al., 2005). Since all the virion morphogenesis homologues are under strong purifying selection in *Mavericks*, this demonstrates they have evolved mainly through formation of viral particles and reinfection, as with the endogenous retroviruses. Overall, we have shown that purifying selection has preserved the functionality of genes from cyprinid *Mavericks* over millions of years.

Inconsistencies between the host and *Maverick* phylogenies within major groups of actinopterygians and tetrapods, indicate that the evolutionary history of *Mavericks* is characterized by a dynamic rather than a stable association with their hosts. Specifically, through our cophylogenetic analyses, we show that these inconsistencies support a history of extensive host switching in cyprinid *Mavericks*. Host switching is a general phenomenon that occurs during viral evolution, and in particular, a strong signal for host switching has been found in the related families of dsDNA viruses *Adenoviridae* and *Poxviridae* (Geoghegan et al., 2017). On the other hand, we have also found evidence for codivergence. Therefore, the evolution of *Mavericks* in vertebrates shows signs of both vertical and cross-species transmissions.

The existence of intact elements in the genomes of vertebrates, together with constrained genome sizes and strong evidence for host switching coincident with an intense purifying selection on core genes (and especially those involved in morphogenesis), are robust

indications that *Mavericks* are an active lineage of integrative viruses that infect vertebrates. These observations call for incorporation of vertebrate *Mavericks* into the classification of viruses. As group-I and group-II *Mavericks* have been previously shown to be paraphyletic, the proposed term 'polintoviruses' (Krupovič and Koonin, 2015) would be taxonomically inadequate. Instead, we propose the creation of the family 'Proteoviridae' and the genus 'Alphaproteovirus' to include vertebrate *Mavericks* and possibly other group-I elements which form a monophyletic group. The family derives its name from the Greek god Proteus, son of Poseidon, who is depicted in the Odyssey as an ancient deity of the sea with the ability to assume the form of different creatures (Homer, 1919). The proposed classification of intact elements into 19 species is shown in Table 2.

It is theoretically possible that some sequences derived from proteoviruses may have been exapted to confer advantageous host phenotypes. Co-option of transposable elements is known to have rewired genetic networks and driven changes in the gene expression patterns of vertebrates (Kunarso et al., 2010; Romanish et al., 2007). For example, the insertion of a proteovirus into the promoter region of the growth hormone gene of *Oncorhynchus tshawytscha*, may be in part responsible for the size difference between this species and the smaller *Salmo salar* (Fischer and Hackl, 2016; von Schalburg et al., 2008). Endogenous proteoviruses might also function as an EVE-derived immunity against their exogenous counterparts, but this remains to be seen since exogenous forms are still unknown.

Table 2. Proposed species-level classification of vertebrate *Mavericks* (family 'Proteoviridae', genus 'Alphaproteovirus'), with designated type sequences.

Species	GenBank acc.	Coordinates
Anabas testudineus alphaproteovirus	NW_020535984.1	16773227–16787566
Anolis carolinensis alphaproteovirus	NC_014778.1	129158247–129172346
Astatotilapia calliptera alphaproteovirus	NC_039306.1	12448581–12465490
Carassius auratus alphaproteovirus	NW_020523509.1	38327–56441
Danio rerio alphaproteovirus 1	NC_007136.7	27671708–27687993
Danio rerio alphaproteovirus 2	NW_001884452.4	163866–182541
Danio rerio alphaproteovirus 3	NC_007134.7	16402434–16421359
Larimichthys crocea alphaproteovirus	NW_017609269.1	1620911–1634436
Mastacembelus armatus alphaproteovirus	OOHQ01000082.1	27018430–27033660
Oreochromis niloticus alphaproteovirus	NC_031977.2	37033790–37051003
Oxygymnocypris stewartii alphaproteovirus 1	QVTF01001200.1	1425766–1443135
Oxygymnocypris stewartii alphaproteovirus 2	QVTF01018600.1	776725–796426
Perca fluviatilis alphaproteovirus 1	QFAT01000030.1	1733268–1750442
Perca fluviatilis alphaproteovirus 2	QFAT01022094.1	1737–17082
Pseudocrenilabrinae alphaproteovirus	NW_020327416.1	45576–61017
Rhinella marina alphaproteovirus	ONZH01012121.1	13566–27538
Terrapene carolina alphaproteovirus	NW_020664598.1	16371040–16384884
Xenopus tropicalis alphaproteovirus	NC_030686.1	6013889–6026611
Xiphophorus alphaproteovirus	QPIH01000028.1	6019497–6036410

An intriguing possibility, is that proteoviruses may function as a virophage-derived immunity against NCLDV infections in vertebrates. The nature of antiviral defense systems based on endogenous virophages has been described in detail for unicellular eukaryotes. Fischer and Hackl (Fischer and Hackl, 2016) discovered that Mavirus provirophages in the genome of *Cafeteria burkhardae* function against infection by *Cafeteria roenbergensis Virus* (CroV). CroV infection triggers virophage synthesis, and although this leads to cell

lysis, the newly produced virophages inhibit replication of the giant virus in cells of the wider host population (Fischer and Hackl, 2016). Therefore, expression of intact endogenous proteoviruses and formation of their virions may depend on activation by a giant virus, thus giving protection at the level of a multicellular organism. The host range of the family *Iridoviridae* is consistent with this hypothesis; iridoviruses are important pathogens of fish, amphibians and non-avian reptiles (Chinchar et al., 2017; Williams et al., 2005). These are precisely the groups where we identified intact elements, while iridoviruses are not known to infect either birds or mammals. Epizootics caused by iridoviruses can reach mortalities of 100% in fish (He et al., 2000) and mass mortality events have been reported for amphibians (Kik et al., 2011). Thus, the extinction of iridoviruses from birds and mammals may have driven the definitive deterioration of proteoviruses in their genomes. On the basis of these dynamics we hypothesise that the rise of adenoviruses, which attain their greatest diversity in mammals and birds, could have coincided with the elimination of proteoviruses in these groups as well.

We have shown that the *Mavericks* of vertebrates represent an ancient lineage of mostly aquatic dsDNA viruses, which have persisted for hundreds of millions of years by integrating into the genomes of hosts and switching species frequently. The existence of elements with an intact coding capacity together with intense signatures of purifying selection on all genes at the level of the virus, strongly indicate that some lineages of vertebrate proteoviruses are functional. Unlike endogenous retroviruses, which are a major component of eukaryotic genomes, it seems proteoviruses are one of the rare instances of endogenous dsDNA viruses in metazoans; the other being the *Teratorns*, large elements found in the genomes of teleosts which originate from a *piggyBac*-alloherpesvirus fusion (Aswad and Katzourakis, 2017; Inoue et al., 2018). Investigating the ecological roles of

proteoviruses remains an important question, especially when they could function as a virophage-based immune system against NCLDVs in a broad range of vertebrates. Molecular characterization of proteoviruses will also be essential, given that the conditions leading to the formation of virions remain unknown. This is particularly relevant since they seem attractive candidates to develop new gene transfer technologies for use in genetic engineering.

References

Abascal F, Zardoya R, Telford MJ. 2010. TranslatorX: Multiple alignment of nucleotide sequences guided by amino acid translations. *Nucleic Acids Research* 38. doi:10.1093/nar/gkq291

Ahi YS, Mittal SK. 2016. Components of adenovirus genome packaging. *Frontiers in Microbiology 7:* 1503. doi: 10.3389/fmicb.2016.01503

Aiewsakun P, Katzourakis A. 2017. Marine origin of retroviruses in the early Palaeozoic Era. *Nature Communications* 8. doi:10.1038/ncomms13954

Aiewsakun P, Katzourakis A. 2016. Time-dependent rate phenomenon in viruses. *Journal of Virology* 90:7184–7195. doi:10.1128/JVI.00593-16

Altschul SF, Gish W, Miller W, Myers EW, Lipman DJ. 1990. Basic local alignment search tool. *Journal of Molecular Biology* 215:403–410. doi:10.1016/S0022-2836(05)80360-2

Aswad A, Katzourakis A. 2018. Cell-derived viral genes evolve under stronger purifying selection in rhadinoviruses. *Journal of Virology* 92. doi:10.1128/jvi.00359-18

Aswad A, Katzourakis A. 2017. A novel viral lineage distantly related to herpesviruses discovered within fish genome sequence data. *Virus Evolution* 3. doi:10.1093/ve/vex016

Bao W, Kojima KK, Kohany O. 2015. Repbase Update, a database of repetitive elements in eukaryotic genomes. *Mobile DNA* 6. doi:10.1186/s13100-015-0041-9

Baudet C, Donati B, Sinaimeri B, Crescenzi P, Gautier C, Matias C, Sagot MF. 2015. Cophylogeny reconstruction via an approximate bayesian computation. *Systematic Biology* 64. doi:10.1093/sysbio/syu129

Bellas CM, Sommaruga R. 2021. Polinton-like viruses are abundant in aquatic ecosystems. *Microbiome* 9. doi:10.1186/s40168-020-00956-0

Belshaw R, Katzourakis A, Pačes J, Burt A, Tristem M. 2005. High copy number in human endogenous retrovirus families is associated with copying mechanisms in addition to reinfection. *Molecular Biology and Evolution* 22:814–817. doi:10.1093/molbev/msi088

Belshaw R, Pereira V, Katzourakis A, Talbot G, Pačes J, Burt A, Tristem M. 2004. Long-term reinfection of the human genome by endogenous retroviruses. *Proc Natl Acad Sci U S A* 101:4894–4899. doi:10.1073/pnas.0307800101

Birney E, Clamp M, Durbin R. 2004. GeneWise and Genomewise. *Genome Research* 14. doi:10.1101/gr.1865504

Bouckaert R, Vaughan TG, Barido-Sottani J, Duchêne S, Fourment M, Gavryushkina A, Heled J, Jones G, Kühnert D, de Maio N, Matschiner M, Mendes FK, Müller NF, Ogilvie HA, du Plessis L, Popinga A, Rambaut A, Rasmussen D, Siveroni I, Suchard MA, Wu CH, Xie D, Zhang C, Stadler T, Drummond AJ. 2019. BEAST 2.5: An advanced software platform for Bayesian evolutionary analysis. *PLoS Computational Biology* 15. doi:10.1371/journal.pcbi.1006650

Caldentey J, Blanco L, Bamford DH, Salas M. 1993. *In vitro* replication of bacteriophage PRD1 DNA. Characterization of the protein-primed initiation site. *Nucleic Acids Research* 21. doi:10.1093/nar/21.16.3725

Camacho C, Coulouris G, Avagyan V, Ma N, Papadopoulos J, Bealer K, Madden TL. 2009. BLAST+: Architecture and applications. *BMC Bioinformatics* 10. doi:10.1186/1471-2105-10-421

Capella-Gutiérrez S, Silla-Martínez JM, Gabaldón T. 2009. trimAl: A tool for automated alignment trimming in large-scale phylogenetic analyses. *Bioinformatics* 25:1972–1973. doi:10.1093/bioinformatics/btp348

Chevenet F, Doyon JP, Scornavacca C, Jacox E, Jousselin E, Berry V. 2016. SylvX: A viewer for phylogenetic tree reconciliations. *Bioinformatics* 32. doi:10.1093/bioinformatics/btv625

Chinchar VG, Hick P, Ince IA, Jancovich JK, Marschang R, Qin Q, Subramaniam K, Waltzek TB, Whittington R, Williams T, Zhang QY. 2017. ICTV virus taxonomy profile: *Iridoviridae*. *Journal of General Virology* 98. doi:10.1099/jgv.0.000818

Conow C, Fielder D, Ovadia Y, Libeskind-Hadas R. 2010. Jane: A new tool for the cophylogeny reconstruction problem. *Algorithms for Molecular Biology* 5. doi:10.1186/1748-7188-5-16

Darling ACE, Mau B, Blattner FR, Perna NT. 2004. Mauve: multiple alignment of conserved genomic sequence with rearrangements. *Genome Research* 14. doi:10.1101/gr.2289704

Darriba Di, Posada D, Kozlov AM, Stamatakis A, Morel B, Flouri T. 2020. ModelTest-NG: a new and scalable tool for the selection of dna and protein evolutionary models. *Molecular Biology and Evolution* 37:291–294. doi:10.1093/molbev/msz189

Deorowicz S, Debudaj-Grabysz A, Gudys A. 2016. FAMSA: fast and accurate multiple sequence alignment of huge protein families. *Scientific Reports* 6. doi:10.1038/srep33964

Doyon JP, Scornavacca C, Gorbunov KY, Szöllosi GJ, Ranwez V, Berry V. 2010. An efficient algorithm for gene/species trees parsimonious reconciliation with losses, duplications and transfers. Lecture Notes in Computer Science (Including Subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics). doi:10.1007/978-3-642-16181-0_9

Drummond AJ, Ho SYW, Phillips MJ, Rambaut A. 2006. Relaxed phylogenetics and dating with confidence. PLoS Biology 4:699–710. doi:10.1371/journal.pbio.0040088

Eddy S. 2002. SQUID - library functions for biological sequence analysis.

Eddy SR. 2009. A new generation of homology search tools based on probabilistic inference. *Genome Inform* 23. doi:10.1142/9781848165632_0019

Eddy SR. 1995. Multiple alignment using hidden Markov models. Proceedings. International Conference on Intelligent Systems for Molecular Biology; ISMB International Conference on Intelligent Systems for Molecular Biology 3.

Elewa A, Wang H, Talavera-López C, Joven A, Brito G, Kumar A, Hameed LS, Penrad-Mobayed M, Yao Z, Zamani N, Abbas Y, Abdullayev I, Sandberg R, Grabherr M, Andersson B, Simon A. 2017. Reading and editing the Pleurodeles waltl genome reveals novel features of tetrapod regeneration. *Nature Communications* 8. doi:10.1038/s41467-017-01964-9

Feschotte C, Pritham EJ. 2005. Non-mammalian c-integrases are encoded by giant transposable elements. *Trends in Genetics*. doi:10.1016/j.tig.2005.07.007

Fischer MG, Hackl T. 2016. Host genome integration and giant virus-induced reactivation of the virophage Mavirus. *Nature* 540:288–291. doi:10.1038/nature20593

Gao X, Voytas DF. 2005. A eukaryotic gene family related to retroelement integrases. *Trends in Genetics*. doi:10.1016/j.tig.2005.01.006

Geoghegan JL, Duchêne S, Holmes EC. 2017. Comparative analysis estimates the relative frequencies of co-divergence and cross-species transmission within viral families. *PLoS Pathogens* 13. doi:10.1371/journal.ppat.1006215

Giberson AN, Davidson AR, Parks RJ. 2012. Chromatin structure of adenovirus DNA throughout infection. *Nucleic Acids Research* 40:2369–2376. doi: 10.1093/nar/gkr1076

Guizard S, Piégu B, Arensburger P, Guillou F, Bigot Y. 2016. Deep landscape update of dispersed and tandem repeats in the genome model of the red jungle fowl, *Gallus gallus*, using a series of *de novo* investigating tools. *BMC Genomics* 17. doi:10.1186/s12864-016-3015-5

Haapa-Paananen S, Wahlberg N, Savilahti H. 2014. Phylogenetic analysis of *Maverick*/*Polinton* giant transposons across organisms. *Molecular Phylogenetics and Evolution* 78. doi:10.1016/j.ympev.2014.05.024

He JG, Wang SP, Zeng K, Huang ZJ, Chan SM. 2000. Systemic disease caused by an iridovirus-like agent in cultured mandarinfish, *Siniperca chuatsi* (Basilewsky), in China. *Journal of Fish Diseases* 23. doi:10.1046/j.1365-2761.2000.00213.x

Homer. 1919. Odyssey. Cambridge: Harvard University Press.

Hubley R, Finn RD, Clements J, Eddy SR, Jones TA, Bao W, Smit AFA, Wheeler TJ. 2016. The Dfam database of repetitive DNA families. *Nucleic Acids Research* 44. doi:10.1093/nar/gkv1272

Inoue Y, Kumagai M, Zhang X, Saga T, Wang D, Koga A, Takeda H. 2018. Fusion of *piggyBac*-like transposons and herpesviruses occurs frequently in teleosts. *Zoological Letters* 4. doi:10.1186/s40851-018-0089-8

Johnson M, Zaretskaya I, Raytselis Y, Merezhuk Y, McGinnis S, Madden TL. 2008. NCBI BLAST: a better web interface. *Nucleic Acids Res* 36. doi:10.1093/nar/gkn201

Jurka J, Klonowski P, Dagman V, Pelton P. 1996. Censor - A program for identification and elimination of repetitive elements from DNA sequences. *Computers and Chemistry* 20. doi:10.1016/S0097-8485(96)80013-1

Kapitonov VV, Jurka J. 2006. Self-synthesizing DNA transposons in eukaryotes. *Proc Natl Acad Sci U S A* 103. doi:10.1073/pnas.0600833103

Katoh K, Misawa K, Kuma KI, Miyata T. 2002. MAFFT: a novel method for rapid multiple sequence alignment based on fast Fourier transform. *Nucleic Acids Research* 30:3059–3066. doi:10.1093/nar/gkf436

Katzourakis A, Gifford RJ. 2010. Endogenous viral elements in animal genomes. *PLoS Genetics* 6:e1001191. doi:10.1371/journal.pgen.1001191

Kik M, Martel A, Sluijs AS van der, Pasmans F, Wohlsein P, Gröne A, Rijks JM. 2011. Ranavirus-associated mass mortality in wild amphibians, the Netherlands, 2010: a first report. *Veterinary Journal* 190. doi:10.1016/j.tvjl.2011.08.031

King AJ, van der Vliet PC. 1994. A precursor terminal protein-trinucleotide intermediate during initiation of adenovirus DNA replication: regeneration of molecular ends *in vitro* by a jumping back mechanism. *EMBO Journal* 13. doi:10.1002/j.1460-2075.1994.tb06917.x

Koonin EV, Krupovič M. 2017. *Polintons*, virophages and transpovirons: a tangled web linking viruses, transposons and immunity. *Current Opinion in Virology* 25:7–15. doi:10.1016/j.coviro.2017.06.008

Kozlov AM, Darriba D, Flouri T, Morel B, Stamatakis A, Wren J. 2019. RAxML-NG: a fast, scalable and user-friendly tool for maximum likelihood phylogenetic inference. *Bioinformatics* 35:4453–4455. doi:10.1093/bioinformatics/btz305

Krupovič M, Bamford DH. 2008. Virus evolution: how far does the double β-barrel viral lineage extend? *Nature Reviews Microbiology* 6:941–948. doi:10.1038/nrmicro2033

Krupovič M, Bamford DH, Koonin EV. 2014. Conservation of major and minor jelly-roll capsid proteins in *Polinton* (*Maverick*) transposons suggests that they are bona fide viruses. *Biology Direct* 9:1–7. doi:10.1186/1745-6150-9-6

Krupovič M, Koonin EV. 2015. *Polintons*: A hotbed of eukaryotic virus, transposon and plasmid evolution. *Nature Reviews Microbiology* 13:105–115. doi:10.1038/nrmicro3389

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

la Scola B, Desnues C, Pagnier I, Robert C, Barrassi L, Fournous G, Merchat M, Suzan-Monti M, Forterre P, Koonin E, Raoult D. 2008. The virophage as a unique parasite of the giant mimivirus. *Nature* 455:100–104. doi:10.1038/nature07218

Lee A, Nolan A, Watson J, Tristem M. 2013. Identification of an ancient endogenous retrovirus, predating the divergence of the placental mammals. *Philosophical Transactions of the Royal Society B: Biological Sciences* 368. doi:10.1098/rstb.2012.0503

McGeoch DJ, Dolan A, Ralph AC. 2000. Toward a comprehensive phylogeny for mammalian and avian herpesviruses. *Journal of Virology* 74. doi:10.1128/jvi.74.22.10401-10406.2000

Mirarab S, Nguyen N, Guo S, Wang LS, Kim J, Warnow T. 2015. PASTA: ultra-large multiple sequence alignment for nucleotide and amino-acid sequences. *Journal of Computational Biology* 22. doi:10.1089/cmb.2014.0156

NCBI Resource Coordinators. 2016. Database resources of the National Center for Biotechnology Information. *Nucleic Acids Res* 44.

Nguyen LT, Schmidt HA, von Haeseler A, Minh BQ. 2015. IQ-TREE: A fast and effective stochastic algorithm for estimating maximum-likelihood phylogenies. *Molecular Biology and Evolution* 32. doi:10.1093/molbev/msu300

Petrov DA, Aminetzach YT, Davis JC, Bensasson D, Hirsh AE. 2003. Size matters: non-LTR retrotransposable elements and ectopic recombination in *Drosophila*. *Molecular Biology and Evolution* 20. doi:10.1093/molbev/msg102

Price MN, Dehal PS, Arkin AP. 2010. FastTree 2 - Approximately maximum-likelihood trees for large alignments. *PLoS ONE* 5. doi:10.1371/journal.pone.0009490

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. doi:10.1016/j.gene.2006.08.008

Rambaut A. 2018. Figtree v1.4.4.

Rambaut A, Drummond AJ, Xie D, Baele G, Suchard MA. 2018. Posterior summarization in Bayesian phylogenetics using Tracer 1.7. *Systematic Biology* 67. doi:10.1093/sysbio/syy032

Romanish MT, Lock WM, van de Lagemaat LN, Dunn CA, Mager DL. 2007. Repeated recruitment of LTR retrotransposons as promoters by the anti-apoptotic locus NAIP during mammalian evolution. *PLoS Genetics* 3. doi:10.1371/journal.pgen.0030010

Ronquist F, Huelsenbeck JP. 2003. MrBayes 3: Bayesian phylogenetic inference under mixed models. *Bioinformatics* 19:1572–1574. doi:10.1093/bioinformatics/btg180

Shannon CE. 1948. A mathematical theory of communication. *Bell System Technical Journal* 27. doi:10.1002/j.1538-7305.1948.tb01338.x

Sievers F, Wilm A, Dineen D, Gibson TJ, Karplus K, Li W, Lopez R, McWilliam H, Remmert M, Söding J, Thompson JD, Higgins DG. 2011. Fast, scalable generation of high-quality protein multiple sequence alignments using Clustal Omega. *Molecular Systems Biology* 7. doi:10.1038/msb.2011.75

Smit A, Hubley R, Grenn P. 2015. RepeatMasker Open-407.

Smith AC, Poulin KL, Parks RJ. 2009. DNA genome size affects the stability of the adenovirus virion. *Journal of Virology* 83. doi:10.1128/jvi.01644-08

Söding J, Biegert A, Lupas AN. 2005. The HHpred interactive server for protein homology detection and structure prediction. *Nucleic Acids Res* 33:244–248. doi:10.1093/nar/gki408

Steinegger M, Söding J. 2017. MMseqs2 enables sensitive protein sequence searching for the analysis of massive data sets. *Nature Biotechnology*. doi:10.1038/nbt.3988

Suh A, Brosius J, Schmitz J, Kriegs JO. 2013. The genome of a Mesozoic paleovirus reveals the evolution of hepatitis B viruses. *Nature Communications* 4. doi:10.1038/ncomms2798

Sun C, López Arriaza JR, Mueller RL. 2012. Slow DNA loss in the gigantic genomes of salamanders. *Genome Biology and Evolution* 4. doi:10.1093/gbe/evs103

The Math Works Inc. 2019. Matlab 2019a. MathWorks Inc, Natick, MA.

von Schalburg KR, Yazawa R, de Boer J, Lubieniecki KP, Goh B, Straub CA, Beetz-Sargent MR, Robb A, Davidson WS, Devlin RH, Koop BF. 2008. Isolation, characterization and

comparison of atlantic and chinook salmon growth hormone 1 and 2. *BMC Genomics* 9. doi:10.1186/1471-2164-9-522

Wang XZ, Gan XN, Li JB, Mayden RL, He SP. 2012. Cyprinid phylogeny based on Bayesian and maximum likelihood analyses of partitioned data: implications for Cyprinidae systematics. *Science China Life Sciences* 55. doi:10.1007/s11427-012-4366-z

Williams T, Barbosa-Solomieu V, Chinchar VG. 2005. A decade of advances in iridovirus research. Advances in Virus Research. doi:10.1016/S0065-3527(05)65006-3

Yang J, Chen Xiaoli, Bai J, Fang D, Qiu Y, Jiang W, Yuan H, Bian C, Lu J, He S, Pan X, Zhang Y, Wang X, You X, Wang Y, Sun Y, Mao D, Liu Y, Fan G, Zhang H, Chen Xiaoyong, Zhang X, Zheng L, Wang Jintu, Cheng L, Chen J, Ruan Z, Li J, Yu H, Peng C, Ma X, Xu J, He Y, Xu Z, Xu P, Wang Jian, Yang H, Wang Jun, Whitten T, Xu X, Shi Q. 2016. The *Sinocyclocheilus* cavefish genome provides insights into cave adaptation. *BMC Biology* 14. doi:10.1186/s12915-015-0223-4

Yang L, Sado T, Vincent Hirt M, Pasco-Viel E, Arunachalam M, Li J, Wang X, Freyhof J, Saitoh K, Simons AM, Miya M, He S, Mayden RL. 2015. Phylogeny and polyploidy: resolving the classification of cyprinine fishes (Teleostei: Cypriniformes). *Molecular Phylogenetics and Evolution* 85. doi:10.1016/j.ympev.2015.01.014

Yang Z. 2007. PAML 4: Phylogenetic analysis by maximum likelihood. *Molecular Biology and Evolution* 24:1586–1591. doi:10.1093/molbev/msm088

Yoshikawa G, Blanc-Mathieu R, Song C, Kayama Y, Mochizuki T, Murata K, Ogata H, Takemura M. 2019. Medusavirus, a novel large DNA virus discovered from hot spring water. *Journal of Virology* 93. doi:10.1128/jvi.02130-18

Yutin N, Shevchenko S, Kapitonov V, Krupovič M, Koonin EV. 2015. A novel group of diverse Polinton-like viruses discovered by metagenome analysis. *BMC Biology* 13:1–14. doi:10.1186/s12915-015-0207-4

Zhou Q, Braasch I, Froschauer A, Böhne A, Schultheis C, Schartl M, Volff JN. 2010. A novel marker for the platyfish (*Xiphophorus maculatus*) W chromosome is derived from a *Polinton* transposon. *Journal of Genetics and Genomics* 37. doi:10.1016/S1673-8527(09)60036-9

Zhu M, Zhao W, Jia L, Lu J, Qiao T, Qu Q. 2009. The oldest articulated osteichthyan reveals mosaic gnathostome characters. *Nature* 458. doi:10.1038/nature07855

Zimmermann L, Stephens A, Nam SZ, Rau D, Kübler J, Lozajic M, Gabler F, Söding J, Lupas AN, Alva V. 2018. A completely reimplemented mpi bioinformatics toolkit with a new HHpred server at its core. *Journal of Molecular Biology* 430. doi:10.1016/j.jmb.2017.12.007

Chapter III

A billion years arms-race between viruses, virophages and eukaryotes

Abstract

Bamfordviruses are arguably the most diverse group of viruses infecting eukaryotes. They include the Nucleocytoplasmic Large DNA viruses (NCLDVs), virophages, adenoviruses, *Mavericks* and Polinton-like viruses. Two main hypotheses for their origins have been proposed: the "nuclear-escape" and "virophage-first" hypotheses. The "nuclear-escape" hypothesis proposes an endogenous, *Maverick*-like ancestor which escaped from the nucleus and gave rise to adenoviruses and NCLDVs. In contrast, the "virophage-first" hypothesis proposes that NCLDVs coevolved with protovirophages; *Mavericks* then evolved from virophages that became endogenous, with adenoviruses escaping from the nucleus at a later stage. Here, we test the predictions made by both models and consider alternative scenarios. We use a data set of the four core virion proteins sampled across the diversity of the lineage, together with Bayesian and maximum-likelihood hypothesis-testing methods and estimate rooted phylogenies. We find strong evidence that adenoviruses and NCLDVs are not sister groups, and that *Mavericks* and Mavirus acquired the rve-integrase independently. We also found strong support for a monophyletic group of virophages (family *Lavidaviridae*) and a most likely root placed between virophages and the other lineages. Our observations support alternative scenarios to the nuclear-escape and a billion years evolutionary arms-race between virophages and NCLDVs.

Introduction

Viruses in the kingdom *Bamfordvirae* comprise one of the most diverse groups in terms of their genome complexity, ecology and morphology. These dsDNA viruses include the largest viruses characterised to date (Legendre et al., 2014; Scheid, 2016), virophages which are virus parasites of other viruses (Fischer and Suttle, 2011; la Scola et al., 2008), endogenous viruses that colonise the genomes of eukaryotes (Kapitonov and Jurka, 2006; Pritham et al., 2007), and a diverse set of viruses that have been identified from metagenomic data (Bellas and Sommaruga, 2021; Moniruzzaman et al., 2020; Paez-Espino et al., 2019; Roux et al., 2017; Schulz et al., 2020; Yutin et al., 2015). More formally, these viruses are classified into the Nucleocytoplasmic Large DNA Viruses (NCLDVs) which comprise families in the phylum *Nucleocytoviricota* (Aylward et al., 2021), the virophage parasites of NCLDVs (family *Lavidaviridae*) (Fischer, 2020), adenoviruses (family *Adenoviridae*) which infect vertebrates (Harrach et al., 2019), the *Maverick/Polinton* endogenous viruses that colonise the genomes of eukaryotes (Kapitonov and Jurka, 2006; Pritham et al., 2007), Polinton-like viruses (PLVs) which are abundant in aquatic metagenomes (Bellas and Sommaruga, 2021; Yutin et al., 2015), and capsidless elements such as transpovirons (Desnues et al., 2012), mitochondrial and cytoplasmic linear plasmids (Krupovič and Koonin, 2015; Meinhardt et al., 1997). The wide taxonomic distribution of NCLDV and *Maverick* hosts, which comprise all major eukaryotic lineages (Kapitonov and Jurka, 2006; Pritham et al., 2007; Schulz et al., 2020), seem to point to an ancient origin of these viral groups.

Despite this remarkable diversity, the elements share an ancestral gene module used for capsid morphogenesis, which is the feature that distinguishes viruses from other mobile

genetic elements (Koonin et al., 2021). The ancestral module is formed by double and single jelly-roll capsid proteins, a family C5 (adenoviral-like) protease and a family FtsK/HerA DNA packaging ATPase (Krupovič and Koonin, 2016; Yutin et al., 2015). In fact, the jelly-roll capsid proteins and DNA packaging ATPase, also occur in phages from the families *Turriviridae*, *Tectiviridae*, *Corticoviridae* and *Autolykiviridae*, which suggests a common viral ancestor for eukaryotic bamfordviruses and their prokaryotic virus relatives (Koonin et al., 2020; Woo et al., 2021). These observations are at odds with the suggestion that NCLDVs evolved from a fourth domain of cellular life by reductive evolution (Colson et al., 2018; Legendre et al., 2012; Patil and Kondabagil, 2021).

Currently two hypotheses have been proposed for the origin of the eukaryotic *Bamfordvirae*: the "nuclear-escape" and "virophage-first" scenarios (Figure 1). According to the "nuclear-escape" scenario, the elements in this lineage evolved from an endogenous virus in the nucleus of an early eukaryote, a *Maverick*-like ancestor (Koonin and Krupovič, 2017; Krupovič and Koonin, 2015). Adenoviruses, cytoplasmic linear plasmids and NCLDVs evolved after their ancestor escaped from the nucleus (Koonin et al., 2015; Krupovič and Koonin, 2015). Therefore, the nuclear-escape hypothesis predicts that adenoviruses form a clade with NCLDVs and cytoplasmic linear plasmids, and that the proteins shared by cytoplasmic linear plasmids and NCLDVs have a single common origin (Koonin and Krupovič, 2017; Krupovič and Koonin, 2015). The nuclear-escape hypothesis is based on phylogenetic trees of the protein-primed DNA polymerase B, which show a paraphyletic group of *Mavericks* at the base of the other lineages (Koonin and Krupovič, 2017; Krupovič, Yutin, et al., 2016; Krupovič and Koonin, 2015). However, the trees are missing NCLDVs which encode a non-homologous nucleotide-primed DNA polymerase, and could therefore

not be included in the analyses, and only show the position for Mavirus-like virophages which encode the protein-primed DNA polymerase.

The "virophage-first" scenario is based on the close similarity in gene content between the Mavirus virophage and *Mavericks*, and the ability of Mavirus to integrate into eukaryotic genomes (Fischer and Hackl, 2016; Fischer and Suttle, 2011). Integrated Mavirus virophages reactivate upon giant virus infection and confer protection against *Cafeteria roenbergensis virus* in the flagellate *Cafeteria burkhardae* (Fischer and Hackl, 2016). The "virophage-first" scenario therefore suggests that an ancestral virophage evolved shortly after the origin of NCLDVs, and was able to parasitise NCLDVs by virtue of their shared promoter and poly-A sequences (Campbell et al., 2017; Fischer and Suttle, 2011). One of these lineages of ancestral virophages would have gained an integrase and endogenised into eukaryotic hosts providing an immune defence system, which gave rise to *Mavericks* and other elements (Fischer and Suttle, 2011; Katzourakis and Aswad, 2014). A nuclear-escape would have still occurred for adenoviruses and cytoplasmic linear plasmids but not NCLDVs, which would belong to a different clade, suggesting the immediate ancestor of NCLDVs was an exogenous virus.

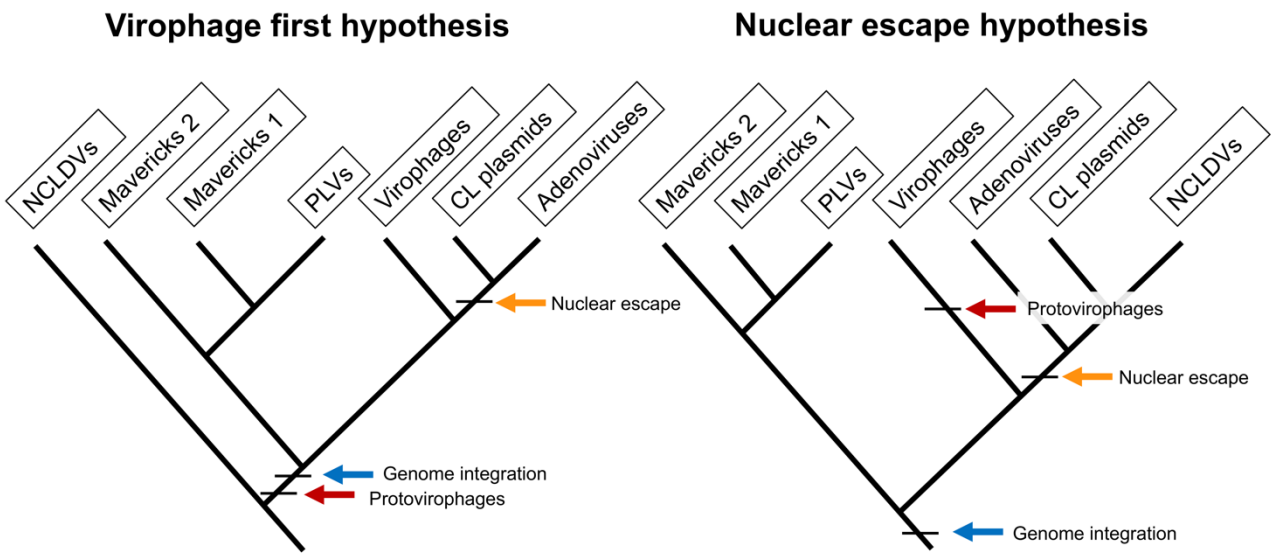


Figure 1. The two main hypotheses for the origin of virophages and NCLDVs. In the virophage first hypothesis, NCLDVs diverge early with its sister lineage evolving into protovirophages. In the nuclear escape hypothesis, NCLDVs descend from endogenous elements (encoding an integrase) that became exogenous, virophages then evolved to become parasitise them.

Phylogenetics can be used to test competing evolutionary models of virus origins in a rigorous statistical framework. Each tree represents an evolutionary hypothesis which is inferred from matrices of aligned homologous characters with different states (the data). In molecular phylogenetics, nucleotide or amino acid alignments are used as the data to obtain the topology of a tree (and also rates or branch lengths). In statistical terms, the data matrices are used to calculate the tree likelihood (probability of the data given the parameters) in both maximum-likelihood and Bayesian frameworks (Schmidt and von Haeseler, 2009; Nascimento et al., 2017). In maximum likelihood, the preferred tree is the one with the

parameter values that maximise the tree likelihood (Schmidt and von Haeseler, 2009). In Bayesian inference, a posterior probability distribution for the parameters is estimated from their prior distributions and the tree likelihood (probability of the parameters given the data) (Nascimento et al., 2017). Together, these approaches can be used to assess the plausibility of competing evolutionary models that make different predictions about tree topologies and the position of the root.

Understanding the diversification of these viruses remains a major open question in virus evolution. To address this problem, we have analysed a data set of virus representatives sampled across the diversity of the eukaryotic bamfordvirus lineage and focused on the 4 core virion proteins: major and minor capsid proteins, protease and ATPase. We use explicit hypothesis-testing methods and estimate the position of the root in our phylogenies, which allows us to infer the direction of evolution followed by the proteins in the morphogenetic module. Our results suggest a new model for the evolutionary origins of these viruses, which is consistent with an exogenous, non-virophage ancestor.

Materials and Methods

We used maximum-likelihood and Bayesian hypothesis-testing methods to compare the plausibility of the nuclear-escape versus alternative evolutionary scenarios. We focused on the 4 core virion proteins shared by viruses in this lineage: major and minor capsid proteins, ATPase and protease. Rooted phylogenies were inferred from the data using a relaxed molecular-clock or an outgroup rooting method. We then assessed whether adenoviruses and NCLDVs formed a monophyletic grouping as predicted by the nuclear-escape scenario.

Finally, we studied the origin of the proteins shared by cytoplasmic linear plasmids and NCLDVs. A more detailed description of methods is presented below.

Selection of viral sequences

We first compiled a data set of viruses belonging to the families *Phycodnaviridae*, *Mimiviridae*, *Ascoviridae*, *Iridoviridae*, *Marseilleviridae*, *Asfarviridae*, *Poxviridae*, *Adenoviridae* and *Lavidaviridae* that were obtained from the NCBI using accession numbers listed in the International Committee for the Taxonomy of Viruses (ICTV) master species list 2020.v1. Previous analyses have shown that *Mavericks/Polintons* fall into two major groups based on the protein-primed DNA polymerase: the group I *Mavericks* (that include elements from vertebrates) and the group II *Mavericks* (which include elements present in insects) (Kapitonov and Jurka, 2006). Group I *Mavericks* can be distinguished from group II by the presence of a ~140 aa insertion similar to a bacterial "very-short-patch" (VSR) repair endonuclease (Kapitonov and Jurka, 2006). To make sure both groups were represented in our analyses, we included the vertebrate *Mavericks* reported in Barreat and Katzourakis (Barreat and Katzourakis, 2021) and added 12 new intact *Maverick* elements that we discovered in the genomes of insects that lack the VSR insertion (Table S1).

We also included the sequences for Polinton-like viruses, virophages and NCLDVs that have been reported in the metagenomic works of Bellas and Sommaruga (Bellas and Sommaruga, 2021), Moniruzzaman et al. (Moniruzzaman et al., 2020), Schultz et al. (Schulz et al., 2020), Paez-Espino et al. (Paez-Espino et al., 2019), Roux et al. (Roux et al., 2017) and Yutin et al. (Yutin et al., 2015). Additional genomes for medusavirus (Yoshikawa et al., 2019), cedratvirus (Jeudy et al., 2020), tupanviruses (Abrahão et al., 2018), pandoraviruses

(Antwerpen et al., 2016; Legendre et al., 2013, 2018), pithovirus (Legendre et al., 2014), mollivirus (Legendre et al., 2015), insectomime virus (Boughalmi et al., 2013), tokyovirus (Takemura, 2016), kaumoebavirus (Bajrai et al., 2016), faustoviruses (Benamar et al., 2016), pacmanvirus, orpheovirus and the Dishui lake virophages (Sayers et al., 2021), were also included. In total, the data set consisted of 9,222 virus genomes.

Protein prediction and subsampling approach

We devised a subsampling approach to arrive at a representative subset of sequences to carry out the phylogenetic analyses and tests of topology. First, open reading frames for all sequences were predicted with getorf in EMBOSS (Rice et al., 2000). Next, we used Hidden Markov Models for each of the 4 core proteins involved in virion morphogenesis (major and minor capsid proteins, ATPase and protease) to extract homologous proteins in 9 groups: genomic/metagenomic virophages, genomic/metagenomic NCLDVs, metagenomic PLVs, PLVs described in Yutin et al. (Yutin et al., 2015), adenoviruses, and *Mavericks 1* and *2* lineages. We used the major capsid protein to make multiple sequence alignments in MAFFT (Katoh et al., 2002) and construct approximate phylogenies for each group in FastTree 2 (Price et al., 2010). A total of 2,803 major capsid proteins could be identified in the genome data, which comprised: adenoviruses (71 sequences), *Mavericks 1/2* (28 sequences), NCLDVs NCBI/metagenomic (2,031 sequences), PLVs Yutin/metagenomic (327 sequences), and virophages NCBI/metagenomic (346 sequences). The trees were used to subsample 6 representatives from each of the 9 groups maintaining the maximal diversity in Treemmer (Menardo et al., 2018). The final data set contained representatives of 54 taxa (Table S1).

Protein multiple sequence alignment

Using the list of representative taxa for the major capsid proteins, we retrieved the corresponding sequences for the ATPase, protease and minor capsid proteins in each genome. We confirmed that each prediction corresponded to its homology group using HHpred (Söding et al., 2005). Sequences were aligned in MAFFT (Katoh et al., 2002) and trimmed using trimAl (Capella-Gutiérrez et al., 2009). We then built a concatenated alignment using a custom script written in Python (van Rossum et al., 2009).

Bayesian phylogenetic inference

We constructed single-protein and concatenated rooted trees in BEAST 2 (Bouckaert et al., 2019). Best models for amino acid substitution were selected for each partition in ModelTest-NG (Darriba et al., 2020). The model LG+G+I (with 4 categories for the gamma rate heterogeneity) was used for the major capsid, protease and ATPase, while the model LG+G (with 4 categories for the gamma rate heterogeneity) was used for the minor capsid protein. We imposed monophyletic constraints at the level of well-established viral clades/families: NCLDVs, *Adenoviridae*, vertebrate *Mavericks* and invertebrate *Mavericks* (however, the monophyly between both groups of *Mavericks* was not constrained). The monophyly of virophages and Polinton-like virophages were not constrained. We conducted a Bayesian MCMC with a chain length of 140 million generations for the concatenated data set and 200 million generations for the single proteins (sampling every 5,000th generation and using relaxed molecular clocks). Convergence was assessed by inspecting the runs for good-mixing and stationarity, and ensuring that Effective Sample Sizes were > 200 for all parameters in Tracer (Rambaut et al., 2018).

Maximum-likelihood phylogenetic inference

We estimated a maximum-likelihood tree in RAxML-NG (Kozlov et al., 2019) using the concatenated data set and the same monophyletic constraints on well-established groups as described above. Proteins of the *Enterobacteria phage PRD1* (ATPase, major and minor capsid) were included in the alignment for outgroup rooting. We chose the best substitution models in ModelTest-NG (Darriba et al., 2020). For the protease and ATPase, we used model LG+G+I+F (with 4 categories for the gamma rate heterogeneity), and for the major and minor capsid proteins we used model LG+G+F (with 4 categories for the gamma rate heterogeneity). The analysis was started from 200 random starting trees and run with 2,200 bootstrap replicates.

Akaike information criterion

To compare the plausibility of the maximum-likelihood trees consistent with each hypothesis, we inferred the best supported phylogenies in RAxML-NG and then performed model-selection based on the Akaike information criterion (Kozlov et al., 2019; Posada and Buckley, 2004). Specifically, we compared a constrained model consistent with nuclear-escape where adenoviruses and NCLDVs are sister groups, to a model where the sister relationship between adenoviruses and NCLDVs was not constrained (alternative model). We included a partitioned analysis of the concatenated data set and 3 core proteins individually (protease, major and minor capsid), using the amino acid substitution models described in the previous section. The analyses were run with 200 random starting trees and we kept the estimate for the log-likelihood of the best tree model for the calculation of the AICs. The AIC for each model was size-corrected given that the alignment (sample) sizes

were small compared to the number of free parameters in each model, or $n/k < 40$ (Posada and Buckley, 2004; Symonds and Moussalli, 2011). For the comparison we calculated the Akaike weights which can be interpreted as the normalised relative likelihood of the model given the data (Posada and Buckley, 2004), or as an analogue of the probability that a model is the best approximating model given the data within a set of candidates (Symonds and Moussalli, 2011). The ATPase was not included in the single-protein analyses since adenoviruses encode a non-homologous ABC family ATPase (Burroughs et al., 2007), and therefore it cannot be used to test their evolutionary relationships to other bamfordviruses.

Posterior model odds

The tree topologies obtained in the Bayesian MCMCs for the single proteins and the concatenated data set, were loaded and filtered in PAUP (Swofford, 1998). Following the method of Bergsten et al. (Bergsten et al., 2013), we counted the number of trees which contained a sister grouping of adenoviruses and NCLDVs (nuclear-escape hypothesis) and compared them to the number of trees inconsistent with this grouping (alternative hypotheses). The posterior model odds were calculated as the ratio of these two quantities (Bergsten et al., 2013). The ATPase was again excluded from the single-protein analysis given that adenoviruses could not be included.

Marginal likelihood and Bayes factors

We estimated the marginal likelihood of each hypothesis using stepping-stone sampling. We ran the analysis on the concatenated data set, for 20 million generations in MrBayes 3 (Ronquist and Huelsenbeck, 2003). For both analyses, we used the monophyletic constraints

described above. In the marginal calculation for the nuclear-escape hypothesis, we used a topological constraint on the sister grouping of adenoviruses and NCLDVs, while we imposed a negative topological constraint on the monophyly of adenoviruses and NCLDVs for the alternatives (adenoviruses and NCLDVs cannot be sister groups). We calculated the Bayes factors as the number e to the power of the difference in log-marginal likelihoods, and compared the resulting value to the table of Kass and Raftery (Kass and Raftery, 1995).

Cytoplasmic linear plasmids

Cytoplasmic linear plasmids have lost the ancestral module of genes involved in formation of the capsid, but they have gained three genes encoding a mRNA capping-enzyme, a helicase and the Rpb2 subunit of the DNA-dependent RNA-polymerase II (Koonin, Krupovič, et al., 2015; Krupovič and Koonin, 2015). We used the proteins in cytoplasmic linear plasmids (Table S3), to search for homologous sequences in the non-redundant protein database using blastp (Altschul et al., 1990) and restricting searches to 'Eukaryota (taxid:2759)' and 'Viruses (taxid:10239)'. The identity of significant matches (evalue < 1e-10) was confirmed with HHpred (Söding et al., 2005), and in the case of eukaryotes, only sequences mapping to chromosome assemblies were used. Virus, eukaryotic and plasmid homologues were aligned in MAFFT (Katoh et al., 2002) and conserved blocks recovered from trimAl (Capella-Gutiérrez et al., 2009). We chose the best models for protein evolution in ModelTest-NG (Darriba et al., 2020) and ran a maximum-likelihood tree search in RAxML-NG (Kozlov et al., 2019) with 1,000 non-parametric bootstrap replicates for each protein separately.

Data availability

The Hidden Markov Models for the four core proteins, concatenated and single-protein alignments have been deposited in Figshare (DOI: 10.6084/m9.figshare.19576117). The scripts used for sequence concatenation and for the AIC analyses are available in GitHub (https://github.com/josegabrielnb/virophage-origins).

Results

The rooted Bayesian and maximum-likelihood trees estimated from the concatenated data had the best overall support and shared important topological similarities (Figure 2 and 3). In the Bayesian tree, adenoviruses belong to a clade with *Mavericks* and most PLVs with a high posterior probability (0.94, Figure 2), and to the exclusion of NCLDVs. This arrangement was also observed in the maximum likelihood topology shown in Figure 3, although with a low bootstrap support (34%). The virophages (family *Lavidaviridae*) emerged as a strongly supported monophyletic group on both trees (posterior probability = 1 in the Bayesian analysis, 95% bootstrap support in the maximum-likelihood analysis), while Polinton-like viruses were polyphyletic.

The position of the root inferred both through the Bayesian relaxed-clock analysis and by outgroup rooting with the *Enterobacteria phage PRD1* (*Tectiviridae*), was placed between virophages and all the other viral lineages. In the Bayesian analysis, the root was placed on this position in 54% of trees (Figure 2), while the clade of (Adenoviruses + *Mavericks* + PLVs + NCLDVs), i.e. to the exclusion of virophages, received a bootstrap support of 73% in the maximum-likelihood analysis (Figure 3). We also examined alternative placements of

the root sampled during the Bayesian Markov-chain Monte Carlo (MCMC) to assess the frequency of other hypotheses (Table S2). Rooting on the clade formed by NCLDVs + PLV BS_539 (ERX2821583_NODE_576) was observed in 27.4% of trees. The root positions at the next highest frequencies were on the branch of PLV BS_539 (5.6%), NCLDVs (5.3%) and PLV BS_539 + NCLDVs + virophages (5%). Roots on other branches were observed at frequencies lower than 1% (Table S2).

The monophyly of virophages and the root on this branch were also observed in the single-protein trees. The major and minor capsid single-protein trees recovered a monophyletic *Lavidaviridae* (0.97 and 0.52 posterior probability) and inferred the root to be placed between virophages and the other viruses (29% and 28%, Figures S1-S2). Although the root was placed at a different position for the ATPase and protease trees; these analyses also failed to recover a monophyletic family of virophages (Figures S3-S4). However, when virophages were constrained to be monophyletic the root was placed again between virophages and other lineages, and thus in agreement with our findings using the concatenated data, major and minor capsid proteins (Figures S5-S6).

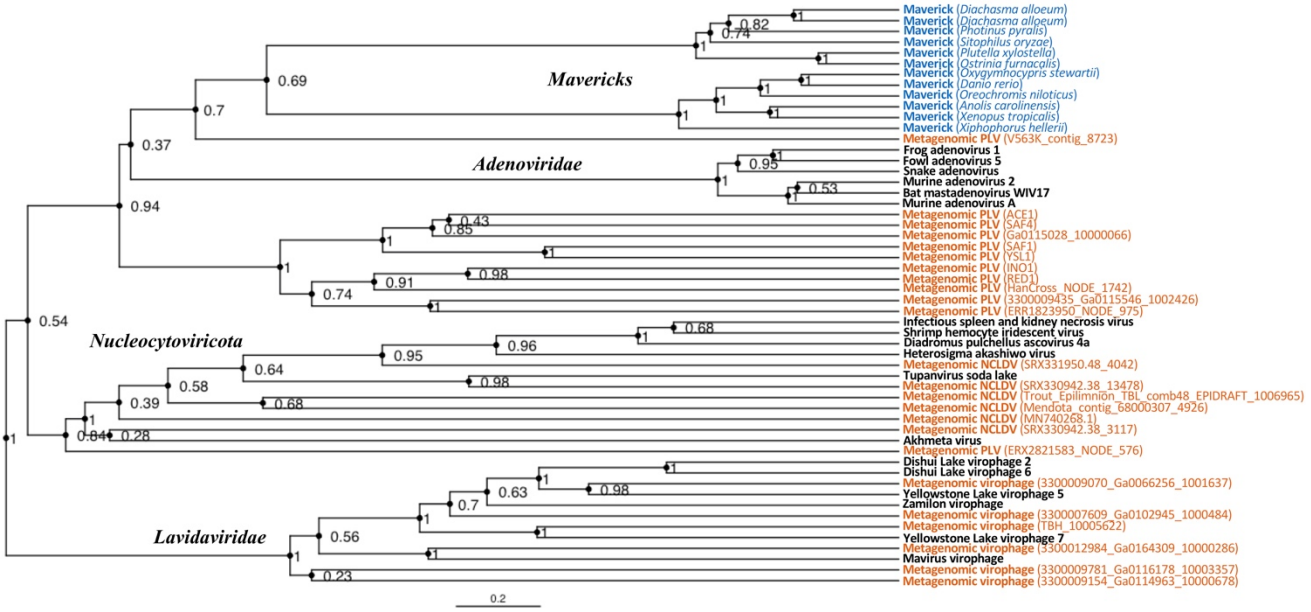


Figure 2. Rooted Bayesian maximum clade credibility tree of the major lineages of eukaryotic *Bamfordvirae*. The tree is based on the concatenated alignment of 4 core proteins involved in virion morphogenesis (major and minor capsid proteins, ATPase and protease). Tree computed in BEAST 2 (Bouckaert et al., 2019) using a relaxed molecular clock and 140 million generations (relative burn-in of 25%). Black: reference viral genomes, Blue: endogenous elements, Orange: metagenomic sequences.

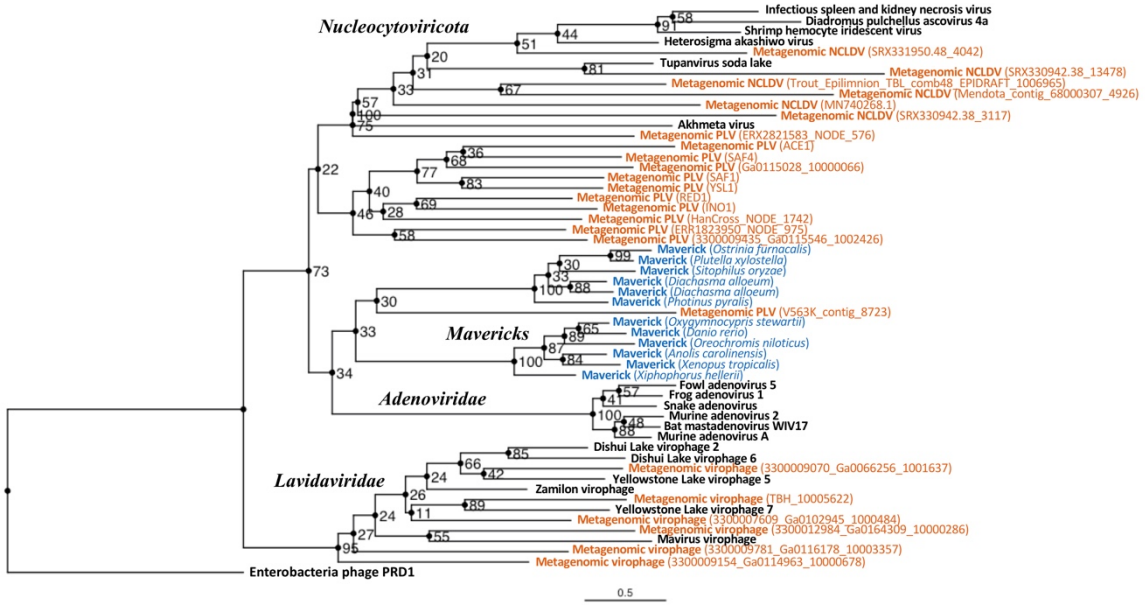


Figure 3. Outgroup-rooted maximum likelihood tree of the major lineages of eukaryotic *Bamfordvirae*. The tree is based on the concatenated alignment of 4 core proteins involved in virion morphogenesis (major and minor capsid proteins, ATPase and protease). *Enterobacteria phage PRD1* (*Tectiviridae*) was used as the outgroup for rooting. Tree computed in RAxML-NG with 200 random starting trees and 2,200 bootstraps (Kozlov et al., 2019). Black: reference viral genomes, Blue: endogenous elements, Orange: metagenomic sequences.

The Bayesian and maximum-likelihood tests of topology also agreed with each other. We used the Akaike information criterion (AIC) to compare the plausibility of the maximum-likelihood topologies consistent with the nuclear-escape and alternative scenarios. All the comparisons preferred a non-sister position for adenoviruses and NCLDVs (Table 1). Indeed, the Akaike weights which can be interpreted as the normalised relative likelihood of the models given the data (Posada and Buckley, 2004), strongly favoured the alternative scenario over the nuclear-escape hypothesis (Concatenated data = 99.7%, Protease = 99.7%, Major capsid protein = 72%, Minor capsid protein = 94.3%). The approach that used the

posterior model odds, calculated following the method of Bergsten et al. (Bergsten et al., 2013), also favoured the alternative scenarios in all of the data sets: concatenated and single-protein (Table 2). This was strongest for the concatenated data (posterior model odds H0/H1= 0.000536), but the same pattern held for the single-protein analyses (posterior model odds = {Protease: 0.00166, Major capsid: 0.11485, Minor capsid: 0.000639}). The Bayesian stepping-stone analysis on the concatenated data also strongly supported the alternative scenario (NCLDVs and adenoviruses not sister groups), over the nuclear-escape hypothesis (NCLDVs and adenoviruses sister groups) (Table 3).

As an independent test of the predictions made by the nuclear-escape hypothesis, we analysed the origin of the transcriptional proteins shared between NCLDVs and cytoplasmic linear plasmids. According to the nuclear-escape hypothesis, these genes were acquired by the most recent common ancestor of cytoplasmic linear plasmids and NCLDVs, and should thus have a single origin. However, our analyses indicate that these proteins were acquired convergently by cytoplasmic linear plasmids and NCLDVs. The maximum-likelihood trees for the Rbp2 subunit of the DNA-directed RNA polymerase, helicase and mRNA-capping proteins all suggest they have independent origins in NCLDVs and cytoplasmic linear plasmids (Figure S7). The homologues specific to cytoplasmic linear plasmids clustered with those of eukaryotes and their distribution agrees with a monophyletic origin in this group. Interestingly, the phylogenetic patterns for NCLDVs were consistent with a single-capture (monophyletic origin) for the mRNA-capping enzyme and the helicase, while the distribution of the Rpb2 was consistent with multiple-captures/exchanges (polyphyletic origin).

Table 1. Comparison between the nuclear-escape and alternative maximum-likelihood models based on the Akaike information criterion (AIC). The log-likelihoods for the models were obtained from the best maximum-likelihood tree consistent with each hypothesis found in RAxML-NG (Kozlov et al., 2019). Results are shown for the concatenated data set of four core proteins (ATPase, protease, major capsid and minor capsid) and for the protease, major and minor capsid proteins, respectively. The AIC was size-corrected given that the alignment (sample) size was small relative to the number of free parameters, i.e. n/k < 40 (Posada and Buckley, 2004; Symonds and Moussalli, 2011). The best model is highlighted in boldface. The "alternatives" refer to non "nuclear-escape" scenarios, that is, models which are not consistent with the predictions made by the nuclear-escape.

Model	Characters	Log-likelihood	AIC[1]	AICc[2]	ΔAICc[3]	Weight[4]
Nuclear-escape (M_0)	Concatenated	-41,618.19	83,616.37	83,864.09	11.41	0.003
Alternatives (M_1)	**Concatenated**	**-41,612.48**	**83,604.96**	**83,852.68**	**0.00**	**0.997**
Nuclear-escape (M_0)	Protease	-6,323.233	12,898.47	10,898.22	11.37	0.003
Alternatives (M_1)	**Protease**	**-6,317.550**	**12,887.10**	**10,886.85**	**0.00**	**0.997**
Nuclear-escape (M_0)	Major capsid	-16,305.52	32,861.04	34,173.54	1.90	0.279
Alternatives (M_1)	**Major capsid**	**-16,304.57**	**32,859.14**	**34,171.64**	**0.00**	**0.721**
Nuclear-escape (M_0)	Minor capsid	-10,234.56	20,719.11	17,855.48	5.62	0.057
Alternatives (M_1)	**Minor capsid**	**-10,231.75**	**20,713.50**	**17,849.86**	**0.00**	**0.943**

[1] $AIC_i = -2 \cdot LogL_i + 2 \cdot K$

[2] $AICc_i = AIC_i + \frac{2K(K+1)}{(n-K-1)}$

[3] $\Delta AICc_i = AICc_i - \min(AICc_i, AICc_j)$

[4] $Weight_i = \frac{e^{-\Delta AICc_i/2}}{\sum_i^j e^{-\Delta AICc_i/2}}$

Table 2. Posterior model odds of the nuclear-escape and alternative hypotheses using concatenated and single-protein data sets. Tree topologies consistent with each hypothesis were filtered and counted from a Bayesian MCMC following the method of Bergsten et al. (2013). All ratios favour the alternative scenarios to the nuclear-escape hypothesis. The best model is highlighted in boldface. The "alternatives" refer to non "nuclear-escape" scenarios, that is, models which are not consistent with the predictions made by the nuclear-escape.

Model	**Characters**	**MCMC tree frequency**	**Posterior model odds $P(M_0 \mid X)/P(M_1 \mid X)$**
Nuclear-escape (M_0)	Concatenated	15/28,001 (0.0536%)	$5.36 \cdot 10^{-4} < 1$
Alternatives (M_1)	**Concatenated**	**27,986/28,001 (99.946%)**	
Nuclear-escape (M_0)	Protease	333/200,001 (0.166%)	$1.66 \cdot 10^{-3} < 1$
Alternatives (M_1)	**Protease**	**199,668/200,001 (99.998%)**	
Nuclear-escape (M_0)	Major capsid	20,605/200,001 (10.302%)	$1.15 \cdot 10^{-1} < 1$
Alternatives (M_1)	**Major capsid**	**179,396/200,001 (89.697%)**	
Nuclear-escape (M_0)	Minor capsid	128/200,001 (0.0639%)	$6.39 \cdot 10^{-4} < 1$
Alternatives (M_1)	**Minor capsid**	**199,873/200,001 (99.936%)**	

Table 3. Bayesian stepping-stone analysis of the nuclear-escape and alternative hypotheses. Each scenario was run on the concatenated data set in MrBayes 3 (Ronquist and Huelsenbeck, 2003) for 20 million generations (average standard deviation of split frequencies < 0.01). The Bayes factor strongly rejects a sister relationship between adenoviruses and NCLDVs (nuclear-escape hypothesis). The best model is highlighted in boldface. The "alternatives" refer to non "nuclear-escape" scenarios, that is, models which are not consistent with the predictions made by the nuclear-escape.

Model	**Likelihood of best state (cold chain)**	**Log-Marginal-likelihood (ln)**	**Mean Log-marginal-likelihood (ln)**	**Bayes factor $P(X \mid M_0)/P(X \mid M_1)$**
Nuclear-escape (M_0)	Run 1: -36,318.81	Run 1: -36,376.42	-36,353.44	$3.5 \times 10^{-94} << 1$***
	Run 2: -36,318.81	Run 2: -36,352.75		
Alternatives (M_1)	**Run 1: -35,965.90**	**Run 1: -36,137.56**	**-36,138.25**	
	Run 2: -35,981.46	**Run 2: -36,450.19**		

***Strong support against M_0

Discussion

To gain a better understanding of the evolutionary history of the eukaryotic bamfordviruses, we analysed a set of four core virion morphogenesis proteins (major and minor capsids, DNA-packaging ATPase and protease) sampled across the diversity of the lineage and using an explicit hypothesis-testing framework. In addition, we analysed the origins of 3 proteins shared by cytoplasmic linear plasmids and NCLDVs. We found strong evidence against the nuclear-escape hypothesis. We also found support for a position of the root between

virophages and the other viral lineages. This position of the root suggests a new evolutionary model for the origin of these viral lineages. However, some alternative root positions were also observed at lower frequencies in the Bayesian analyses which could be interpreted as "virophage-first"-like scenarios.

We have shown that adenoviruses and NCLDVs are not sister groups as proposed by the nuclear-escape hypothesis (Koonin and Krupovič, 2017; Krupovič and Koonin, 2015). Instead, adenoviruses belong to a clade with *Mavericks*, to the exclusion of NCLDVs. The phylogenies of the proteins shared by cytoplasmic linear plasmids and NCLDVs also disagree with a nuclear escape. Indeed, phylogenies of the protein-primed DNA polymerase place cytoplasmic linear plasmids as the sister-group to adenoviruses (Krupovič, Yutin, et al., 2016), implying that adenoviruses, cytoplasmic linear plasmids and NCLDVs form a clade. If a single nuclear-escape had occurred we would expect the Rbp2, helicase and mRNA capping enzymes shared by cytoplasmic linear plasmids and NCLDVs to have a monophyletic origin (since they descend from the same "escaped" ancestor). However, their distribution agrees with a more recent origin in cytoplasmic linear plasmids. These observations agree with the more restricted taxonomic distribution of cytoplasmic linear plasmids, which are known only to infect fungi in the Order Saccharomycetales (Division Ascomycota) (Meinhardt et al., 1997), in contrast to the taxonomically diverse eukaryotic hosts of NCLDVs.

Virophages emerged as a highly supported monophyletic group. This supports the validity of the family *Lavidaviridae* which has been based on shared gene contents of virophages and their unique parasitic (or commensal) lifestyles (Fischer, 2020; Krupovič, Kuhn, et al., 2016). By using two independent rooting methods, relaxed-molecular clocks and outgroup

rooting, we found that the root was placed most confidently between virophages and the other lineages; making them the most basal lineage of eukaryotic bamfordviruses. The basal position of virophages has also been found in an independent work that looked at phylogenies based on the ATPase and a concatenation of the ATPase and the major capsid protein (Woo et al., 2021). This basal position of virophages and their divergence prior to the origin of NCLDVs, suggests that the protovirophages were not parasites of other viruses. Instead, they would have become parasites of other viruses at a later point once the most recent common ancestor of NCLDVs had evolved (see Ideas and speculation).

The tree topologies that we estimated (Figure 2 and Figure 3), also suggest that the most recent common ancestor of *Mavericks*, adenoviruses and PLVs was not a virophage, which is at odds with a virophage-first scenario. Indeed, the basal position of virophages is inconsistent with the virophage-first hypothesis which proposes that protovirophages coevolved with NCLDVs, acquired an integrase and then gave rise to *Mavericks* and other elements in the lineage (Campbell et al., 2017; Fischer and Suttle, 2011). However, we found that the second-highest frequency root sampled in the Bayesian MCMC was on the branch of PLV BS_539 + NCLDVs. Rooting on this branch would be consistent with a "virophage-first"-like scenario, where the lineage of NCLDVs and associated viruses evolved early and in parallel with protovirophages. Therefore, we cannot definitely rule out a "virophage-first"-like scenario. In contrast, root placements on the branch of *Mavericks* and consistent with nuclear-escape were observed in 0% of the trees, which seems to rule out this possibility (Table S2).

The relative timing of events suggests that the most recent common ancestor of eukaryotic bamfordviruses existed more than a billion years ago. There is general agreement based on

paleontological and molecular evidence that the Last Eukaryotic Common Ancestor (LECA) existed at least 1 billion years ago (Porter, 2020). Analyses of genetic exchanges between the DNA-dependent RNA polymerases of NCLDVs and eukaryotes have determined that the two super-clades of NCLDVs (PAM and MAPI) already existed before the origin of LECA (Guglielmini et al., 2019). Moreover, discovery and phylogenetic analysis of actin and actin-related proteins encoded in NCLDVs, also suggest that early gene transfers occurred between these viruses and proto-eukaryotes before the emergence of LECA (da Cunha et al., 2022). The basal placement that we observed for virophages, or alternatively, their early origin after the divergence of the lineage of NCLDVs ("virophage-first"-like scenario), suggest that virophages evolved after the origin of the first NCLDVs during the diversification of the early eukaryotes. These observations, together with the existence of various phage relatives that infect bacteria and archaea, point to a very early origin of the kingdom *Bamfordvirae*, perhaps extending to the initial stages of cellular life. Studies that consider the time-dependent rate phenomenon in reconstructing viral evolution may be able to shed light on these issues (Aiewsakun and Katzourakis, 2016; Ghafari et al., 2021).

Ideas and speculation

Our analyses favour a new model for the evolution of the major lineages of eukaryotic *Bamfordvirae*; we call this model the "exogenous, non-virophage scenario". The model is presented in Figure 4, where we have mapped shared character states onto a diagram of the concatenated Bayesian phylogeny (which has the best overall support). According to this idea, the most recent common ancestor of the eukaryotic bamfordviruses was a small exogenous dsDNA virus with a linear genome, flanked by inverted repeats. The virus had an autonomous, i.e. non-virophage, lifestyle. It used a protein-primed DNA polymerase B

for replication and its morphogenetic module was formed by the major and minor jelly-roll capsid proteins, an adenoviral-like protease and the FtsK/HerA family DNA-packaging ATPase. The gain of the adenoviral-like protease differentiated this eukaryotic lineage from their phage relatives (Krupovič and Koonin, 2015). The first divergence from this ancestor led to the emergence of the protovirophages and another lineage that would give rise to *Mavericks*, adenoviruses, PLVs and NCLDVs (Figure 4).

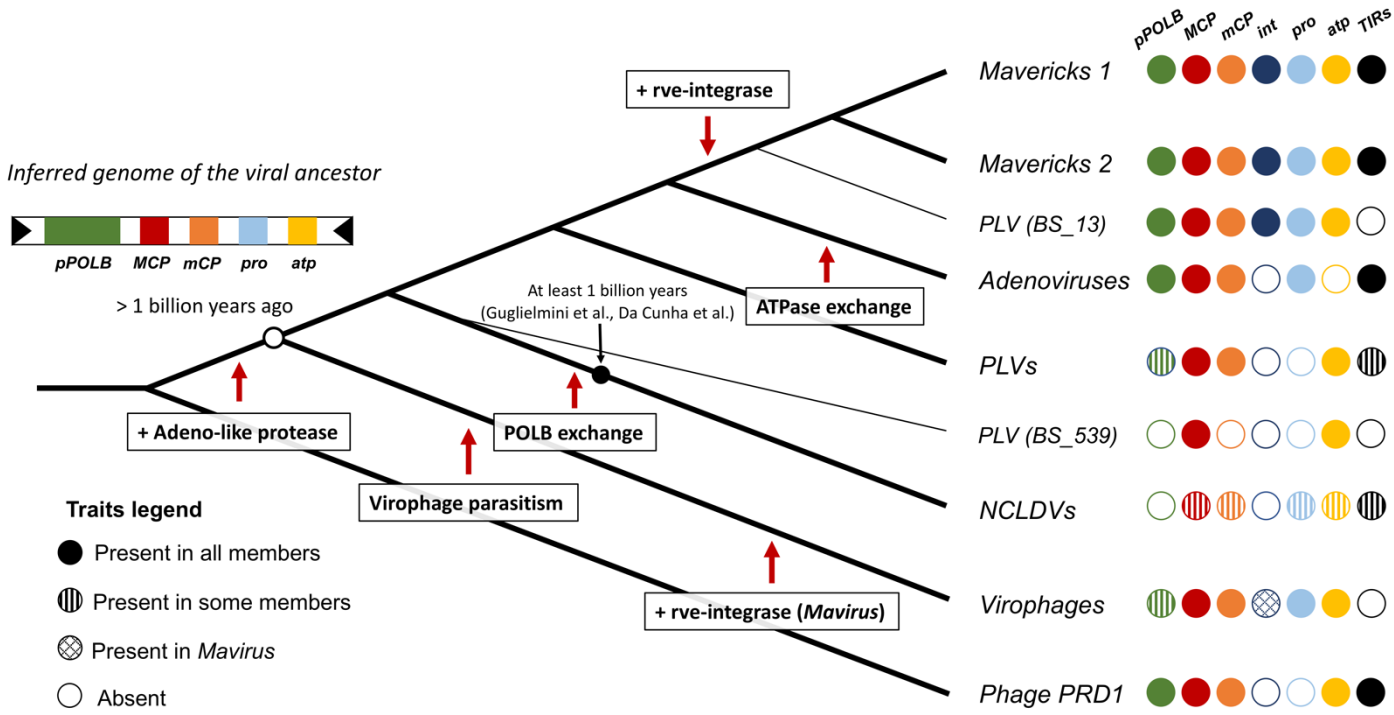


Figure 4. Evolutionary model for the origin of the major lineages of eukaryotic *Bamfordvirae*. The viral ancestor is inferred to have been an exogenous virus, while the rve-integrase was captured independently by the clade of *Mavericks* + Polinton-like virus BS_13 and Mavirus, possibly by horizontal gene transfer. Virophages evolved from an autonomous virus that became specialised to parasitise the ancestor of NCLDVs. The vertical cross-hatching indicates that the trait is found in some but not all members of the group.

The second divergence gave rise to the lineage from which NCLDVs evolved. The most recent common ancestor of NCLDVs is believed to have already had a complex genome (~40 genes), which involved the gain of numerous genes from their eukaryotic hosts, other viruses and bacteria (Iyer et al., 2006; Koonin and Yutin, 2010; Yutin and Koonin, 2012). Once this virus evolved the capacity to exploit significant cell resources, there would have been a window of opportunity for protovirophages to evolve into specialised parasites of NCLDVs, occupying this new ecological niche. The close match between virophage and NCLDV regulatory sequences, which underlie the capacity of virophages to parasitise NCLDVs, may have evolved by parallel evolutionary changes on the sequences of their shared virus ancestor. It is plausible that functional regulatory sequences in virophages could evolve *de novo* by a few changes in the ancestral promoters/terminators. For example, functional *lac* operon promoters have been evolved successfully even from random sequences in the presence of lactose in *E. coli* (Yona et al., 2018). A deletion mutant in the ORF 8 gene of the Guarani virophage expanded its host range to previously non-permissive giant viruses, so the acquisition and mutation of proteins involved in virus-virophage interactions may also be critical for virophage adaptation (Mougari et al., 2020). Interestingly, the exchange of the protein-primed DNA polymerase B by a nucleotide-primed DNA polymerase B (Mönttinen et al., 2021; Yutin and Koonin, 2012), may have been an early counter-measure of NCLDVs to decrease the parasitic burden imposed by virophages. The acquisition of this new polymerase, may be in part responsible for the considerable increases in genome size that we see in NCLDVs relative to other eukaryotic bamfordviruses. Therefore, we can hypothesise that the large genomes of NCLDVs, one of their most distinctive features, could have evolved as a result of their ancient interaction with virophages.

The distribution of the rve-integrase on the phylogeny suggests it was acquired independently on two separate occasions. Indeed, the rve-integrase is a universal feature of *Mavericks* + PLV BS_13, while it does not seem to be present in any of their closest relatives. Therefore, the rve-integrase seems to be a unique derived feature of the clade of *Mavericks* + PLV BS_13. Mavirus also encodes a rve-integrase but it appears to be the only virophage with this gene (no other hits found in a blastp search of the nr database using the labels "*Lavidaviridae*" + "unclassified *Lavidaviridae*"). Since Mavirus is firmly nested within the phylogenies for the morphogenetic module, it seems likely that this is a unique derived feature of Mavirus. The most parsimonious explanation for these observations is that the integrase gene was acquired on two separate occasions and that the common ancestor of Mavirus and *Mavericks* was an exogenous virus without an integrase (# character state changes = 2 gains). This is in line with a previous proposal that Mavirus evolved by recombination of an ancestral virophage with a *Maverick* (Yutin et al., 2013).

Concluding remarks

Our findings strengthen the view of a viral origin for NCLDVs and other eukaryotic elements in the kingdom *Bamfordvirae*. We have decisively shown that adenoviruses and NCLDVs are not sister groups, and that adenoviruses, virophages, PLVs and NCLDVs did not escape from the nucleus. The rve-integrase appears to have been captured independently by *Mavericks* + PLV BS_13 on the one hand, and in Mavirus on the other. These observations support a new evolutionary model that is different from the nuclear-escape and virophage-first scenarios. This new model proposes that the ancestor of the eukaryotic *Bamfordvirae* was an exogenous, non-virophage DNA virus with a small genome. The lifestyle of virophages would have evolved at a later stage as these became specialised

parasites of the ancestral NCLDVs. Darwin's closing words in the *Origin of Species* seem well-suited for this group of viruses: "from so simple a beginning endless forms most beautiful and most wonderful have been, and are being, evolved" (Darwin, 1859).

References

Abrahão J, Silva L, Silva LS, Khalil JYB, Rodrigues R, Arante T, Assis F, Boratt P, Andrade M, Kroon, EG, Ribeiro B, Bergier I, Seligmann H, Ghigo E, Colson P, Levasseur A, Kroemer G, Raoult D and la Scola B. (2018). Tailed giant Tupanvirus possesses the most complete translational apparatus of the known virosphere. *Nature Communications* 9(1). https://doi.org/10.1038/s41467-018-03168-1

Aiewsakun P and Katzourakis A. (2016). Time-dependent rate phenomenon in viruses. *Journal of Virology*, 90(16), 7184–7195. doi: 10.1128/JVI.00593-16

Altschul SF, Gish W, Miller W, Myers EW and Lipman DJ. (1990). Basic local alignment search tool. *Journal of Molecular Biology*, 215(3), 403–410. doi: 10.1016/S0022-2836(05)80360-2

Antwerpen MH, Georgi E, Zoeller L, Woelfel R, Stoecker K and Scheid P. (2016). Whole-genome sequencing of a Pandoravirus isolated from keratitis-inducing *Acanthamoeba*. *Genome Announcements*, 3(2). https://doi.org/10.1128/genomeA.00136-15

Aylward FO, Moniruzzaman M, Ha AD and Koonin, EV. (2021). A phylogenomic framework for charting the diversity and evolution of giant viruses. *PLoS Biology*. doi: 10.1371/JOURNAL.PBIO.3001430

Bajrai LH, Benamar S, Azhar EI, Robert C, Levasseur A, Raoult D, and la Scola B. (2016). Kaumoebavirus, a new virus that clusters with faustoviruses and *Asfarviridae*. Viruses 8(11). doi: 10.3390/v8110278

Barreat JGN and Katzourakis A. (2021). Phylogenomics of the *Maverick* virus-like mobile genetic elements of vertebrates. *Molecular Biology and Evolution*. doi: 10.1093/molbev/msaa291

Bellas, CM and Sommaruga R. (2021). Polinton-like viruses are abundant in aquatic ecosystems. *Microbiome* 9(1). doi: 10.1186/s40168-020-00956-0

Benamar S, Reteno DGI, Bandaly V, Labas N, Raoult D and la Scola B. (2016). Faustoviruses: comparative genomics of new megavirales family members. *Frontiers in Microbiology* 7(FEB). doi: 10.3389/fmicb.2016.00003

Bergsten J, Nilsson AN and Ronquist F. (2013). Bayesian tests of topology hypotheses with an example from diving beetles. *Systematic Biology* 62(5). doi: 10.1093/sysbio/syt029

Bouckaert R, Vaughan TG, Barido-Sottani J, Duchêne S, Fourment M, Gavryushkina A, Heled J, Jones G, Kühnert D, de Maio N, Matschiner M, Mendes FK, Müller NF, Ogilvie HA, du Plessis L, Popinga A, Rambaut A, Rasmussen D, Siveroni I, … Drummond AJ. (2019). BEAST 2.5: An advanced software platform for Bayesian evolutionary analysis. *PLoS Computational Biology* 15(4). doi: 10.1371/journal.pcbi.1006650

Boughalmi M, Pagnier I, Aherfi S, Colson P, Raoult D and la Scola B. (2013). First isolation of a marseillevirus in the diptera syrphidae eristalis tenax. *Intervirology* 56(6). doi: 10.1159/000354560

Burroughs A, Iyer L and Aravind L. (2007). Comparative genomics and evolutionary trajectories of viral ATP dependent DNA-packaging systems. *Genome Dynamics* (Vol. 3). doi: 10.1159/000107603

Campbell S, Aswad A and Katzourakis A. (2017). Disentangling the origins of virophages and *Polintons*. *Current Opinion in Virology*. doi: 10.1016/j.coviro.2017.07.011

Capella-Gutiérrez S, Silla-Martínez JM and Gabaldón T. (2009). trimAl: A tool for automated alignment trimming in large-scale phylogenetic analyses. *Bioinformatics* 25(15). doi: 10.1093/bioinformatics/btp348

Colson P, Levasseur A, la Scola B, Sharma V, Nasir A, Pontarotti P, Caetano-Anollés G and Raoult D. (2018). Ancestrality and mosaicism of giant viruses supporting the definition of the fourth TRUC of microbes. *Frontiers in Microbiology* 9. doi: 10.3389/fmicb.2018.02668

da Cunha V, Gaia M, Ogata H, Jaillon O, Delmont TO and Forterre P. (2022). Giant viruses encode actin-related proteins. *Molecular Biology and Evolution*, 39(2). doi: 10.1093/molbev/msac022

Darriba D, Posada D, Kozlov AM, Stamatakis A, Morel B and Flouri T. (2020). ModelTest-NG: a new and scalable tool for the selection of dna and protein evolutionary models. *Molecular Biology and Evolution*, 37(1). doi: 10.1093/molbev/msz189

Darwin, CR. (1859). On the origin of species by means of natural selection, or the preservation of favoured races in the struggle for life. (1st edition). John Murray.

Desnues C, la Scola B, Yutin N, Fournous G, Robert C, Azza S, Jardot P, Monteil S, Campocasso A, Koonin EV and Raoult D. (2012). Provirophages and transpovirons as the diverse mobilome of giant viruses. *Proc Natl Acad Sci U S A* 109(44). doi: 10.1073/pnas.1208835109

Fischer MG. (2020). The virophage family *Lavidaviridae*. *Current Issues in Molecular Biology*, 40. doi: 10.21775/cimb.040.001

Fischer MG and Hackl T. (2016). Host genome integration and giant virus-induced reactivation of the virophage Mavirus. *Nature* 540(7632). doi: 10.1038/nature20593

Fischer MG and Suttle CA. (2011). A virophage at the origin of large DNA transposons. *Science*. 10.1126/science.1199412

Ghafari M, Simmonds P, Pybus OG and Katzourakis A. (2021). A mechanistic evolutionary model explains the time-dependent pattern of substitution rates in viruses. *Current Biology* 31(21). doi: 10.1016/j.cub.2021.08.020

Guglielmini J, Woo AC, Krupovič M, Forterre P and Gaia M. (2019). Diversification of giant and large eukaryotic dsDNA viruses predated the origin of modern eukaryotes. *Proc Natl Acad Sci U S A*. doi: 10.1073/pnas.1912006116

Harrach B, Tarján ZL and Benkő M. (2019). Adenoviruses across the animal kingdom: a walk in the zoo. *FEBS Letters* (Vol. 593, Issue 24). doi: 10.1002/1873-3468.13687

Iyer LM, Balaji S, Koonin EV and Aravind L. (2006). Evolutionary genomics of nucleo-cytoplasmic large DNA viruses. *Virus Research* 117(1). doi: 10.1016/j.virusres.2006.01.009

Jeudy S, Rigou S, Alempic JM, Claverie JM, Abergel C and Legendre M. (2020). The DNA methylation landscape of giant viruses. *Nature Communications* 11(1). doi: 10.1038/s41467-020-16414-2

Kapitonov VV and Jurka, J. (2006). Self-synthesizing DNA transposons in eukaryotes. *Proc Natl Acad Sci U S A* 103(12). doi: 10.1073/pnas.0600833103

Kass R. E and Raftery AE. (1995). Bayes factors. *Journal of the American Statistical Association* 90(430). doi: 2291091

Katoh K, Misawa K, Kuma KI and Miyata T. (2002). MAFFT: a novel method for rapid multiple sequence alignment based on fast Fourier transform. *Nucleic Acids Research* 30(14). doi: 10.1093/nar/gkf436

Katzourakis A and Aswad A. (2014). The origins of giant viruses, virophages and their relatives in host genomes. *BMC Biology*. doi: 10.1186/s12915-014-0051-y

Koonin EV, Dolja VV and Krupovič M. (2015). Origins and evolution of viruses of eukaryotes: the ultimate modularity. *Virology*. doi: 10.1016/j.virol.2015.02.039

Koonin EV, Dolja VV, Krupovič M and Kuhn, JH. (2021). Viruses defined by the position of the virosphere within the replicator space. *Microbiology and Molecular Biology Reviews* 85(4). doi: 10.1128/mmbr.00193-20

Koonin EV, Dolja VV, Krupovič M, Varsani A, Wolf YI, Yutin N, Zerbini FM and Kuhn JH. (2020). Global organization and proposed megataxonomy of the virus world. *Microbiology and Molecular Biology Reviews* 84(2). doi: 10.1128/mmbr.00061-19

Koonin EV and Krupovič M. (2017). *Polintons*, virophages and transpovirons: a tangled web linking viruses, transposons and immunity. *Current Opinion in Virology* 25. doi: 10.1016/j.coviro.2017.06.008

Koonin EV, Krupovič M and Yutin N. (2015). Evolution of double-stranded DNA viruses of eukaryotes: From bacteriophages to transposons to giant viruses. *Annals of the New York Academy of Sciences* 1341(1). doi: 10.1111/nyas.12728

Koonin EV and Yutin N. (2010). Origin and evolution of eukaryotic large nucleo-cytoplasmic DNA viruses. *Intervirology*. doi: 10.1159/000312913

Kozlov AM, Darriba D, Flouri T, Morel B, Stamatakis A and Wren J. (2019). RAxML-NG: A fast, scalable and user-friendly tool for maximum likelihood phylogenetic inference. *Bioinformatics* 35(21). doi: 10.1093/bioinformatics/btz305

Krupovič M and Koonin, EV. (2015). *Polintons*: A hotbed of eukaryotic virus, transposon and plasmid evolution. *Nature Reviews Microbiology* 13(2). doi: 10.1038/nrmicro3389

Krupovič M and Koonin EV. (2016). Self-synthesizing transposons: unexpected key players in the evolution of viruses and defense systems. *Current Opinion in Microbiology* 31. doi: 10.1016/j.mib.2016.01.006

Krupovič M, Kuhn JH and Fischer MG. (2016). A classification system for virophages and satellite viruses. *Archives of Virology* 161(1). doi: 10.1007/s00705-015-2622-9

Krupovič M, Yutin N and Koonin EV. (2016). Fusion of a superfamily 1 helicase and an inactivated DNA polymerase is a signature of common evolutionary history of *Polintons*, Polinton-like viruses, *Tlr1* transposons and transpovirons. *Virus Evolution* 2(1). doi: 10.1093/ve/vew019

la Scola B, Desnues C, Pagnier I, Robert C, Barrassi L, Fournous G, Merchat M, Suzan-Monti M, Forterre P, Koonin E and Raoult D. (2008). The virophage as a unique parasite of the giant mimivirus. *Nature* 455(7209). doi: 10.1038/nature07218

Legendre M, Arslan D, Abergel C and Claverie JM. (2012). Genomics of Megavirus and the elusive fourth domain of life. *Communicative and Integrative Biology* 5(1). doi: 10.4161/cib.18624

Legendre M, Bartoli J, Shmakova L, Jeudy S, Labadie K, Adrait A, Lescot M, Poirot O, Bertaux L, Bruley C, Couté Y, Rivkina E, Abergel C and Claverie JM. (2014). Thirty-thousand-year-old distant relative of giant icosahedral DNA viruses with a pandoravirus morphology. *Proc Natl Acad Sci U S A* 111(11). doi: 10.1073/pnas.1320670111

Legendre M, Doutre G, Poirot O, Lescot M, Arslan D, Seltzer V, Bertaux L, Bruley C and Claverie J. (2013). Pandoraviruses: amoeba viruses with genomes up to 2.5 Mb reaching that of parasitic eukaryotes. *Science* 341. doi: 10.1126/science.1239181

Legendre M, Fabre E, Poirot O, Jeudy S, Lartigue A, Alempic JM, Beucher L, Philippe N, Bertaux L, Christo-Foroux E, Labadie K, Couté Y, Abergel C and Claverie JM. (2018). Diversity and evolution of the emerging Pandoraviridae family. *Nature Communications* 9(1). doi: 10.1038/s41467-018-04698-4

Legendre M, Lartigue A, Bertaux L, Jeudy S, Bartoli J, Lescot M, Alempic JM, Ramus C, Bruley C, Labadie K, Shmakova L, Rivkina E, Couté Y, Abergel C and Claverie JM. (2015). In-depth

study of *Mollivirus sibericum*, a new 30,000-yold giant virus infecting *Acanthamoeba*. *Proc Natl Acad Sci U S A* 112(38). doi: 10.1073/pnas.1510795112

Meinhardt F, Schaffrath R and Larsen M. (1997). Microbial linear plasmids. *Applied Microbiology and Biotechnology*. doi: 10.1007/s002530050936

Menardo F, Loiseau C, Brites D, Coscolla M, Gygli SM, Rutaihwa LK, Trauner A, Beisel C, Borrell S and Gagneux S. (2018). Treemmer: a tool to reduce large phylogenetic datasets with minimal loss of diversity. *BMC Bioinformatics* 19(1). doi: 10.1186/s12859-018-2164-8

Moniruzzaman M, Martinez-Gutierrez CA, Weinheimer AR and Aylward FO. (2020). Dynamic genome evolution and complex virocell metabolism of globally-distributed giant viruses. *Nature Communications* 11(1). doi: 10.1038/s41467-020-15507-2

Mönttinen HAM, Bicep C, Williams TA and Hirt RP. (2021). The genomes of nucleocytoplasmic large DNA viruses: viral evolution writ large. *Microbial Genomics*. doi: 10.1099/MGEN.0.000649

Mougari S, Chelkha N, Sahmi-Bounsiar D, di Pinto F, Colson P, Abrahao J and la Scola B. (2020). A virophage cross-species infection through mutant selection represses giant virus propagation, promoting host cell survival. *Communications Biology*. doi: 10.1038/s42003-020-0970-9

Nascimento, F.F., Reis, M.d. & Yang, Z. (2017). A biologist's guide to Bayesian phylogenetic analysis. Nature Ecology and Evolution 1. doi: 10.1038/s41559-017-0280-x

Paez-Espino D, Zhou J, Roux S, Nayfach S, Pavlopoulos GA, Schulz F, McMahon KD, Walsh D, Woyke T, Ivanova NN, Eloe-Fadrosh EA, Tringe SG and Kyrpides NC. (2019). Diversity, evolution, and classification of virophages uncovered through global metagenomics. *Microbiome* 7(1). doi: 10.1186/s40168-019-0768-5

Patil S and Kondabagil K. (2021). Coevolutionary and phylogenetic analysis of mimiviral replication machinery suggest the cellular origin of mimiviruses. *Molecular Biology and Evolution* 38(5). doi: 10.1093/molbev/msab003

Porter SM. (2020). Insights into eukaryogenesis from the fossil record. *Interface Focus* 10(4). doi: 10.1098/rsfs.2019.0105

Posada D and Buckley TR. (2004). Model selection and model averaging in phylogenetics: advantages of Akaike information criterion and Bayesian approaches over likelihood ratio tests. *Systematic Biology*. doi: 10.1080/10635150490522304

Price MN, Dehal PS and Arkin AP. (2010). FastTree 2 - Approximately maximum-likelihood trees for large alignments. *PLoS ONE* 5(3). doi: 10.1371/journal.pone.0009490

Pritham EJ, Putliwala T and Feschotte C. (2007). *Mavericks*, a novel class of giant transposable elements widespread in eukaryotes and related to DNA viruses. *Gene* 390. doi: 10.1016/j.gene.2006.08.008

Rambaut A, Drummond AJ, Xie D, Baele G and Suchard MA. (2018). Posterior summarization in Bayesian phylogenetics using Tracer 1.7. *Systematic Biology* 67(5). doi: 10.1093/sysbio/syy032

Rice P, Longden L and Bleasby A. (2000). EMBOSS: The european molecular biology open software suite. *Trends in Genetics*. doi: 10.1016/S0168-9525(00)02024-2

Ronquist F and Huelsenbeck JP. (2003). MrBayes 3: Bayesian phylogenetic inference under mixed models. *Bioinformatics* 19(12). doi: 10.1093/bioinformatics/btg180

Roux S, Chan LK, Egan R, Malmstrom RR, McMahon KD and Sullivan MB. (2017). Ecogenomics of virophages and their giant virus hosts assessed through time series metagenomics. *Nature Communications* 8(1). doi: 10.1038/s41467-017-01086-2

Sayers EW, Beck J, Bolton EE, Bourexis D, Brister JR, Canese K, Comeau DC, Funk K, Kim S, Klimke W, Marchler-Bauer A, Landrum M, Lathrop S, Lu Z, Madden TL, O'Leary N, Phan L, Rangwala SH, Schneider VA, … Sherry ST. (2021). Database resources of the National Center for Biotechnology Information. *Nucleic Acids Res* 49(D1). doi: 10.1093/nar/gkaa892

Scheid P. (2016). A strange endocytobiont revealed as largest virus. *Current Opinion in Microbiology*. doi: 10.1016/j.mib.2016.02.005

Schulz F, Roux S, Paez-Espino D, Jungbluth S, Walsh DA, Denef VJ, McMahon KD, Konstantinidis KT, Eloe-Fadrosh EA, Kyrpides NC and Woyke T. (2020). Giant virus diversity and host interactions through global metagenomics. *Nature* 578(7795). doi: 10.1038/s41586-020-1957-x

Schmidt, HA and von Haeseler A. (2009). Phylogenetic inference using maximum likelihood methods. The Phylogenetic Handbook: a practical approach to phylogenetic analysis and hypothesis testing. 2nd Edition. Cambridge University Press, Cambridge, 181-209.

Söding J, Biegert A and Lupas AN. (2005). The HHpred interactive server for protein homology detection and structure prediction. *Nucleic Acids Res*. doi: 10.1093/nar/gki408

Swofford DL. (1998). PAUP*. Phylogenetic Analysis Using Parsimony (*and other methods).. Sinauer Associates.

Symonds MRE and Moussalli A. (2011). A brief guide to model selection, multimodel inference and model averaging in behavioural ecology using Akaike's information criterion. *Behavioral Ecology and Sociobiology*. doi: 10.1007/s00265-010-1037-6

Takemura M. (2016). Draft genome sequence of Tokyovirus, a member of the family *Marseilleviridae* isolated from the Arakawa River of Tokyo, Japan. *Genome Announcements* 4(3). doi: 10.1128/genomeA.00429-16

van Rossum G, Drake FL, Harris CR, Millman KJ, van der Walt SJ, Gommers R, Virtanen P, Cournapeau D, Wieser E, Taylor J, Berg S, Smith NJ, Kern R, Picus M, Hoyer S, van Kerkwijk MH, Brett M, Haldane A, del Río JF, … Oliphant TE. (2009). Python 3 Reference Manual.

Woo AC, Gaia M, Guglielmini J, da Cunha V and Forterre P. (2021). Phylogeny of the Varidnaviria morphogenesis module: congruence and incongruence with the tree of life and viral taxonomy. *Frontiers in Microbiology* 12. doi: 10.3389/fmicb.2021.704052

Yona AH, Alm EJ and Gore J. (2018). Random sequences rapidly evolve into de novo promoters. *Nature Communications* 9(1). doi: 10.1038/s41467-018-04026-w

Yoshikawa G, Blanc-Mathieu R, Song C, Kayama Y, Mochizuki T, Murata K, Ogata H and Takemura M. (2019). Medusavirus, a novel large dna virus discovered from hot spring water. *Journal of Virology* 93(8). doi: 10.1128/jvi.02130-18

Yutin N and Koonin EV. (2012). Hidden evolutionary complexity of Nucleo-Cytoplasmic Large DNA viruses of eukaryotes. *Virology Journal* 9. doi: 10.1186/1743-422X-9-161

Yutin N, Raoult D and Koonin EV. (2013). Virophages, *Polintons*, and transpovirons: A complex evolutionary network of diverse selfish genetic elements with different reproduction strategies. *Virology Journal* 10. doi: 10.1186/1743-422X-10-158

Yutin N, Shevchenko S, Kapitonov V, Krupovič M and Koonin EV. (2015). A novel group of diverse Polinton-like viruses discovered by metagenome analysis. *BMC Biology* 13(1). doi: 10.1186/s12915-015-0207-4

Chapter IV

Ecological and evolutionary dynamics of cell-virus-virophage systems

Abstract

Microbial eukaryotes can be infected by giant viruses, which may be infected by virophages. Virophages are parasites of the virus transcription machinery and can interfere with virus replication, resulting in a benefit to the eukaryotic host population. Surprisingly, virophages can integrate into the genomes of their cell or virus hosts, and have been shown to reactivate during coinfection. This raises interesting questions about the role of integration in the dynamics of cell-virus-virophage systems. Here, we use mathematical models and computational simulations to understand the effect of virophage integration on populations of cells and viruses. We also investigate programmed-cell death and multicellularity as potential antiviral defence strategies used by cells. We found that virophages which enter the cell independently of the host virus, such as Mavirus, are expected to integrate commonly into the genomes of their cell hosts. In contrast, virophages like Sputnik which form a complex with the giant virus, should rarely be found integrated in cell genomes. Alternatively, we found that Sputnik virophages can stably persist integrated in the virus genome, as long as they do not completely inhibit virus replication. We also show that increasing virophage inhibition can stabilise oscillatory dynamics, which may explain the long-term persistence of viruses and virophages in the environment. Our results demonstrate that inhibition by virophages, programmed cell-death and multicellularity are effective antiviral strategies that may act in synergy against viral infection in microbial species.

Introduction

Giant viruses are large dsDNA viruses from multiple families that belong to the clade of Nucleocytoplasmic Large DNA Viruses (NCLDVs) (Brandes and Linial, 2019). They are known to infect a wide variety of eukaryotes including algae, protozoans and vertebrates (Brandes and Linial, 2019), and some are amongst the largest viruses known in terms of genome size and physical dimensions; which can be easily seen under the light microscope (la Scola et al., 2003; Legendre et al., 2013; Raoult et al., 2004). The striking sizes of giant viruses, which rival those of bacteria, are believed to have evolved independently from smaller viral ancestors via extensive horizontal gene transfer (Katzourakis and Aswad, 2014; Koonin and Yutin, 2018). A distinctive feature in the infectious cycle of giant viruses is the formation of electron-dense areas in the cell cytoplasm where viral transcription and assembly take place (Fridmann-Sirkis et al., 2016). These viral factories concentrate significant cell resources and are the sites where virally-encoded components of the transcription and translation machineries are expressed (Abrahão et al., 2018). The newly assembled viruses exit by lysis or budding, which results in the death of the infected cell (Mackinder et al., 2009).

Virophages are viral parasites of giant viruses which use their viral factories for replication. The first virophage to be described was Sputnik, found in association with *Acanthamoeba polyphaga mimivirus* (la Scola et al., 2008). Sputnik could not replicate in the absence of its host virus and when grown together, caused the appearance of abortive and misassembled virus particles (la Scola et al., 2008). Another class of virophages was then discovered in the flagellate *Cafeteria burkhardae* (formerly *Cafeteria roenbergensis*) infected by *Cafeteria roenbergensis virus* (CroV) (Fischer and Suttle, 2011). Similar to Spuntik, Mavirus required

coinfection of the cell with CroV and interfered with CroV replication (Fischer and Suttle, 2011). The Guarani virophage was also shown to have this inhibitory effect on the replication of its mimivirus host (Mougari et al., 2019). In general, it seems that virophages benefit their eukaryotic hosts by lowering the giant virus progeny and increasing the survival of the cell population (Fischer and Hackl, 2016; Mougari et al., 2019). However, this might not be universal as the Zamilon virophage does not seem to have a significant impact on its host virus (Gaia et al., 2014), and so it is a "neutral" virophage.

Virophages can have different biological properties which are illustrated by Mavirus and Sputnik. Mavirus can infect the cell independently of its host virus via receptor-mediated endocytosis (Fischer and Suttle, 2011), while Sputnik enters the cell forming a complex with its host mimivirus by attaching to glycosylated protein fibres of the outer capsid (Boyer et al., 2011; Desnues and Raoult, 2010). Virophages can also integrate using different enzymes: a tyrosine recombinase in Sputnik or a rve-integrase in Mavirus (Berjón-Otero et al., 2019). Interestingly, only Mavirus-like virophages have been found in the genomes of their eukaryotic hosts (Blanc et al., 2015; Fischer and Hackl, 2016). Sputnik-like virophages have been found integrated in the genomes of the giant virus but not in the eukaryotic host, despite attempts to find integrated virophages in several *Acanthamoeba* species (Chelkha et al., 2018; Desnues et al., 2012). In the flagellate *Cafeteria*, reactivation of integrated Mavirus virophages works as an inducible defence system against CroV infection (Fischer and Hackl, 2016). Although the integrated virophage does not seem to inhibit the virus host and infection leads to lysis, exogenous virophages are produced which can interfere with viral replication in new rounds of coinfection, protecting the wider host population (Fischer and Hackl, 2016).

Giant viruses and virophages are emerging as important players in the function of aquatic ecosystems, as evidenced by their sheer abundance and diversity across globally distributed metagenomes (Endo et al., 2020; Roux et al., 2017; Schulz et al., 2020; Yau et al., 2011). They also seem to be ancient components of the eukaryotic virome (da Cunha et al., 2022; Guglielmini et al., 2019; Woo et al., 2021), and could thus be implicated in the evolution of antiviral strategies used by the early eukaryotes. Here we use mathematical models and computational simulations to study the ecological and evolutionary dynamics of this tripartite parasitic system. Importantly, we consider the role of virophage integration on the system dynamics which had not been studied in previous theoretical work. Our models explain why different virophages follow different integration strategies, they show that virus inhibition by virophages can stabilise the dynamics of the system and that inhibitory virophages, multicellularity and programmed-cell death may be used as effective antiviral strategies. Taken together, our findings shed light on various aspects of the ecology and evolutionary genomics of virophages and their hosts, and provide a framework of testable hypotheses to inspire future experimental work.

Materials and Methods

We used two main approaches to study the dynamic interactions between cells, viruses and virophages in the presence of virophage integration. Our first approach was to design and study the qualitative behaviour of coupled systems of Ordinary Differential Equations (ODEs). These systems considered the different virophage infection mechanisms and integration into the cell or host virus genomes. The second approach was to use an Agent-based model (ABM) which allowed us to consider spatial effects and analyse the effect of multicellularity on the dynamics of the system. The codes and scripts developed to

implement these models are available in GitHub (https://github.com/josegabrielnb/virophage-dynamics).

ODE models

To study the impact different infection mechanisms and virophage integration, we developed three systems of coupled ODEs. Our systems of ODEs explicitly model the populations of cells, viruses and virophages, and assume that the system is well-mixed. The first system is based on the independent-entry mechanism of Mavirus and integration into the host cell genome. The second system describes the paired-entry mechanism of Sputnik (the virophage enters the cell in a complex with the virus), and integration into the cell genome. The third model is also based on the paired-entry mechanism but considers integration of Sputnik into the host virus genome, which has been observed experimentally (Desnues et al., 2012).

Mavirus model

The Mavirus model assumes that the virophage can enter the cell independently of its host virus (Figure 1). Naïve cells (C_x) can be infected by viruses (G) or virophages (V). Cells infected by a virus (C_g), die at a certain rate and produce viral progeny, or they can be infected by a virophage. Cells that are infected by a virus and a virophage (C_{gv}), produce viral progeny and release virophages when the cell dies. The amount of viral progeny produced is determined by the inhibition parameter (**f**), where **f** = 0 is total inhibition and **f** = 1 is no inhibition (neutral virophages). In the absence of a host virus, a virophage may infect and integrate into the genome of a naïve cell (C_i). When this cell encounters a giant

virus (C_{ig}), the integrated virophage is reactivated but does not lead to virus inhibition, which is in line with experimental observations (Fischer and Hackl, 2016). The virus-infected cell with a provirophage may also be infected by an exogenous virophage (C_{igv}), which can have an inhibitory effect on viral replication. The model makes the simplifying assumption that virophages which infect a cell but which do not integrate are degraded. A similar assumption was made by Wodarz in model 1a (Wodarz, 2013). This was done to arrive at a simpler model and to allow a more direct comparison with the Sputnik mechanism.

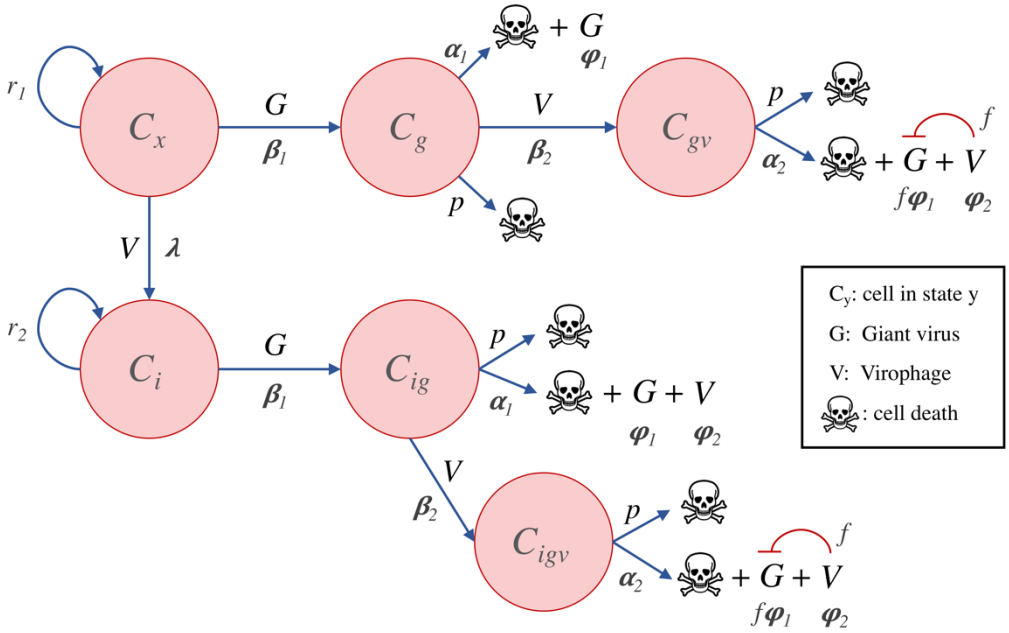


Figure 1. Model of virophage integration into the cell genome following the Mavirus mechanism. Cells with (C_i) or without an integrated virophage (C_x) can be infected by a virus. Infection of C_x cells by a virus results in lysis and release of viruses, while infection of C_i cells by a virus leads to virophage reactivation and release of viruses without inhibition. Cells infected by a virus (C_g or C_{ig}) can also be infected by a virophage, which results in lysis, release of virophages and virus inhibition.

The Mavirus model is described by a system of 8 coupled ODEs and 14 parameters:

$$\begin{cases} C'_x = \boldsymbol{r_1} C_x(1 - \boldsymbol{C_T/K}) - \boldsymbol{\beta_1} C_x G - \boldsymbol{\lambda} C_x V \\ C'_g = \boldsymbol{\beta_1} C_x G - \boldsymbol{\beta_2} C_g V - \boldsymbol{\alpha_1} C_g - \boldsymbol{p} C_g \\ C'_{gv} = \boldsymbol{\beta_2} C_g V - \boldsymbol{\alpha_2} C_{gv} - \boldsymbol{p} C_{gv} \\ C'_i = \boldsymbol{r_2} C_i(1 - \boldsymbol{C_T/K}) + \boldsymbol{\lambda} C_x V - \boldsymbol{\beta_1} C_i G \\ C'_{ig} = \boldsymbol{\beta_1} C_i G - \boldsymbol{\beta_2} C_{ig} V - \boldsymbol{\alpha_1} C_{ig} - \boldsymbol{p} C_{ig} \\ C'_{igv} = \boldsymbol{\beta_2} C_{ig} V - \boldsymbol{\alpha_2} C_{igv} - \boldsymbol{p} C_{igv} \\ G' = \boldsymbol{\alpha_1 \varphi_1} C_g + \boldsymbol{\alpha_1 f \varphi_1} C_{gv} + \boldsymbol{\alpha_1 \varphi_1} C_{ig} + \boldsymbol{\alpha_2 f \varphi_1} C_{igv} - \boldsymbol{\beta_1} C_x G - \boldsymbol{\beta_1} C_i G - \boldsymbol{\gamma_1} G \\ V' = \boldsymbol{\alpha_2 \varphi_2} C_{gv} + \boldsymbol{\alpha_1 \varphi_2} C_{ig} + \boldsymbol{\alpha_2 \varphi_2} C_{igv} - \boldsymbol{\beta_2} C_g V - \boldsymbol{\beta_2} C_{ig} V - \boldsymbol{\gamma_2} V \end{cases}$$

Where $\mathbf{r_1}$ and $\mathbf{r_2}$ are the intrinsic growth rates of naïve cells and cells carrying a provirophage, **K** is the carrying capacity of cells in the population, $\boldsymbol{\beta_1}$ and $\boldsymbol{\beta_2}$ are the infection rates of viruses and virophages, $\boldsymbol{\alpha_1}$ and $\boldsymbol{\alpha_2}$ are the death rates of cells infected by a virus or a virus and virophage, $\boldsymbol{\varphi_1}$ and $\boldsymbol{\varphi_2}$ are the virus and virophage burst sizes, $\boldsymbol{\gamma_1}$ and $\boldsymbol{\gamma_2}$ are the decay rates for viruses and virophages, $\boldsymbol{\lambda}$ is the integration rate of virophages into the cell genome, **f** is the degree of virophage inhibition and **p** is the PCD parameter. Cells can commit suicide in response to a viral infection if $\mathbf{p} > 0$.

Sputnik model 1a: integration into the cell genome

In contrast to Mavirus, Sputnik enters the cell by forming a complex with its host virus (Figure 2). In this model, naïve cells (C_x) can be infected by a virus (G) or virus-virophage complex (GV). Cells infected by the virus (C_g) will die at a certain rate and produce viral progeny. Alternatively, cells may be infected by a complex (C_{gv}), release virophages and viral progeny at a level determined by the inhibition parameter (**f**). Since virophages can only enter the cell as a complex, the only way for a virophage to integrate into the cell genome is as a result of an abortive infection. Thus, the integration rate $\boldsymbol{\lambda}$ is a compound

rate of integration and abortive infection. Cells with a virophage integration (C_i), can be infected by a virus (leading to reactivation but no inhibition), or by a virus-virophage complex (leading to inhibition).

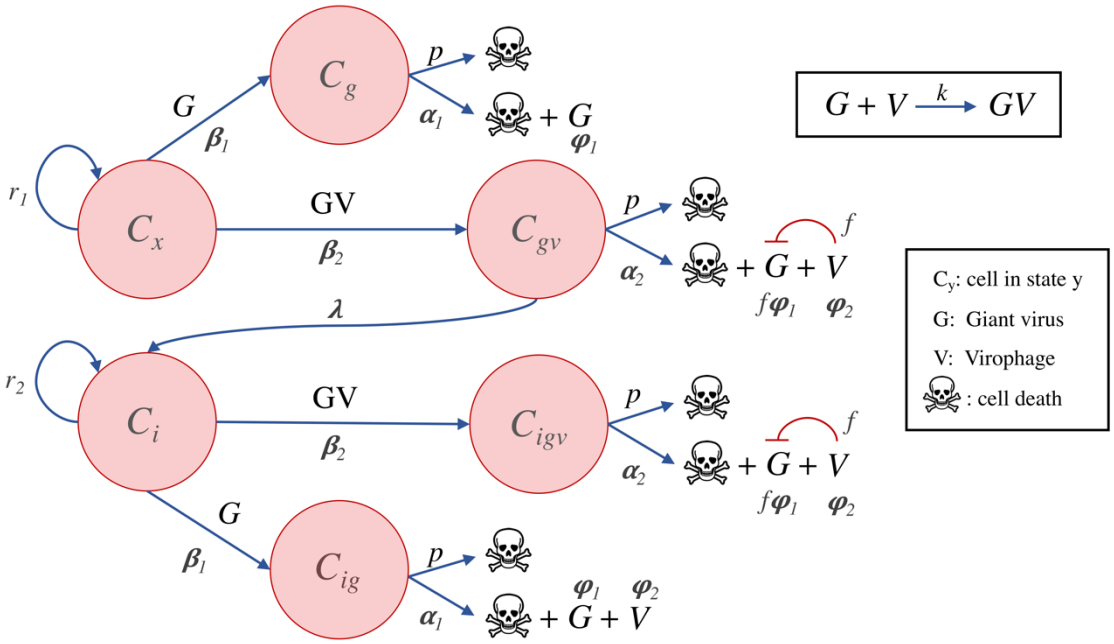


Figure 2. Model of virophage integration into the cell genome following the Sputnik mechanism. Cells with (C_i) or without an integrated virophage (C_x) can be infected by a virus or by a virus-virophage complex (GV). Infection of C_x cells by a virus results in lysis and virus release, while infection of C_i cells by a virus leads to virophage reactivation and release of viruses without inhibition. Infection of C_x or C_i cells by a complex, results in lysis, release of virophages and virus inhibition.

The Sputnik model with integration into the cell genome is described by a system of 9 coupled ODEs and 15 parameters:

$$
\begin{cases}
C'_x = \boldsymbol{r_1} C_x(1 - \boldsymbol{C_T/K}) - \boldsymbol{\beta_1} C_x G - \boldsymbol{\beta_2} C_x GV \\
C'_g = \boldsymbol{\beta_1} C_x G - \boldsymbol{\alpha_1} C_g - \boldsymbol{p} C_g \\
C'_{gv} = \boldsymbol{\beta_2} C_x GV - \boldsymbol{\lambda} C_{gv} - \boldsymbol{\alpha_2} C_{gv} - \boldsymbol{p} C_{gv} \\
C'_i = \boldsymbol{r_2} C_i(1 - \boldsymbol{C_T/K}) + \boldsymbol{\lambda} C_{gv} - \boldsymbol{\beta_2} C_i GV - \boldsymbol{\beta_1} C_i G \\
C'_{ig} = \boldsymbol{\beta_1} C_i G - \boldsymbol{\alpha_1} C_{ig} - \boldsymbol{p} C_{ig} \\
C'_{igv} = \boldsymbol{\beta_2} C_i GV - \boldsymbol{\alpha_2} C_{igv} - \boldsymbol{p} C_{igv} \\
G' = \boldsymbol{\alpha_1 \varphi_1} C_g + \boldsymbol{\alpha_2 f \varphi_1} C_{gv} + \boldsymbol{\alpha_1 \varphi_1} C_{ig} + \boldsymbol{\alpha_1 f \varphi_1} C_{igv} - \boldsymbol{\beta_1} C_x G - \boldsymbol{\beta_1} C_i G - \boldsymbol{k} GV - \boldsymbol{\gamma_1} G + \boldsymbol{\gamma_2} GV \\
V' = \boldsymbol{\alpha_2 \varphi_2} C_{gv} + \boldsymbol{\alpha_1 \varphi_2} C_{ig} + \boldsymbol{\alpha_2 \varphi_2} C_{igv} - \boldsymbol{k} GV - \boldsymbol{\gamma_2} V + \boldsymbol{\gamma_1} GV \\
GV' = \boldsymbol{k} GV - \boldsymbol{\beta_2} C_x GV - \boldsymbol{\beta_2} C_i GV - \boldsymbol{\gamma_1} GV - \boldsymbol{\gamma_2} GV
\end{cases}
$$

The parameters are the same as outlined for the Mavirus model, except for β_2 which refers to the infection rate of the complex (assumed to be equal to β_1), and **k** which is the rate of complex formation.

Sputnik model 1b: integration into the virus genome

We examined a model in which Sputnik can integrate into the virus genome (Figure 3), given that Sputnik provirophages have been found in the genome of *Acanthamoeba polyphaga* Lentille *virus* (Desnues et al., 2012). In this model, naïve cells can be infected by a virus (*G*), by a complex (*GV*) or a virus with an integrated virophage (G_v). Infection by a virus carrying a provirophage results in production of the virus of the same genotype (G_v) according to the inhibition parameter (**f**), and release of virophages. Alternatively, infection of a cell by a complex result in production of viral progeny according to **f**, release of virophages and a proportion of viruses with an integrated virophage (the proportion is λ*). A virus may also infect a naïve cell and replicate without inhibition.

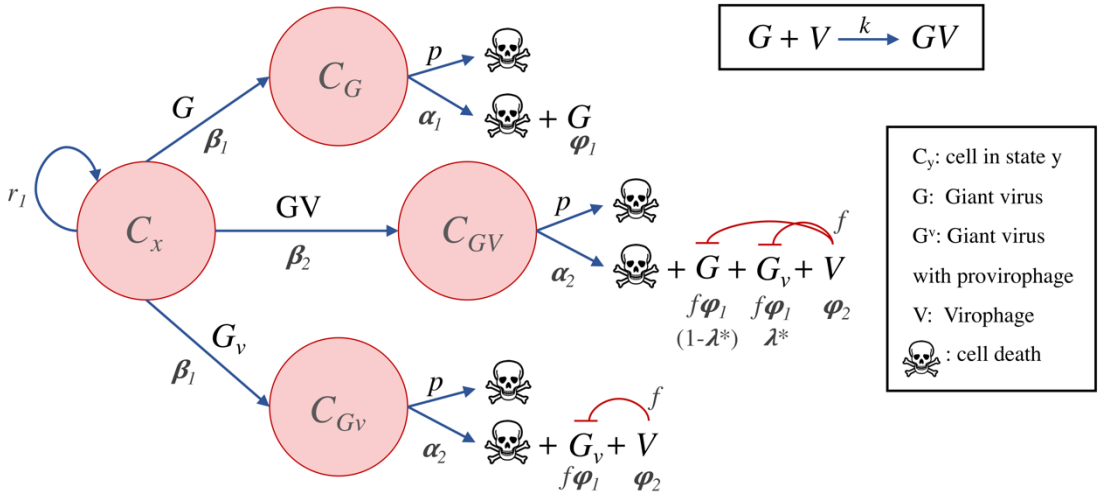


Figure 3. Model of virophage integration into the host virus genome following the Sputnik mechanism. Cells may be infected by a virus (*G*), a virus-virophage complex (*GV*) or a virus carrying a provirophage (G_v). During coinfection by a complex (*GV*), a proportion $\boldsymbol{\lambda^*}$ of the virus progeny incorporates a virophage into their genomes. Alternatively, when a cell is infected by a virus carrying a provirophage, the virus will produce progeny that also carry the integration, and at the same time will produce expression of virophages.

The Sputnik model of integration into the virus genome is described by a system of 8 coupled ODEs and 14 parameters:

$$\begin{cases} C_x' = \boldsymbol{r_1} C_x(1 - \boldsymbol{C_T/K}) - \boldsymbol{\beta_1} C_x G - \boldsymbol{\beta_1} C_x G_v - \boldsymbol{\beta_2} C_x GV \\ C_G' = \boldsymbol{\beta_1} C_x G - \boldsymbol{a_1} C_G - \boldsymbol{p} C_G \\ C_{GV}' = \boldsymbol{\beta_2} C_x GV - \boldsymbol{a_2} C_{GV} - \boldsymbol{p} C_{GV} \\ C_{Gv}' = \boldsymbol{\beta_1} C_x G_v - \boldsymbol{a_2} C_{Gv} - \boldsymbol{p} C_{Gv} \\ G' = \boldsymbol{\alpha_2 f \varphi_1}(\boldsymbol{1} - \boldsymbol{\lambda^*}) C_{GV} + \boldsymbol{\alpha_1 \varphi_1} C_G - \boldsymbol{\beta_1} C_x G - \boldsymbol{k} GV - \boldsymbol{\gamma_1} G + \boldsymbol{\gamma_2} GV \\ G'_v = \boldsymbol{\alpha_2 f \varphi_1 \lambda^*} C_{GV} + \boldsymbol{\alpha_2 \varphi_1} C_{Gv} - \boldsymbol{\beta_1} C_x G_v - \boldsymbol{\gamma_1} G_v \\ GV' = \boldsymbol{k} GV - \boldsymbol{\beta_2} C_x GV - \boldsymbol{\gamma_1} GV - \boldsymbol{\gamma_2} GV \\ V' = \boldsymbol{\alpha_2 \varphi_2} C_{GV} + \boldsymbol{\alpha_2 \varphi_2} C_{Gv} - \boldsymbol{k} GV - \boldsymbol{\gamma_2} V + \boldsymbol{\gamma_1} GV \end{cases}$$

The parameters have the same meanings as those we indicate for the model of Sputnik integration into the cell genome, except for λ^* which is the proportion of viruses with an integrated virophage that are produced during infection with a complex.

Numerical integration and choice of parameters

We integrated the models numerically in MATLAB 2020a using the ode45 solver for non-stiff differential equations (MATLAB, 2021; Shampine and Reichelt, 1997). We simulated the models under different scenarios by varying the integration rate of virophages (λ, range: [10^{-16}, 10^{-4}]) and the degree of virophage inhibition (**f**, range: [0 = total inhibition, 1 = no inhibition]). The values for the cell carrying capacity (**K**), cell intrinsic growth rate (**r**), virus burst size (φ_1), virophage burst size (φ_2) and the rate of complex formation (**k**) have been measured experimentally or estimated from biophysical first principles (Taylor et al., 2014). Therefore, we set our parameters to reflect the order of magnitude of these measurements. Cells carrying a provirophage incurred in a 20% cost of the intrinsic growth rate of naïve cells ($\mathbf{r_2} = 0.8\mathbf{r_1}$). Our choices for the infection rates of viruses (β_1) and virophages (β_2), also comprised values in the order of magnitude of those that have been previously used to model this system (Taylor et al., 2014; Wodarz, 2013). We adjusted the death rate of hosts infected by a virus (α_1) or by a virus and virophage (α_2), and the decay rates of viruses (γ_1) and virophages (γ_2) so that they led to predator-prey oscillations in the presence of neutral virophages; this is our null hypothesis. We used this setup as a baseline scenario to examine the effect of varying the virophage integration rates (λ) and degrees of inhibition (**f**) on the system dynamics, which allowed comparison across the different models.

Agent-based model

We developed an ABM in Julia (Bezanson et al., 2017), to study the effect of multicellularity, programmed cell-death and virophage inhibition on the dynamics of the system. Cell, virus and virophage agents were modelled as mutable structures in Julia, with interactions in 3D-space. Cells changed their internal states as they interacted with viruses and virophages (analogous to ODE compartments). Viruses were assumed to follow an independent entry mechanism as in Mavirus. We did not consider dividing cells in this case since this considerably reduces the complexity of the simulations. Therefore, the model follows a single epidemic wave in a population of susceptible hosts.

We started the simulation with 1024 cells, 512 viruses and 2048 virophages. Each agent was assigned a random position in space by sampling a uniform distribution. During the simulation, each agent underwent Gaussian random walks. The size of the displacement for a particle depends on its diffusion coefficient, which according to the Einstein-Stokes equation is a function of the particle size (Einstein, 1956). Since virophages are the smallest particles in the system, we set an arbitrary displacement of 1 cell diameter per time step. The displacements of cells and viruses were expressed as a function of the virophage displacement (see Text S1 for derivation).

After every time step, we calculated the distances between cells and viruses, as well as the distance between cells and virophages. If the distance to a cell was less than 1 cell radius, we considered the cell to be infected. Cells infected by viruses died and released viral progeny after the incubation period. If the cell was infected by a virus and a virophage, the cell still died but released virophages and viral progeny according to the inhibition

parameter. A cell only infected by a virophage could clear it after a certain number of time steps or alternatively, the virophage could integrate into the cell genome. A virus that infected a cell with an integrated virophage, killed the cell but released virophages and viruses without inhibition.

Multicellular organisms were modelled as cells that were attached to each other in a specified geometry and used the same random displacement vectors. We analysed the dynamics of systems with single-cell, 2-celled, 4-celled, 8-celled and 16-celled agents. Virus induced cell-death was modelled by a constant probability of death per time step, given that a cell is infected by a virus. We analysed three treatments for each population of cell agents: neutral virophages, inhibitory virophages and PCD.

We ran stochastic simulations in 100 replicates for each cell number group/treatment and analysed the effects of treatments and multicellularity by fitting Generalised Additive Models (GAMs). We decided to use GAMs since we observed non-linear relationships in the data and these models allowed testing for multiple effects using smooth functions. Fitting was done using the mgcv package in R (R Core Team, 2020; Wood, 2011).

Results

ODE models

We first examined the system dynamics in the absence of virophages (Figure 4, upper row). In both models, cells and viruses engaged in oscillatory predator-prey interactions. This oscillatory behaviour has also been observed in other models where cells are killed by lytic

viruses (Lauro et al., 2011; Weitz, 2015; Yau et al., 2011). In contrast to Lotka-Volterra predator-prey models where trajectories converge to different orbits depending on initial conditions (neutral oscillations), our models show trajectories that converge to a limit cycle attractor in 2 dimensions (a stable oscillatory regime). This is similar to the behaviour observed in Holling-Tanner predator-prey models where population growth is also controlled by a logistic term (Garfinkel et al., 2017; J. D. Murray, 1993). The appearance of an attractor is a desirable property since the system outcome is robust to small perturbations in the initial conditions (Garfinkel et al., 2017).

When we added neutral virophages to the system (which do not inhibit virus replication), all three populations showed a stable oscillatory regime (Figure 4, middle row). This can be seen in the appearance in state-space of a 3-dimensional limit cycle attractor. Interestingly, when we set the inhibition parameter equal to zero (**f** = 0, virophages completely inhibit virus replication), the trajectories converge to a point equilibrium where the populations of cells, viruses and virophages remain unchanged (Figure 4, bottom row). This is an example of a Hopf bifurcation, where a fixed-point equilibrium gives rise to a limit cycle attractor (Garfinkel et al., 2017; J. D. Murray, 1993), in this case as a function of the inhibition parameter (**f**).

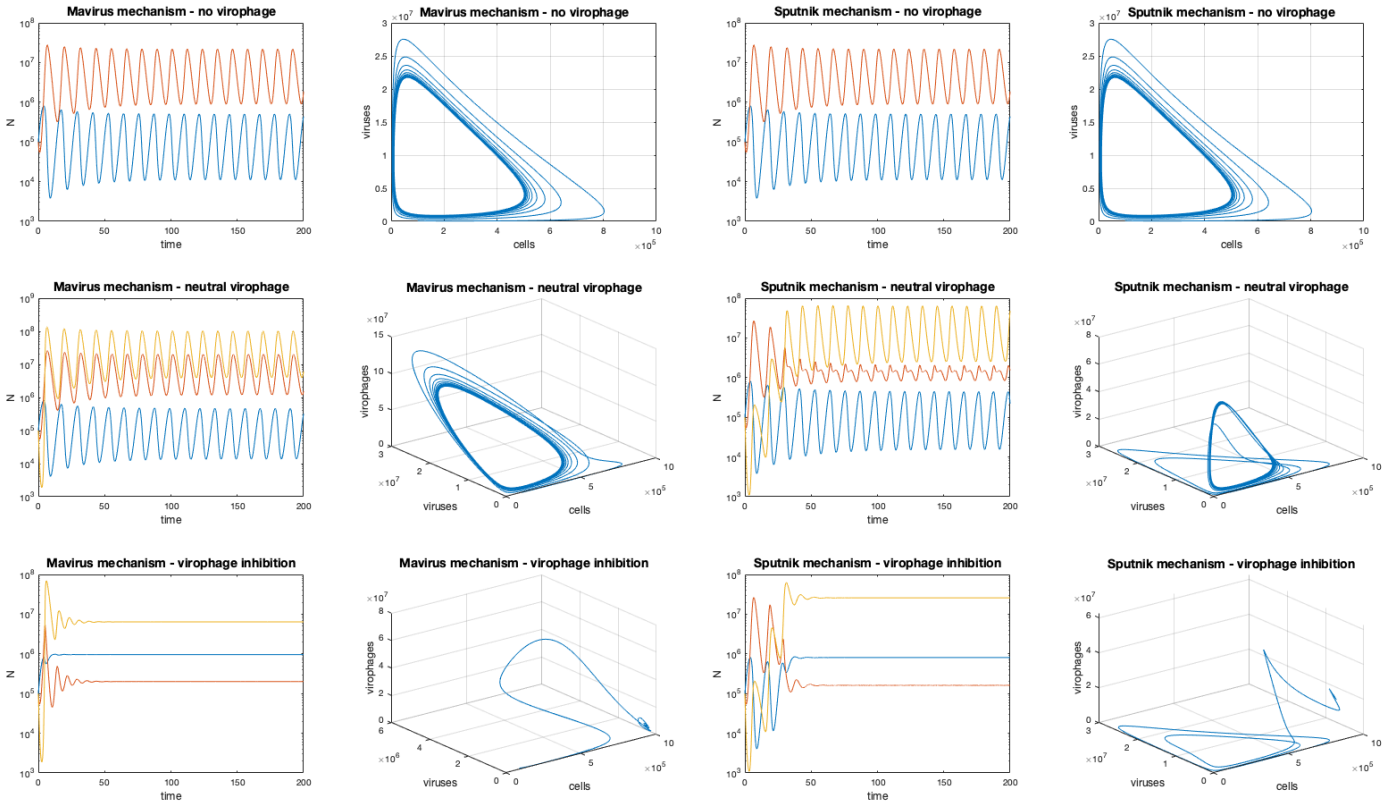


Figure 4. Dynamics of the Mavirus and Sputnik 1a models in the absence of virophage integration (baseline scenario). In the time-course diagrams, the blue line represents the population of cells, the orange represents viruses and the yellow line represents virophages. **Upper row:** cell-virus interactions produce Holling-Tanner oscillations, trajectories converge to a 2-dimensional limit cycle in state space. **Middle row:** when neutral virophages are added, the dynamics are still oscillatory, trajectories converge to a 3-dimensional limit cycle. **Bottom row:** total inhibition by virophages stabilises the system dynamics, trajectories now converge to a point equilibrium. Parameters for model 1: $\alpha_1 = 1$, $\alpha_2 = 0.9$, $\beta_1 = 10^{-7}$, $\beta_2 = 10^{-6}$, $\gamma_1 = 1.2$, $\gamma_2 = 2.6$, $\lambda = 0$, $\varphi_1 = 100$, $\varphi_2 = 1000$, $\mathbf{f} = \{1, 0\}$, $\mathbf{K} = 10^6$, $\mathbf{r_1} = 1$, $\mathbf{r_2} = 0.8$, $\mathbf{p} = 0$. Parameters for model 2 are the same except for $\beta_2 = 10^{-7}$ and $\mathbf{k} = 8 \cdot 10^{-7}$. Initial conditions: $\mathbf{C_{x,0}} = 10^5$, $\mathbf{G_0} = 10^5$, $\mathbf{V_0} = 10^5$.

The effect of virophage integration into the cell genome

Inclusion of integrating virophages affected dynamics of the system differently depending on whether virophages followed the Mavirus or Sputnik infection mechanisms. When virophages are neutral and the rate of integration is low ($\mathbf{f} = 1$, $\lambda = 10^{-9}$), we observed the appearance of cells with an integrated virophage which oscillated together with naïve cells, virophages and viruses in the Mavirus model (Figure 5, upper row). In contrast, we did not observe cells with an integrated virophage emerging in the model for Sputnik under these conditions (Figure 5, upper row). If we then increased the integration rate ($\mathbf{f} = 1$, $\lambda = 10^{-4}$), cells with an integrated virophage were observed in the Sputnik model, while these cells replaced the naïve cell population in the Mavirus model (Figure 5, upper middle row). Keeping the integration rate constant but in presence of moderate virophage inhibition ($\mathbf{f} = 0.6$, $\lambda = 10^{-4}$), led to the appearance of a point equilibrium with an "inward spiral" pattern in state-space (Figure 5, lower middle row). In the Mavirus model, naïve cells were replaced by cells with an integrated virophage, while in the Sputnik model both cell populations were present. Under total virophage inhibition and low integration rate ($\mathbf{f} = 0$, $\lambda = 5 \cdot 10^{-8}$), both models showed damped oscillations converging to a point equilibrium, with replacement of naïve cells in the Mavirus model and stable coexistence in the Sputnik model (Figure 5, lower panel).

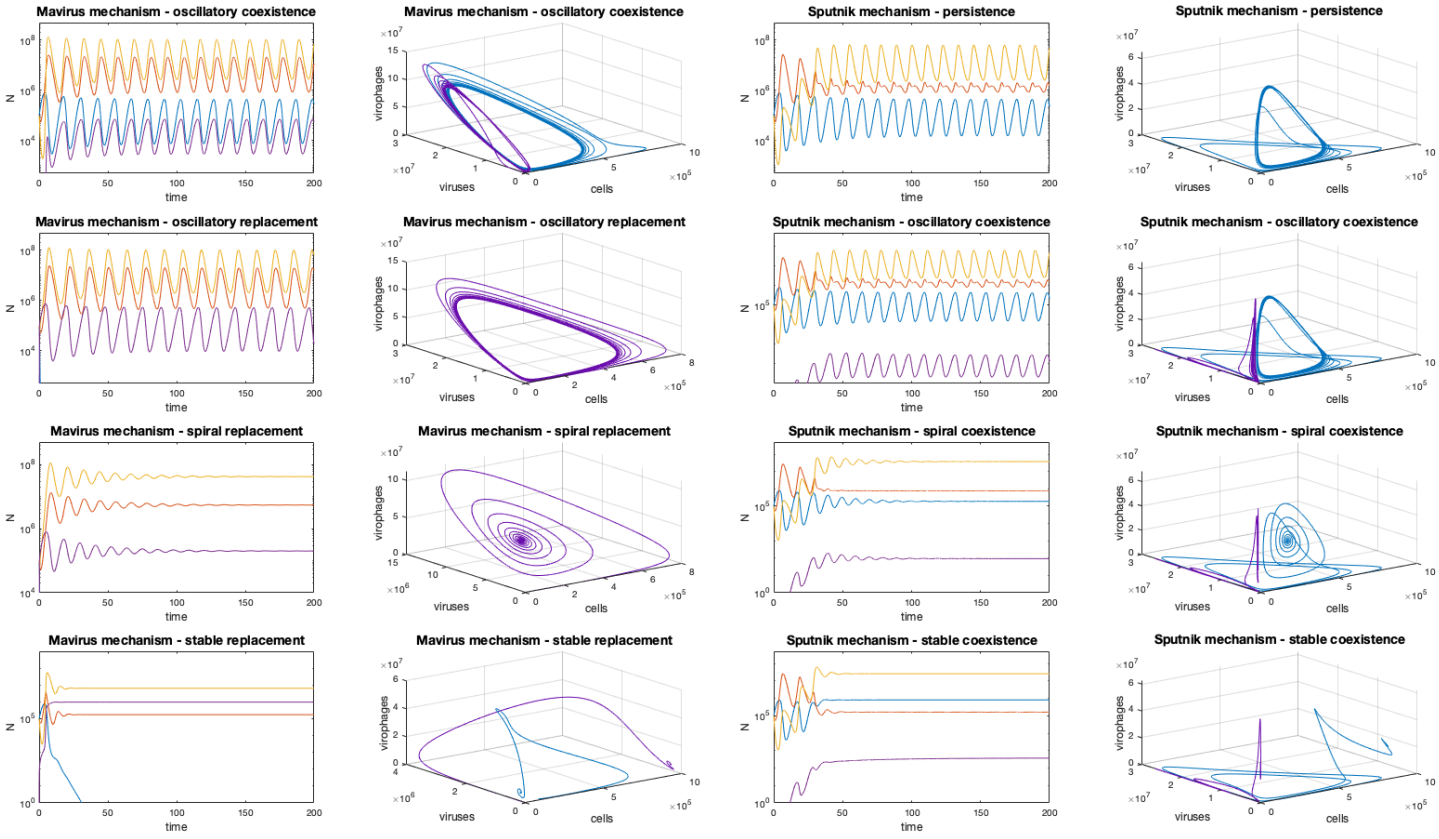


Figure 5. Dynamics of the Mavirus and Sputnik 1a models in the presence of virophage integration. In the time-course diagrams, the blue line represents the population of naïve cells, the purple line represents cells with an integrated virophage, the orange represents viruses and the yellow line represents virophages. **Upper row:** in the presence of neutral virophages and low rates of integration ($\mathbf{f} = 1$, $\boldsymbol{\lambda} = 10^{-9}$), we observe oscillatory coexistence of all populations in the Mavirus model while cells with an integrated virophage cannot establish in the Sputnik mechanism. The Mavirus trajeries converge to a 4-dimensional limit cycle in state-space (shown as two nested trajectories), while the Sputnik trajectories converge to a 3-dimensional limit cycle. **Upper middle row:** in the presence of neutral virophages and higher integration rates ($\mathbf{f} = 1$, $\boldsymbol{\lambda} = 10^{-4}$), cells with an integration become fixed in the population in the Mavirus model (3-dimensional limit cycle), while coexistence of all four populations was observed in the Sputnik model (4-dimensional limit cycle). **Bottom middle row:** in the presence of moderate inhibition by virophages ($\mathbf{f} = 0.6$, $\boldsymbol{\lambda} = 10^{-4}$), damping-wave

oscillations are observed in both systems but fixation occurs in the Mavirus mechanism while coexistence occurs with Sputnik. Populations converge to a point equilibrium with spiral trajectories. **Bottom row:** total inhibition leads to stabilisation in both systems ($\mathbf{f} = 0$, $\boldsymbol{\lambda} = 5e^{-8}$), with fixation of cells with an integrated virophage for Mavirus and coexistence for Sputnik. Parameters for model 1: $\boldsymbol{\alpha_1} = 1$, $\boldsymbol{\alpha_2} = 0.9$, $\boldsymbol{\beta_1} = 10^{-7}$, $\boldsymbol{\beta_2} = 10^{-6}$, $\boldsymbol{\gamma_1} = 1.2$, $\boldsymbol{\gamma_2} = 2.6$, $\boldsymbol{\varphi_1} = 100$, $\boldsymbol{\varphi_2} = 1000$, $\mathbf{K} = 10^6$, $\mathbf{r_1} = 1$, $\mathbf{r_2} = 0.8$, $\mathbf{p} = 0$. Parameters for model 2 are the same except for $\boldsymbol{\beta_2} = 10^{-7}$ and $\mathbf{k} = 8 \cdot 10^{-7}$. Initial conditions: $\mathbf{C_{x,0}} = 10^5$, $\mathbf{G_0} = 10^5$, $\mathbf{V_0} = 10^5$.

We examined the effects on the system outcomes more systematically by sampling the inhibition parameter and the integration rate over multiple orders of magnitude (Figure 6). For both models, we observed that the main parameter determining system outcome is the integration rate, although it does interact to an extent with the inhibition parameter. In the Mavirus model, moderate rates of integration led to coexistence of naïve cells with cells carrying an integrated virophage (green region), while higher integration rates led to replacement of naïve cells with cells carrying an integrated virophage (red region). In contrast, the outcomes of the Sputnik model were dominated by naïve cells alone (blue region), while coexistence of cell populations was only observed at high rates of integration; higher than those required to observe replacement in the Mavirus model (green region).

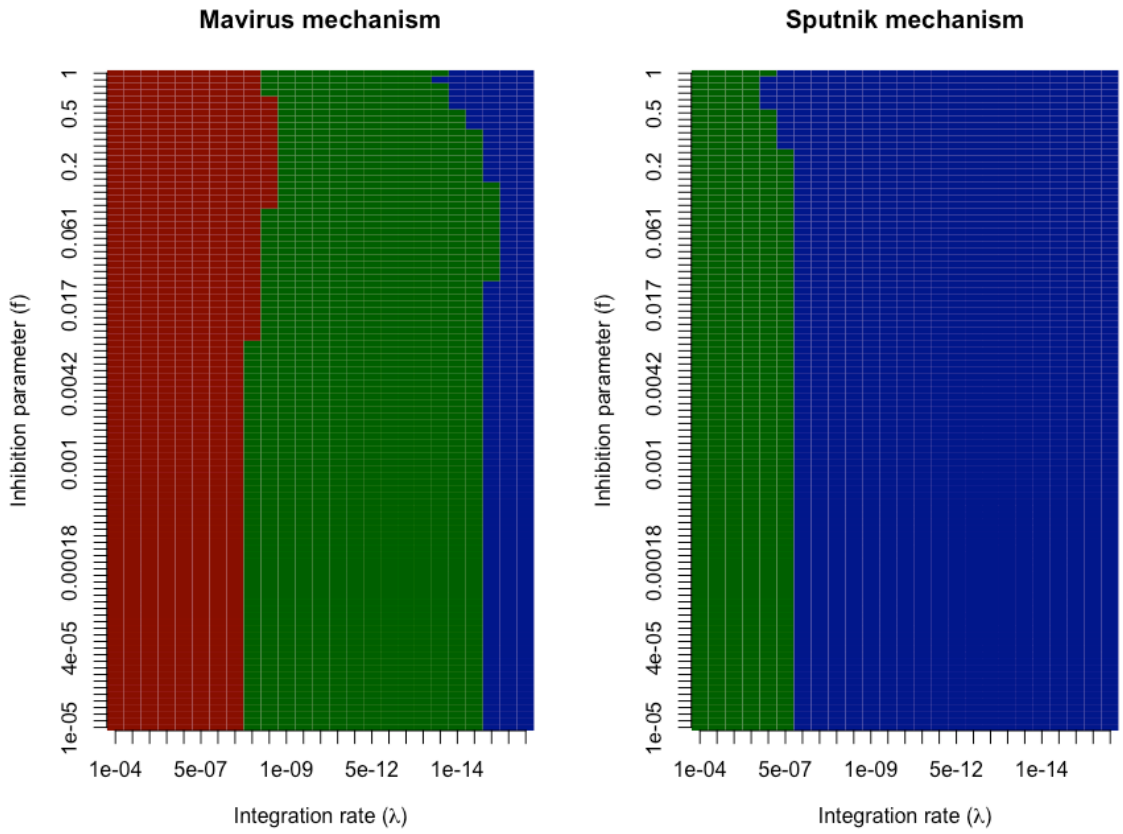


Figure 6. System outcomes depending on the virophage infection mechanism. We explored the outcome of the Mavirus and Sputnik models by varying the integration rates (**λ**) and the inhibition parameter (**f**) and mapping the dynamical outcomes. **Mavirus model (left):** we observed three possible outcomes, fixation of cells with an integrated virophage (red), coexistence of cells with and without an integrated virophage (green) and persistence of cells without a virophage integration (blue). **Sputnik model (right):** two outcomes were observed for the same parameter ranges, coexistence of cells with and without and integrated virophage (green) and persistence of naïve cells (blue). Parameters for model 1: $\alpha_1 = 1$, $\alpha_2 = 0.9$, $\beta_1 = 10^{-7}$, $\beta_2 = 10^{-6}$, $\gamma_1 = 1.2$, $\gamma_2 = 2.6$, $\varphi_1 = 100$, $\varphi_2 = 1000$, $\mathbf{K} = 10^6$, $\mathbf{r_1} = 1$, $\mathbf{r_2} = 0.8$, $\mathbf{p} = 0$. Parameters for model 2 are the same except for $\beta_2 = 10^{-7}$ and $\mathbf{k} = 8 \cdot 10^{-7}$. Initial conditions: $\mathbf{C_{x,0}} = 10^5$, $\mathbf{G_0} = 10^5$, $\mathbf{V_0} = 10^5$.

The effect of virus-induced PCD (Programmed cell death)

The introduction of PCD triggered by viral infection had a similar stabilising effect on the system dynamics as observed for virophage inhibition (Figure S1). We considered this case by setting the PCD parameter to be greater than zero (**p** > 0). Compared to a model where Mavirus shows oscillatory replacement (**f** = 1, $\lambda = 10^{-7}$, **p** = 0), inclusion of the PCD parameter (**p** = 6) led to stabilisation and convergence to a fixed-point equilibrium (Figure S1, upper row). If the PCD was increased (**p** = 8), the cells reached a fixed-point equilibrium with an integration polymorphism while viruses and virophages went extinct (Figure S1, upper row). Compared to an oscillatory persistence regime for the Sputnik model (**f** = 1, $\lambda = 10^{-7}$, **p** = 0), setting **p** = 6 also led to stabilisation of the dynamics and convergence to a point equilibrium but virophages went extinct (Figure S1, lower row). Increasing the PCD parameter to 8 led to the extinction of viruses and virophages, while naïve cells grew to carrying capacity (Figure S1, lower row). The extinction of viruses and virophages occurred in both models with **p** > 6.6. In the Mavirus model, the time to extinction of viruses and virophages decreases as the PCD parameter is increased (Figure S1, third column). The time to extinction of virophages was independent of increases in the PCD parameter in the Sputnik model, while the time to extinction of viruses also decreased as the PCD parameter was increased.

Integration of Sputnik into the virus genome

We explored integration of Sputnik into the host virus genome since it has been observed experimentally (Desnues et al., 2012). In the absence of integrating virophages ($\lambda^* = 0$), the system showed the same oscillatory dynamics with neutral virophages and damping

oscillations under virophage inhibition, as described for the previous two models (Figure 7, upper row). An integration frequency of 1/10,000 ($\boldsymbol{\lambda}^* = 10^{-4}$) led to the appearance of viruses with an integrated virophage under no to moderate inhibition ($\mathbf{f} = \{1, 0.7\}$; Figure 7, middle row). However, in the presence of total virophage inhibition ($\mathbf{f} = 0$), viruses with an integrated virophage did not establish in the population (Figure 7, middle row). This makes intuitive sense since a virus carrying a totally inhibitory virophage would not be able to replicate. Interestingly, when 1/10 viruses produced during a mixed infection carried an integrated virophage ($\boldsymbol{\lambda}^* = 10^{-1}$), these viruses outcompeted the ones without an integration if the virophages were neutral or moderately inhibitory ($\mathbf{f} = \{1, 0.7\}$), but they were not observed if inhibition was total (Figure 7, lower row).

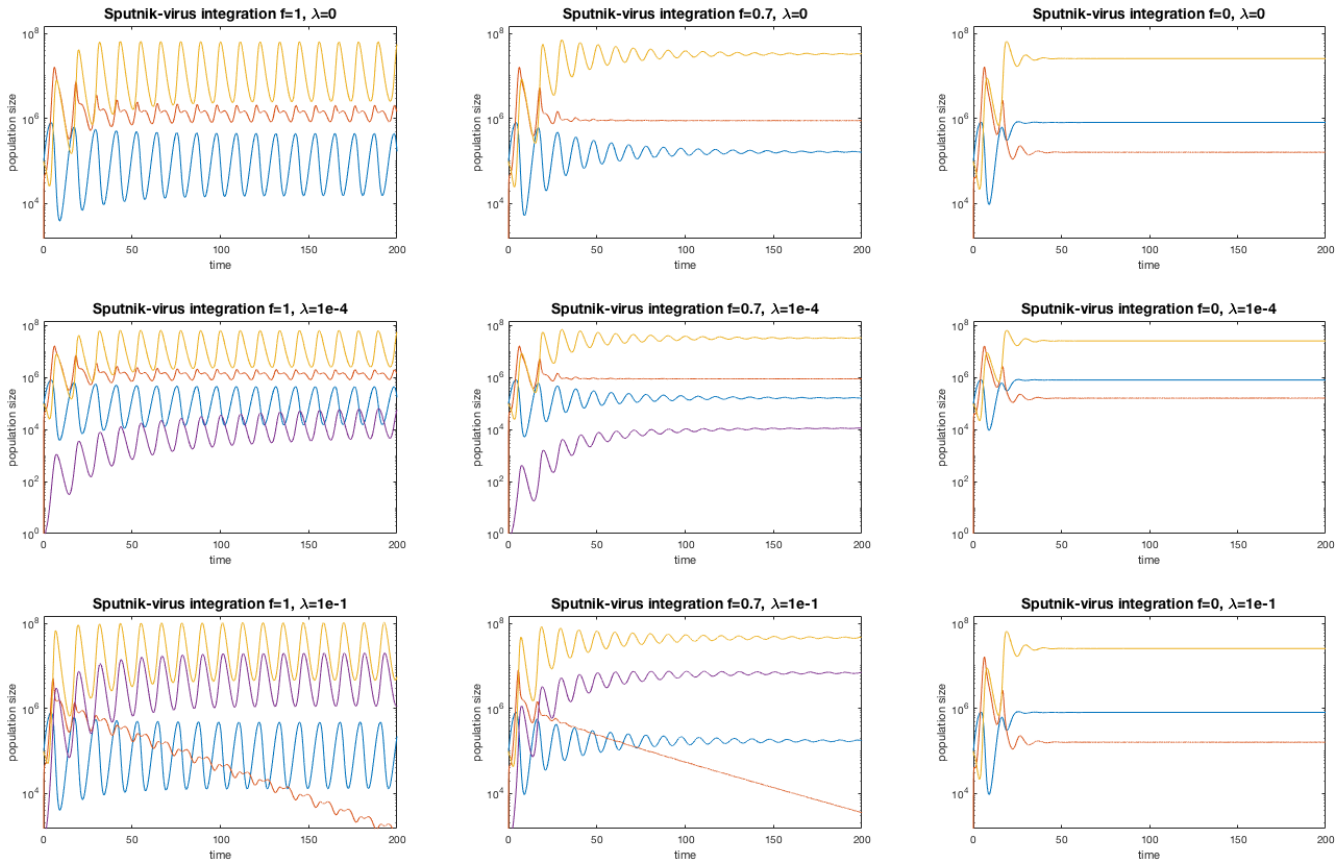


Figure 7. Dynamics of the Sputnik 1b model with integration into the host virus genome. In the time-course diagrams, the blue line represents the population of cells, the orange represents viruses, the purple line represents viruses with an integrated virophage and the yellow line represents virophages. Virophage inhibition leads to stabilisation of the dynamics. **Upper row:** in the absence of integration ($\lambda^* = 0$), the cell, virus and virophage populations coexist in time. **Middle row:** in the presence of moderate integration ($\lambda^* = 10\text{-}4$), viruses with an integrated virophage can establish in the population and are maintained as a polymorphism ($\mathbf{f} = \{1, 0.7\}$), unless there is total virophage inhibition ($\mathbf{f} = 0$). **Bottom row:** in the presence of high integration ($\lambda^* = 0.1$), the original virus population is replaced by viruses carrying an integrated virophage ($\mathbf{f} = \{1, 0.7\}$), except when inhibition is total ($\mathbf{f} = 0$).

Agent-based model

We developed an agent-based model (ABM) in 3-dimensions to understand the impact of spatial effects in a stochastic simulation of cell-virus-virophage interactions. In particular, we were interested in understanding the effect of multicellularity on the system dynamics and how it interacted with virophage inhibition and PCD. For this, we considered 15 different scenarios in which populations of organisms with different cell numbers (single-, 2-, 4-, 8- and 16-celled organisms) were exposed to neutral virophages, inhibitory virophages or PCD. Our ABM considered a population of non-dividing cells in order to reduce the computational complexity of the simulations, and thus centres on a single-epidemic wave.

The effect of virophage inhibition, multicellularity and PCD

In the simulations we can see how the cell population was infected by viruses producing clouds of infection and bursts of virophages whenever they happened to coinfect a cell (Videos S1-S15, https://figshare.com/s/b88b9d60f5ba24e3adbc, legends in Appendix). An example time-course gives an idea of the underlying dynamics (Figures S1-S3). In a population of single-celled organisms, the neutral virophage treatment produced the largest waves of viruses and the lowest numbers of cell survivors. Both virophage inhibition and PCD appeared to have a protective effect on the cell populations. Increasing the number of cells per organism (which introduces spatial clustering of cells), also had a protective effect since the waves for viruses were lower and it increased the number of cell survivors compared to the single-celled case.

To assess these effects systematically, we carried out 100 replicates of the stochastic simulations per treatment and then analysed them using Generalised Additive Models (GAMs, https://figshare.com/s/8a2e9f9948199eedfeda). Compared to a population of single-celled organisms, increasing the cell number led to a general decrease in the maximum amplitude of the virus wave up to 8-celled organisms, although an increase was then observed from the 8- to 16-celled populations (Figure 8). The treatments with virophage inhibition had lower maximum amplitudes of the virus wave compared to neutral virophages, and the most effective treatment was PCD (Figure 8). Multicellularity also lowered the time to virus extinction across treatments, and the protective effects of virophage inhibition and PCD were also observed (Figure 9). Finally, the number of surviving cells was highest for the PCD treatment, followed by the treatment with inhibitory virophages, and increased as a function of the number of cells per organism (Figure 10).

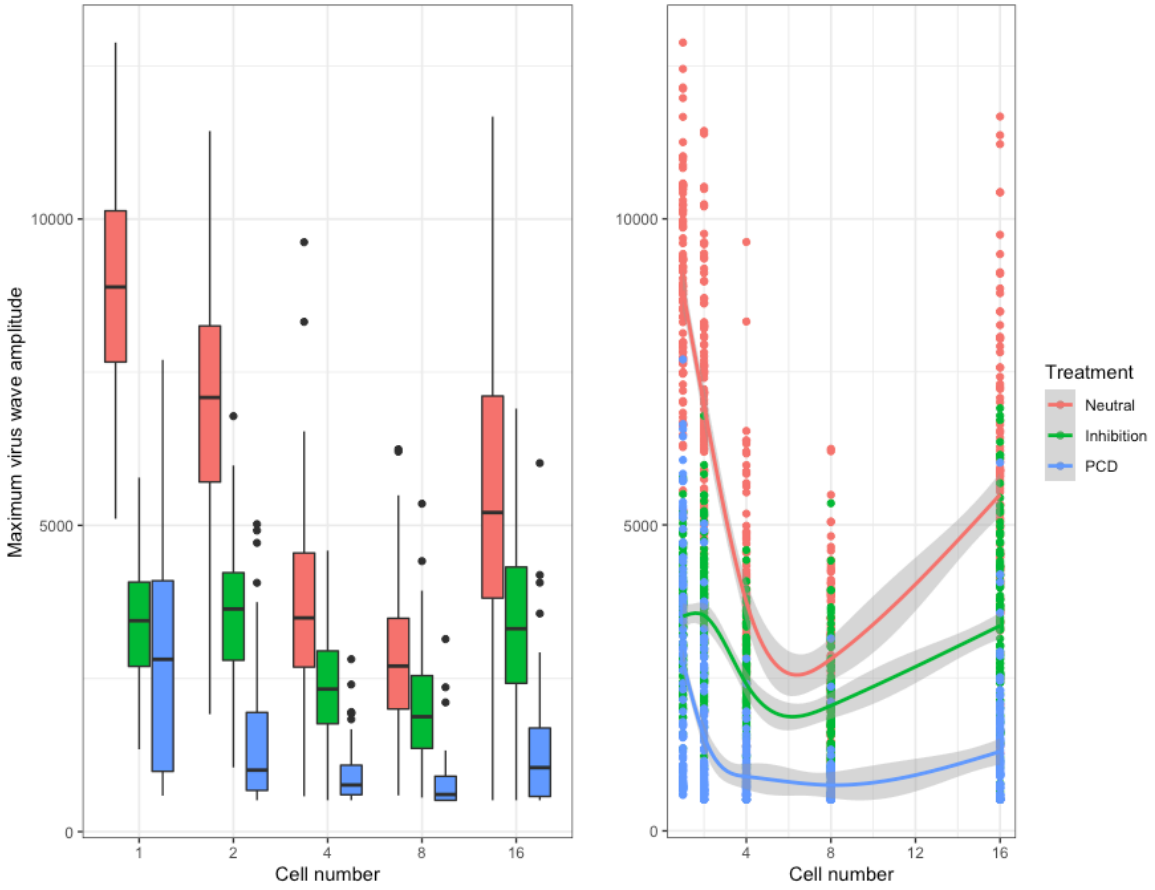


Figure 8. Maximum amplitude of the virus wave. The maximum amplitude of the virus wave decreases as a function of the number of cells and the treatment in the simulations. Boxplots of the resulting distributions are shown to the left and the fitted GAMs are shown to the right. However, note the increase in the maximum wave amplitude across all treatments from 8- to 16-celled organisms. The three intercepts are significantly different and the smoothing terms statistically significant to the level of $p < 10^{-15}$. The normalised root mean square error (RMSE) is 11% and the deviance explained is 72%.

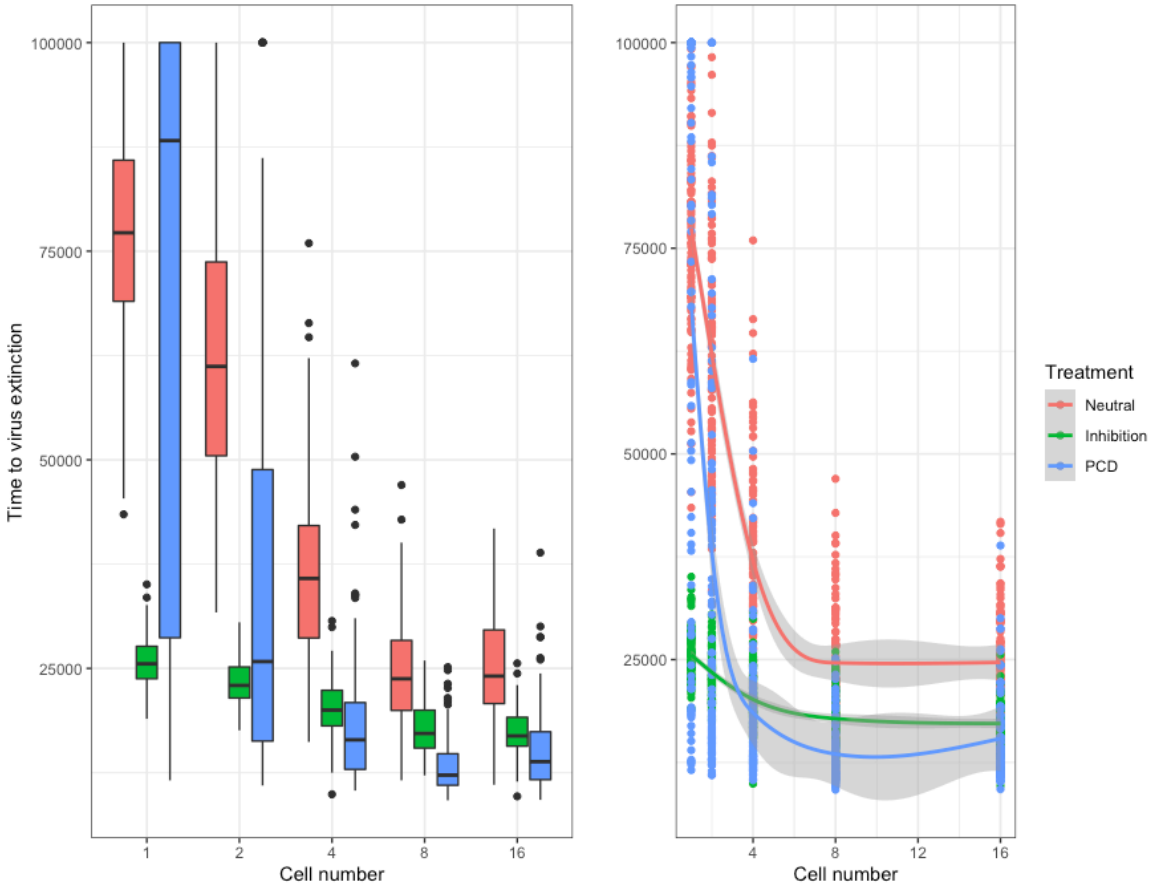


Figure 9. Time to virus extinction. The time to virus extinction also decreases as a function of the cell number and treatment in the simulations. Boxplots of the resulting distributions are shown to the left and the fitted GAMs are shown to the right. For single-celled organisms, virophage inhibition was more effective than PCD, since it led to faster extinction of viruses. The treatments with neutral and inhibitory virophages appear to level-off with increasing cell number, while there was an increase in the time to virus extinction in the PCD treatment from 8- to 16-celled organisms (despite the errors being larger). The three intercepts are significantly different and the smoothing terms statistically significant for the neutral virophages and PCD to the level of $p < 10^{-15}$. The smoothing term for the treatment with inhibitory virophages was significant to the level of $p < 10^{-5}$. The normalised root mean squared error (RMSE) is 14.6% and the deviance explained is 69.3%.

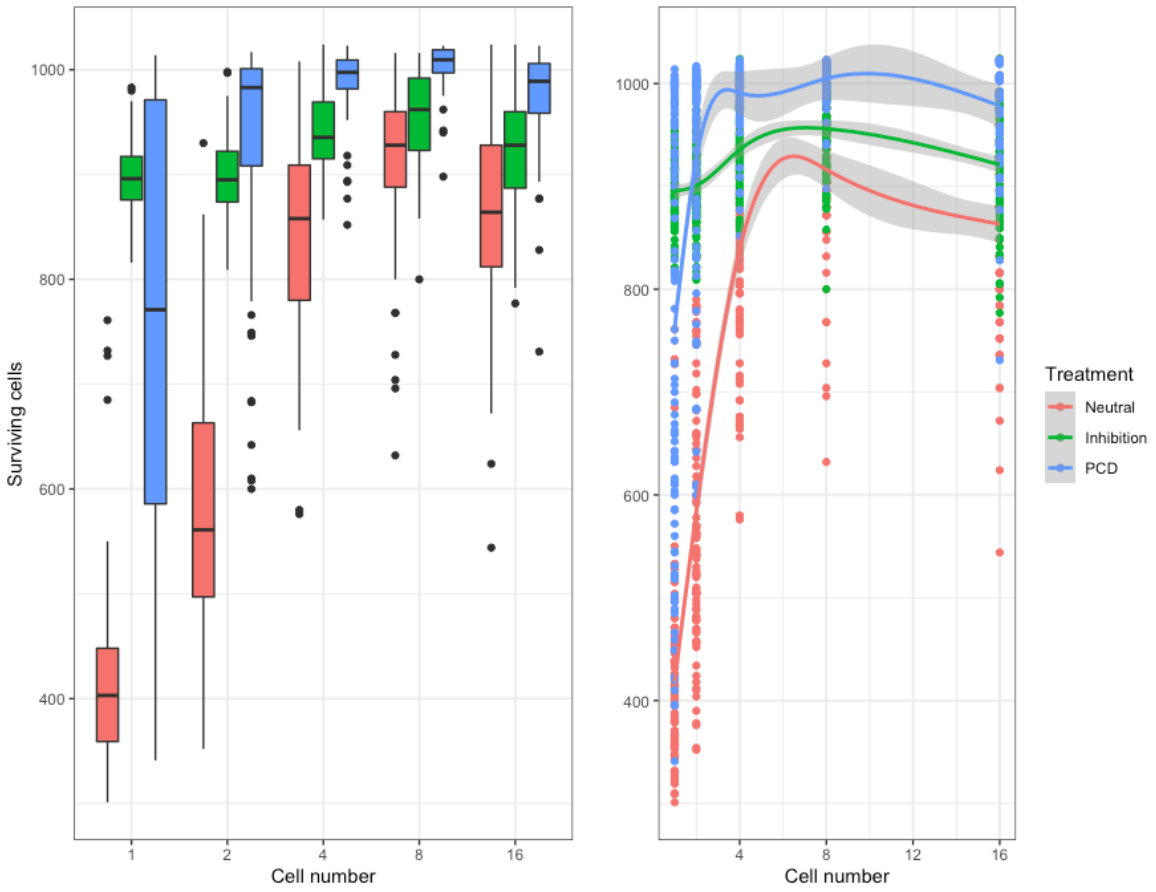


Figure 10. Number of surviving cells. The number of surviving cells at the end of the simulations increased as a function of the cell number and the treatment. Boxplots of the resulting distributions are shown to the left and the fitted GAMs are shown to the right. The highest levels of survival were observed for the PCD treatment, followed by virophage inhibition. The three intercepts are significantly different and the smoothing terms statistically significant for the neutral virophages and PCD to the level of $p < 10^{-15}$. The smoothing term for the treatment with inhibitory virophages was significant to the level of $p < 10^{-5}$. The normalised root mean squared error (RMSE) is 0.7% and the deviance explained is 76.8%.

Discussion

Our findings reveal several key aspects of the natural history of virophages and the interactions with their viral and cellular hosts. We have shown that increasing the inhibition of virophages on virus replication can have a stabilising effect on the oscillatory dynamics. This is important since non-oscillatory stable systems are more robust to stochastic variations that can lead to collapse of the populations (Wodarz, 2013). A previous model that did not consider the populations of viruses and virophages explicitly, suggested that increasing inhibition would be destabilising, leading to oscillatory dynamics and making the populations prone to extinction (Wodarz, 2013). To reconciliate this with the observed abundance and persistence of virophages in nature (Paez-Espino et al., 2019; Roux et al., 2017), the existence of a viral reservoir or meta-population dynamics were proposed (Wodarz, 2013). However, our models show that the long-term stability of virophages and giant viruses in nature is possible without the need of external intervening factors, although it is likely that multiple factors may be at play.

Another important observation that emerged from our analyses is that integration of virophages into the cell or virus genomes is potentially determined by the virophage infection mechanism. Virophages that follow an independent-entry mechanism such as Mavirus would be expected to be more commonly found in the genomes of their eukaryotic hosts, while viruses that enter the cells as a complex would be harder to be found integrated in cells. These predictions are consistent with experimental observations: Mavirus and Mavirus-like virophages have been found in the genomes of *Cafeteria burkhardae* and *Bigelowiella natans* (Blanc et al., 2015; Fischer and Hackl, 2016; Hackl et al., 2021), while Sputnik and Sputnik-like virophages have only been found integrated in the genomes of

their virus hosts (Berjón-Otero et al., 2019; Chelkha et al., 2018; Desnues et al., 2012). One important aspect about the natural history of the Sputnik system is that a virophage integration in a cell can only survive if it occurs together with an abortive viral infection, otherwise the cell is lysed and its descendants will not inherit the integration (Berjón-Otero et al., 2019).

In addition, we observed that when Sputnik virophages integrated into a high proportion of viral progeny (>1/10,000), viruses with integrated virophages could persist stably in the population. High rates of integration into the viral genome could indicate that Sputnik has evolved a greater affinity for integration into the virus compared to the host cell, probably as a consequence of its unique infection mechanism. However, tyrosine recombinases do seem capable of integrating into the host cell genome, as shown by a Polinton-like virus that was discovered integrated in the genome of the cryptophyte alga *Guillardia theta* (Yutin et al., 2015). It would be interesting to test if changing the infection mechanisms of Mavirus and Sputnik leads to different integration dynamics as predicted by our models. For example, Mavirus and Sputnik may be genetically modified to use each other's outer capsid proteins and host virus (i.e. Mavirus/APMV and Sputnik/CroV), and it could then be seen which integration pattern they follow.

Our analyses also indicate that virophage inhibition, virus-induced PCD and the transition to multicellularity can be effective antiviral strategies in microbial eukaryotes. PCD has been classically studied in the context of complex multicellular organisms, where is has important functions in development and homeostasis (Fuchs and Steller, 2011), but PCD also occurs in microbes (Durand et al., 2016). This gives rise to a paradox since the suicide of a single cell ends its reproductive potential. However, artificial life models have shown that kin

selection is sufficient to explain the evolution of suicide in single-celled organisms (Vostinar et al., 2019). In *E. coli*, cell-suicide has been shown to be advantageous in the presence of lytic phage T4rII even when relatedness is low: the benefit to the closest kin greatly exceeds the cost to the infected cell since it will die anyway (Refardt et al., 2013). In the simulations, we observed that a multicellular population of organisms gains protection against viral infection compared to a population of single-celled organisms. Our explanation for this spatial effect is that as long as viruses decay over time, hosts that are more widely dispersed in space will be harder to reach. Therefore, it is more likely that the virus will decay if it has to traverse a longer diffusing distance, while in the single-celled case, it can use hosts as stepping-stones to travel through space more easily. Our results are consistent with previous modelling work that used a Markov process to show that cell-clustering and PCD are optimal strategies in the presence of high viral loads and imperfect immunity (Iranzo et al., 2014).

We have provided mechanistic insights into the interactions of virophages with their virus and cellular hosts, and at the same time explored antiviral defence strategies that are at the disposal of microbial eukaryotes. These results are relevant to the evolution of antiviral immunity in early eukaryotes and to the evolution of multicellularity. Indeed, there is evidence to support horizontal gene transfers between NCLDVs and proto-eukaryotes, which indicates these viral lineages already existed more than a billion years ago (da Cunha et al., 2022; Guglielmini et al., 2019). Virophages are also believed to be an ancient group of viruses, given their diversity, abundance and association with multiple NCLDV host families (Fischer, 2020; Paez-Espino et al., 2019; Roux et al., 2017). In light of this ancient arms-race, the early eukaryotes may have used inhibitory virophages to protect themselves against the onslaught from NCLDVs. The interaction with viruses may also have contributed to the evolution of PCD in eukaryotes and to the initial transition to multicellularity. In this

regard, it would be interesting to test if simple colonial forms of unicellular eukaryotes can evolve in the presence of a lytic virus, as they can do when exposed to a predator (Boraas et al., 1998; Herron et al., 2019). Other processes, such as the sharing of public goods, may have also played major roles in the evolution of multicellularity and PCD (Biernaskie and West, 2015; Vostinar et al., 2019). Since they do not appear to be mutually exclusive, it seems likely that the interaction with viruses and the sharing of public goods may have acted together to favour the appearance of multicellular eukaryotes on multiple occasions. The extent to which viruses may be a driving or reinforcing factor during these transitions can be explored in future work.

References

Abrahão J, Silva L, Silva LS, Khalil JYB, Rodrigues R, Arantes T, Assis F, Boratto P, Andrade M, Kroon EG, Ribeiro B, Bergier I, Seligmann H, Ghigo E, Colson P, Levasseur A, Kroemer G, Raoult D, la Scola B. 2018. Tailed giant Tupanvirus possesses the most complete translational apparatus of the known virosphere. *Nature Communications* 9. doi:10.1038/s41467-018-03168-1

Berjón-Otero M, Koslová A, Fischer MG. 2019. The dual lifestyle of genome-integrating virophages in protists. *Ann N Y Acad Sci*. doi:10.1111/nyas.14118

Bezanson J, Edelman A, Karpinski S, Shah VB. 2017. Julia: A fresh approach to numerical computing. *SIAM Review* 59:65–98. doi:10.1137/141000671

Biernaskie JM, West SA. 2015. Cooperation, clumping and the evolution of multicellularity. *Proceedings of the Royal Society B: Biological Sciences* 282. doi:10.1098/rspb.2015.1075

Blanc G, Gallot-Lavallée L, Maumus F. 2015. Provirophages in the *Bigelowiella* genome bear testimony to past encounters with giant viruses. *Proc Natl Acad Sci U S A* 112:E5318–E5326. doi:10.1073/pnas.1506469112

Boraas ME, Seale DB, Boxhorn JE. 1998. Phagotrophy by flagellate selects for colonial prey: A possible origin of multicellularity. *Evolutionary Ecology* 12. doi:10.1023/A:1006527528063

Boyer M, Azza S, Barrassi L, Klose T, Campocasso A, Pagnier I, Fournous G, Borg A, Robert C, Zhang X, Desnues C, Henrissat B, Rossmann MG, la Scola B, Raoult D. 2011. Mimivirus shows dramatic genome reduction after intraamoebal culture. *Proc Natl Acad Sci U S A* 108. doi:10.1073/pnas.1101118108

Brandes N, Linial M. 2019. Giant viruses-big surprises. *Viruses*. doi:10.3390/v11050404

Chelkha N, Colson P, Levasseur A, la Scola B. 2018. Deciphering the genomes of 16 *Acanthamoeba* species does not provide evidence of integration of known giant virus-associated mobile genetic elements. *Virus Research* 251:14–16. doi: 10.1016/j.virusres.2018.04.018

da Cunha V, Gaia M, Ogata H, Jaillon O, Delmont TO, Forterre P. 2022. Giant viruses encode actin-related proteins. *Molecular Biology and Evolution* 39:msac022. doi:10.1093/molbev/msac022

Desnues C, la Scola B, Yutin N, Fournous G, Robert C, Azza S, Jardot P, Monteil S, Campocasso A, Koonin EV, Raoult D. 2012. Provirophages and transpovirons as the diverse mobilome of giant viruses. *Proc Natl Acad Sci U S A* 109:18078–18083. doi:10.1073/pnas.1208835109

Desnues C, Raoult D. 2010. Inside the lifestyle of the virophage. *Intervirology* 53. doi:10.1159/000312914

Durand PM, Sym S, Michod RE. 2016. Programmed cell death and complexity in microbial systems. *Current Biology* 26:R587–R593. doi:10.1016/j.cub.2016.05.057

Einstein A. 1956. Investigations on the theory of the Brownian movement. Mineola, NY: Dover. pp. 1–18. doi:10.1016/j.chroma.2008.07.021

Endo H, Blanc-Mathieu R, Li Y, Salazar G, Henry N, Labadie K, de Vargas C, Sullivan MB, Bowler C, Wincker P, Karp-Boss L, Sunagawa S, Ogata H. 2020. Biogeography of marine giant viruses reveals their interplay with eukaryotes and ecological functions. *Nature Ecology and Evolution* 4. doi:10.1038/s41559-020-01288-w

Fischer MG. 2020. The virophage family *Lavidaviridae*. *Current Issues in Molecular Biology* 40. doi:10.21775/cimb.040.001

Fischer MG, Hackl T. 2016. Host genome integration and giant virus-induced reactivation of the virophage Mavirus. *Nature* 540:288–291. doi:10.1038/nature20593

Fischer MG, Suttle CA. 2011. A virophage at the origin of large DNA transposons. Science (1979) 231–234. doi: 10.1126/science.1199412

Fridmann-Sirkis Y, Milrot E, Mutsafi Y, Ben-Dor S, Levin Y, Savidor A, Kartvelishvily E, Minsky A. 2016. Efficiency in complexity: composition and dynamic nature of *Mimivirus* replication factories. *Journal of Virology* 90. doi:10.1128/jvi.01319-16

Fuchs Y, Steller H. 2011. Programmed cell death in animal development and disease. *Cell*. doi:10.1016/j.cell.2011.10.033

Gaia M, Benamar S, Boughalmi M, Pagnier I, Croce O, Colson P, Raoult D, la Scola B. 2014. Zamilon, a novel virophage with *Mimiviridae* host specificity. *PLoS ONE* 9. doi:10.1371/journal.pone.0094923

Garfinkel A, Shevtsov J, Guo Y. 2017. Modeling Life, 1st ed, Modeling Life. Springer. doi:10.1007/978-3-319-59731-7

Guglielmini J, Woo AC, Krupovič M, Forterre P, Gaia M. 2019. Diversification of giant and large eukaryotic dsDNA viruses predated the origin of modern eukaryotes. *Proc Natl Acad Sci U S A*. doi:10.1073/pnas.1912006116

Hackl T, Duponchel S, Barenhoff K, Weinmann A, Fischer MG. 2021. Virophages and retrotransposons colonize the genomes of a heterotrophic flagellate. *eLife* 10. doi:10.7554/eLife.72674

Herron MD, Borin JM, Boswell JC, Walker J, Chen ICK, Knox CA, Boyd M, Rosenzweig F, Ratcliff WC. 2019. *De novo* origins of multicellularity in response to predation. *Scientific Reports* 9. doi:10.1038/s41598-019-39558-8

Iranzo J, Lobkovsky AE, Wolf YI, Koonin EV. 2014. Virus-host arms race at the joint origin of multicellularity and programmed cell death. *Cell Cycle* 13:3083–3088. doi:10.4161/15384101.2014.949496

J. D. Murray. 1993. Mathematical biology, 2nd ed. Berlin: Springer.

Katzourakis A, Aswad A. 2014. The origins of giant viruses, virophages and their relatives in host genomes. *BMC Biology*. doi:10.1186/s12915-014-0051-y

Koonin EV, Yutin N. 2018. Multiple evolutionary origins of giant viruses. *F1000Res*. doi:10.12688/f1000research.16248.1

la Scola B, Audic S, Robert C, Jungang L, de Lamballerie X, Drancourt M, Birtles R, Claverie JM, Raoult D. 2003. A giant virus in amoebae. *Science* (1979). doi:10.1126/science.1081867

la Scola B, Desnues C, Pagnier I, Robert C, Barrassi L, Fournous G, Merchat M, Suzan-Monti M, Forterre P, Koonin E, Raoult D. 2008. The virophage as a unique parasite of the giant mimivirus. *Nature* 455:100–104. doi:10.1038/nature07218

Lauro FM, Demaere MZ, Yau S, Brown M v., Ng C, Wilkins D, Raftery MJ, Gibson JA, Andrews-Pfannkoch C, Lewis M, Hoffman JM, Thomas T, Cavicchioli R. 2011. An integrative study of a meromictic lake ecosystem in Antarctica. *ISME Journal* 5. doi:10.1038/ismej.2010.185

Legendre M, Doutre G, Poirot O, Lescot M, Arslan D, Seltzer V, Bertaux L, Bruley C, Claverie J. 2013. Pandoraviruses: amoeba viruses with genomes up to 2.5 Mb reaching that of parasitic eukaryotes. *Science* (1979) 341:281–286. doi:10.1126/science.1239181

Mackinder LCM, Worthy CA, Biggi G, Hall M, Ryan KP, Varsani A, Harper GM, Wilson WH, Brownlee C, Schroeder DC. 2009. A unicellular algal virus, *Emiliania huxleyi virus* 86, exploits an animal-like infection strategy. *Journal of General Virology* 90. doi:10.1099/vir.0.011635-0

MATLAB. 2021. Matlab 2020. The MathWorks Inc.

Mougari S, Bekliz M, Abrahao J, Pinto F di, Levasseur A, la Scola B. 2019. Guarani virophage, a new Sputnik-like isolate from a Brazilian lake. *Frontiers in Microbiology*. doi:10.3389/fmicb.2019.01003

Paez-Espino D, Zhou J, Roux S, Nayfach S, Pavlopoulos GA, Schulz F, McMahon KD, Walsh D, Woyke T, Ivanova NN, Eloe-Fadrosh EA, Tringe SG, Kyrpides NC. 2019. Diversity, evolution, and classification of virophages uncovered through global metagenomics. *Microbiome* 7. doi:10.1186/s40168-019-0768-5

R Core Team. 2020. R: A language and environment for statistical computing.

Raoult D, Audic S, Robert C, Abergel C, Renesto P, Ogata H, la Scola B, Suzan M, Claverie JM. 2004. The 1.2-megabase genome sequence of *Mimivirus*. *Science* (1979) 306. doi:10.1126/science.1101485

Refardt D, Bergmiller T, Kümmerli R. 2013. Altruism can evolve when relatedness is low: Evidence from bacteria committing suicide upon phage infection. *Proceedings of the Royal Society B: Biological Sciences* 280. doi:10.1098/rspb.2012.3035

Roux S, Chan LK, Egan R, Malmstrom RR, McMahon KD, Sullivan MB. 2017. Ecogenomics of virophages and their giant virus hosts assessed through time series metagenomics. *Nature Communications* 8. doi:10.1038/s41467-017-01086-2

Schulz F, Roux S, Paez-Espino D, Jungbluth S, Walsh DA, Denef VJ, McMahon KD, Konstantinidis KT, Eloe-Fadrosh EA, Kyrpides NC, Woyke T. 2020. Giant virus diversity and host interactions through global metagenomics. *Nature* 578. doi:10.1038/s41586-020-1957-x

Shampine LF, Reichelt MW. 1997. The MATLAB ODE suite. SIAM 18:1–22.

Taylor BP, Cortez MH, Weitz JS. 2014. The virus of my virus is my friend: Ecological effects of virophage with alternative modes of coinfection. *Journal of Theoretical Biology* 354:124–136. doi:10.1016/j.jtbi.2014.03.008

Vostinar AE, Goldsby HJ, Ofria C. 2019. Suicidal selection: Programmed cell death can evolve in unicellular organisms due solely to kin selection. *Ecology and Evolution* 9:9129–9136. doi:10.1002/ece3.5460

Weitz JS. 2015. Quantitative Viral Ecology: Dynamics of Viruses and Their Microbial Hosts. Princeton Univeristy Press.

Wodarz D. 2013. Evolutionary dynamics of giant viruses and their virophages. *Ecology and Evolution* 3:2103–2115. doi:10.1002/ece3.600

Woo AC, Gaia M, Guglielmini J, da Cunha V, Forterre P. 2021. Phylogeny of the Varidnaviria Morphogenesis Module: Congruence and Incongruence With the Tree of Life and Viral Taxonomy. *Frontiers in Microbiology* 12. doi:10.3389/fmicb.2021.704052

Wood S. 2011. Fast stable restricted maximum likelihood and marginal likelihood estimation of semiparametric generalized linear models. *Journal of the Royal Statistical Society Series B* (Statistical Methodology) 73:3–36.

Yau S, Lauro FM, DeMaere MZ, Brown M v., Thomas T, Raftery MJ, Andrews-Pfannkoch C, Lewis M, Hoffman JM, Gibson JA, Cavicchioli R. 2011. Virophage control of antarctic algal host-virus dynamics. *Proc Natl Acad Sci U S A* 108:6163–6168. doi:10.1073/pnas.1018221108

Yutin N, Shevchenko S, Kapitonov V, Krupovič M, Koonin EV. 2015. A novel group of diverse Polinton-like viruses discovered by metagenome analysis. *BMC Biology* 13:1–14. doi:10.1186/s12915-015-0207-4

Chapter V

An ancient endogenous DNA virus in the human genome

Abstract

The genomes of eukaryotes preserve a vast diversity of ancient viruses in the form of endogenous viral elements (EVEs). Study of this genomic fossil record provides insights into the diversity, origin and evolution of viruses across geological timescales. In particular, *Mavericks* have emerged as one of the oldest groups of endogenous viruses infecting vertebrates (≥419 My). They have been found in the genomes of fish, amphibians and non-avian reptiles but had been overlooked in mammals. Thus, their evolutionary history and the causes of their demise in mammals remain puzzling questions. Here, we conduct a detailed evolutionary study of two *Maverick* integrations found on human chromosomes 7 and 8. We performed a comparative analysis of the integrations and determined their orthology across placental mammals (Eutheria) via the syntenic arrangement of neighbouring genes. The integrations were absent at the orthologous sites in the genomes of marsupials and monotremes. These observations allowed us to reconstruct a time-calibrated phylogeny and infer the age of their most recent common ancestor at 262 (200–335) My. In addition, we estimate the age of the individual integrations at ~102 My which represent the oldest non-retroviral EVEs found in the human genome. Our findings suggest that active *Mavericks* still existed in the ancestors of modern mammals ~172 My ago (Jurassic Period) and potentially to the end of the Early Cretaceous. We hypothesise *Mavericks* could have gone extinct in mammals from the evolution of an antiviral defence system or from reduced opportunities for transmission in terrestrial hosts.

Introduction

Viruses are highly diverse and abundant infectious agents that persist in hosts by exploiting two main strategies: 1) horizontal transmission between cells in the host population or 2) vertical transmission between dividing cells, either by integrating into the host genome or by persisting as an extrachromosomal nucleic acid (episome) (Barreat and Katzourakis, 2021a; Katzourakis and Gifford, 2010; McBride et al., 2006). Vertical transmission involves a state of stable association with the host cell in which viral particles (virions) are not produced. In the integrated state endogenous viruses resemble transposons, which also colonise the host genome and can be transferred vertically and horizontally (Schaack et al., 2010). However, endogenous viruses are distinguished from transposons by the capacity to reactivate and form virions. Therefore, endogenous viruses retain a conserved gene module involved in the formation of virions, the capsid morphogenesis module, which is common to all viruses and absent in transposable elements (Barreat and Katzourakis, 2021a; Koonin et al., 2021).

The genomes of eukaryotes preserve a large amount of viral sequences derived both from viruses that actively integrate into the host genome, and from viruses which do not have this capacity but have become integrated incidentally by other means (e.g. non-homologous recombination or by interaction with retroelements) (Barreat and Katzourakis, 2021a; Katzourakis and Gifford, 2010). Over time, the host genome becomes populated by lineages of functional viruses that persist by cycles of reactivation, horizontal transmission and integration (‘intact endogenous viruses’), defective viral sequences that have accumulated inactivating mutations (‘viral genomic fossils’), and in some cases, viral sequences that have been repurposed to benefit the host (‘viral exaptations’) (Barreat and Katzourakis, 2021a).

Study of viral sequences found in host genomes provides insights into the deep origins and evolution of viruses as well as into the ancient interactions between viruses and hosts.

Although sequences derived from every type of viral genome (Baltimore groups) have been discovered integrated in the genomes of animals, retroviruses remain the most thoroughly studied group of endogenous viruses (Johnson, 2019). Indeed, endogenous retroviruses account for significant proportions of vertebrate genomes and their study has led to key evolutionary insights, for example, shedding light on the evolutionary history of lentiviruses which include the important human pathogen HIV (Feschotte and Gilbert, 2012; Holmes, 2011). The evolutionary success of retroviruses has been linked to their capacity of integration, latency and reactivation in host cells. Another factor contributing to their success is that replication involves the transcription of genomic RNAs (after integration) which are capped and polyadenylated, making them indistinguishable from host mRNAs and able to evade pattern-recognition receptors (PRRs) that target dsRNA viruses or uncapped RNAs (Beachboard and Horner, 2016; Goff, 2013; Chen et al., 2017; Peter Simmonds, personal communication). Retroviruses are also able to infect germ cells allowing them to infiltrate the host germline. In addition, they may increase in copy number by intragenomic amplification (retrotransposition without formation of capsids).

However, retroviruses are not the only group of animal viruses with this ability; several families of DNA viruses can also integrate into the host chromosome and are known to transmit vertically. Human herpesviruses 6 and 7 stably integrate into the telomeric regions and have been passed vertically in the human population (Aimola et al., 2020; Aswad et al., 2020; Liu et al., 2020; Prusty et al., 2017). Other less well-known families of endogenous viruses are the *Teratorns* which have been found in the genomes of fish (Aswad and

Katzourakis, 2017; Inoue et al., 2018, 2017), and *Mavericks*, which are widely but patchily distributed across eukaryotes and are commonly found in the genomes of aquatic animals (Barreat and Katzourakis, 2021b; Haapa-Paananen et al., 2014; Kapitonov and Jurka, 2006; Pritham et al., 2007).

Mavericks had been widely considered transposons given that they appear in multiple copies in host genomes and because the details of their molecular biology remain obscure (Kapitonov and Jurka, 2006; Pritham et al., 2007). However, both evolutionary analyses and computational modelling of protein structure have shown that they encode a conserved set of genes involved in the formation of virions: major and minor capsid proteins, a DNA packaging ATPase and an adenoviral-like protease (Barreat and Katzourakis, 2021b; Krupovič et al., 2014; Krupovič and Bamford, 2008). In addition, *Mavericks* encode a protein-primed family B DNA polymerase (PolB) and a retroviral-like integrase (Kapitonov and Jurka, 2006; Pritham et al., 2007). In fact, genome content and phylogenetic analyses firmly nest *Mavericks* within the viral kingdom *Bamfordvirae*, together with other viral lineages such as adenoviruses, virophages and the Nucleocytoplasmic Large DNA viruses (Krupovič et al., 2014; Krupovič and Bamford, 2008; Krupovič and Koonin, 2016, 2015). These observations strongly suggest that *Mavericks* are a lineage of endogenous DNA viruses that can occasionally form virions and undergo exogenous transmission (Barreat and Katzourakis, 2021b; Starrett et al., 2020).

In vertebrates, *Mavericks* have been found in the genomes of teleost fish, coelacanths, amphibians, birds and non-avian reptiles. They are an ancient lineage of viruses that have infected vertebrates for the last 419 million years. Most of the intact endogenous viruses occur in teleost fish (97%) while only 4 (3%) have been found in tetrapod genomes (Barreat

and Katzourakis, 2021b). All the elements found so far in birds seem to be degraded and they were believed to be altogether absent from the genomes of mammals (Barreat and Katzourakis, 2021b; Kapitonov and Jurka, 2006; Pritham et al., 2007). It was therefore surprising when a *Maverick*-like integration was recently reported on human chromosome 7 (Starrett et al., 2020; Yao and Frith, 2021). The sequence contained hits to the DNA polymerase, packaging ATPase and major capsid protein, but other markers such as the integrase and minor capsid were not detected (Starrett et al., 2020). Furthermore, the authenticity of the integration had not been verified and thus the possibility of spurious sequence contamination had not been ruled out (Aswad and Katzourakis, 2016). Authenticity can be confirmed by a comparative genomic analysis of the integration locus demonstrating orthology in different species; this shows the integration was ancestral and suggests it reached fixation at some point in the past (Aswad and Katzourakis, 2016; Katzourakis and Gifford, 2010).

Here we conduct a detailed comparative genomic and evolutionary analysis of the chromosome 7 integration locus across mammals and verify the authenticity of the integration via the syntenic arrangement of neighbouring protein-coding genes in multiple species. We find additional hits to conserved *Maverick* genes, including the minor capsid protein and integrase, which serve to better understand the genome of the ancestor and the origin of the integration. We also discover and authenticate a new element on human chromosome 8 with hits to the DNA polymerase, PM, PZ and major capsid proteins. Our findings allow us to reconstruct the evolutionary history of the two human elements and infer their timescales of evolution. We also discover numerous PolB-like sequences in marsupials and monotremes, which suggests that *Mavericks* had not gone extinct, and were still circulating in the genomes of the stem mammals. Taken together, these results provide

new insights into the evolutionary history of *Mavericks* in mammals and the forces that have shaped their evolution. We also discuss several hypotheses which could explain the demise of the elements in birds and mammals.

Materials and Methods

We used the protein-primed polymerase B (pPOLB) of the common box-turtle (*Terrapene carolina*) *Maverick* (Barreat and Katzourakis, 2021b; Starrett et al., 2020) as a probe to screen the human genome with tblastn (Altschul et al., 1990) (assembly GRCh38.p13, BLOSUM 45 substitution matrix, word size = 2). We obtained two hits with e-values < 1e-7: one on chromosome 7 (chr7:37601124–37601615, query cover = 15%, e-value = 8e-8, percent identity = 31.14%) and another on chromosome 8 (chr8:12858023–12858613, query cover = 25%, e-value = 5e-8, percent identity = 29.95%). In order to assess orthology of the integrations in the genomes of other mammals, we initially tried to use nucleotide flanks and find pairs of hits in other genomes but this approach was not successful, due to extensive sequence degeneration. To overcome this limitation, we identified the three most proximal protein-coding genes upstream and downstream of the integrations, using the Ensembl genome browser (Howe et al., 2021). The sequence conservation of the genes allowed us to analyse synteny and orthology across mammals.

The protein-coding genes used as genetic landmarks for the chromosome 7 integration were: *NME8*, *EPDR1*, *GPR141*, *ELMO1*, *AOAH* and *ANLN*, and for chromosome 8: *C8orf48*, *DLC1*, *TRMT9B*, *LONRF1*, *FAM86B2* and *DEFB130A* (NCBI protein database (Sayers et al., 2021)). We used these sequences as queries in tblastn (set to the default) to screen the NCBI RefSeq genome database using the following taxon labels: ‘Monotremata (taxid:

9255)', 'Metatheria (taxid: 9263)', 'Afrotheria (taxid: 311790)', 'Xenarthra (taxid: 9348)', 'Laurasiatheria (taxid: 314145)' and 'Euarchontoglires (taxid: 314146)'. To verify the presence or absence of the integration, we screened the regions flanked by the genetic landmarks with the full set of proteins encoded by the 4 tetrapod *Mavericks* reported previously (Barreat and Katzourakis, 2021b) again with tblastn (e-value < 1e-5).

As an independent test for the presence of the integrations, we used the whole-genome alignment of 120 mammals published by Hecker and Hiller (Hecker and Hiller, 2020). We extracted the regions corresponding to the human pPOLB integrations and extended them assuming the ancestral gene coded for 1053 amino acids as in the box-turtle *Maverick* (chr7:37600749–37603904, chr8:12856919–12860047). The regions were extracted from the chr7 and chr8 whole-genome alignments in MAF format (85.16 GB and 81.07 GB, respectively) with custom *.bed files and using the mafsInRegion utility (Kuhn et al., 2013). Subalignments were transformed into fasta format with the MAF to FASTA programme (version 1.0.1) in Galaxy (Afgan et al., 2018). The taxa present in these subalignments were then compared to those obtained with the blast-hit method.

We merged the chromosome 7 and 8 subalignments corresponding to the *ppolb* gene using the –merge function in MAFFT (Katoh et al., 2002). The alignment was trimmed in trimAl version 1.4.rev22 (Capella-Gutiérrez et al., 2009) with the –automated1 option (which selects the optimal trimming method for the alignment). We then selected the best nucleotide substitution model in ModelTest-NG version 0.1.7 (Darriba et al., 2020) (TVM+Γ4 under the AIC, BIC and cAIC criteria). After determining the orthology of the integrations, we conducted a phylogenetic analysis under a relaxed-molecular clock using divergence-time calibrations in BEAST 2 version 2.6.7 (Bouckaert et al., 2019). For this, we used 13

calibration points obtained from TimeTree (Kumar et al., 2017) (Table S1) using normal distributions, and then ran a Bayesian MCMC chain for 100,000,000 generations, sampling every 5,000th generation. We confirmed that the analysis had converged by inspecting the mixing and stationarity of posterior samples and ensuring that the Estimated Sample Sizes (ESSs) were greater than 200 for all parameters (burn-in of 25%).

Finally, we tested the possibility of selective constraints acting on the integrations. To this end, we calculated the pairwise genetic distances (measured as the observed proportion of nucleotide differences) between the taxa in the chromosome 7 and chromosome 8 subalignments separately, and compared these to 100 randomly-sampled, surrounding non-coding genomic regions of the same size in each case. We built two empirical distributions from these, and then tested the hypothesis that the distribution of *Maverick* genetic distances was distinct from the non-coding distribution. This was done with a one-tailed, non-parametric Kolmogorov-Smirnov test. An R script was developed for this purpose using functionality from the Ape package (Paradis and Schliep, 2019; R Core Team, 2020). We further checked if the human integrations fell within piRNA clusters reported for chromosomes 7 and 8 (Rosenkranz et al., 2021), but this resulted in no matches (Supplementary Excel File 1).

Data availability

The supplementary materials for this study have been deposited in Figshare (https://dx.doi.org/10.6084/m9.figshare.17819708). The R Code written to compare distributions of pair-wise genetic distances is available on GitHub (https://github.com/josegabrielnb/pair-wise_distributions).

Results

We investigated the evolutionary history of the two human *Maverick* insertions initially discovered by local similarity to the DNA polymerase of the endogenous virus found in the genome of the common-box turtle, *Terrapene carolina*. We first identified the closest protein-coding genes surrounding the integrations in the human genome, which then allowed us to find the orthologous regions across mammals and assess whether the integrations were present or absent in these species. With a detailed knowledge of orthology, we used temporal information about the divergence of different host clades to reconstruct a time-calibrated phylogeny of the elements in mammals and uncover aspects of their evolutionary history.

We could determine that the *Maverick*-like integrations on chromosomes 7 and 8 are orthologous across the clade of placental mammals (Eutheria). The orthology of the regions was validated by the syntenic arrangement of the closest protein-coding genes, presence of the taxa in the whole-genome alignment of 120 mammals, and the relative arrangement of BLAST hits to different *Maverick* proteins which were consistent with the most common genetic organisation observed in vertebrate *Mavericks* (Barreat and Katzourakis, 2021b). In addition, the taxonomic distributions resulting from the blast-hit and whole-genome alignment methods were consistent with each other (Supplementary Excel File 1 and multiple sequence alignments). However, some taxa were present in the whole-genome alignment corresponding to the pPOLB marker which were not detected with the blast method, probably because of high divergence of the sequences (producing higher e-values than the cut-off). We could not detect orthologous integrations in the genomes of monotremes (*Ornithorhynchus anatinus*) or marsupials neither through blast nor in the whole-genome alignments.

Orthologues to the human chromosome 7 element were found in the genomes of primates, rodents, afrotherians (elephants, manatees and aardvark) and xenarthrans (sloths and armadillos) (Figure 1). In addition, the pPOLB marker was present in the whole genome-alignment for scandentians (tree shrews), dermopterans (flying lemurs), lagomorphs (hares) and two laurasiatherians: the star-nosed mole (*Condylura cristata*) and the European hedgehog (*Erinaceus europaeus*). This finding is consistent with the presence of hits to the major capsid, minor capsid, PW and integrase in other laurasiatherians, namely carnivores and cetaceans (Figure 1, Supplementary Excel File 1). As expected for a *Maverick*, the hits to the integrase are on the same strand as the major capsid, and did not show similarity to the integrases of retroelements in Censor (Kohany et al., 2006). In total, we found hits to 8 of the 9 conserved core proteins reported for vertebrate *Mavericks* in the orthologues of the human chromosome 7 integration (Figure 1). Orthologues of the human chromosome 8 integration were found in the genomes of primates and xenarthrans (blast-hit method). The pPOLB marker again had a broader taxonomic distribution in the whole-genome alignment and also included scandentians, dermopterans, lagomorphs and chiropterans (bats). The chromosome 8 element is more degraded than the chromosome 7 *Maverick*, we could only detect 4 hits to the set of 9 conserved core proteins (Figure 2). Most elements preserve fragments of the DNA polymerase, PM and PZ genes, while a single element in the nine-banded armadillo (*Dasypus novemcinctus*) showed a hit to the major capsid protein (Figure 2).

All the integrations were highly degraded showing strong signs of being non-functional at the protein level. The open-reading frames are interrupted with multiple early stop-codon and frameshift mutations (Figure 1, Figure 2). The elements also lack the terminal inverted

repeats characteristic of intact *Maverick* elements. In addition, the genetic distances of the elements are not significantly different from the distances between non-coding regions in their surrounding genomic neighbourhood (Kolmogorov-Smirnov test, $p > 0.99$, Figures S1 and S2), which suggests the elements have been evolving neutrally. Interestingly, the chromosome 7/8 *Maverick* orthologues appear to have been lost in some mammals since they could not be detected by the blast or whole-genome alignment methods but were still present in related species (however, this could also be caused by a gap in the assemblies of these species which would need to be re-sequenced). For example, the chromosome 7 integration appears to be absent from the genome of the Ugandan red colobus (*Piliocolobus tephrosceles*) while it is present in other colobine monkeys (*Trachypithecus francoisi*, *Rhinopithecus roxellana*). Similarly, the chromosome 8 integration seems to be absent from the genomes of the grey mouse lemur (*Microcebus murinus*) and Coquerel's sifaka (*Propithecus coquereli*), despite the presence of the pPOLB marker in the northern greater galago (*Otolemur garnettii*) (this sequence is present in the whole genome alignment and shows a blast hit to the pPOLB of the common-box turtle with an e-value of 5e-4).

The presence of the pPOLB marker on both the chromosome 7 and 8 orthologues allowed us to estimate a joint phylogeny with the common ancestor of the elements represented by the root of the tree (Figure 3). The Bayesian estimate for the age of this common ancestor suggests it most likely existed in the Permian Period, 262 to 257 million years ago (Mya), which correspond to the mean and median of the posterior probability distribution (Figure 3, Table 1). However, the 95% highest probability densities for this estimate span the Carboniferous to Triassic Periods, 335 to 200 Mya, so the ancestor may have existed from the end of the Palaeozoic to the start of the Mesozoic Eras. The minimum dates of the human insertions were estimated at 103 (100-105) Mya for the chromosome 7 element, and 102

(99-105) Mya for the chromosome 8 element, which correspond to the transition between the Early and Late Cretaceous. The evolutionary rates we inferred in the analysis (mean = 0.0026, IQR = [0.0021, 0.0029]) agree with the overall pan-genomic substitution rates reported previously for mammals of 0.0027 nucleotide substitutions/site/million years (Zhang et al., 2014), which confirms that the timescales were calibrated properly.

In addition to characterisation of these two placental-wide orthologues, we discovered multiple hits to the box-turtle *Maverick* polymerase in non-orthologous regions in the genomes of marsupials and egg-laying mammals (monotremes). We found 40 hits to the *Maverick* DNA polymerase in the genomes of 7 marsupials, both from the American and Australian continents (Table S2). Three hits were found in the genomes of egg-laying mammals, one in the platypus (*Ornithorhynchus anatinus*) and two in the short-beaked echidna (*Tachyglossus aculeatus*) (Table S2).

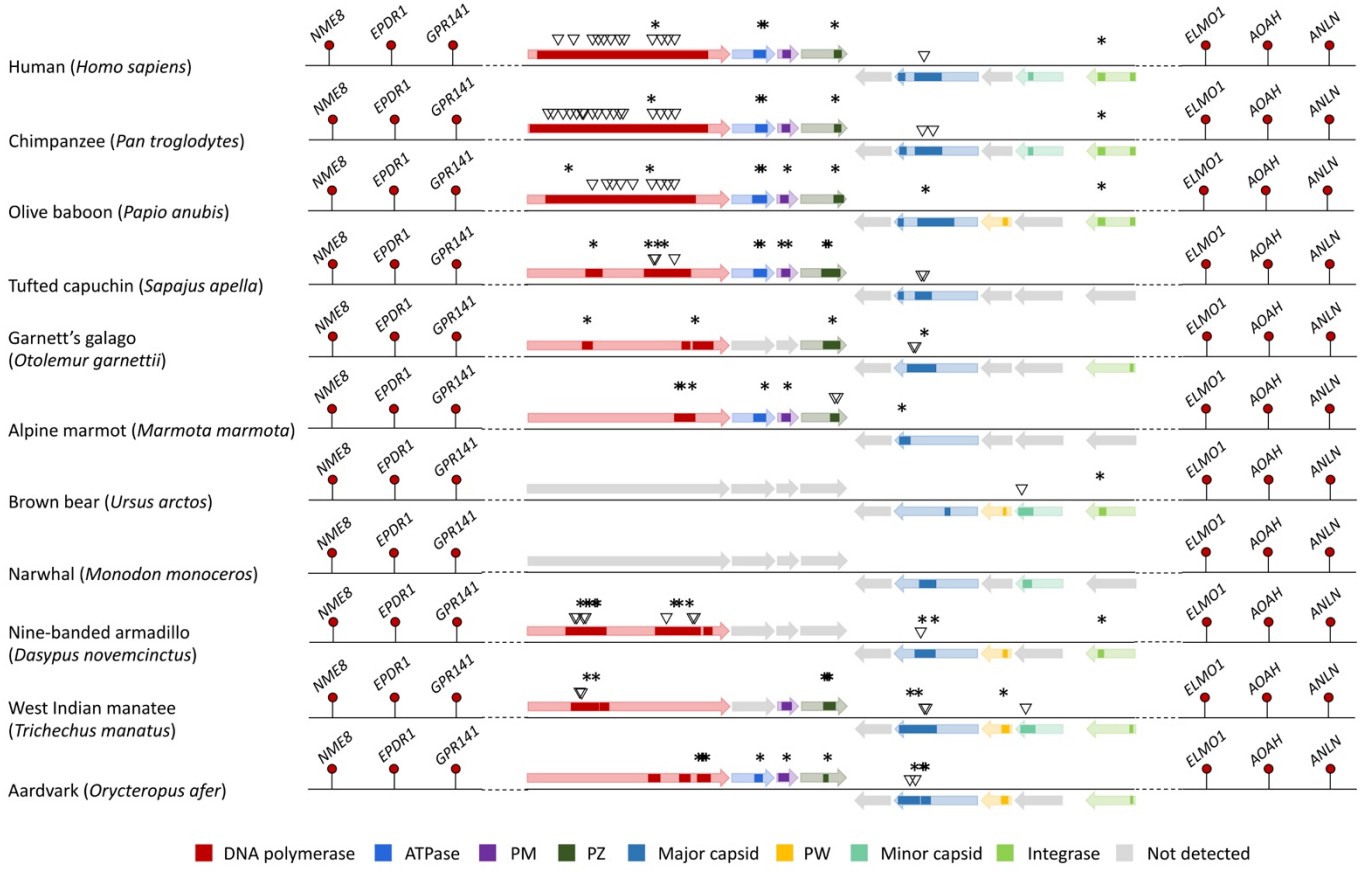


Figure 1. Comparison of the regions orthologous to the chromosome 7 *Maverick* integration in humans (*Homo sapiens*). Coloured rectangles denote regions of local similarity to the proteins encoded by the common box-turtle *Terrapene carolina* (tblastn, e-value < 0.05). Asterisks represent predicted early stop codons, and inverted triangles show frameshift mutations (inferred from the blast local alignments and modelling of the gene structure in Genewise (Birney et al., 2004)). The arrow outlines represent the open reading frames found in the intact *Maverick* of the turtle. Hits to eight of the nine (8/9) conserved proteins found in vertebrate *Mavericks* could be detected across the orthologues (missing hits to the protease).

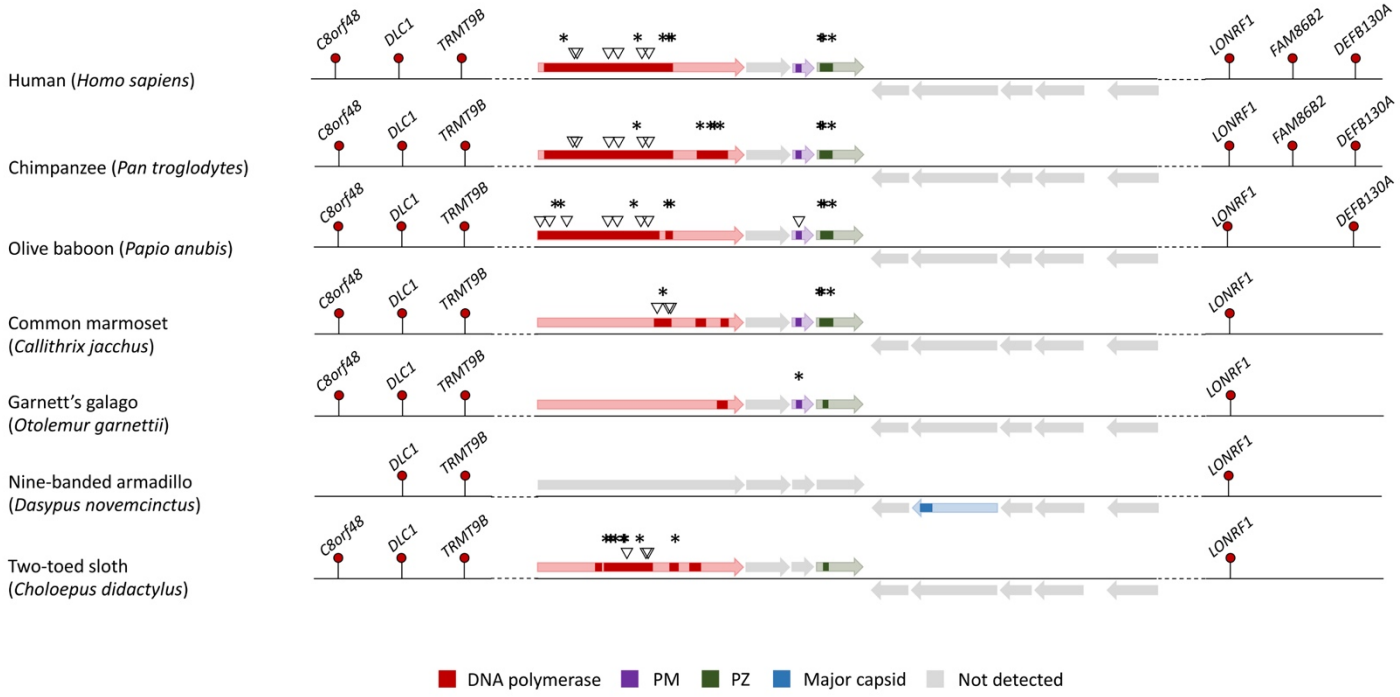


Figure 2. Comparison of the regions orthologous to the chromosome 8 *Maverick* integration in humans (*Homo sapiens*). Coloured rectangles denote regions of local similarity to the proteins encoded by the common box-turtle *Terrapene carolina* (tblastn, e-value < 0.05). Asterisks represent predicted early stop codons, and inverted triangles show frameshift mutations (inferred from the blast local alignments and modelling of the gene structure in Genewise (Birney et al., 2004)). The arrow outlines represent the open reading frames found in the intact *Maverick* of the turtle. Only four of the nine (4/9) conserved proteins in vertebrate *Mavericks* could be detected, indicating more extensive degeneration compared to the chromosome 7 group of orthologues.

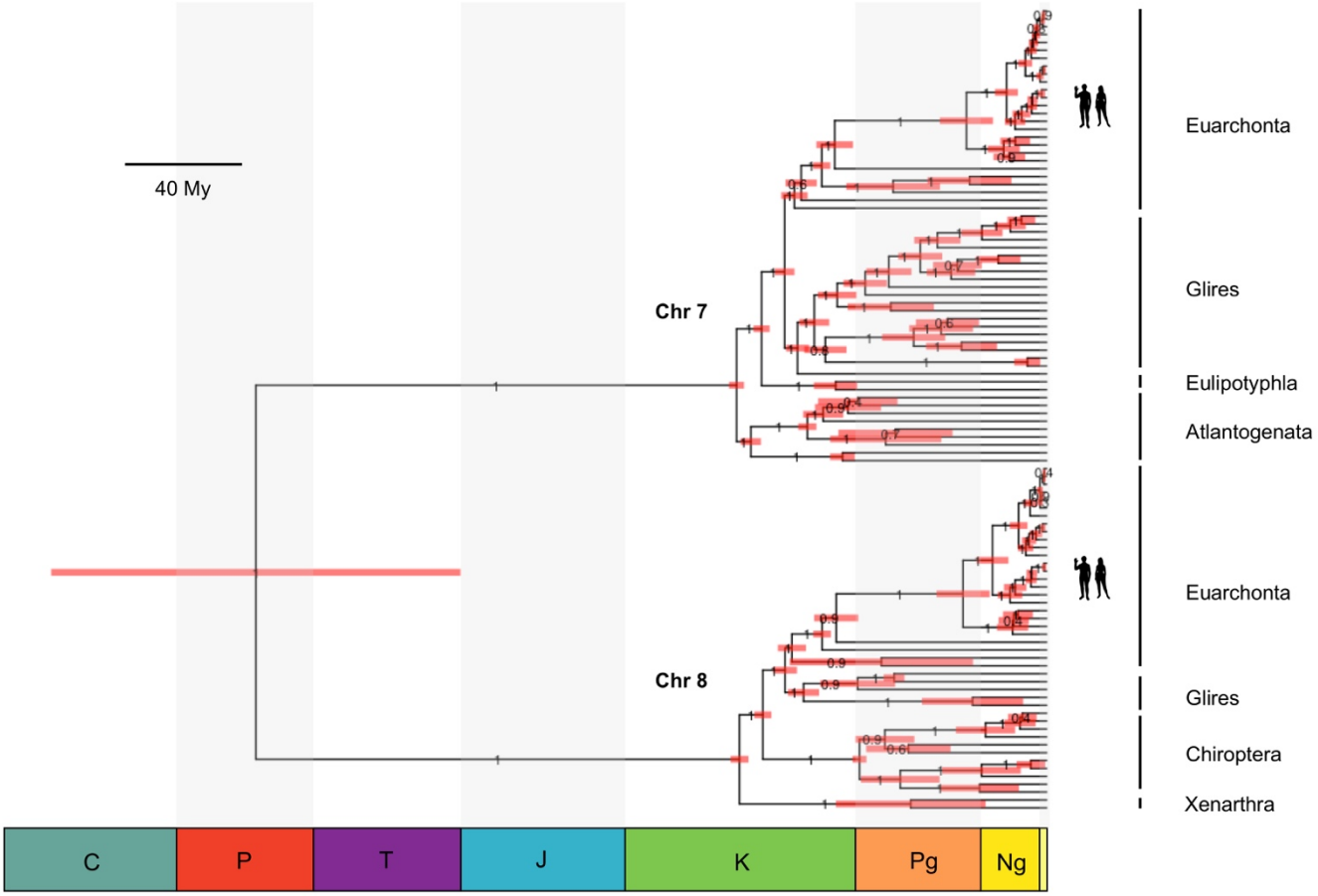


Figure 3. Phylogenetic tree of the mammalian polymerase B-like sequences orthologous to the viral fossils in human chromosomes 7 and 8. The horizontal node bars represent the 95% highest probability densities (HPD) for the age of the ancestor. Scale bar in million years (My). C: Carboniferous, P: Permian, T: Triassic, J: Jurassic, K: Cretaceous, Pg: Paleogene, Ng: Neogene.

Table 1. Posterior estimates for the age of the root, and the minimum ages of the chromosome 7/8 integrations (in million years from the present).

Node	Mean	Median	95% HPD
Root	262	257	[200–335]
Chromosome 7 CA	103	103	[100–105]
Chromosome 8 CA	102	102	[99–105]

Discussion

Our analyses have revealed two ancient integrations from *Mavericks* in the human genome that are shared across the placental mammals. Critically, we have established orthology of the two integrations across placentals by describing the conserved arrangement of protein-coding genes around the integration loci (conserved synteny). The discovery of orthologous integrations to the human chromosome 7 and 8 elements in other mammals is strong evidence for their authenticity, and rules out possible sequence contamination and misassembly (Aswad and Katzourakis, 2016). Therefore, we were able to authenticate the two human *Maverick* elements and confirm the existence of *Maverick* integrations in the genomes of placental mammals. This changes our view of the evolutionary history of *Mavericks* in vertebrates which were thought to be entirely absent from the mammalian lineage (Barreat and Katzourakis, 2021b; Haapa-Paananen et al., 2014; Kapitonov and Jurka, 2006; Pritham et al., 2007).

There is a blurry boundary between viruses and transposons (Barreat and Katzourakis, 2021a; Koonin et al., 2021), and one possibility is that the two integrations we describe originated from a transposable element that evolved from *Mavericks*. Before it was realised that they encode all the genes necessary for capsid formation (Krupovič et al., 2014; Krupovič and Bamford, 2008), and that these genes are under strong purifying selection (Barreat and Katzourakis, 2021b), *Mavericks* were considered transposons (Kapitonov and Jurka, 2006; Pritham et al., 2007). In the human chromosome 7 group of orthologues, we found remnants of the DNA-packaging ATPase, major and minor capsid proteins, in addition to the DNA polymerase and integrase, and other 3 core proteins found in the *Mavericks* of vertebrates. These findings suggest that the ancestor of the element encoded

the genes necessary to form a virion and integrated as an endogenous virus. Similarly, we found a fragment of the major capsid in the chromosome 8 orthologue in the genome of the nine-banded armadillo. If these two integrations had been derived from a *Maverick*-like transposon, we would expect natural selection to have deleted the genes responsible for capsid morphogenesis since these would serve no purpose. We would have expected also to find a greater copy number of the elements, and perhaps an intact element with inverted repeats required for mobilisation. However, our findings are more consistent with an integrated viral genome that has progressively degenerated in their hosts. Therefore, these two integrations are clearly highly degraded endogenous viral elements, not transposon sequences.

We estimated a minimum age for the chromosome 7 and 8 integrations of 103 (100–105) and 102 (99–105) My, respectively, which represent to our knowledge the oldest non-retroviral EVEs found in the human genome. An endogenous retrovirus-L element (ERV-L) of a comparable age (at least 104–110 My) has also been identified in humans and shown to be orthologous across the placental mammals (Lee et al., 2013). Comparison of the sequence divergence between the 5' and 3' LTRs of elephant and human (the 5' and 3' LTRs would be identical upon integration in the common ancestor), further suggests that the element may have integrated 24-36 My prior to the initial split between afrotherians and boreoeutherians (Lee et al., 2013). In terms of DNA viruses, an endogenous parvovirus-like element discovered in an intron of the *Ellis van Creveld syndrome 2* gene was shown to be present in primates, carnivores, ungulates and dolphins but not in afrotherians, giving it a minimum age of 98 My (Liu et al., 2011). The human genome thus preserves remnants of multiple viruses that infected our ancestors during the Mesozoic Era.

Discovery of the human element found on chromosome 8, together with the one found on chromosome 7, allowed us to gain a more detailed understanding of the evolutionary history of the elements before the diversification of placentals. Our time-calibrated phylogeny with both elements allowed us to infer the age of the root. This analysis suggests that *Mavericks* were circulating in the ancestors of mammals 262 (200–335) Mya, which corresponds to the Carboniferous to Triassic periods. Further evidence is the discovery of multiple pPOLB-like sequences at non-orthologous positions in the genomes of marsupials and egg-laying mammals, which indicate that active *Mavericks* infected the most recent common ancestor of mammals. Clear orthology of the chromosome 7 and 8 integrations across placental mammals together with their absence in marsupials and monotremes, suggest that the viruses continued to be active and integrated into the genome of the placental ancestor after the split with marsupials. Molecular estimates place this divergence around 172 (168–178) Mya (dos Reis et al., 2012), which is consistent with the age of the earliest fossil eutherian, *Juramaia sinensis*, from the Late Jurassic (160 Mya) of China (Bi et al., 2018; Luo et al., 2011). Thus, it seems highly likely that active *Mavericks* persisted in the genomes of the early eutherian ancestors to the end of the Jurassic Period.

The extensive degeneration of the chromosome 7 and 8 integrations also suggest that they had already been inactivated by the time of the most recent common ancestor of placentals (~102 Mya) (Kumar et al., 2017), pointing to an older age of integration. This is evidenced by deletion of the genes downstream of *pz* in most chromosome 8 orthologues, and the absence of the *pro* (protease) gene in the chromosome 7 insertions. The two *Maverick* insertions thus resemble those found in birds which are highly degraded and whose genes contain multiple inactivating mutations (Barreat and Katzourakis, 2021b). We show that the two insertions found in humans and placental mammals are likely to be non-functional since

they do not seem to be under selection, they do not localise to piRNA clusters and have been lost on several occasions in several species of mammals.

The reasons for the demise of *Mavericks* in mammals and birds remains a mystery. Here we hypothesise several plausible scenarios. It has been suggested that *Mavericks* may function as a virophage-induced defence against the infection of large DNA viruses (Katzourakis and Aswad, 2014; Koonin and Krupovič, 2017), in particular iridoviruses (Barreat and Katzourakis, 2021b). Iridoviruses are important pathogens of fish, amphibians and non-avian reptiles but they do not seem to infect either birds or mammals (Chinchar et al., 2017; Williams et al., 2005). One possibility is that iridoviruses went extinct in mammals and birds as a result of this *Maverick* defence system. Once their viral hosts went extinct, endogenous virophages (which depend on a host virus for replication) would not have been able to mobilise and this would have led to their degeneration. This hypothesis could be tested by performing iridovirus infection experiments on cell cultures of hosts which carry intact *Maverick* elements (teleost fish, amphibians or non-avian reptiles). Alternatively, endogenous *Mavericks* could have been co-opted as a defence against exogenous counterparts. Exaptation of endogenous viruses to serve this purpose has been reported extensively, for example in case of the *ev3*, *ev6* and *ev9* genes in chicken (Robinson et al., 1981), *Fv1* and *Fv4* genes in mice (Ikeda and Sugimura, 1989; Jolicoeur and Baltimore, 1976; Jolicoeur and Rassart, 1980; Kozak et al., 1984) and the Jaagsiekte endogenous retrovirus (enJS56A1) in sheep which are able to restrict exogenous viruses (Murcia et al., 2007). Evolution of an effective antiviral defence system would lead to the extinction of the exogenous virus and degeneration of the defence locus once the selective pressure from the pathogen has been lifted (Aswad and Katzourakis, 2012). Therefore, the *Maverick* integrations found in humans and other placentals could represent an antiviral defence

system that has served its purpose and is now decaying. However, *Mavericks* have not been linked to any specific pathology and their exogenous stage has yet to be observed. Still another possibility is that these *Mavericks* were incidentally inactivated by host innate immune genes that evolved to fight other viral pathogens. For example, cytidine deaminases of the APOBEC3 family function as an antiviral defence in eutherian mammals and have been shown to target both DNA viruses and the DNA stage of retroviruses (Conticello et al., 2005; Ito et al., 2020; Krishnan et al., 2018). APOBEC3 enzymes leave a characteristic signal of C-to-T or G-to-A hypermutation in the genome of the virus or provirus (Jern et al., 2007; Lee et al., 2008), which could be tested by reconstructing the genome of the viral ancestor and comparing it to the sequences of the endogenous elements.

In addition to the viral arms-race scenarios, other hypotheses relating to changes in host biology could also explain the demise of *Mavericks* in the genomes of mammals and birds. It is possible that the receptor used for entry has been lost or has acquired resistance mutations in these animals and the elements could no longer amplify in the germline by reinfection or cross-species transmissions. Support for this idea would require identification of the host cell receptor used for entry of *Mavericks* and assessment of its present state across vertebrates, which are both still unknown. Finally, the demise of *Mavericks* in these lineages might be linked to terrestrialisation and the origin of the amniotic egg, which may have limited the opportunities for the spread of these viruses and led to their ultimate degeneration. Although the natural history of *Mavericks* is still not understood, it seems they are most successful in aquatic organisms which suggests a water-borne mode of transmission (Barreat and Katzourakis, 2021b; Haapa-Paananen et al., 2014), perhaps they become active at the early stages when sperm and ova are released into the water by animals with external fertilisation. In fact, potentially active elements reach their highest diversity

and copy numbers in fish (where cross-species transmissions have also been detected), while only two intact elements have been discovered in amniotes (in turtles and lizards) (Barreat and Katzourakis, 2021b; Starrett et al., 2020). It therefore seems likely that the reduced opportunities for transmission could have led to a decrease in copy numbers which eventually made *Mavericks* prone to extinction.

Our results support a model where *Maverick* endogenous viruses were circulating in the genomes of the stem mammals at the end of the Palaeozoic and remained active in eutherian ancestors into the Jurassic Period. The elements found integrated on human chromosomes 7 and 8 represent the relics of ancient viruses that infected our ancestors >102 million years ago. Several hypotheses could explain the demise of *Mavericks* in mammals, which range from exaptation for an antiviral defence system followed by decay, inactivation by an innate immune mechanism, loss/mutation of the host cell receptor or changes to the biology of hosts which lowered their chances of transmission. Experimental studies focusing on the molecular virology of *Mavericks* and the virus-host interactions are needed to shed light on these issues.

References

Afgan E, Baker D, Batut B, van den Beek M, Bouvier D, Ech M, Chilton J, Clements D, Coraor N, Grüning BA, Guerler A, Hillman-Jackson J, Hiltemann S, Jalili V, Rasche H, Soranzo N, Goecks J, Taylor J, Nekrutenko A, Blankenberg D. 2018. The Galaxy platform for accessible, reproducible and collaborative biomedical analyses: 2018 update. *Nucleic Acids Res* 46. doi:10.1093/nar/gky379

Aimola G, Beythien G, Aswad A, Kaufer BB. 2020. Current understanding of human herpesvirus 6 (HHV-6) chromosomal integration. *Antiviral Research*. doi:10.1016/j.antiviral.2020.104720

Altschul SF, Gish W, Miller W, Myers EW, Lipman DJ. 1990. Basic local alignment search tool. *Journal of Molecular Biology* 215:403–410. doi:10.1016/S0022-2836(05)80360-2

Aswad A, Aimola G, Wight D, Roychoudhury P, Zimmermann C, Hill J, Lassner D, Xie H, Huang M-L, Parrish NF, Schultheiss H-P, Venturini C, Lager S, Smith GCS, Charnock-Jones DS, Breuer J, Greninger AL, Kaufer BB. 2020. Evolutionary history of endogenous *Human Herpesvirus 6* reflects human migration out of Africa. *Molecular Biology and Evolution*. doi:10.1093/molbev/msaa190

Aswad A, Katzourakis A. 2017. A novel viral lineage distantly related to herpesviruses discovered within fish genome sequence data. *Virus Evolution* 3. doi:10.1093/ve/vex016

Aswad A, Katzourakis A. 2016. Paleovirology: the study of endogenous viral elements. Virus Evolution: Current Research and Future Directions. doi:10.21775/9781910190234.10

Aswad A, Katzourakis A. 2012. Paleovirology and virally derived immunity. *Trends in Ecology and Evolution*. doi:10.1016/j.tree.2012.07.007

Barreat JGN, Katzourakis A. 2021a. Paleovirology of the DNA viruses of eukaryotes. *Trends in Microbiology*. doi:10.1016/j.tim.2021.07.004

Barreat JGN, Katzourakis A. 2021b. Phylogenomics of the *Maverick* virus-like mobile genetic elements of vertebrates. *Molecular Biology and Evolution*. doi:10.1093/molbev/msaa291

Beachboard DC, Horner SM. (2016). Innate immune evasion strategies of DNA and RNA viruses. *Current Opinion in Microbiology* 32. doi: 10.1016/j.mib.2016.05.015

Bi S, Zheng X, Wang X, Cignetti NE, Yang S, Wible JR. 2018. An Early Cretaceous eutherian and the placental-marsupial dichotomy. *Nature* 558. doi:10.1038/s41586-018-0210-3

Birney E, Clamp M, Durbin R. 2004. GeneWise and Genomewise. *Genome Research* 14. doi:10.1101/gr.1865504

Bouckaert R, Vaughan TG, Barido-Sottani J, Duchêne S, Fourment M, Gavryushkina A, Heled J, Jones G, Kühnert D, de Maio N, Matschiner M, Mendes FK, Müller NF, Ogilvie HA, du Plessis L, Popinga A, Rambaut A, Rasmussen D, Siveroni I, Suchard MA, Wu CH, Xie D, Zhang C, Stadler T, Drummond AJ. 2019. BEAST 2.5: An advanced software platform for Bayesian evolutionary analysis. *PLoS Computational Biology* 15. doi:10.1371/journal.pcbi.1006650

Capella-Gutiérrez S, Silla-Martínez JM, Gabaldón T. 2009. trimAl: A tool for automated alignment trimming in large-scale phylogenetic analyses. *Bioinformatics* 25:1972–1973. doi:10.1093/bioinformatics/btp348

Chen N, Xia P, Li S, Zhang T, Wang TT, Zhu J. (2017). RNA sensors of the innate immune system and their detection of pathogens. *IUBMB Life* 69. doi: 10.1002/iub.1625

Chinchar VG, Hick P, Ince IA, Jancovich JK, Marschang R, Qin Q, Subramaniam K, Waltzek TB, Whittington R, Williams T, Zhang QY. 2017. ICTV virus taxonomy profile: *Iridoviridae*. Journal of General Virology 98. doi:10.1099/jgv.0.000818

Conticello SG, Thomas CJF, Petersen-Mahrt SK, Neuberger MS. 2005. Evolution of the AID/APOBEC family of polynucleotide (deoxy)cytidine deaminases. *Molecular Biology and Evolution* 22. doi:10.1093/molbev/msi026

Darriba Di, Posada D, Kozlov AM, Stamatakis A, Morel B, Flouri T. 2020. ModelTest-NG: a new and scalable tool for the selection of dna and protein evolutionary models. *Molecular Biology and Evolution* 37:291–294. doi:10.1093/molbev/msz189

dos Reis M, Inoue J, Hasegawa M, Asher RJ, Donoghue PCJ, Yang Z. 2012. Phylogenomic datasets provide both precision and accuracy in estimating the timescale of placental mammal phylogeny. *Proceedings of the Royal Society B: Biological Sciences* 279. doi:10.1098/rspb.2012.0683

Feschotte C, Gilbert C. 2012. Endogenous viruses: Insights into viral evolution and impact on host biology. *Nature Reviews Genetics*. doi:10.1038/nrg3199

Goff, S. P. (2013). *Retroviridae. Fields virology*, 2, 1424-1473.

Haapa-Paananen S, Wahlberg N, Savilahti H. 2014. Phylogenetic analysis of *Maverick*/*Polinton* giant transposons across organisms. *Molecular Phylogenetics and Evolution* 78. doi:10.1016/j.ympev.2014.05.024

Hecker N, Hiller M. 2020. A genome alignment of 120 mammals highlights ultraconserved element variability and placenta-associated enhancers. *Gigascience* 9. doi:10.1093/gigascience/giz159

Holmes EC. 2011. The evolution of endogenous viral elements. *Cell Host and Microbe*. doi:10.1016/j.chom.2011.09.002

Howe KL, Achuthan P, Allen James, Allen Jamie, Alvarez-Jarreta J, Ridwan Amode M, Armean IM, Azov AG, Bennett R, Bhai J, Billis K, Boddu S, Charkhchi M, Cummins C, da Rin Fioretto L, Davidson C, Dodiya K, el Houdaigui B, Fatima R, Gall A, Giron CG, Grego T, Guijarro-Clarke C, Haggerty L, Hemrom A, Hourlier T, Izuogu OG, Juettemann T, Kaikala V, Kay M, Lavidas I, Le T, Lemos D, Martinez JG, Marugán JC, Maurel T, McMahon AC, Mohanan S, Moore B, Muffato M, Oheh DN, Paraschas D, Parker A, Parton A, Prosovetskaia I, Sakthivel MP, Abdul Salam AI, Schmitt BM, Schuilenburg H, Sheppard D, Steed E, Szpak M, Szuba M, Taylor K, Thormann A, Threadgold G, Walts B, Winterbottom A, Chakiachvili M, Chaubal A, de Silva N, Flint B, Frankish A, Hunt SE, IIsley GR, Langridge N, Loveland JE, Martin FJ, Mudge JM, Morales J, Perry E, Ruffier M, Tate J, Thybert D, Trevanion SJ, Cunningham F, Yates AD, Zerbino DR, Flicek P. 2021. Ensembl 2021. *Nucleic Acids Res* 49. doi:10.1093/nar/gkaa942

Ikeda H, Sugimura H. 1989. *Fv-4* resistance gene: a truncated endogenous murine leukemia virus with ecotropic interference properties. *Journal of Virology* 63. doi:10.1128/jvi.63.12.5405-5412.1989

Inoue Y, Kumagai M, Zhang X, Saga T, Wang D, Koga A, Takeda H. 2018. Fusion of *piggyBac*-like transposons and herpesviruses occurs frequently in teleosts. *Zoological Letters* 4. doi:10.1186/s40851-018-0089-8

Inoue Y, Saga T, Aikawa T, Kumagai M, Shimada A, Kawaguchi Y, Naruse K, Morishita S, Koga A, Takeda H. 2017. Complete fusion of a transposon and herpesvirus created the *Teratorn* mobile element in medaka fish. *Nature Communications* 8. doi:10.1038/s41467-017-00527-2

Ito J, Gifford RJ, Sato K. 2020. Retroviruses drive the rapid evolution of mammalian APOBEC3 genes. *Proc Natl Acad Sci U S A* 117. doi:10.1073/pnas.1914183116

Jern P, Stoye JP, Coffin JM. 2007. Role of APOBEC3 in genetic diversity among endogenous murine leukemia viruses. *PLoS Genetics* 3. doi:10.1371/journal.pgen.0030183

Johnson WE. 2019. Origins and evolutionary consequences of ancient endogenous retroviruses. *Nature Reviews Microbiology*. doi:10.1038/s41579-019-0189-2

Jolicoeur P, Baltimore D. 1976. Effect of *Fv-1* gene product on proviral DNA formation and integration in cells infected with murine leukemia viruses. *Proc Natl Acad Sci U S A* 73. doi:10.1073/pnas.73.7.2236

Jolicoeur P, Rassart E. 1980. Effect of *Fv-1* gene product on synthesis of linear and supercoiled viral DNA in cells infected with murine leukemia virus. *Journal of Virology* 33. doi:10.1128/jvi.33.1.183-195.1980

Kapitonov VV, Jurka J. 2006. Self-synthesizing DNA transposons in eukaryotes. *Proc Natl Acad Sci U S A* 103. doi:10.1073/pnas.0600833103

Katoh K, Misawa K, Kuma KI, Miyata T. 2002. MAFFT: A novel method for rapid multiple sequence alignment based on fast Fourier transform. *Nucleic Acids Res* 30:3059–3066. doi:10.1093/nar/gkf436

Katzourakis A, Aswad A. 2014. The origins of giant viruses, virophages and their relatives in host genomes. *BMC Biology*. doi:10.1186/s12915-014-0051-y

Katzourakis A, Gifford RJ. 2010. Endogenous viral elements in animal genomes. *PLoS Genetics* 6. doi:10.1371/journal.pgen.1001191

Kohany O, Gentles AJ, Hankus L, Jurka J. 2006. Annotation, submission and screening of repetitive elements in Repbase: RepbaseSubmitter and Censor. *BMC Bioinformatics* 7. doi:10.1186/1471-2105-7-474

Koonin EV, Dolja VV, Krupovič M, Kuhn JH. 2021. Viruses defined by the position of the virosphere within the replicator space. *Microbiology and Molecular Biology Reviews* 85. doi:10.1128/mmbr.00193-20

Koonin EV, Krupovič M. 2017. *Polintons*, virophages and transpovirons: a tangled web linking viruses, transposons and immunity. *Current Opinion in Virology* 25:7–15. doi:10.1016/j.coviro.2017.06.008

Kozak CA, Gromet NJ, Ikeda H, Buckler CE. 1984. A unique sequence related to the ecotropic murine leukemia virus is associated with the *Fv-4* resistance gene. *Proc Natl Acad Sci U S A* 81. doi:10.1073/pnas.81.3.834

Krishnan A, Iyer LM, Holland SJ, Boehm T, Aravind L. 2018. Diversification of AID/APOBEC-like deaminases in metazoa: Multiplicity of clades and widespread roles in immunity. *Proc Natl Acad Sci U S A* 115. doi:10.1073/pnas.1720897115

Krupovič M, Bamford DH. 2008. Virus evolution: how far does the double β-barrel viral lineage extend? *Nature Reviews Microbiology* 6:941–948. doi:10.1038/nrmicro2033

Krupovič M, Bamford DH, Koonin EV. 2014. Conservation of major and minor jelly-roll capsid proteins in *Polinton* (*Maverick*) transposons suggests that they are bona fide viruses. *Biology Direct* 9:1–7. doi:10.1186/1745-6150-9-6

Krupovič M, Koonin EV. 2016. Self-synthesizing transposons: Unexpected key players in the evolution of viruses and defense systems. *Current Opinion in Microbiology* 31:25–33. doi:10.1016/j.mib.2016.01.006

Krupovič M, Koonin EV. 2015. *Polintons*: A hotbed of eukaryotic virus, transposon and plasmid evolution. *Nature Reviews Microbiology* 13:105–115. doi:10.1038/nrmicro3389

Kuhn RM, Haussler D, James Kent W. 2013. The UCSC genome browser and associated tools. *Briefings in Bioinformatics* 14. doi:10.1093/bib/bbs038

Kumar S, Stecher G, Suleski M, Hedges SB. 2017. TimeTree: a resource for timelines, timetrees, and divergence times. *Molecular Biology and Evolution* 34. doi:10.1093/molbev/msx116

Lee A, Nolan A, Watson J, Tristem M. 2013. Identification of an ancient endogenous retrovirus, predating the divergence of the placental mammals. *Philosophical Transactions of the Royal Society B: Biological Sciences* 368. doi:10.1098/rstb.2012.0503

Lee YN, Malim MH, Bieniasz PD. 2008. Hypermutation of an ancient human retrovirus by APOBEC3G. *Journal of Virology* 82. doi:10.1128/jvi.00751-08

Liu H, Fu Y, Xie J, Cheng J, Ghabrial SA, Li G, Peng Y, Yi X, Jiang D. 2011. Widespread endogenization of densoviruses and parvoviruses in animal and human genomes. *Journal of Virology* 85. doi:10.1128/jvi.00828-11

Liu X, Kosugi S, Koide R, Kawamura Y, Ito J, Miura H, Matoba N, Matsuzaki M, Fujita M, Kamada AJ, Nakagawa H, Tamiya G, Matsuda K, Murakami Y, Kubo M, Aswad A, Sato K, Momozawa Y, Ohashi J, Terao C, Yoshikawa T, Parrish NF, Kamatani Y. 2020. Endogenization and excision of human herpesvirus 6 in human genomes. *PLoS Genetics* 16. doi:10.1371/JOURNAL.PGEN.1008915

Luo ZX, Yuan CX, Meng QJ, Ji Q. 2011. A Jurassic eutherian mammal and divergence of marsupials and placentals. *Nature* 476. doi:10.1038/nature10291

McBride AA, Oliveira JG, McPhillips MG. 2006. Partitioning viral genomes in mitosis: Same idea, different targets. *Cell Cycle*. doi:10.4161/cc.5.14.3094

Murcia PR, Arnaud F, Palmarini M. 2007. The transdominant endogenous retrovirus enJS56A1 associates with and blocks intracellular trafficking of jaagsiekte sheep retrovirus Gag. *Journal of Virology* 81. doi:10.1128/jvi.01859-06

Paradis E, Schliep K. 2019. Ape 5.0: An environment for modern phylogenetics and evolutionary analyses in R. *Bioinformatics* 35. doi:10.1093/bioinformatics/bty633

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. doi:10.1016/j.gene.2006.08.008

Prusty BK, Gulve N, Rasa S, Murovska M, Hernandez PC, Ablashi D v. 2017. Possible chromosomal and germline integration of human herpesvirus 7. *Journal of General Virology* 98. doi:10.1099/jgv.0.000692

R Core Team. 2020. R: A language and environment for statistical computing.

Robinson HL, Astrin SM, Senior AM, Salazar FH. 1981. Host Susceptibility to endogenous viruses: defective, glycoprotein-expressing proviruses interfere with infections. *Journal of Virology* 40. doi:10.1128/jvi.40.3.745-751.1981

Rosenkranz D, Zischler H, Gebert D. 2021. piRNAclusterDB 2.0: update and expansion of the piRNA cluster database. *Nucleic Acids Res*. doi:10.1093/nar/gkab622

Sayers EW, Beck J, Bolton EE, Bourexis D, Brister JR, Canese K, Comeau DC, Funk K, Kim S, Klimke W, Marchler-Bauer A, Landrum M, Lathrop S, Lu Z, Madden TL, O'Leary N, Phan L, Rangwala SH, Schneider VA, Skripchenko Y, Wang J, Ye J, Trawick BW, Pruitt KD, Sherry ST. 2021. Database resources of the National Center for Biotechnology Information. *Nucleic Acids Res* 49. doi:10.1093/nar/gkaa892

Schaack S, Gilbert C, Feschotte C. 2010. Promiscuous DNA: Horizontal transfer of transposable elements and why it matters for eukaryotic evolution. *Trends in Ecology and Evolution*. doi:10.1016/j.tree.2010.06.001

Starrett GJ, Tisza MJ, Welch NL, Belford AK, Peretti A, Pastrana DV, Buck CB. 2020. Adintoviruses: a proposed animal-tropic family of midsize eukaryotic linear dsDNA (MELD) viruses. *Virus Evolution*. doi:10.1093/ve/veaa055

Williams T, Barbosa-Solomieu V, Chinchar VG. 2005. A decade of advances in iridovirus research. Advances in Virus Research. doi:10.1016/S0065-3527(05)65006-3

Yao Y, Frith MC. 2021. Improved DNA-versus-protein homology search for protein fossils. Lecture Notes in Computer Science (Including Subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics). doi:10.1007/978-3-030-74432-8_11

Zhang, G. et al. 2014. Comparative genomics reveals insights into avian genome evolution and adaptation. *Science*. doi: 10.1126/science.1251385.

Discussion

In this thesis, various aspects of the evolutionary biology of eukaryoviruses in the kingdom *Bamfordvirae* have been explored. The paleovirological and phylogenetic analyses presented in chapters II and III, are in line with recent findings suggesting that this is an ancient lineage of viruses. It is likely these viruses were, and continue to be, a major force shaping the ecology of microbial eukaryotes. The modelling work of chapter IV highlighted the interactions of these viruses with unicellular organisms, and the potential role they may have had influencing antiviral immunity and the evolution of multicellularity. Finally, in chapter V, comparative genomic and phylogenetic evidence was described for the existence of members of this lineage integrated in the human genome, which were active in the genome of the most recent common ancestor of mammals. Together, these insights have shed light on the evolutionary history of the viral kingdom *Bamfordvirae* and their ancient interaction with eukaryotes. In the sections below, I will discuss the main findings and implications of this thesis, describe the limitations and suggest future prospects for study in this field.

Main findings and implications

Mavericks are endogenous viruses, not transposons

The ability to form a virion is what distinguishes viruses from capsidless genetic replicators such as transposons or plasmids (Koonin et al., 2021). In *Mavericks*, the predicted PY protein was found to be homologous to the major capsid protein of *Paramecium bursaria Chlorella virus*, by 3-dimensional modelling guided by a resolved crystal structure (Krupovič et al., 2014). The PX protein of *Mavericks* was hypothesised to be a homologue

of the penton protein based on conservation of the predicted secondary structure to that of virophages (Krupovič et al., 2014). In chapter I, evolutionary evidence for the existence of *Maverick* virions was presented. In particular, hundreds of elements that are intact were identified (suggesting they are replication-competent), and which encode the full set of four core proteins involved in capsid formation (major and minor capsid, ATPase and protease). These genes were found to be under strong purifying selection showing evidence of cross-species transmissions (horizontal transfers), which suggests that *Mavericks* infect new hosts by forming virions. This is also consistent with the constrained genome sizes that were found in intact elements, which indicates the genome size is under selection; this agrees with the observation that changes in the genome size can affect the stability of the capsid in the related adenoviruses (Smith et al., 2009). The phylogenetic analyses presented in chapter III, which were based on the four core proteins involved in capsid formation, also revealed that *Mavericks* are entirely nested within the lineages of the kingdom *Bamfordvirae* which are known to form virions; probably as a sister group to adenoviruses.

To obtain additional evidence for the existence of *Maverick* virions, I predicted the structures of the major and minor capsid proteins, and their capsomeres, *de novo* using AlphaFold (Jumper et al., 2021). Viruses in kingdom *Bamfordvirae* form capsids with two main building-blocks: a trimer of the major capsid protein which assemble at the faces of the icosahedron, and a pentamer of the minor capsid protein which forms the vertices (Figure 1, Figure 2) (Krupovič and Bamford, 2008a).

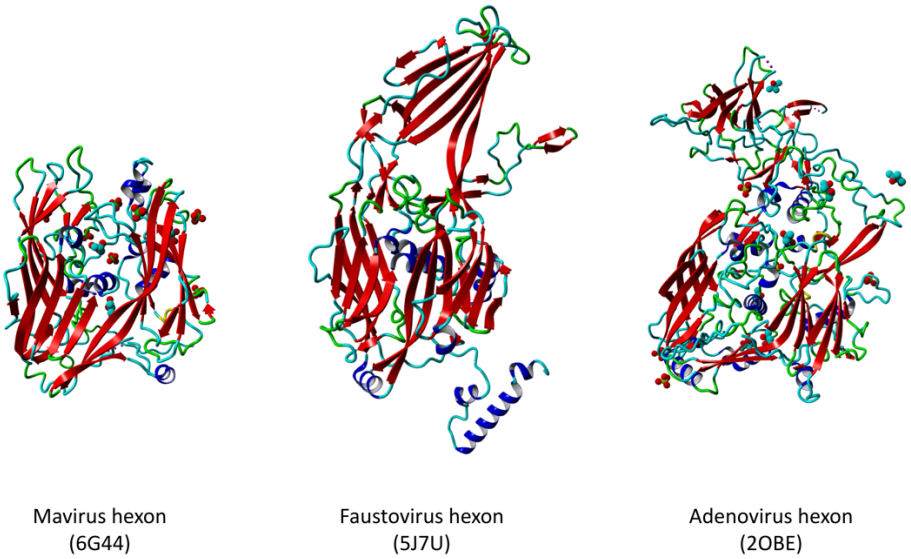


Figure 1. Monomeric structures of the major capsid proteins (hexons) of Mavirus (*Lavidaviridae*), Faustovirus (NCLDV) and Chimpanzee adenovirus type 68 (*Adenoviridae*). The double jelly-roll fold is conserved.

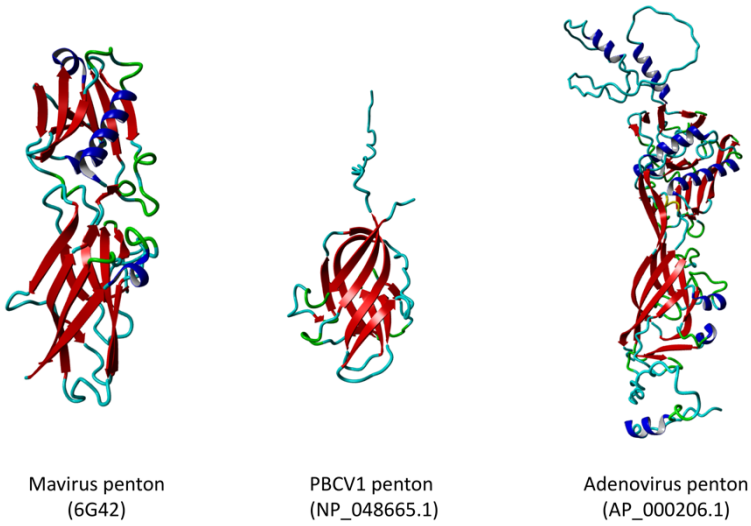


Figure 2. Monomeric structures of the minor capsid proteins (pentons) from Mavirus (*Lavidaviridae*), *Paramecium bursaria Chlorella virus 1* (PBCV1, NCLDV) and Human adenovirus 5 (*Adenoviridae*). The single jelly-roll fold (lower module) is conserved. The PBCV1 and HAdV5 protein structures were predicted in AlphaFold (pLDDT = 82.8, 76.9).

The structures for the major capsid trimer and minor capsid pentamer were predicted in ColabFold (Mirdita et al., 2022), an implementation of the AlphaFold algorithm using MMseqs2 and the "UniRef+Environmental" protein database (Steinegger and Söding, 2017). The models were refined using 12 recycles followed by relaxation of the structure via energy minimisation using AMBER (Case et al., 2005; Eastman et al., 2017). In Figure 3, we can see how the major capsid proteins adopt the characteristic vertical double-jelly roll, while the minor capsid proteins adopt a single vertical jelly-roll fold, as expected. Figures 4 and 5 show how the major and minor capsid proteins assemble into a trimer or pentamer, respectively. The structures for the major capsid trimers were predicted to a high accuracy, i.e. plDDT > 90 (mean plDDT, *Anolis carolinensis* = 94.24; mean plDDT, *Terrapene carolina* = 94.53). The structure of the pentamers had a lower quality score, but their values suggest that the protein backbone was modelled to a good accuracy, i.e. plDDT > 84 (mean plDDT, *Anolis carolinensis* = 86.92; mean plDDT, *Terrapene carolina* = 84.61).

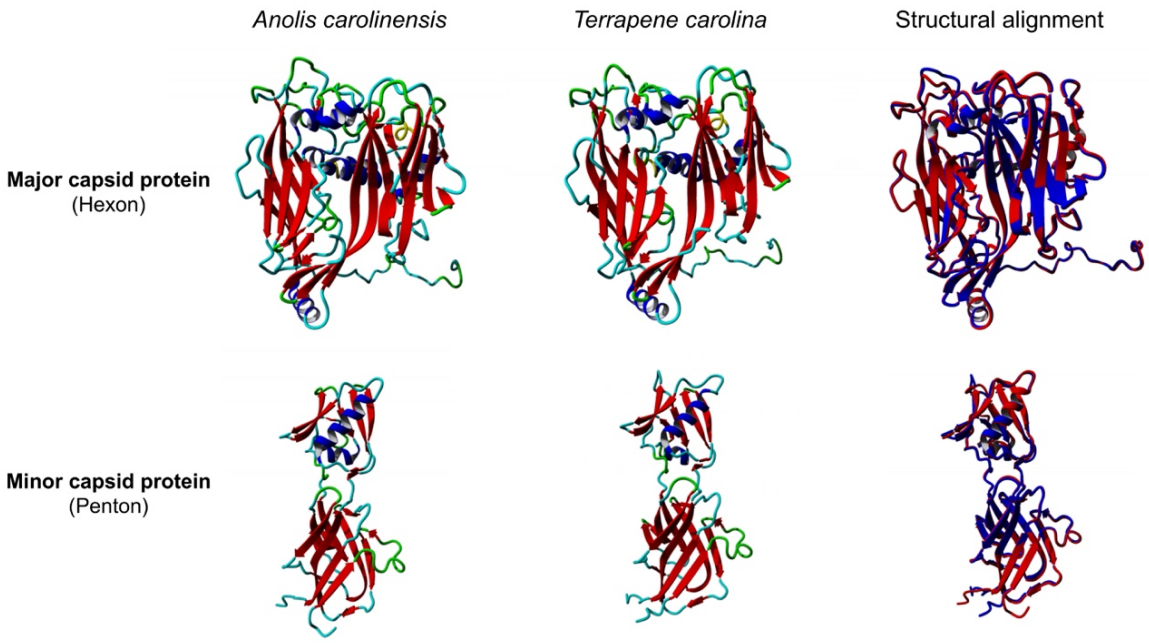


Figure 3. *De novo* predicted structures of the major capsid and minor capsid monomers for the intact *Mavericks* in *Anolis carolinensis* and in *Terrapene carolina*. **Top row:** the major capsid proteins adopt the vertical, double jelly-roll fold (β-sheets in red). **Bottom row:** the minor capsid proteins adopt a single vertical jelly-roll fold (lower β-sheets, also in red). The structural alignments show the close correspondence in the predicted protein structures: the structure from *Anolis carolinensis* in red, the one from *Terrapene carolina* in blue.

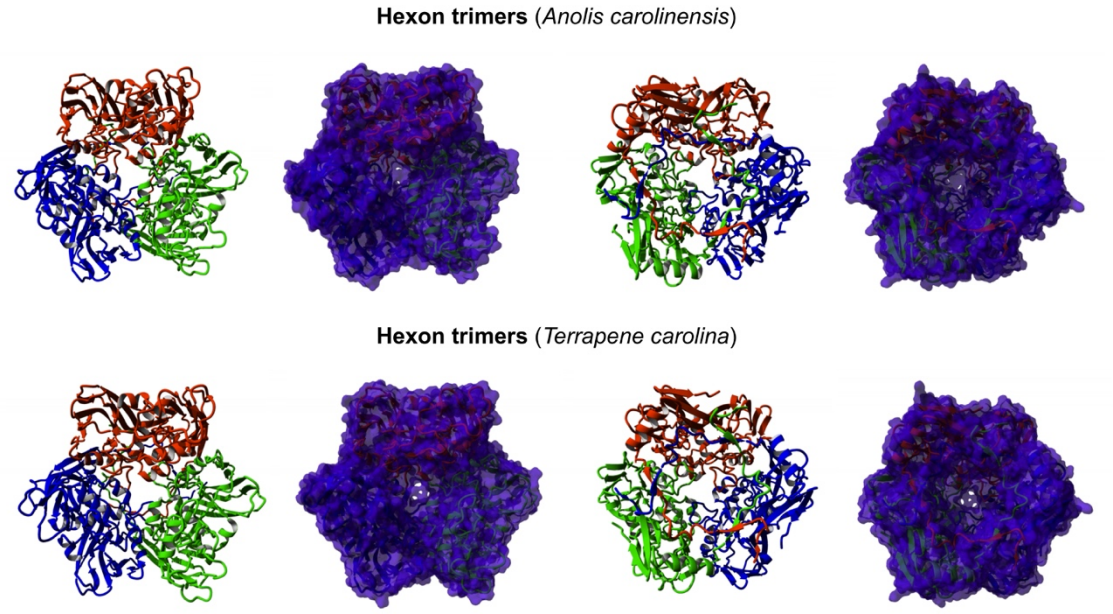


Figure 4. *De novo* predicted structure for the trimer of hexon proteins in the *Mavericks* of *Anolis carolinensis* and *Terrapene carolina*. The diagrams show the top and bottom views of the proteins in ribbon-style, next to a depiction with an outline of the molecular surface (calculated from the atomic radii and by rolling a water molecule over the model (Connolly, 1983)).

These structural predictions together with the evolutionary evidence presented in this thesis leave little doubt that *Mavericks* are endogenous viruses capable of forming virions. The virion would be predicted to be particle with icosahedral symmetry and an internal lipid membrane (given that *Mavericks* encode the FtsK-HerA packaging ATPase observed in viruses with internal lipid membranes). Given the allometric scaling between virus genome length and the size of the virion, *Mavericks* with genome sizes comprised between 12.7 and 24.6 kb, are expected to form particles 48-67 nm in diameter (Cui et al., 2014) (see Appendix). By comparison, native virions of the Mavirus virophage (with a genome size of 19 kb, expected size 58 nm), have a diameter between 70-80 nm (Born et al., 2018; Fischer and Suttle, 2011); the inferred size of *Maverick* virions may represent a lower limit.

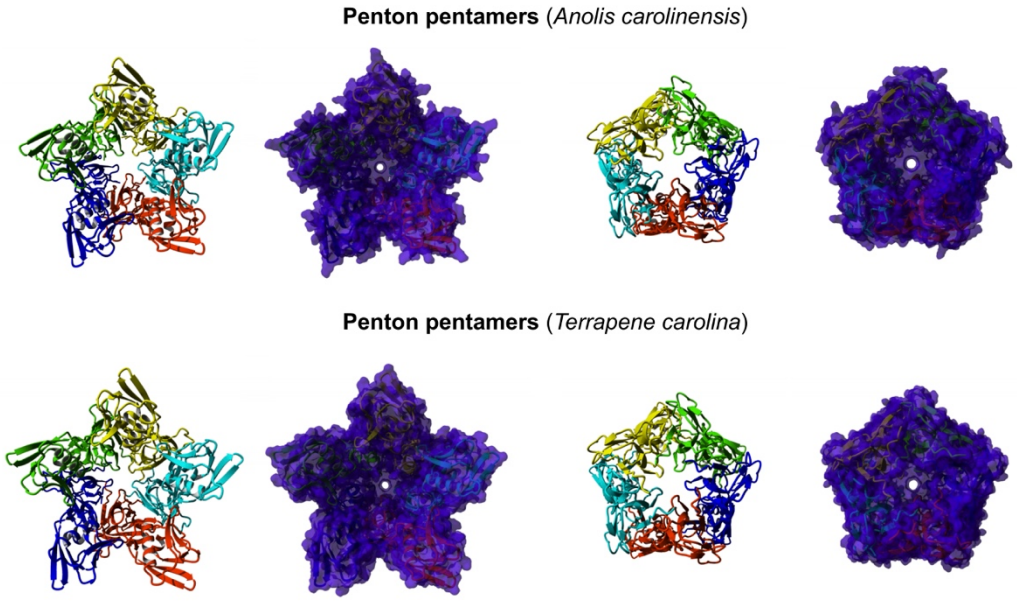


Figure 5. *De novo* predicted structure of the minor capsid pentamers in the *Mavericks* of *Anolis carolinensis* and *Terrapene carolina*. The diagrams show the top and bottom views of the proteins in ribbon-style, next to a depiction with an outline of the molecular surface.

Bamfordviruses have ancient origins and probably infected the first eukaryotes

Strong evidence for the stable coexistence of *Mavericks* with vertebrates for hundreds of millions of years (at least 419 My), was shown in chapter II. In chapter III, the phylogenetic reconstructions also showed that the *Mavericks* of vertebrates and insects likely form a monophyletic group, and so it is possible that *Mavericks* infected the common ancestor of arthropods and vertebrates ~560 million years ago (Cartwright and Collins, 2007). Also, the placement of *Mavericks* as one of the more derived lineages is consistent with their unique gain of an integrase, and the older estimated ages for the common ancestor of NCLDVs. The existence of the NCLDV ancestor more than a billion years ago is supported by several early gene exchanges involving subunits of the DNA-dependent RNA polymerase, the nucleotide-primed DNA polymerase and actins (da Cunha et al., 2022; Guglielmini et al.,

2019; Takemura et al., 2015). The basal position observed for virophages in these rooted phylogenies further suggests that bamfordviruses have infected eukaryotes for an even longer period of time. These observations imply that bamfordviruses were probably an integral part of the virome of proto-eukaryotes and could have played important roles in eukaryogenesis. Furthermore, the existence of families of bamfordviruses that infect bacteria and archaea, suggest that the evolution of this viral kingdom may extend back to the origins of the first cellular organisms, as has been hypothesised for the realm *Duplodnaviria* (Krupovič et al., 2020).

Mavericks infected the most recent common ancestor of mammals

Mavericks were thought to be altogether absent from the genomes of mammals (Barreat and Katzourakis, 2021; Haapa-Paananen et al., 2014; Kapitonov and Jurka, 2006; Pritham et al., 2007). However, the report for a *Maverick*-like integration into human chromosome 7 (Starrett et al., 2020) was authenticated in chapter V, and an additional integration was discovered on human chromosome 8 via comparative genomic analyses. The orthology across the placental mammals together with the phylogenetic dating, suggest these are some of the most ancient viral integrations so far described in the human genome (>100 million years old) (Lee et al., 2013; Liu et al., 2011). In both cases, the integrations are highly degraded and they preserve remnants of the genes involved in capsid formation, so it is most likely they represent degraded genomic fossils from endogenous viruses. Furthermore, discovery of non-orthologous *Maverick*-like sequences in monotreme and marsupial genomes, indicates that active *Mavericks* were circulating in the most recent common ancestor of mammals. These findings show that bamfordviruses have also interacted for extended periods of time with mammals, including the direct ancestors of humans.

Theory provides insights into the dynamic interactions between cells, virophages and NCLDVs

The mathematical and computational models presented in chapter IV, have provided mechanistic insights into the role of virophage integration in the ecological interactions and evolution of cell-virus-virophage systems. In particular, it was shown that virophages that can enter cells independently are expected to be found more commonly integrated in eukaryotic host genomes, than virophages which enter cells by forming a complex. The models also show that very strong inhibition of virus replication by virophages can stabilise the system dynamics, which would make the system more robust to stochastic collapses (Wodarz, 2013); and explain the abundance and persistence of virophages and NCLDVs across globally spread metagenomes (Endo et al., 2020; Paez-Espino et al., 2019; Roux et al., 2017; Schulz et al., 2020). The potential effect of programmed cell-death, inhibition by virophages and multicellularity as antiviral defence mechanisms were also explored. These analyses revealed that the presence of inhibitory virophages and the suicide of an infected cell can benefit the host population. Similarly, the spatial clustering of cells that can result from a transition to multicellularity also had a protective effect. At least from a theoretical perspective, it seems plausible that viruses may have contributed to the evolution of cell-death and to the establishment of multicellularity (Iranzo et al., 2014), which are so ubiquitous amongst eukaryotes and enabled the evolution of increasingly complex organisms (Fisher et al., 2020; Klim et al., 2018). These ideas are consistent with the ancient origins of virophages and NCLDVs, which existed at the time of proto-eukaryotes and interacted with them closely (as evidence by the ancestral NCLDV-eukaryote gene exchanges).

Limitations

The works presented here have been limited by several factors, including computational resources, biases in sampling, and the lack of experimental systems to test some of our hypotheses. For example, the phylogenomic analysis of *Mavericks* presented in chapter II, were obtained by a search of sequences based on similarity to the pPOLB and integrase of a previously reported *Maverick* in the zebrafish, *Danio rerio*. Although this approach succeeded in finding the degraded *Mavericks* of birds and other reptiles, it failed to detect those of mammals (which were later recovered using a *Maverick* from *Terrapene carolina*). Similarly, major groups of vertebrates such as the cartilaginous fish did not have many genome assemblies available and it is still unclear whether these animals truly lack *Mavericks*. In the study about the evolution of the major lineages of eukaryotic bamfordviruses (chapter III), only a sample from a reduced number of sequences (54) could be used, compared to the number of genomic and metagenomic assemblies available (~9,000), due to the high computational expense of calculations. Therefore, a small subset of genomes had to be chosen, trying to maximise the diversity of sequences. Although the main interest was to understand the deepest branches in the topology (for which a small but diverse dataset is sufficient), a more detailed understanding of the evolutionary biology of the lineage will depend on a greater sample of sequences and more powerful computational methods/resources. Chapter V was limited to focus on just a few integrations, but to understand the history of *Mavericks* in mammals will probably require a systematic search at a broader scale.

The models presented in chapter IV were designed to draw qualitative, but not quantitative, conclusions. Some parameters were chosen based on experimental measurements or

biophysical principles (Taylor et al., 2014), but others were chosen to fit our null expectation for the baseline scenario (predator-prey dynamics in the presence of neutral virophages). The models further assumed an idealised, well-mixed system of only three interacting species with no external intervening factors. It will be important to test our predictions experimentally in a system with a unicellular eukaryote infected by a giant virus and its virophage. This would allow to measure key model parameters and to develop quantitative predictions of the system dynamics. The models also ignored the effect that kin selection could have on these dynamics, which would add another layer of complexity, but may be of relevance to explain the evolution of the altruistic phenotypes analysed in this system (Bourke, 2011).

Future perspectives

Elucidating the natural history and molecular virology of *Mavericks* remains a major open question in the field. Despite the strong evolutionary and structural evidence supporting the capacity of *Mavericks* to form virions, their viral particles have still not been observed. Bioinformatic analysis of transcriptomic data sets could provide indications as to which species, tissues or life stages of host organisms express *Maverick* genes. Since *Mavericks* appear in the germline of many organisms, it is possible that expression coincides with the formation of gametes which undergo reprogramming of the epigenome and activation of repressed mobile genetic elements (Deniz et al., 2019). Ultimately, experimental studies will be needed to understand the life cycle of the *Maverick* viruses and how they interact with their hosts. These could be used in combination with synthetic biology approaches to study the viruses in cell cultures. For example, the gene coding for the major and minor capsid proteins of the Mavirus virophage have been synthesised, cloned into plasmids and

expressed in *E. coli*, leading to the formation of virus-like particles (Born et al., 2018). The properties of the capsid proteins from ancient endogenous retroviruses (RELIK/PSIV) have also been studied by gene-synthesis, cloning into plasmids and expression in cell cultures (Goldstone et al., 2010). In this case, chimeric constructs of the capsid genes in an *Equine infectious anaemia virus* background, allowed to follow the incorporation of these ancient capsid proteins into mature virions (Goldstone et al., 2010). More generally, entire viral genomes may be constructed synthetically and used as infectious clones; this approach has been used to synthesise and study the virology of the 1918/19 influenza virus (which caused the "Spanish flu"), poliovirus, retroviruses, a bat SARS-like coronavirus and the phage ΦX174 (Wimmer et al., 2009). In addition, α-capsid protein antisera could be produced to look for *in vivo* expression of *Maverick* capsid genes in hosts such as the zebrafish (*Danio rerio*) (Peter Simmonds, personal communication).

One of the most intriguing possibilities, is that vertebrate *Mavericks* may have adopted a virophage lifestyle and protect hosts against NCLDVs, probably against iridoviruses which infect animals where intact *Mavericks* are found (fish, amphibians, non-avian reptiles). For example, the only two reptiles in which we found intact *Mavericks* are also known hosts of ranaviruses which cause skin lesions, anorexia, lethargy and are potentially deadly (de Voe et al., 2004; Stöhr et al., 2013). However, another possibility is that *Mavericks* may be susceptible to the higher body temperatures of mammals and birds, which are endotherms, and so decayed in these animals. Indeed, iridoviruses seem not to be able to replicate at temperatures > 32 °C, which seems to explain their restricted distribution to ectothermic vertebrates (Chinchar et al., 2009).

The theoretical models presented in chapter IV have also provided hypotheses that can be tested in the laboratory. In particular, it could be confirmed whether the virophage entry mechanism is the main driving factor of the system dynamics by genetically manipulating the virophage genomes. Viruses can be modified with the CRISPR-Cas9 system, which has been used successfully to introduce mutations in the genomes of dsDNA viruses such as *Vaccinia virus* (*Poxviridae*) and Epstein-Barr virus (*Herpesviridae*) (Yuan et al., 2015; Yuen et al., 2015). Therefore, the capsids of Mavirus/Sputnik virophages could be exchanged which would be expected to reverse the integration patterns we observe in the genomes of eukaryotes. Genetic exchanges could also determine whether the different integrases of Mavirus or Sputnik-like virophages have different affinities for integrating into specific sequences (e.g. eukaryote vs. virus genome). In addition, it could be tested to what extent virophage inhibition and programmed cell-death occur as antiviral defences in microbial eukaryotes. Testing the effect of a transition to multicellularity in the presence of a pathogenic virus might prove more difficult, but could be explored by using a species that can alternate between unicellular/clonal multicellular states, such as some choanoflagellates (Brunet and King, 2017), and exposing them to a virus isolated from this species.

The immense diversity of viruses in the kingdom *Bamfordvirae*, presents many opportunities for future investigations as well. In particular, answering whether they existed at the time of the Last Universal Common Ancestor (LUCA) is an important issue. So far, the root of the bamfordvirus tree has been assumed to lie between the prokaryotic and eukaryotic groups (Woo et al., 2021). However, the root could potentially lie between the families of bacterial viruses (phages) and the archaeo/eukaryotic viruses which would reinforce this view. Although only two species of archaeoviruses are recognised in the family *Turriviridae*, additional viruses have been discovered as proviruses integrated in the

genomes of archaea (Krupovič and Bamford, 2008b). A systematic study of the position of the root, which considers the diversity of archaeal viruses, may shed light on the early evolution of this viral kingdom.

Conclusion

The viral kingdom *Bamfordvirae* constitutes one of the most diverse groups of viruses on planet Earth, and probably one of the most ancient ones, whose origins possibly extend to the beginnings of cellular life (~4 billion years ago). In eukaryotes, they diversified to an extent not seen anywhere else in the virosphere; evolving into giant viruses and their parasites, endogenous elements that populate the genomes of eukaryotes and highly abundant viruses in the oceans and freshwaters. In this thesis, I provided evidence that *Mavericks* are endogenous viruses of eukaryotes and not transposons. Discovery and characterisation of their virions will open opportunities to investigate a new group of viruses, whose biology remains entirely unknown. The evolutionary history of viruses in this kingdom was also explored, which provided new interpretations on their evolution and support the existence of this viral lineage at the time of the first eukaryotic organisms. The theoretical models presented have given insights into the dynamic interactions of virophages and NCLDVs, the patterns of integration of virophages into eukaryotic genomes, and the potential contribution of viruses to the evolution of programmed cell-death and multicellularity. As the number of genomes and metagenomes continues to increase, we are expected to uncover a greater diversity of viruses in this lineage, which will refine our understanding of their evolutionary history, the impact they have had on their hosts and on our planet's ecosystems.

References

Barreat JGN, Katzourakis A. 2021. Phylogenomics of the *Maverick* virus-like mobile genetic elements of vertebrates. *Molecular Biology and Evolution*. doi:10.1093/molbev/msaa291

Born D, Reuter L, Mersdorf U, Mueller M, Fischer MG, Meinhart A, Reinstein J. 2018. Capsid protein structure, self-assembly, and processing reveal morphogenesis of the marine virophage Mavirus. *Proc Natl Acad Sci U S A* 115. doi:10.1073/pnas.1805376115

Bourke AFG. 2011. The validity and value of inclusive fitness theory. *Proceedings of the Royal Society B: Biological Sciences*. doi:10.1098/rspb.2011.1465

Brunet T, King N. 2017. The origin of animal multicellularity and cell differentiation. *Developmental Cell*. doi:10.1016/j.devcel.2017.09.016

Cartwright P, Collins A. 2007. Fossils and phylogenies: integrating multiple lines of evidence to investigate the origin of early major metazoan lineages. *Integrative and Comparative Biology* 47. doi:10.1093/icb/icm071

Case DA, Cheatham TE, Darden T, Gohlke H, Luo R, Merz KM, Onufriev A, Simmerling C, Wang B, Woods RJ. 2005. The Amber biomolecular simulation programs. *Journal of Computational Chemistry*. doi:10.1002/jcc.20290

Chinchar VG, Hyatt A, Miyazaki T, Williams T. 2009. Family *Iridoviridae*: poor viral relations no longer. *Current Topics in Microbiology and Immunology*. doi:10.1007/978-3-540-68618-7_4

Connolly ML. 1983. Analytical molecular surface calculation. *Journal of Applied Crystallography* 16. doi:10.1107/s0021889883010985

Cui J, Schlub TE, Holmes EC. 2014. An allometric relationship between the genome length and virion volume of viruses. *Journal of Virology* 88. doi:10.1128/jvi.00362-14

da Cunha V, Gaia M, Ogata H, Jaillon O, Delmont TO, Forterre P. 2022. Giant viruses encode actin-related proteins. *Molecular Biology and Evolution* 39. doi:10.1093/molbev/msac022

de Voe R, Geissler K, Elmore S, Rotstein D, Lewbart G, Guy J. 2004. Ranavirus-associated morbidity and mortality in a group of captive eastern box turtles (*Terrapene carolina carolina*). *Journal of Zoo and Wildlife Medicine* 35. doi:10.1638/03-037

Deniz Ö, Frost JM, Branco MR. 2019. Regulation of transposable elements by DNA modifications. *Nature Reviews Genetics*. doi:10.1038/s41576-019-0106-6

Eastman P, Swails J, Chodera JD, McGibbon RT, Zhao Y, Beauchamp KA, Wang LP, Simmonett AC, Harrigan MP, Stern CD, Wiewiora RP, Brooks BR, Pande VS. 2017. OpenMM 7: Rapid development of high performance algorithms for molecular dynamics. *PLoS Computational Biology* 13. doi:10.1371/journal.pcbi.1005659

Endo H, Blanc-Mathieu R, Li Y, Salazar G, Henry N, Labadie K, de Vargas C, Sullivan MB, Bowler C, Wincker P, Karp-Boss L, Sunagawa S, Ogata H. 2020. Biogeography of marine giant viruses reveals their interplay with eukaryotes and ecological functions. *Nature Ecology and Evolution* 4. doi:10.1038/s41559-020-01288-w

Fischer MG, Suttle CA. 2011. A virophage at the origin of large DNA transposons. *Science* (1979) 231–234. doi: 10.1126/science.1199412

Fisher RM, Shik JZ, Boomsma JJ. 2020. The evolution of multicellular complexity: The role of relatedness and environmental constraints: Evolution of multicellular complexity. *Proceedings of the Royal Society B: Biological Sciences* 287. doi:10.1098/rspb.2019.2963rspb20192963

Goldstone DC, Yap MW, Robertson LE, Haire LF, Taylor WR, Katzourakis A, Stoye JP, Taylor IA. 2010. Structural and functional analysis of prehistoric lentiviruses uncovers an ancient molecular interface. *Cell Host and Microbe* 8. doi:10.1016/j.chom.2010.08.006

Guglielmini J, Woo AC, Krupovič M, Forterre P, Gaia M. 2019. Diversification of giant and large eukaryotic dsDNA viruses predated the origin of modern eukaryotes. *Proc Natl Acad Sci U S A*. doi:10.1073/pnas.1912006116

Haapa-Paananen S, Wahlberg N, Savilahti H. 2014. Phylogenetic analysis of *Maverick*/*Polinton* giant transposons across organisms. *Molecular Phylogenetics and Evolution* 78. doi:10.1016/j.ympev.2014.05.024

Iranzo J, Lobkovsky AE, Wolf YI, Koonin EV. 2014. Virus-host arms race at the joint origin of multicellularity and programmed cell death. *Cell Cycle* 13:3083–3088. doi:10.4161/15384101.2014.949496

Jumper J, Evans R, Pritzel A, Green T, Figurnov M, Ronneberger O, Tunyasuvunakool K, Bates R, Žídek A, Potapenko A, Bridgland A, Meyer C, Kohl SAA, Ballard AJ, Cowie A, Romera-Paredes B, Nikolov S, Jain R, Adler J, Back T, Petersen S, Reiman D, Clancy E, Zielinski M, Steinegger M, Pacholska M, Berghammer T, Bodenstein S, Silver D, Vinyals O, Senior AW, Kavukcuoglu K, Kohli P, Hassabis D. 2021. Highly accurate protein structure prediction with AlphaFold. *Nature* 596. doi:10.1038/s41586-021-03819-2

Kapitonov VV, Jurka J. 2006. Self-synthesizing DNA transposons in eukaryotes. *Proc Natl Acad Sci U S A* 103. doi:10.1073/pnas.0600833103

Klim J, Gladki A, Kucharczyk R, Zielenkiewicz U, Kaczanowski S. 2018. Ancestral state reconstruction of the apoptosis machinery in the common ancestor of eukaryotes. *G3: Genes, Genomes, Genetics* 8. doi:10.1534/g3.118.200295

Koonin EV, Dolja VV, Krupovič M, Kuhn JH. 2021. viruses defined by the position of the virosphere within the replicator space. *Microbiology and Molecular Biology Reviews* 85. doi:10.1128/mmbr.00193-20

Krupovič M, Bamford DH. 2008a. Virus evolution: how far does the double β-barrel viral lineage extend? *Nature Reviews Microbiology* 6:941–948. doi:10.1038/nrmicro2033

Krupovič M, Bamford DH. 2008b. Archaeal proviruses TKV4 and MVV extend the PRD1-adenovirus lineage to the phylum Euryarchaeota. *Virology* 375. doi:10.1016/j.virol.2008.01.043

Krupovič M, Bamford DH, Koonin EV. 2014. Conservation of major and minor jelly-roll capsid proteins in *Polinton* (*Maverick*) transposons suggests that they are *bona fide* viruses. *Biology Direct* 9:1–7. doi:10.1186/1745-6150-9-6

Krupovič M, Dolja VV, Koonin EV. 2020. The LUCA and its complex virome. *Nature Reviews Microbiology* 18. doi:10.1038/s41579-020-0408-x

Lee A, Nolan A, Watson J, Tristem M. 2013. Identification of an ancient endogenous retrovirus, predating the divergence of the placental mammals. *Philosophical Transactions of the Royal Society B: Biological Sciences* 368. doi:10.1098/rstb.2012.0503

Liu H, Fu Y, Xie J, Cheng J, Ghabrial SA, Li G, Peng Y, Yi X, Jiang D. 2011. Widespread endogenization of densoviruses and parvoviruses in animal and human genomes. *Journal of Virology* 85. doi:10.1128/jvi.00828-11

Mirdita M, Schütze K, Moriwaki Y, Heo L, Ovchinnikov S, Steinegger M. 2022. ColabFold: making protein folding accessible to all. *Nature Methods*. doi:10.1038/s41592-022-01488-1

Paez-Espino D, Zhou J, Roux S, Nayfach S, Pavlopoulos GA, Schulz F, McMahon KD, Walsh D, Woyke T, Ivanova NN, Eloe-Fadrosh EA, Tringe SG, Kyrpides NC. 2019. Diversity, evolution, and classification of virophages uncovered through global metagenomics. *Microbiome* 7. doi:10.1186/s40168-019-0768-5

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. doi:10.1016/j.gene.2006.08.008

Roux S, Chan LK, Egan R, Malmstrom RR, McMahon KD, Sullivan MB. 2017. Ecogenomics of virophages and their giant virus hosts assessed through time series metagenomics. *Nature Communications* 8. doi:10.1038/s41467-017-01086-2

Schulz F, Roux S, Paez-Espino D, Jungbluth S, Walsh DA, Denef VJ, McMahon KD, Konstantinidis KT, Eloe-Fadrosh EA, Kyrpides NC, Woyke T. 2020. Giant virus diversity and host interactions through global metagenomics. *Nature* 578. doi:10.1038/s41586-020-1957-x

Smith AC, Poulin KL, Parks RJ. 2009. DNA Genome size affects the stability of the adenovirus virion. *Journal of Virology* 83. doi:10.1128/jvi.01644-08

Starrett GJ, Tisza MJ, Welch NL, Belford AK, Peretti A, Pastrana D v, Buck CB. 2020. Adintoviruses: a proposed animal-tropic family of midsize eukaryotic linear dsDNA (MELD) viruses. *Virus Evolution*. doi:10.1093/ve/veaa055

Steinegger M, Söding J. 2017. MMseqs2 enables sensitive protein sequence searching for the analysis of massive data sets. *Nature Biotechnology*. doi:10.1038/nbt.3988

Stöhr AC, Blahak S, Heckers KO, Wiechert J, Behncke H, Mathes K, Günther P, Zwart P, Ball I, Rüschoff B, Marschang RE. 2013. Ranavirus infections associated with skin lesions in lizards. *Veterinary Research* 44. doi:10.1186/1297-9716-44-84

Takemura M, Yokobori SI, Ogata H. 2015. Evolution of eukaryotic DNA polymerases via interaction between cells and large DNA viruses. *Journal of Molecular Evolution* 81. doi:10.1007/s00239-015-9690-z

Taylor BP, Cortez MH, Weitz JS. 2014. The virus of my virus is my friend: ecological effects of virophage with alternative modes of coinfection. *Journal of Theoretical Biology* 354:124–136. doi:10.1016/j.jtbi.2014.03.008

Wimmer E, Mueller S, Tumpey TM, Taubenberger JK. 2009. Synthetic viruses: a new opportunity to understand and prevent viral disease. *Nature Biotechnology*. doi:10.1038/nbt.1593

Wodarz D. 2013. Evolutionary dynamics of giant viruses and their virophages. *Ecology and Evolution* 3:2103–2115. doi:10.1002/ece3.600

Woo AC, Gaia M, Guglielmini J, da Cunha V, Forterre P. 2021. Phylogeny of the *Varidnaviria* morphogenesis module: congruence and incongruence with the tree of life and viral taxonomy. *Frontiers in Microbiology* 12. doi:10.3389/fmicb.2021.704052

Yuan M, Zhang W, Wang J, al Yaghchi C, Ahmed J, Chard L, Lemoine NR, Wang Y. 2015. Efficiently editing the vaccinia virus genome by using the CRISPR-Cas9 system. *Journal of Virology* 89. doi:10.1128/jvi.00339-15

Yuen KS, Chan CP, Wong NHM, Ho CH, Ho TH, Lei T, Deng W, Tsao SW, Chen H, Kok KH, Jin DY. 2015. CRISPR/Cas9-mediated genome editing of Epstein–Barr virus in human cells. *Journal of General Virology* 96. doi:10.1099/jgv.0.000012

Appendix to Chapter I

Supplementary material: Paleovirology of the DNA viruses of eukaryotes

Glossary

Virus: a DNA or RNA genetic element which encodes a capsid for horizontal transmission between cells; viruses are obligate parasites of cells.

Reverse transcription: the process by which an RNA molecule is transcribed into a complementary DNA copy.

Virion: infectious viral particle that can successfully parasitise a susceptible and permissive host cell.

Prophage: the endogenous form of a bacteriophage integrated into a bacterial chromosome.

Provirus: the endogenous form of a retrovirus integrated into a eukaryotic chromosome.

Provirophage: the endogenous form of a virophage integrated into the genome of a eukaryotic or giant virus host.

Endogenous viral elements (EVEs): a viral sequence integrated into the genome of a cell in the germline, which has been vertically inherited (from parent to offspring). EVEs may

retain the ability to form infectious viral particles, they may be degraded (fossils) or have acquired a new function.

Baltimore groups: classification of viruses based on the nature of their genome and mRNA expression strategy. Viruses are classified into 7 Baltimore groups: dsDNA viruses (Group I), ssDNA viruses (Group II), dsRNA viruses (Group III), plus-sense ssRNA viruses (Group IV), minus-sense ssRNA viruses (Group V), reverse transcribing ssRNA viruses (Group VI) and reverse transcribing dsDNA viruses (Group VII).

Exaptation: biological trait that has been repurposed by natural selection to perform a new function.

Virophage: a virus satellite that has a negative (parasitic) effect on the replication of the helper virus. The virophages described so far are all parasites of giant viruses and belong to the family *Lavidaviridae*.

Giant virus: a virus in the phylum *Nucleocytoviricota* with a genome > 500 kb. All giant viruses discovered so far belong to this clade, but this does not exclude the possibility of other uncharacterized lineages of giant viruses.

Transposon (or transposable element): a sequence that can be copied into a new genomic location, either by encoding the enzymes required for transposition (autonomous) or by encoding the replication signals to use the enzymes from another element (non-autonomous). Unlike viruses, transposons do not encode capsids for horizontal transmission between cells.

Virus satellite: a virus that depends on coinfection of a cell by a helper virus in order to replicate.

Highlights

a) Paleovirology is informed by comparisons of virus protein structure, genome contents and architecture as well as the phylogenetic study of endogenous viral elements.

b) In particular, endogenous viral elements are increasing our knowledge of the exogenous and past viral diversity, extending the likely host ranges of a number of viruses and recalibrating our timescales of viral evolution.

c) Phylogenetic comparisons between endogenous viral elements and their hosts point to the widespread occurrence of cross-species transmissions during the evolutionary history of viruses.

d) *Mavericks* and *Teratorns* have emerged as new endogenous viral lineages of eukaryotes, probably representing at least three new viral families.

e) Geminivirus and parvovirus-like transposons as well as transpovirons have probably derived from viral ancestors.

Outstanding questions

a) How did the main viral superlineages evolve and in which hosts?

b) Have viral molecular fossils been preserved in the geological fossil record? Are there more instances of indirect viral fossils?

c) Can we recover more ancient viral DNA from exceptionally well-preserved samples?

d) What is the reason for the apparent absence of oncogenic viruses (except for hepadnaviruses) in the genomic fossil record?

e) Why are the rep genes of CRESS DNA viruses enriched in eukaryotic genomes?

f) What are the conditions for reactivation and viral particle formation for *Mavericks* and *Teratorns*?

g) How large is the diversity of virophages and how many integrate into the genomes of their hosts?

h) Do *Mavericks* constitute an inducible immune defence against NCLDV infection in eukaryotes?

i) What is the experimental evidence for the specific integration mechanisms that produced DNA virus EVEs?

j) What is the evolutionary history of the recently discovered fossil *Mavericks* in mammals?

k) Are the siRNAs and piRNAs derived from DNA virus EVEs actually used to restrict exogenous DNA viruses?

l) What is the extent of eukaryote-to-eukaryote horizontal gene transfer mediated by viruses?

m) Does the *MOV10L1* locus integration of HHV6A/B actually confer protection against integration by exogenous human herpesviruses?

n) Are there more viral lineages remaining to be discovered in the genomes of eukaryotes?

Appendix to Chapter II

Supplementary material: Phylogenomics of the *Maverick* virus-like mobile genetic elements of vertebrates

List S1. Taxonomy of host species whose genomes harbour *Mavericks*.

Fossil elements are indicated with a cross (†), while potentially active elements are depicted with 3 asterisks (⁂).

SUBPHYLUM VERTEBRATA

SUPERCLASS ACTINOPTERYGII

CLASS TELEOSTEI

Order ANGUILLIFORMES

Family Anguillidae

Species *Anguilla japonica* (†)

Species *Anguilla rostrata* (†)

Order BELONIFORMES

Family Adrianichthyidae

Species *Oryzias melastigma* (†)

Order BERYCIFORMES

Family Berycidae

Species *Beryx splendens* (†)

Family Holocentridae

Species *Neoniphon sammara* (†)

Family Monocentridae

Species *Monocentris japonica* (†)

Order CHARACIFORMES

Family Characidae

Species *Astyanax mexicanus* (†)

Species *Pygocentrus nattereri* (†)

Order CLUPEIFORMES

Family Clupeidae

Species *Clupea harengus* (†)

Species *Tenualosa ilisha* (†)

Order CYPRINIFORMES

Family Cyprinidae

Species *Carassius auratus* (†⁂)

Species *Cyprinus carpio* (†)

Species *Danionella dracula* (†)

Species *Danio rerio* (†⁂)

Species *Leuciscus waleckii* (†)

Species *Oxygymnocypris stewartii* (†⁂)

Species *Pimephales promelas* (†)

Species *Sinocyclocheilus anshuiensis* (†)

Species *Sinocyclocheilus grahami* (†)

Species *Sinocyclocheilus rhinocerous* (†)

Order CYPRINODONTIFORMES

Family Aplocheilidae

Species *Austrofundulus limnaeus* (†)

Family Cyprinodontidae

Species *Cyprinodon variegatus* (†)

Family Fundulidae

Species *Fundulus heteroclitus* (†)

Family Nothobranchiidae

Species *Nothobranchius furzeri* (†)

Family Poeciliidae

Species *Poecilia formosa* (†)

Species *Poecilia reticulata* (†)

Species *Xiphophorus couchianus* (⁂)

Species *Xiphophorus hellerii* (†⁂)

Species *Xiphophorus maculatus* (†⁂)

Family Rivulidae

Species *Kryptolebias marmoratus* (†)

Order ESOCIFORMES

Family Esocidae

Species *Esox lucius* (†)

Order GADIFORMES

Family Gadidae

Species *Gadus morhua* (†)

Species *Lota lota* (†)

Family Merlucciidae

Species *Merluccius capensis* (†)

Order GASTEROSTEIFORMES

Family Gasterosteidae

Species *Gasterosteus aculeatus* (†)

Order GYMNOTIFORMES

Family Electrophoridae

Species *Electrophorus electricus* (†)

Order LAMPRIDIFORMES

Family Stylephoridae

Species *Stylephorus chordatus* (†)

Order LOPHIIFORMES

Family Antennariidae

Species *Antennarius striatus* (†)

Order MYCTOPHIFORMES

Family Myctophidae

Species *Benthosema glaciale* (†)

Order OPHIDIIFORMES

Family Carapidae

Species *Carapus acus* (†)

Order OSTEOGLOSSIFORMES

Family Mormyridae

Species *Paramormyrops kingsleyae* (†)

Order PERCIFORMES

Family Anabantidae

Species *Anabas testudineus* (†⁂)

Family Blenniidae

Species *Parablennius parvicornis* (†)

Family Carangidae

Species *Seriola dumerili* (†)

Species *Seriola lalandi* (†)

Species *Seriola quinqueradiata* (†)

Species *Seriola rivoliana* (†)

Family Centrarchidae

Species *Micropterus floridanus* (†)

Family Centropomidae

Species *Lates calcarifer* (†)

Family Cichlidae

Species *Amphilophus citrinellus* (†)

Species *Astatotilapia calliptera* (†⁂)

Species *Maylandia zebra* (†)

Species *Neolamprologus brichardi* (†)

Species *Oreochromis niloticus* (†⁂)

Species *Simochromis diagramma* (†⁂)

Family Helostomatidae

Species *Helostoma temmincki* (†)

Family Labridae

Species *Labrus bergylta* (†)

Species *Symphodus melops* (†)

Family Oplegnathidae

Species *Oplegnathus fasciatus* (†)

Family Osphronemidae

Species *Betta splendens* (†)

Family Percichthyidae

Species *Maccullochella peelii* (†)

Family Percidae

Species *Perca fluviatilis* (⁂)

Family Pomacentridae

Species *Acanthochromis polyacanthus* (†)

Species *Amphiprion ocellaris* (†)

Species *Amphiprion percula* (†)

Family Sciaenidae

Species *Larimichthys crocea* (†⁂)

Species *Miichthys miiuy* (†)

Family Scombridae

Species *Thunnus albacares* (†)

Species *Thunnus orientalis* (†)

Species *Thunnus thynnus* (†)

Family Sparidae

Species *Pagrus major* (†)

Species *Sparus aurata* (†)

Order PERCOPSIFORMES

Family Amblyopsidae

Species *Typhlichthys subterraneus* (†)

Family Percopsidae

Species *Percopsis transmontana* (†)

Order PLEURONECTIFORMES

Family Cynoglossidae

Species *Cynoglossus semilaevis* (†)

Family Paralichthyidae

Species *Paralichthys olivaceus* (†)

Family Scophthalmidae

Species *Scophthalmus maximus* (†)

Order SALMONIFORMES

Family Salmonidae

Species *Hucho hucho* (†)

Species *Oncorhynchus kisutch* (†)

Species *Oncorhynchus mykiss* (†)

Species *Oncorhynchus tshawytscha* (†)

Species *Salmo salar* (†)

Species *Salvelinus alpinus* (†)

Order SCORPAENIFORMES

Family Scorpaenidae

Species *Sebastes nigrocinctus* (†)

Species *Sebastes rubrivinctus* (†)

Order SILURIFORMES

Family Auchenipteridae

Species *Ageneiosus marmoratus* (†)

Family Ictaluridae

Species *Ictalurus punctatus* (†)

Family Pangasiidae

Species *Pangasianodon hypophthalmus* (†)

Order STEPHANOBERYCIFORMES

Family Stephanoberycidae

Species *Acanthochaenus leutkenii* (†)

Order SYNBRANCHIFORMES

Family Mastacembelidae

Species *Mastacembelus armatus* (†⁂)

Family Synbranchidae

Species *Monopterus albus* (†)

Order TETRAODONTIFORMES

Family Molidae

Species *Mola mola* (†)

Family Tetraodontidae

Species *Tetraodon nigroviridis* (†)

Species *Takifugu rubripes* (†)

Order ZEIFORMES

Family Parazenidae

Species *Cyttopsis rosea* (†)

SUPERCLASS SARCOPTERYGII

CLASS COELACANTHI

Order COELACANTHIFORMES

Family Latimeridae

Species *Latimeria chalumnae* (†)

CLASS AMPHIBIA

Order ANURA

Family Bufonidae

Species *Rhinella marina* (†⁂)

Family Pipidae

Species *Xenopus laevis* (†)

Species *Xenopus tropicalis* (†⁂)

Order URODELA

Family Ambystomatidae

Species *Ambystoma mexicanum* (†)

Family Salamandridae

Species *Pleurodeles waltl* (†)

CLASS REPTILIA (paraphyletic)

Order CROCODILIA

Family Alligatoridae

Species *Alligator mississippiensis* (†)

Species *Alligator sinensis* (†)

Family Crocodilidae

Species *Crocodylus porosus* (†)

Family Gavialidae

Species *Gavialis gangeticus* (†)

Order RHYNCOCEPHALIA

Family Sphenodontidae

Species *Sphenodon punctatus* (†)

Order SQUAMATA

Family Agamidae

Species *Pogona vitiiceps* (†)

Family Colubridae

Species *Thamnophis sirtalis* (†)

Family Dactyloidae

Species *Anolis carolinensis* (†⁂)

Family Elapidae

Species *Notechis scutatus* (†)

Species *Ophiophagus hannah* (†)

Family Gekkonidae

Species *Gekko japonicus* (†)

Species *Paroedura picta* (†)

Family Pythonidae

Species *Python bivitattus* (†)

Family Viperidae

Species *Crotalus horridus* (†)

Species *Crotalus viridis* (†)

Species *Protobothrops flavoviridis* (†)

Species *Protobothrops mucrosquamatus* (†)

Order TESTUDINES

Family Cheloniidae

Species *Chelonia mydas* (†)

Family Emydidae

Species *Chrysemys picta* (†)

Species *Malaclemys terrapin* (†)

Species *Terrapene carolina* (†⁂)

Family Testudinidae

Species *Gopherus agassizii* (†)

Family Trionychidae

Species *Apalone spinifera* (†)

Species *Pelodiscus sinensis* (†)

CLASS AVES

Order APODIFORMES

Family Apodidae

Species *Chaetura pelagica* (†)

Family Trochilidae

Species *Calypte anna* (†)

Order APTERYGIFORMES

Family Apterygidae

Species *Apteryx australis* (†)

Species *Apteryx haastii* (†)

Species *Apteryx owenii* (†)

Species *Apteryx rowi* (†)

Order CASUARIIFORMES

Family Casuariidae

Species *Casuarius casuarius* (†)

Order OPISTHOCOMIFORMES

Family Opisthocomidae

Species *Opisthocomus hoazin* (†)

Order PELECANIFORMES

Family Threskiornithidae

Species *Nipponia nippon* (†)

Order SPHENISCIFORMES

Family Spheniscidae

Species *Pygoscelis antarticus* (†)

Species *Pygoscelis papua* (†)

Species *Spheniscus humboldti* (†)

Species *Spheniscus mendelicus* (†)

Species *Spheniscus magellanicus* (†)

Order STRIGIFORMES

Family Strigidae

Species *Strix occidentalis* (†)

Order STRUTHIONIFORMES

Family Struthionidae

Species *Struthio camelus* (†)

Family Rheidae

Species *Rhea pennata* (†)

Order SULIFORMES

Family Phalacrocoracidae

Species *Phalacrocorax auritus* (†)

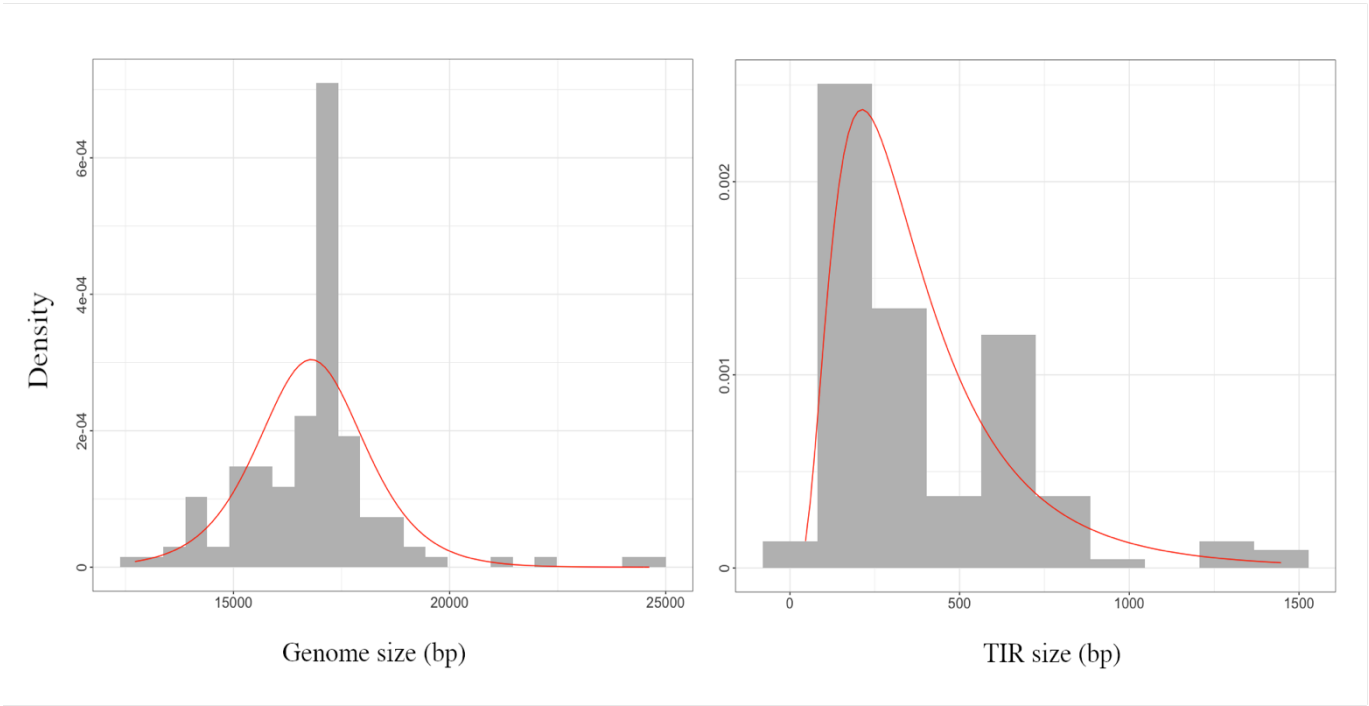


Figure S1. Genome and TIR size distributions of intact *Mavericks* found in the genomes of 17 species of vertebrates. Best-fit distributions are logistic for the genome sizes (Loglikelihood = -1,170.65) and lognormal for the TIR sizes (Loglikelihood = -907.90), shown in red. N = 134.

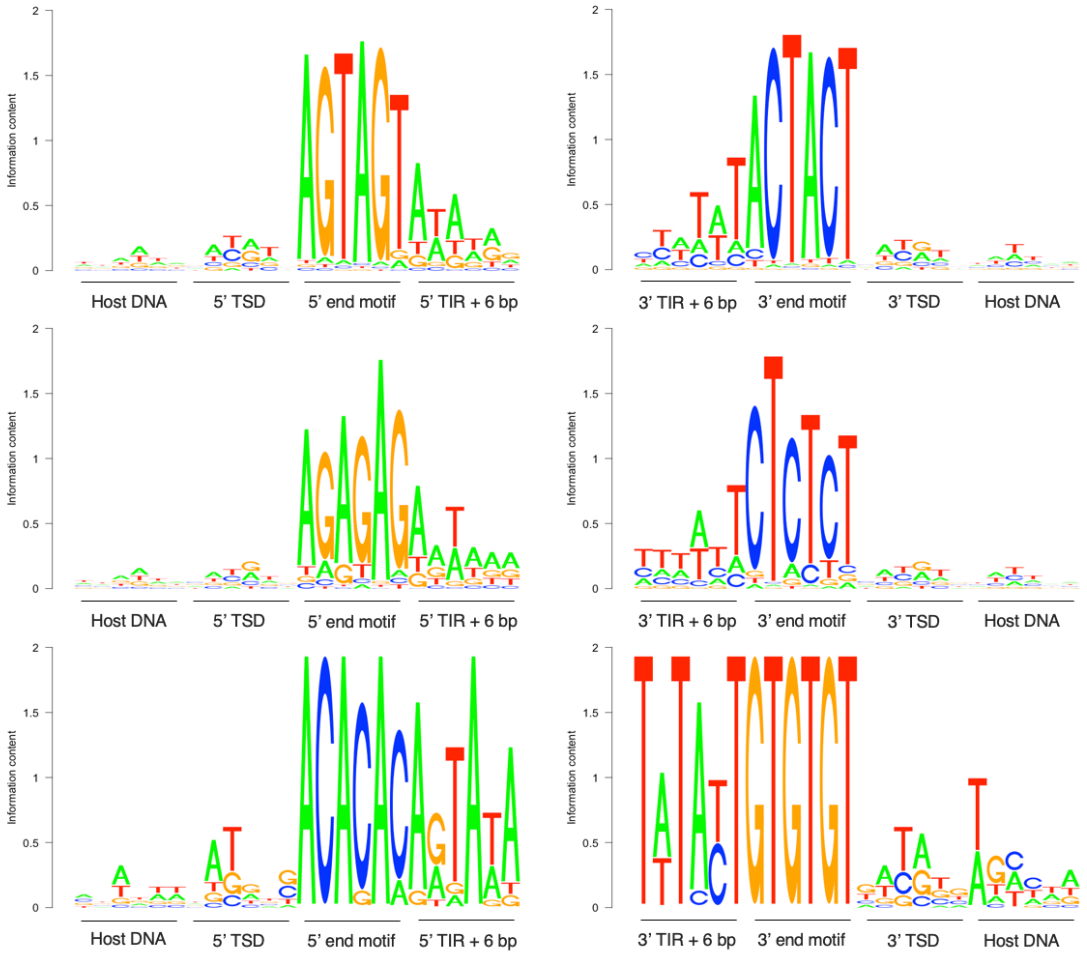


Figure S2. Sequence logos of the conserved motifs at the termini of vertebrate *Mavericks*. The upper panel shows the 5'-$(AGT)_2$//$(ACT)_2$-3' motif (N = 878), the 5'-$(AG)_3$//$(CT)_3$-3' motif (N = 460) in the middle and the 5'-$(AC)_3$A//T$(GT)_3$-3' motif (N = 14) in the lower panel. The height of each nucleotide is proportional to its conservation at that position. Information content shown in bits (2 bits = invariable site, 0 bits = random site).

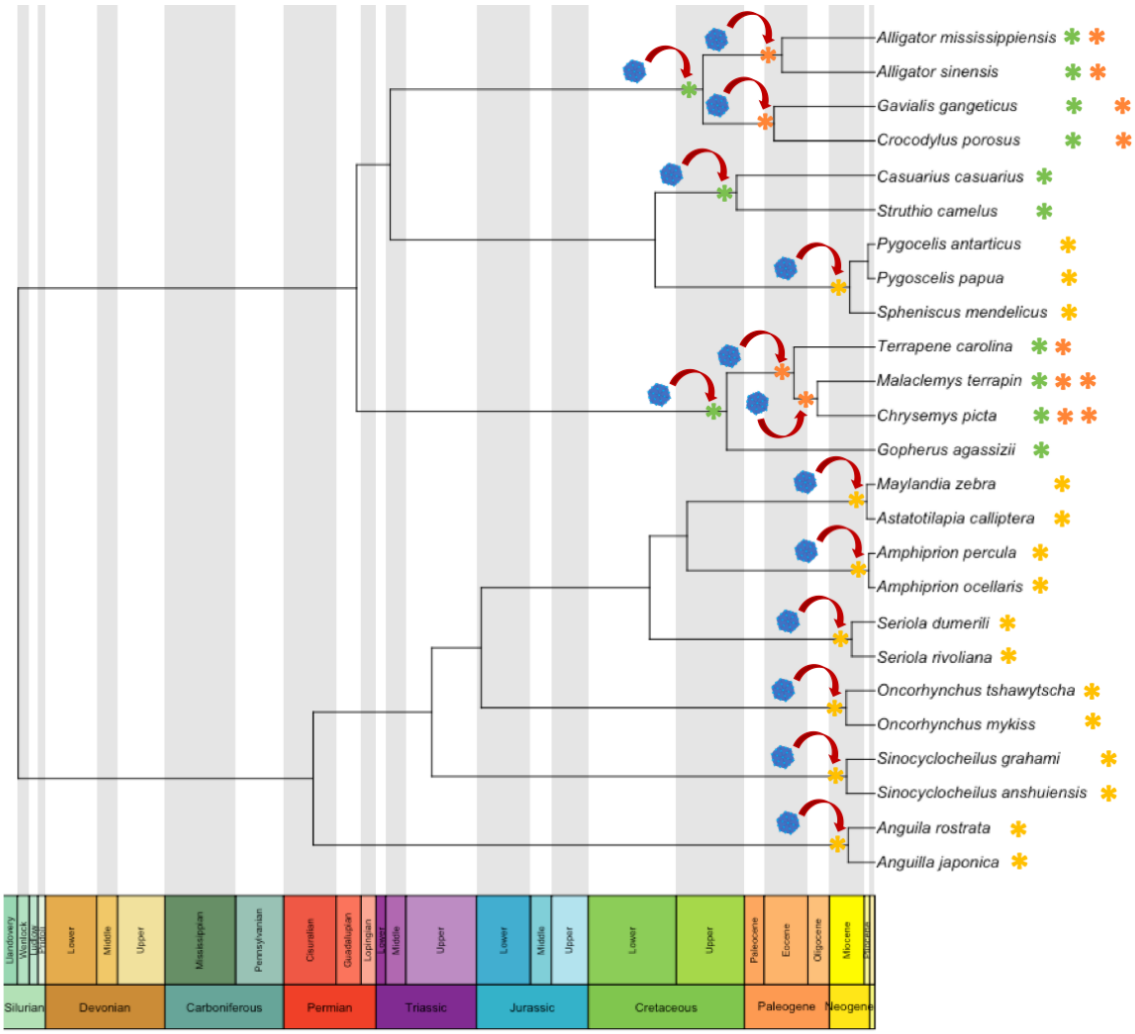


Figure S3. Time tree showing the integration events that gave rise to different orthologue groups as inferred from the host phylogeny. Fourteen event classes are represented which comprise all 54 groups of orthologues. Asterisk colours indicate the minimum age of the integration. A geological timescale is shown below.

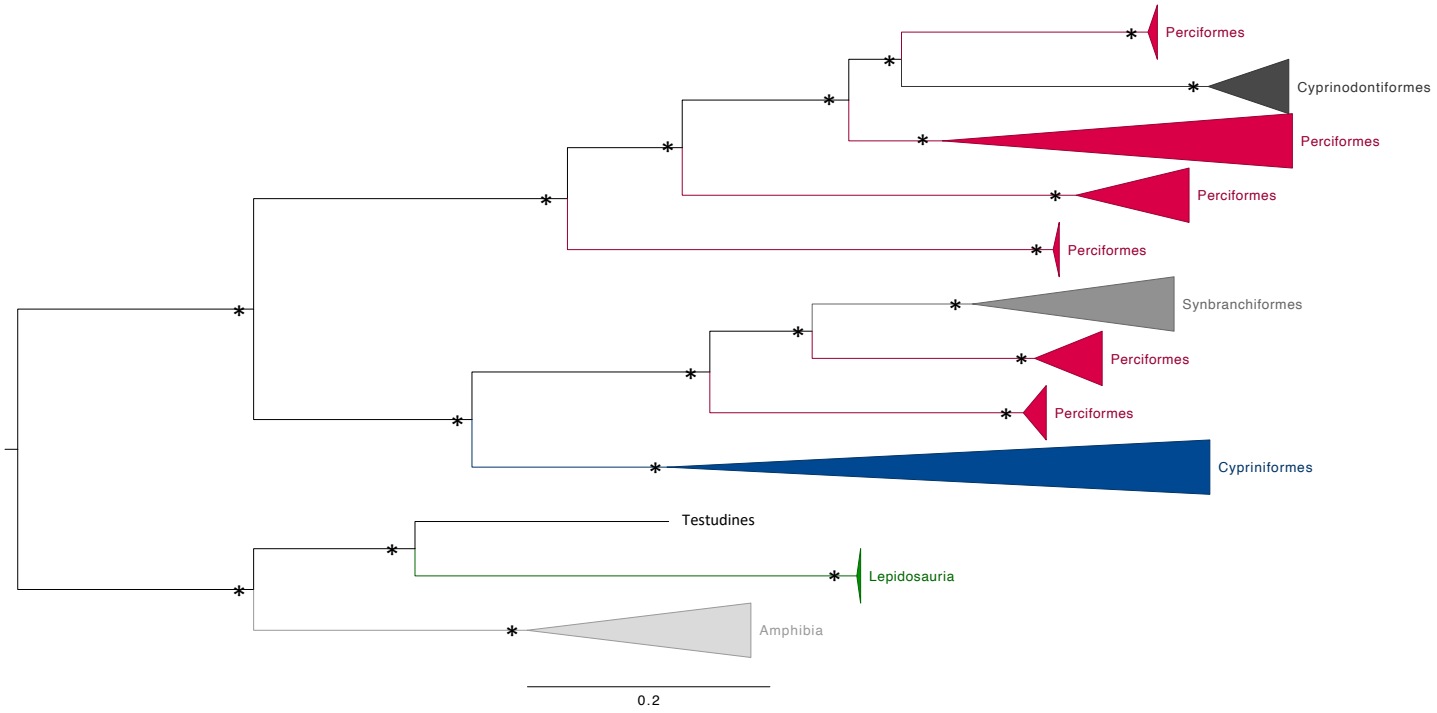


Figure S4. Majority-rule consensus tree of the intact vertebrate *Mavericks* based on the protein-primed DNA polymerase B, asterisks indicate a posterior probability > 0.99. Scale bar = 0.2 amino acid substitutions per site. Note the topology is equivalent to Figure 4.

Table S1. Copy numbers of intact *Mavericks* in the genomes of 17 vertebrate species. The genomes of teleosts host a greater number of intact elements than tetrapod genomes (Mann-Whitney U test: p-value < 0.05).

Clade	**Order**	**Family**	**Species**	**Nº copies**
Actinopterygii	Cypriniformes	Cyprinidae	*Carassius auratus*	4
Actinopterygii	Cypriniformes	Cyprinidae	*Danio rerio*	15
Actinopterygii	Cypriniformes	Cyprinidae	*Oxygymnocypris stewartii*	38
Actinopterygii	Cyprinodontiformes	Poeciliidae	*Xiphophorus couchianus*	2
Actinopterygii	Cyprinodontiformes	Poeciliidae	*Xiphophorus hellerii*	12
Actinopterygii	Cyprinodontiformes	Poeciliidae	*Xiphophorus maculatus*	5
Actinopterygii	Perciformes	Anabantidae	*Anabas testudineus*	4
Actinopterygii	Perciformes	Cichlidae	*Astatotilapia calliptera*	11
Actinopterygii	Perciformes	Cichlidae	*Oreochromis niloticus*	5
Actinopterygii	Perciformes	Cichlidae	*Simochromis diagramma*	1
Actinopterygii	Perciformes	Percidae	*Perca fluviatilis*	19
Actinopterygii	Perciformes	Sciaenidae	*Larimichthys crocea*	2
Actinopterygii	Synbranchiformes	Mastacembelidae	*Mastacembelus armatus*	11
Lepidosauria	Squamata	Dactyloidae	*Anolis carolinensis*	2
Lissamphibia	Anura	Bufonidae	*Rhinella marina*	1
Lissamphibia	Anura	Pipidae	*Xenopus tropicalis*	1
Testudinata	Testudines	Emydidae	*Terrapene carolina*	1
Total				**134**

Table S2. Root ages (in millions of years from the present) and evolutionary rates (in amino acid substitutions per site per year x 10^{-9}) estimated from the Bayesian evolutionary analyses. Rate summary statistics correspond to the averages, standard deviations, minima and maxima of the posterior rate median for all branches.

Protein	Root age	Root age 95% HPD	Rate ± SD	Rate min–max
PZ	454.73	347.17–568.59	2.151 ± 1.448	0.451–17.052
Minor capsid protein	407.39	276.38–559.05	2.098 ± 1.511	0.430–11.771
pPOLB (polymerase)	381.69	279.61–489.40	2.465 ± 1.777	0.255–28.339
Protease	377.11	295.42–464.04	2.294 ± 1.236	0.614–11.104
ATPase	351.11	269.04–437.96	2.072 ± 1.566	0.522–15.955
PW	278.38	210.61–362.56	2.892 ± 1.381	0.285–11.721
Integrase	271.42	209.84–335.69	3.290 ± 2.943	0.369–27.245
Major capsid protein	231.86	176.95–294.89	3.669 ± 3.595	0.247–78.172

Appendix to Chapter III

Supplementary material: A billion years arms race between viruses, virophages and eukaryotes

Table S1. Viral species included in the final multiple sequence alignments (54 taxa) with their accession numbers and source reference.

Species	**ID**	**GenBank/metagenomic accessions**	**Source**
Maverick (*Danio rerio*)	Proteo5	NC_007136.7: 27671708-27687993	(Barreat and Katzourakis, 2021)
Maverick (*Xenopus tropicalis*)	Proteo18	NC_030686.1: 6013889-6026611	(Barreat and Katzourakis, 2021)
Maverick (*Anolis carolinensis*)	Proteo2	NC_014778.1: 129158247-129172346	(Barreat and Katzourakis, 2021)
Maverick (*Xiphophorus hellerii*)	Proteo19	QPIH01000028.1: 6019497-6036410	(Barreat and Katzourakis, 2021)

Maverick (*Oreochromis niloticus*)	Proteo15	NW_020327416.1: 45576-61017	(Barreat and Katzourakis, 2021)
Maverick (*Oxygymnocypris stewartii*)	Proteo11	QVTF01001200.1: 1425766-1443135	(Barreat and Katzourakis, 2021)
Maverick (*Plutella xylostella*)	Group2_9	NW_011952032.1: 283307-321039	This work
Maverick (*Ostrinia furnacalis*)	Group2_10	NW_021132744.1: 18414-41178	This work
Maverick (*Diachasma alloeum*)	Group2_5	NW_021681489.1: 105940-131438	This work
Maverick (*Diachasma alloeum*)	Group2_6	NW_021680987.1: 30857-54497	This work
Maverick (*Photinus pyralis*)	Group2_7	NW_022171288.1: 44859-60568	This work
Maverick (*Sitophilus oryzae*)	Group2_8	NW_022147237.1: 411147-463473	This work
Metagenomic NCLDV	NCLDV_Roux3157	Trout_Epilimnion_TBL_comb48_EPIDRAFT_1006965	(Roux et al., 2017)
Metagenomic NCLDV	NCLDV_Mo545	SRX331950.48.dc.fa_4042	(Moniruzzaman et al., 2020)
Metagenomic NCLDV	NCLDV_Mo338	SRX330942.38.dc.fa_3117	(Moniruzzaman et al., 2020)
Metagenomic NCLDV	NCLDV_Mo338	SRX330942.38.dc.fa_13478	(Moniruzzaman et al., 2020)
Metagenomic NCLDV	NCLDV_Roux1456	Mendota_contig-68000307_4926_nucleotides	(Roux et al., 2017)

Metagenomic NCLDV	NCLDV_Sch668	MN740268.1	(Schulz et al., 2020)
Shrimp hemocyte iridescent virus	—	NC_055165.1	NCBI
Infectious spleen and kidney necrosis virus	—	NC_003494.1	NCBI
Heterosigma akashiwo virus 01	—	NC_038553.1	NCBI
Akhmeta virus	—	NC_055230.1	NCBI
Tupanvirus soda lake	—	KY523104.2	NCBI
Diadromus pulchellus ascovirus 4a	—	NC_011335.1	NCBI
Metagenomic PLV	PLV_BS940	3300009435_____Ga0115546_1002426	(Bellas and Sommaruga, 2021)
Metagenomic PLV	PLV_BS718	Ga0115028_10000066	(Bellas and Sommaruga, 2021)
Metagenomic PLV	PLV_BS13	V563K_contig_8723_len_16581_bp	(Bellas and Sommaruga, 2021)

Metagenomic PLV	PLV_BS395	ERR1823950_NODE_975_length_11499_cov_287.344855	(Bellas and Sommaruga, 2021)
Metagenomic PLV	PLV_BS539	ERX2821583_NODE_576_length_19830_cov_33.560977	(Bellas and Sommaruga, 2021)
Metagenomic PLV	PLV_BS262	HanCross_NODE_1742_length_13974_cov_48.888145	(Bellas and Sommaruga, 2021)
Metagenomic PLV	Plike_4	SAF4	(Yutin et al., 2015)
Metagenomic PLV	Plike_7	INO1	(Yutin et al., 2015)
Metagenomic PLV	Plike_17	RED1	(Yutin et al., 2015)
Metagenomic PLV	Plike_1	SAF1	(Yutin et al., 2015)
Metagenomic PLV	Plike_19	YSL1	(Yutin et al., 2015)
Metagenomic PLV	Plike_25	ACE1	(Yutin et al., 2015)
Metagenomic virophage	Virophage_PE125	3300009070_____Ga0066256_1001637	(Paez-Espino et al., 2019)
Metagenomic virophage	Virophage_PE85	3300009781_____Ga0116178_10003357	(Paez-Espino et al., 2019)
Metagenomic virophage	Virophage_PE47	3300012984_____Ga0164309_10000286	(Paez-Espino et al., 2019)

Metagenomic virophage	Virophage_PE169	3300009154_____Ga0114963_10000678	(Paez-Espino et al., 2019)
Metagenomic virophage	Virophage_PE184	3300007609_____Ga0102945_1000484	(Paez-Espino et al., 2019)
Metagenomic virophage	Virophage_Roux27	TBH_10005622	(Roux et al., 2017)
Maverick-related virus strain Spezl	—	NC_015230.1	NCBI
Yellowstone Lake virophage 7	—	KM502591.1	NCBI
Zamilon virus	—	NC_022990.1	NCBI
Yellowstone Lake virophage 5	—	NC_028269.1	NCBI
Dishui Lake virophage 6	—	MN940573.1	NCBI
Dishui Lake virophage 2	—	MN940570.1	NCBI
Bat mastadenovirus WIV17	—	NC_034626.1	NCBI
Murine mastadenovirus A	—	AC_000012.1	NCBI
Murine adenovirus 2	—	NC_014899.1	NCBI
Fowl aviadenovirus 5	—	NC_021221.1	NCBI
Frog adenovirus 1	—	NC_002501.1	NCBI

Snake adenovirus 1	—	NC_009989.1	NCBI

Table S2. Distribution of root positions calculated from the MCMC posterior tree sample. The best supported position of the root was on the branch leading to virophages (53.9%), followed by NCLDVs and metagenomic PLV BS539 (27.4%). Other root positions received < 6% support. The frequencies of trees with a certain position of the root were estimated by filtering different topologies in PAUP. Number of generations = 140 million.

Root position	**Number of trees**	**Percentage**	**Cumulative sum**
Virophages	**15099**	**53.923**	**53.923**
NCLDVs + PLV_BS539	7672	27.399	81.322
PLV_BS539	1572	5.614	86.936
NCLDVs	1489	5.318	92.254
Virophages + NCLDVs + PLV_BS539	1388	4.957	97.211
PLV_BS13	185	0.661	97.872
Virophages + PLV_BS13	159	0.568	98.439
Adenoviruses	107	0.382	98.821
Virophages + NCLDVs	44	0.157	98.979
Virophages + NCLDVs + PLV_BS539 + PLV_BS13	35	0.125	99.104
Adenovirus + NCLDVs + PLV_BS539	29	0.104	99.207
Adenoviruses + NCLDVs	4	0.014	99.221
Virophages + Adenoviruses	3	0.011	99.232
NCLDVs + PLV_BS539 + most PLVs	1	0.004	99.236

Virophages + PLV_BS13 + Adenoviruses	1	0.004	99.239
Adenovirus + PLV_BS13	1	0.004	99.243
most PLVs	0	0.000	99.243
Mavericks 1 + Mavericks 2	0	0.000	99.243
Mavericks 1	0	0.000	99.243
Mavericks 2	0	0.000	99.243
Virophages + most PLVs	0	0.000	99.243
Adenoviruses + most PLVs	0	0.000	99.243
Mavericks 1 + Mavericks 2 + PLV_BS13	0	0.000	99.243
Mavericks 1 + Mavericks 2 + Adenoviruses + PLV_BS13	0	0.000	99.243
Others	212	0.757	100

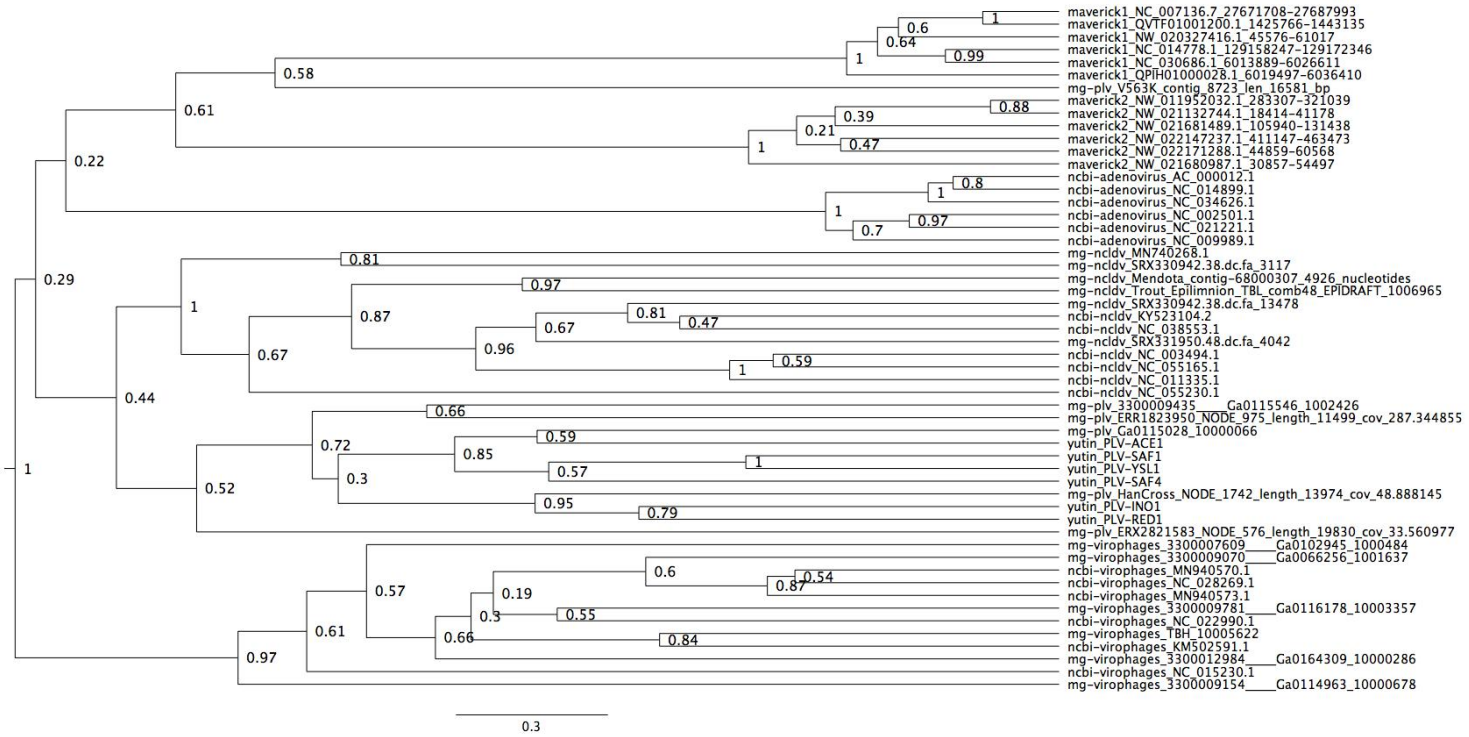


Figure S1. Bayesian maximum credibility tree of the major capsid protein inferred with a relaxed molecular clock. Virophages are monophyletic with a high posterior support (0.97). The root falls between virophages and all other elements. Tree estimated from 200 million MCMC generations and a 25% relative burn-in.

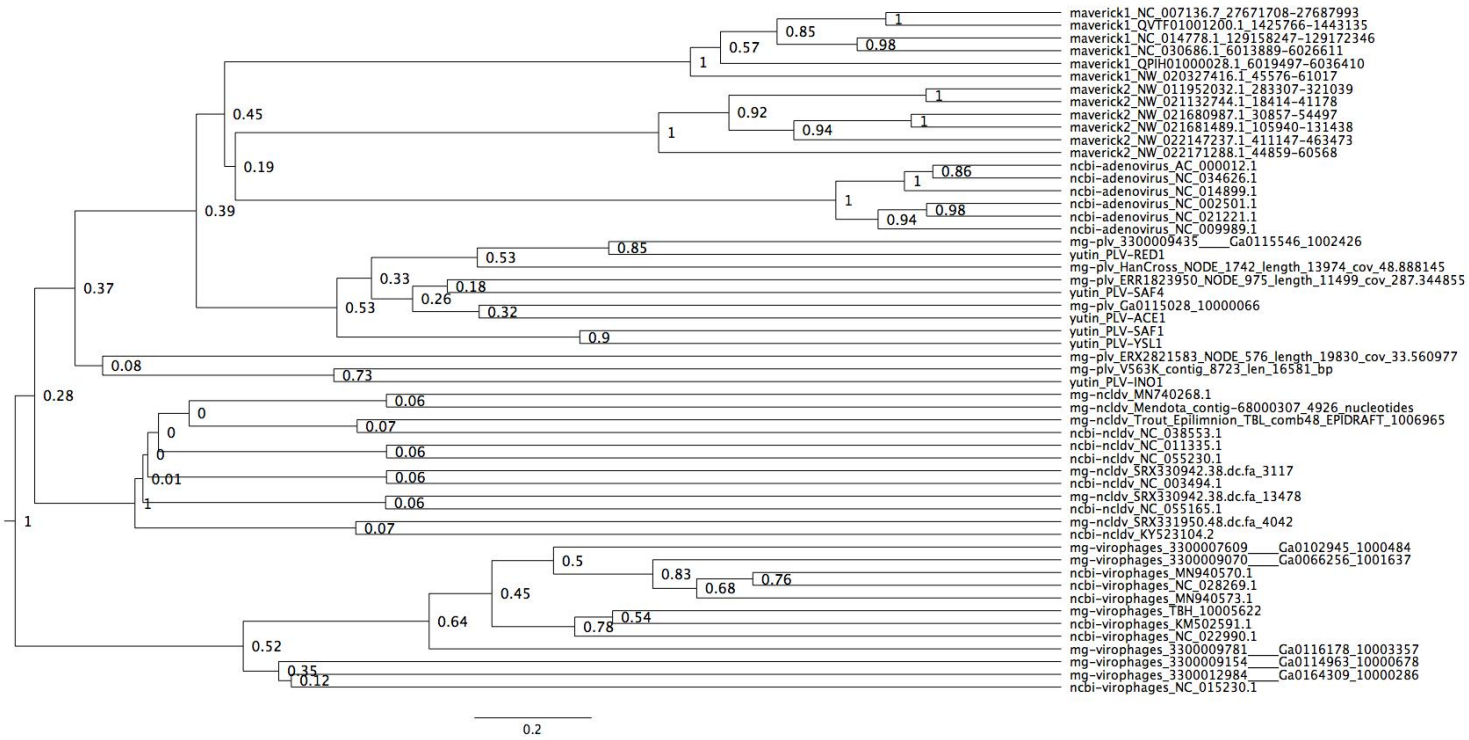


Figure S2. Bayesian maximum credibility tree of the minor capsid protein inferred with a relaxed molecular clock. Virophages are monophyletic with a moderate posterior support (0.52). The root falls between virophages and all other elements. Tree estimated from 200 million MCMC generations and a 25% relative burn-in.

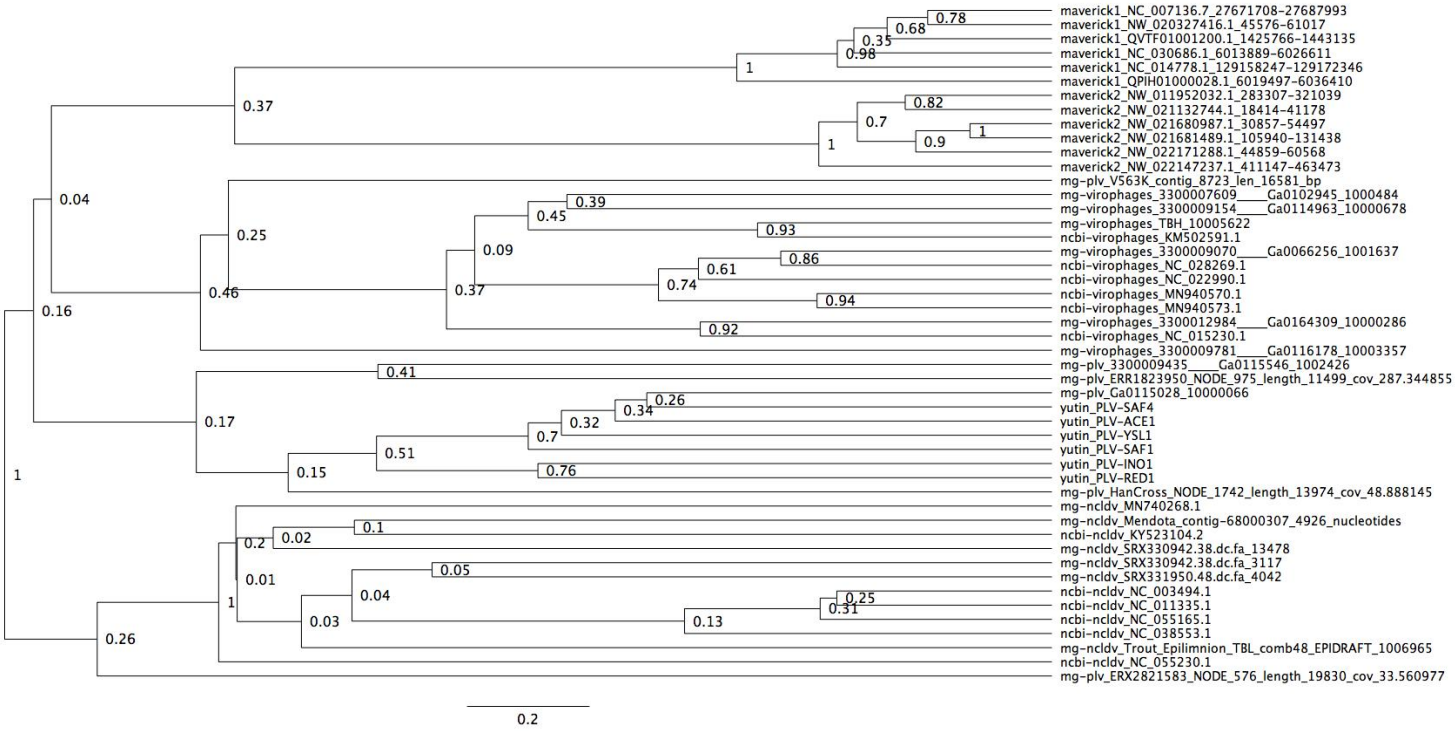


Figure S3. Bayesian maximum credibility tree of the ATPase inferred with a relaxed molecular clock. We did not use a monophyletic constraint on virophages in this tree. The root falls between a clade formed by NCLDVs plus a metagenomic PLV (0.26 posterior probability), and all other elements. According to this tree, virophages are not monophyletic. Tree estimated from 200 million MCMC generations and a 25% relative burn-in. Adenoviruses are not included in this tree since they encode a non-homologous ATPase of the ABC superfamily (Burroughs, et al., 2007).

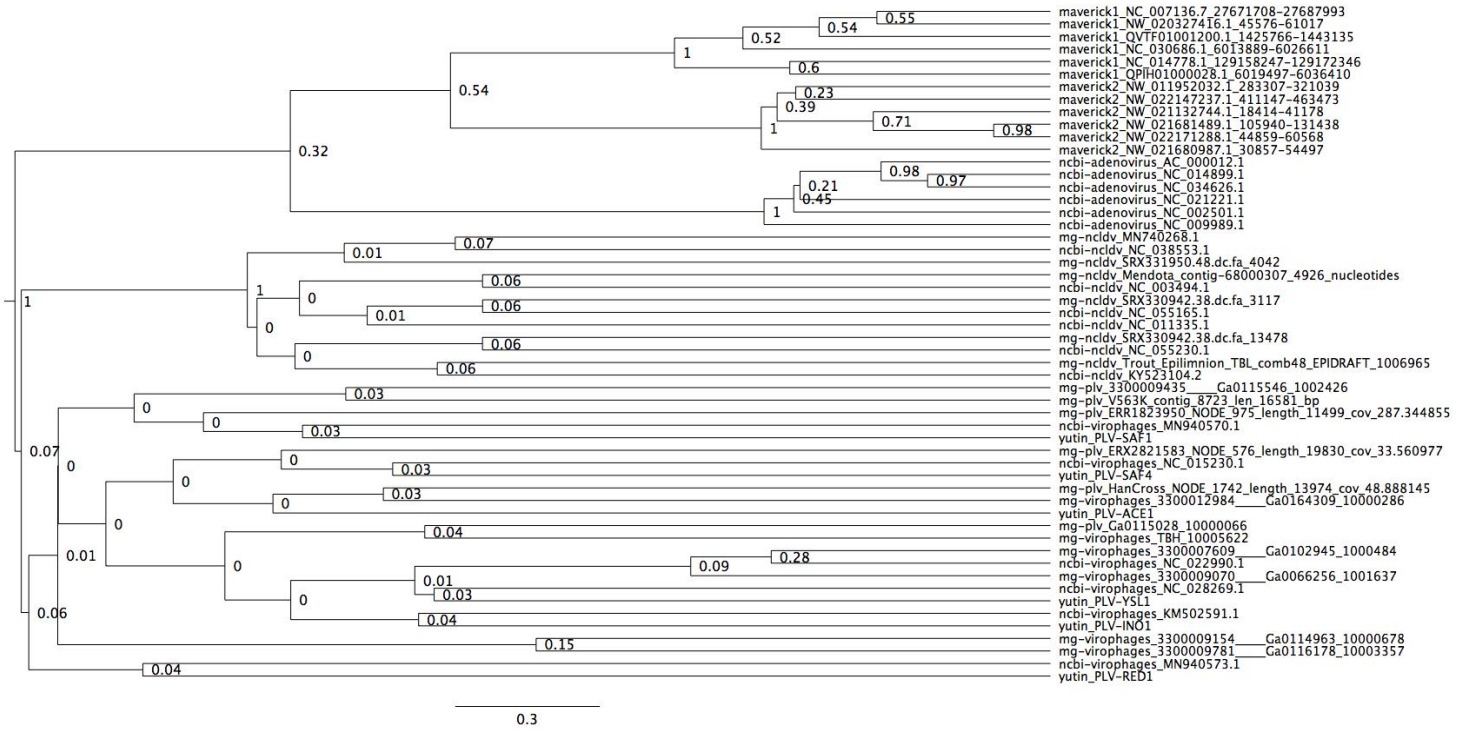


Figure S4. Bayesian maximum credibility tree of the protease inferred with a relaxed molecular clock. We did not use a monophyletic constraint on virophages in this tree. The root falls between a clade formed by adenoviruses and *Mavericks* and all other elements. However, notice how virophages do not form a monophyletic group. Tree estimated from 200 million MCMC generations and a 25% relative burn-in.

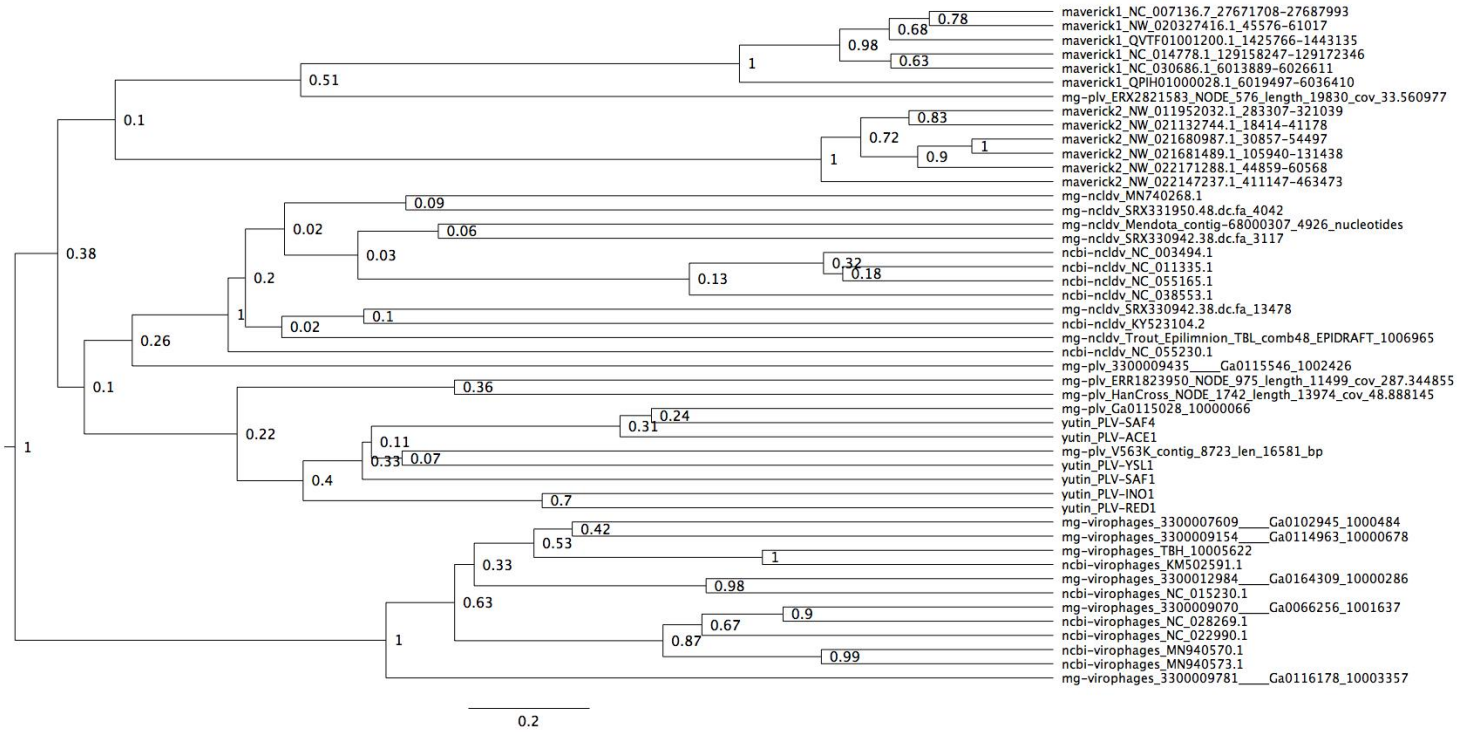


Figure S5. Bayesian maximum credibility tree of the ATPase inferred with a relaxed molecular clock, and using a monophyletic constraint on virophages. The root of this tree falls between virophages and all other elements. Tree estimated from 200 million MCMC generations and a 25% relative burn-in. Adenoviruses are not included in this tree since they encode a non-homologous ATPase of the ABC superfamily (Burroughs, et al., 2007).

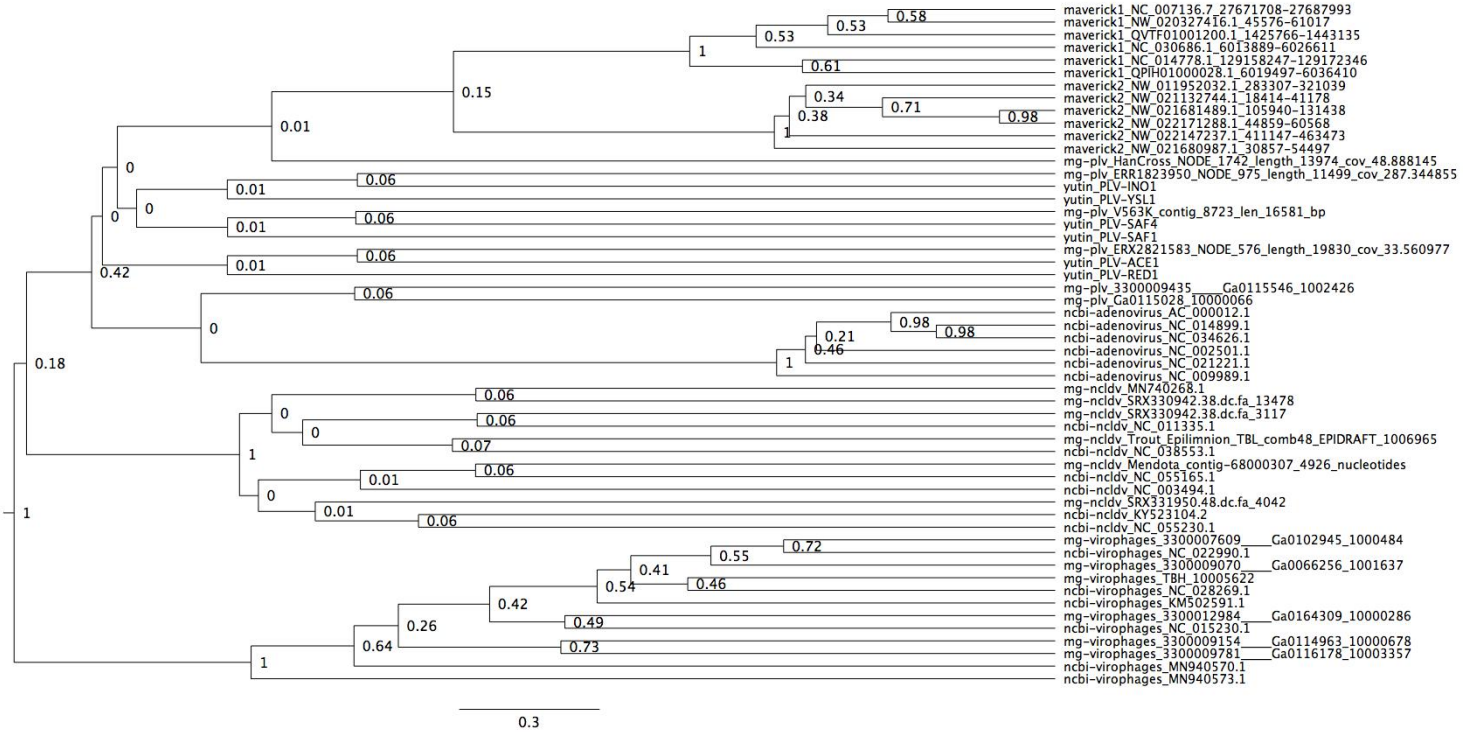


Figure S6. Bayesian maximum credibility tree of the protease inferred with a relaxed molecular clock, and using a monophyletic constraint on virophages. The root falls between virophages and all other elements. Tree estimated from 200 million MCMC generations and a 25% relative burn-in.

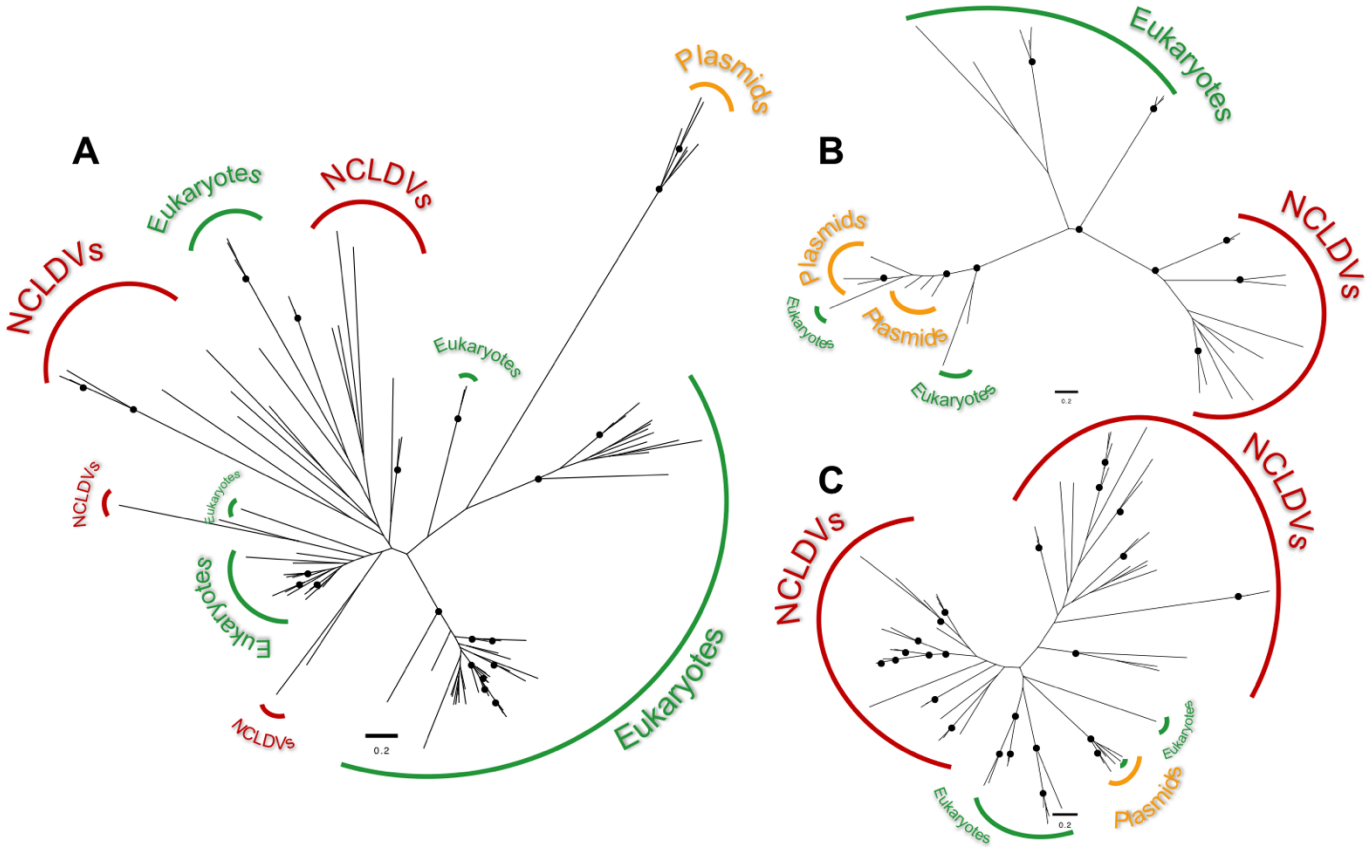


Figure S7. Maximum-likelihood unrooted phylogenetic trees of the transcriptional homologues encoded in cytoplasmic linear plasmids. (A) Trees for the DNA-directed RNA polymerase II, subunit Rpb2, (B) mRNA capping enzyme, and (C) helicase. In all cases, the topologies rule out a sister grouping of the NCLDV and cytoplasmic linear plasmid homologues, suggesting they were acquired independently. Black circles indicate bootstrap support ≥ 0.94 after 1,000 replicates.

Table S3. Cytoplasmic linear plasmids used for querying the databases in search for protein homologues.

Plasmid	**GenBank accession**	**Host organism** (Order Saccharomycetales)
pDH4C	MF795093.1	*Debaryomyces hansenii* (Debaryomycetaceae)
pGK12	X07776.1	*Kluyveromyces lactis* (Saccharomycetaceae)
pKPGS115	CP014724.1	*Komagataella phaffii* (Phaffomycetaceae)
pKP	CP014714.1	*Komagataella phaffii* (Phaffomycetaceae)
pKPCBS743	MG491503.1	*Komagataella phaffii* (Phaffomycetaceae)
pPac1-1	AM180622.1	*Millerozyma acacia* (Debaryomycetaceae)
pPE1B	AJ278986.2	*Schwanniomyces etchellsi* (Debaryomycetaceae)
pSKL	X54850.1	*Lachancea kluyveri* (Saccharomycetaceae)

Appendix to Chapter IV

Supplementary material: Ecological and evolutionary dynamics of cell-virus-virophage systems

Text S1. Derivation of the displacements of cells and viruses relative to virophages. These results were used to set the displacements per time step in the ABM simulations.

The Mean Squared Displacement (MSD) is defined as the deviation in the position of a particle with respect to a reference position with time. In statistical mechanics, it is expressed as an ensemble average:

$$(1)\ \text{MSD} = \frac{1}{N}\sum_{i=0}^{N}\left|x^{i}_{(t)} - x^{i}_{(t=0)}\right|^{2}$$

Since we will consider a single particle moving from the origin, we can set N = 1 and x^i at time zero equal to zero. Thus, we have the following expression for the MSD:

$$(2)\ \text{MSD} = (x^{i}_{(t)})^{2}$$

For a random walk in *n* dimensions the MSD is equal to:

$$(3)\ \text{MSD} = 2nDt$$

where D is the diffusion coefficient for the particle and t is time. Since the particles are moving in 3 spatial-dimensions we then have:

$$(4)\ \text{MSD} = 6Dt$$

Now we use the Einstein-Stokes equation for the diffusion coefficient of spherical particles at low Reynolds numbers:

$$(5)\ D = \frac{k_b T}{6\pi\eta r}$$

where k_b is Boltzmann's constant, T is the temperature in Kelvin, η is the viscosity of the fluid and r is the radius of a suspended particle.

To relate the displacements to the radius of the particle we substitute (5) in (4), giving:

$$(6)\ \text{MSD} = \frac{k_b T t}{\pi\eta r}$$

To find the mean displacement we solve for x in equation number (2):

$$(7)\ x = \sqrt{\text{MSD}}$$

Now we are interested in finding the mean displacement of the virus with respect to the virophage:

$$(8)\quad \frac{x_{virus}}{x_{virophage}} = \sqrt{\frac{\frac{k_b T t}{\pi \eta r_{virus}}}{\frac{k_b T t}{\pi \eta r_{virophage}}}}$$

$$(9)\quad x_{virus} = x_{virophage}\sqrt{\frac{r_{virophage}}{r_{virus}}}$$

Similarly, the magnitude of the mean displacement of a cell relative to the virophage will be:

$$(10)\quad x_{cell} = x_{virophage}\sqrt{\frac{r_{virophage}}{r_{cell}}}$$

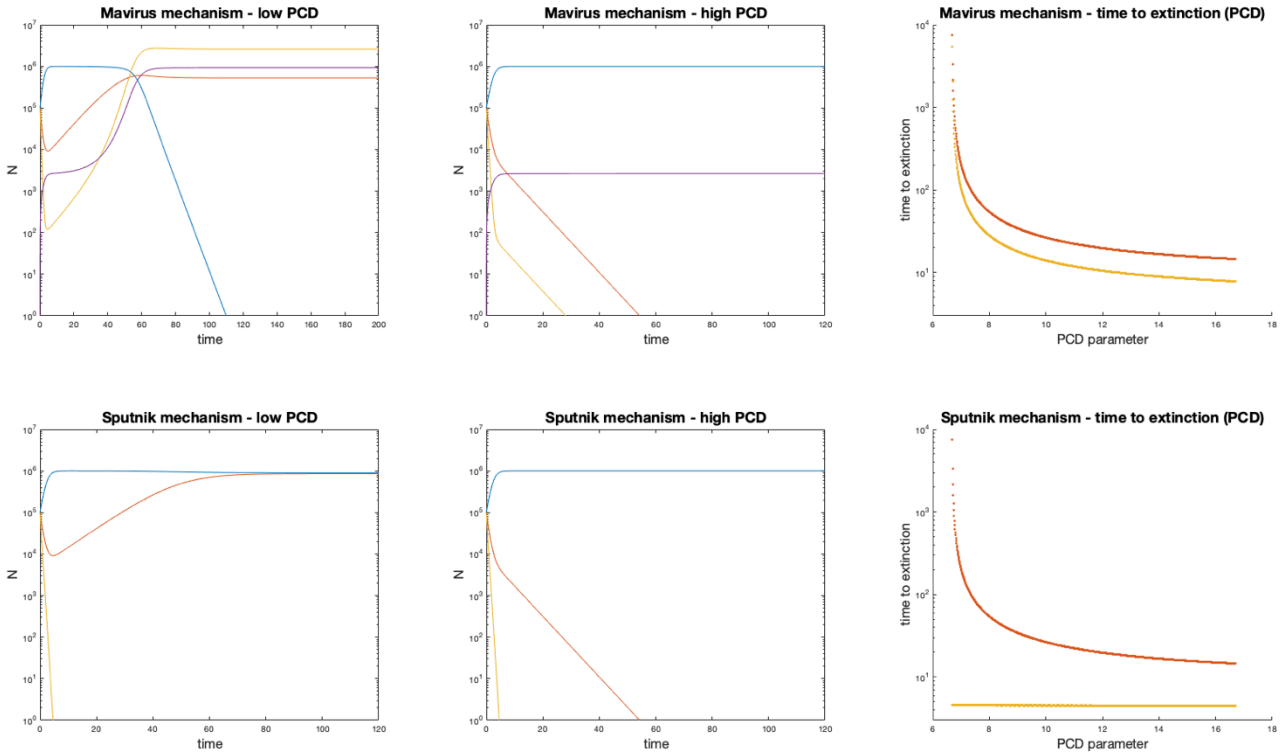


Figure S1. Dynamics of the Mavirus and Sputnik mechanisms in the presence of programmed cell-death (virophages assumed to be neutral). Blue: naïve celles, purple: cells with integrated virophage, orange: viruses, yellow: virophages. **First column:** for the Mavirus model and a low PCD parameter (**p** = 6), we observe fixation of cells with the integrated virophage and stable coexistence of viruses and neutral virophages. In the Sputnik model, virophages go extinct and naïve cells persist in the population. **Second column:** increasing the PCD parameter (**p** = 8) leads to extinction of viruses and virophages in both cases. An integration polymorphism is observed in the Mavirus model, while naïve cells grow to carrying capacity in the Sputnik model. **Third column:** transition to the extinction regime occurs with a PCD parameter greater than 6.6. Time to virus extinction decreases as a function of the PCD parameter in both models, and for virophages in the Mavirus model. Parameters for model 1: $\alpha_1 = 1$, $\alpha_2 = 0.9$, $\beta_1 = 10^{-7}$, $\beta_2 = 10^{-6}$, $\gamma_1 = 1.2$, $\gamma_2 = 2.6$, $\lambda = 10^{-7}$, $\varphi_1 = 100$, $\varphi_2 = 1000$, $f = 1$, $K = 10^6$, $r_1 = 1$, $r_2 = 0.8$. Parameters for model 2 are the same except for $\beta_2 = 10^{-7}$ and $k = 8 \cdot 10^{-7}$. Initial conditions: $C_{x,0} = 10^5$, $G_0 = 10^5$, $V_0 = 10^5$.

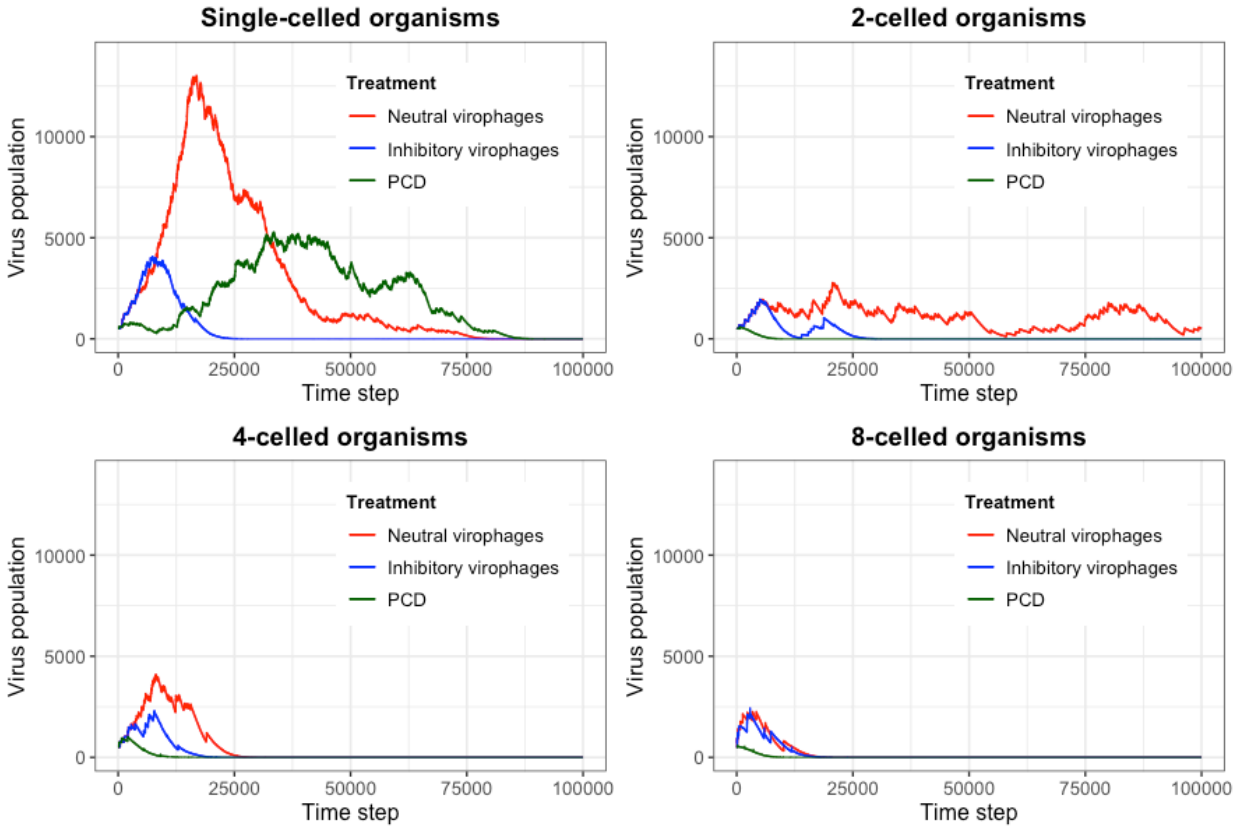


Figure S2. Time-course of the population of viruses during an ABM simulation. The total virus population during a stochastic simulation is shown as a function of time and grouped by the number of cells per organism and treatment. The highest virus wave is observed in the population of single-celled organisms in the presence of neutral virophages. By comparison, treatment with inhibitory virophages and PCD show lower maxima of the virus waves. The effect of multicellularity can also be observed in the lower virus waves from the single-celled to 8-celled cases. Random seed = 1234.

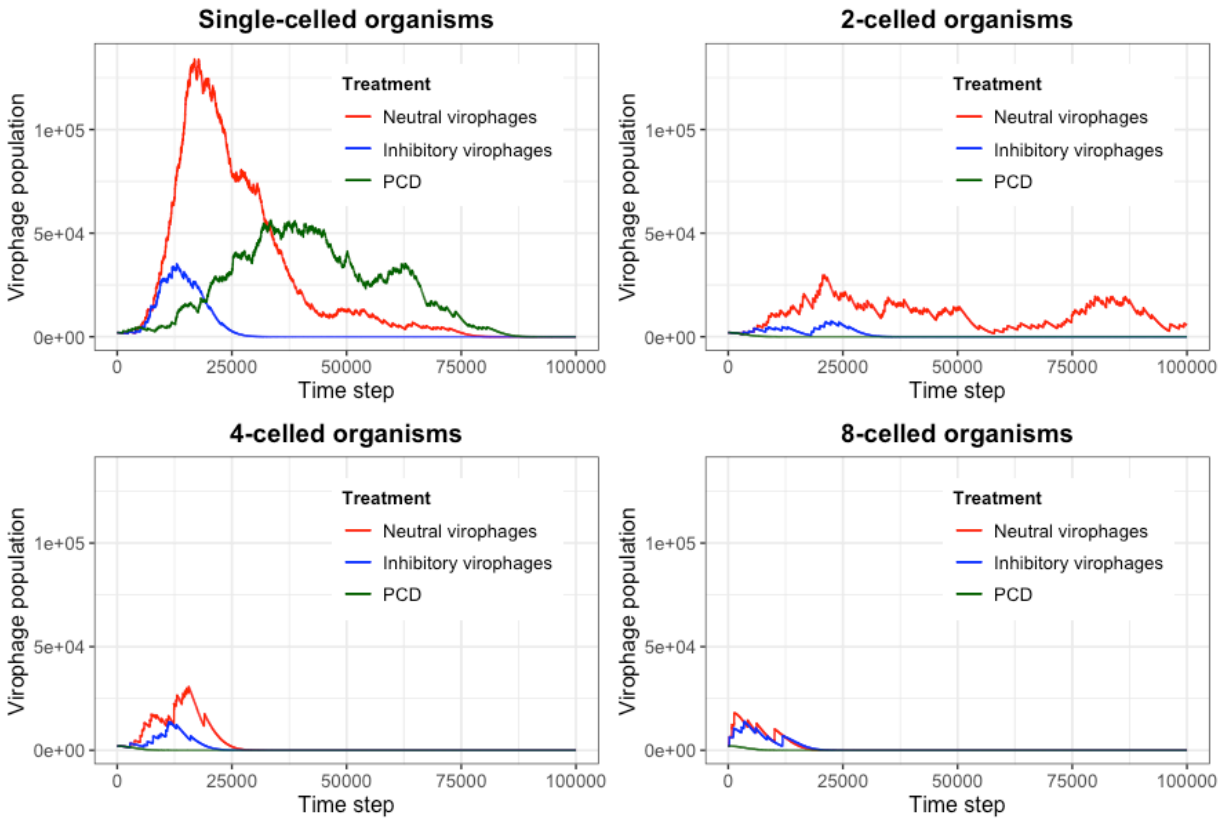


Figure S3. Time-course of the population of virophages during an ABM simulation. The total virophage population during a stochastic simulation is plotted as a function of time and grouped by the number of cells per organism and treatment. We observe the same pattern as for the virus population, where the highest virophage numbers were attained in the simulation with neutral virophages and lower for inhibitory virophages and PCD. In contrast to viruses, virophages attain much larger population sizes. Multicellularity also has an effect on reducing the maximum size of the virophage waves. Random seed = 1234.

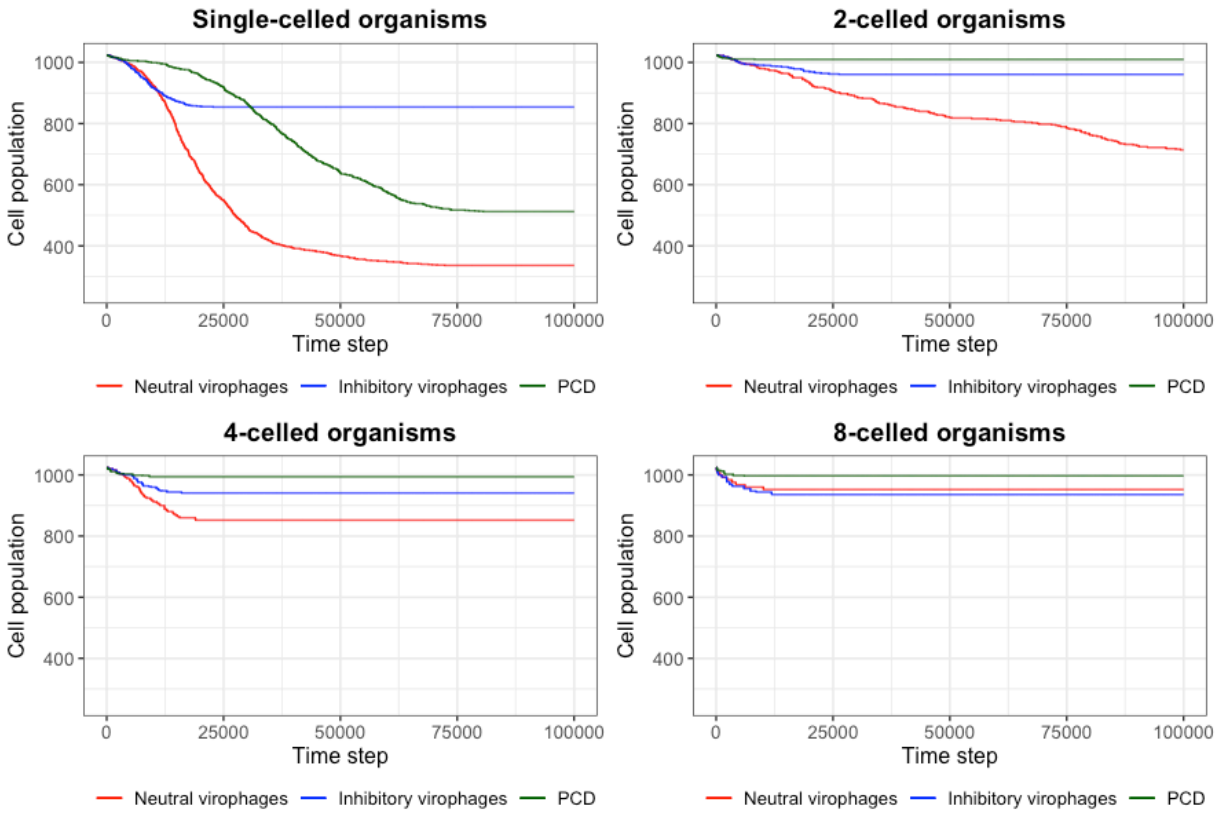


Figure S4. Time-course of the number of surviving cells during an ABM simulation. The total number of cells during a stochastic simulation is shown as a function of time and grouped by the number of cells per organism and treatment. We can observe a greater survival of cells in treatments with PCD and inhibitory virophages, while the lowest survival occurred in the treatments with neutral virophages. Grouping cells together also provided an advantage since the survival of cells increases with increasing number of cells per organism. Random seed = 1234.

Videos available at: https://figshare.com/s/b88b9d60f5ba24e3adbc

Video S1 (in figshare). Simulation of a population of single-celled organisms interacting with viruses and neutral virophages. The simulation was started with 1024 cells, 512 viruses and 2048 virophages. Cells are represented as blue, viruses are red and virophages are green. The axes are in micrometres. Random seed = 1234.

Video S2 (in figshare). Simulation of a population of single-celled organisms interacting with viruses and inhibitory virophages. The simulation was started with 1024 cells, 512 viruses and 2048 virophages. Cells are represented as blue, viruses are red and virophages are green. The axes are in micrometres. Random seed = 1234.

Video S3 (in figshare). Simulation of a population of single-celled organisms interacting with viruses and neutral virophages, which can undergo virus-induced PCD. The simulation was started with 1024 cells, 512 viruses and 2048 virophages. Cells are represented as blue, viruses are red and virophages are green. The axes are in micrometres. Random seed = 1234.

Video S4 (in figshare). Simulation of a population of 2-celled organisms interacting with viruses and neutral virophages. The simulation was started with 1024 cells, 512 viruses and 2048 virophages. Cells are represented as blue, viruses are red and virophages are green. The axes are in micrometres. Random seed = 1234.

Video S5 (in figshare). Simulation of a population of 2-celled organisms interacting with viruses and inhibitory virophages. The simulation was started with 1024 cells, 512

viruses and 2048 virophages. Cells are represented as blue, viruses are red and virophages are green. The axes are in micrometres. Random seed = 1234.

Video S6 (in figshare). Simulation of a population of 2-celled organisms interacting with viruses and neutral virophages, which can undergo virus-induced PCD. The simulation was started with 1024 cells, 512 viruses and 2048 virophages. Cells are represented as blue, viruses are red and virophages are green. The axes are in micrometres. Random seed = 1234.

Video S7 (in figshare). Simulation of a population of 4-celled organisms interacting with viruses and neutral virophages. The simulation was started with 1024 cells, 512 viruses and 2048 virophages. Cells are represented as blue, viruses are red and virophages are green. The axes are in micrometres. Random seed = 1234.

Video S8 (in figshare). Simulation of a population of 4-celled organisms interacting with viruses and inhibitory virophages. The simulation was started with 1024 cells, 512 viruses and 2048 virophages. Cells are represented as blue, viruses are red and virophages are green. The axes are in micrometres. Random seed = 1234.

Video S9 (in figshare). Simulation of a population of 4-celled organisms interacting with viruses and neutral virophages, which can undergo virus-induced PCD. The simulation was started with 1024 cells, 512 viruses and 2048 virophages. Cells are represented as blue, viruses are red and virophages are green. The axes are in micrometres. Random seed = 1234.

Video S10 (in figshare). Simulation of a population of 8-celled organisms interacting with viruses and neutral virophages. The simulation was started with 1024 cells, 512 viruses and 2048 virophages. Cells are represented as blue, viruses are red and virophages are green. The axes are in micrometres. Random seed = 1234.

Video S11 (in figshare). Simulation of a population of 8-celled organisms interacting with viruses and inhibitory virophages. The simulation was started with 1024 cells, 512 viruses and 2048 virophages. Cells are represented as blue, viruses are red and virophages are green. The axes are in micrometres. Random seed = 1234.

Video S12 (in figshare). Simulation of a population of 8-celled organisms interacting with viruses and neutral virophages, which can undergo virus-induced PCD. The simulation was started with 1024 cells, 512 viruses and 2048 virophages. Cells are represented as blue, viruses are red and virophages are green. The axes are in micrometres. Random seed = 1234.

Video S13 (in figshare). Simulation of a population of 16-celled organisms interacting with viruses and neutral virophages. The simulation was started with 1024 cells, 512 viruses and 2048 virophages. Cells are represented as blue, viruses are red and virophages are green. The axes are in micrometres. Random seed = 1234.

Video S14 (in figshare). Simulation of a population of 16-celled organisms interacting with viruses and inhibitory virophages. The simulation was started with 1024 cells, 512 viruses and 2048 virophages. Cells are represented as blue, viruses are red and virophages are green. The axes are in micrometres. Random seed = 1234.

Video S15 (in figshare). Simulation of a population of 16-celled organisms interacting with viruses and neutral virophages, which can undergo virus-induced PCD. The simulation was started with 1024 cells, 512 viruses and 2048 virophages. Cells are represented as blue, viruses are red and virophages are green. The axes are in micrometres. Random seed = 1234.

Appendix to Chapter V

Supplementary material: An ancient endogenous DNA virus in the human genome

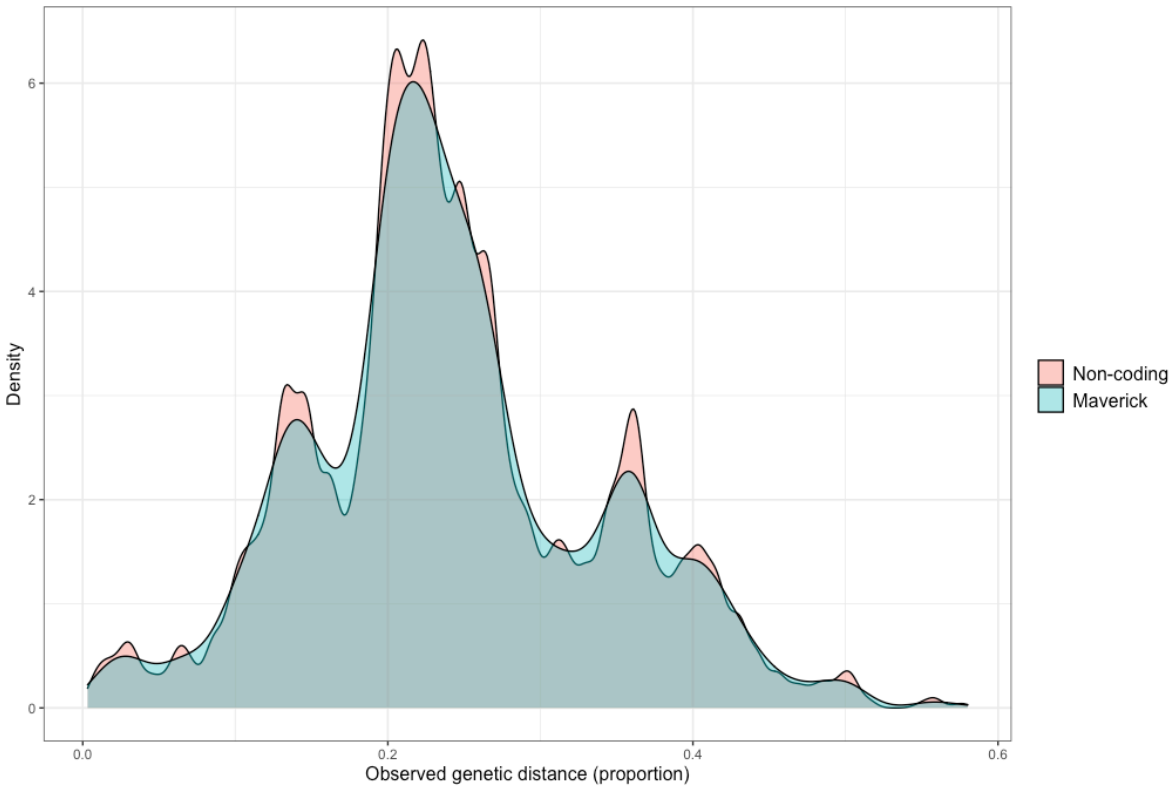


Figure S1. Distribution of observed genetic distances in the chromosome 7 *Maverick* integration and in neighbouring noncoding regions (#taxa = 120, #pair-wise comparisons: *Maverick* = 7,140; non-coding = 714,000). The two distributions are not significantly different (Kolmogorov-Smirnov test, $D = 5.6 \cdot 10^{-4}$, $p = 0.9956$). Percentage overlap of empirical distributions = 92%.

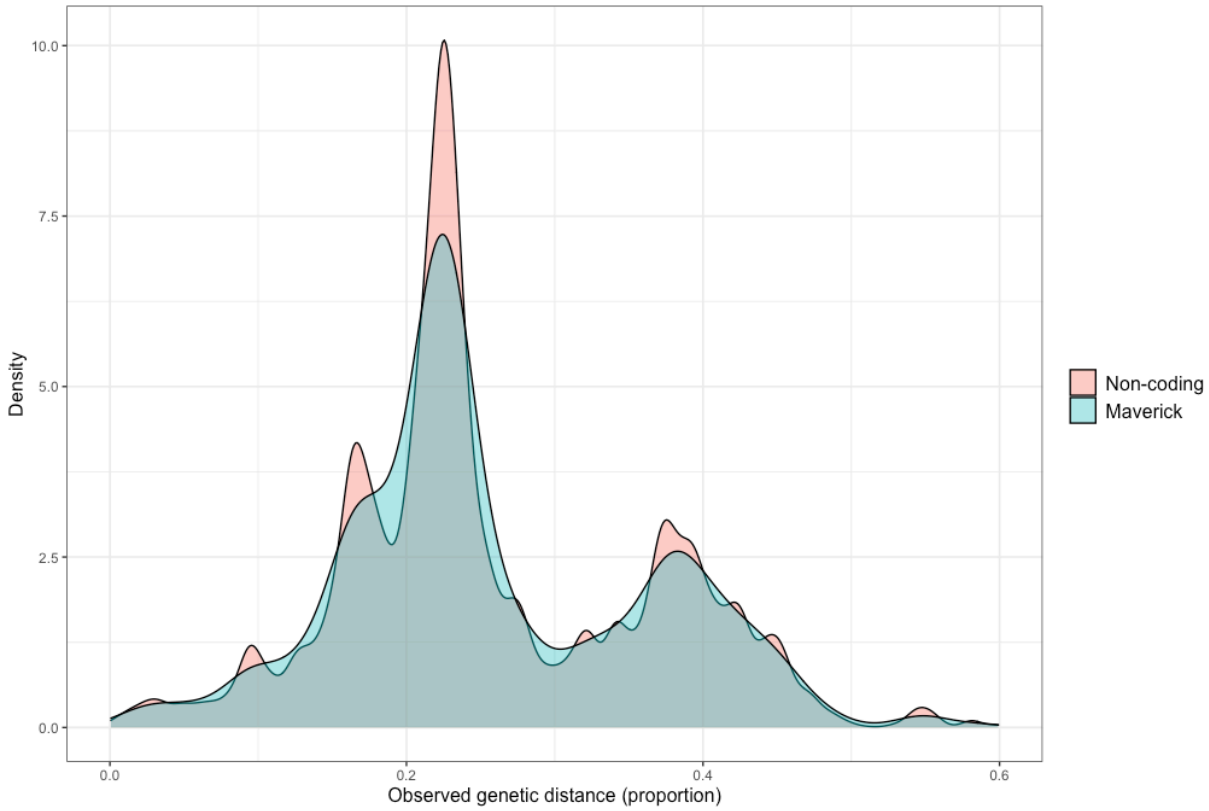


Figure S2. Distribution of observed genetic distances in the chromosome 8 *Maverick* integration and in the neighbouring noncoding regions (#taxa = 120, #pair-wise comparisons: *Maverick* = 7,140; non-coding = 714,000). The two distributions are not significantly different (Kolmogorov-Smirnov test, $D = 4.2 \cdot 10^{-4}$, $p = 0.9975$). Percentage overlap of empirical distributions = 85%.

Table S1. Calibration points from TimeTree used for the divergence-time analysis in BEAST.

Clade	Median	Estimated	Range	BEAST mean prior	BEAST deviation prior
Hominini	6.4	6.7	5.1–11.8	6.4	1.2
Ochotona/*Oryctolagus*	50.6	51.4	47.3–54	50.6	1.8
Chiroptera	62	61.7	60–64.6	62	1.2
Xenarthra	68	66	61–72	68	3
Condylura/*Erinaceus*	70	67	60–72	70	4
Primates	74.1	73.8	70.9–76.9	74.1	1.5
Glires	80	82	77–88	80	3
Afrotheria	80	83	77–86	80	1.6
Euarchonta	85	82	76–88	85	4
Euarchontoglires	89	90	85–97	89	3
Boreoeutheria	94	96	91–101	94	1.5
Atlantogenata	99	101	95–104	99	2
Eutheria	102	105	99–109	102	1.5

Table S2. Significant matches to the *Maverick*-PolB protein of *Terrapene carolina* in marsupial and monotreme genomes.

Clade	Species	Chromosome/ Scaffold	Accession	Query cover	e-value	% identity
Monotremata	*Tachyglossus aculeatus*	Chr X4	NC_052098.1	15%	7e-13	41.49%
Monotremata	*Tachyglossus aculeatus*	Chr X1	NC_052101.1	8%	3e-08	39.56%
Monotremata	*Ornithorhynchus anatinus*	Chr X5	NC_041753.1	8%	3e-07	41.76%
Marsupialia	*Trichosurus vulpecula*	Chr 3	NC_050575.1	44%	1e-34	40.51%
Marsupialia	*Trichosurus vulpecula*	Chr 4	NC_050576.1	17%	6e-21	42.53%
Marsupialia	*Trichosurus vulpecula*	Chr 1	NC_050573.1	31%	5e-08	39.45%
Marsupialia	*Trichosurus vulpecula*	Chr 8	NC_050580.1	35%	6e-10	41.46%
Marsupialia	*Trichosurus vulpecula*	Chr 2	NC_050574.1	19%	9e-10	44.44%
Marsupialia	*Trichosurus vulpecula*	Chr 5	NC_050577.1	20%	2e-07	32.56%
Marsupialia	*Phascolarctos cinereus*	scaf00042	NW_018343993.1	20%	1e-28	42.86%
Marsupialia	*Phascolarctos cinereus*	scaf00105	NW_018344056.1	24%	1e-16	32.21%

Marsupialia	*Phascolarctos cinereus*	scaf00322	NW_018344273.1	42%	4e-10	32.53%
Marsupialia	*Phascolarctos cinereus*	scaf00360	NW_018344311.1	33%	1e-14	46.30%
Marsupialia	*Phascolarctos cinereus*	scaf00046	NW_018343997.1	13%	2e-11	36.99%
Marsupialia	*Phascolarctos cinereus*	scaf00057	NW_018344008.1	18%	4e-09	29.85%
Marsupialia	*Phascolarctos cinereus*	scaf00062	NW_018344013.1	12%	4e-08	37.80%
Marsupialia	*Phascolarctos cinereus*	scaf00250	NW_018344201.1	8%	8e-07	36.96%
Marsupialia	*Phascolarctos cinereus*	scaf00147	NW_018344098.1	6%	9e-06	41.43%
Marsupialia	*Vombatus ursinus*	UNPS02014838.1	NW_020954616.1	33%	3e-10	37.93%
Marsupialia	*Vombatus ursinus*	UNPS02014830.1	NW_020954608.1	40%	3e-07	31.03%
Marsupialia	*Vombatus ursinus*	UNPS02014158.1	NW_020953936.1	25%	1e-11	49.38%
Marsupialia	*Vombatus ursinus*	UNPS02014864.1	NW_020954642.1	8%	2e-11	53.19%
Marsupialia	*Vombatus ursinus*	UNPS02014945.1	NW_020954723.1	14%	2e-10	36.97%
Marsupialia	*Vombatus ursinus*	UNPS02014809.1	NW_020954587.1	13%	4e-09	37.14%
Marsupialia	*Vombatus ursinus*	UNPS02013287.1	NW_020953065.1	6%	6e-07	44.62%
Marsupialia	*Vombatus ursinus*	UNPS02013304.1	NW_020953082.1	6%	6e-07	44.62%
Marsupialia	*Vombatus ursinus*	UNPS02014599.1	NW_020954377.1	6%	6e-07	44.62%
Marsupialia	*Dromiciops gliroides*	Chr 2	NC_057862.1	25%	1e-20	44.36%
Marsupialia	*Dromiciops gliroides*	Chr 1	NC_057861.1	16%	8e-12	37.08%

Marsupialia	*Dromiciops gliroides*	Chr 3	NC_057863.1	14%	2e-08	50.00%
Marsupialia	*Gracilinanus agilis*	Chr 2	NC_058131.1	29%	8e-20	49.60%
Marsupialia	*Gracilinanus agilis*	Chr 3	NC_058132.1	10%	2e-10	32.98%
Marsupialia	*Gracilinanus agilis*	Chr 4	NC_058133.1	25%	7e-08	29.79%
Marsupialia	*Gracilinanus agilis*	Chr 1	NC_058130.1	21%	8e-07	28.22%
Marsupialia	*Gracilinanus agilis*	scaffold31845	NW_025364518.1	9%	1e-06	38.78%
Marsupialia	*Gracilinanus agilis*	Chr 6	NC_058135.1	7%	3e-06	38.96%
Marsupialia	*Sarcophilus harrisii*	Chr 4	NC_045429.1	13%	9e-17	37.50%
Marsupialia	*Sarcophilus harrisii*	Chr 2	NC_045427.1	12%	1e-11	50.91%
Marsupialia	*Sarcophilus harrisii*	Chr 1	NC_045426.1	6%	1e-08	48.65%
Marsupialia	*Monodelphis domestica*	Chr 1	NC_008801.1	38%	4e-15	38.22%
Marsupialia	*Monodelphis domestica*	Chr 4	NC_008804.1	10%	4e-12	42.11%
Marsupialia	*Monodelphis domestica*	Chr 7	NC_008807.1	8%	3e-08	46.97%

Marsupialia	*Monodelphis domestica*	Chr 2	NC_008802.1	17%	3e-08	46.43%

Appendix to the Discussion

Calculation of the predicted virion size of *Mavericks* and adenoviruses

This estimate is based on the allometric relationship between genome size and virion volume reported by Cui, Schlub and Holmes (2014). The allometric equation they describe is of the form:

$$V = b \cdot L^a$$

Where *V* is the volume of the virion in cube nanometers (nm^3) and *L* is the size of the genome in thousands of base pairs (kb). For dsDNA viruses they find that **a** = 1.52 and **b** = 1182. By approximating the volume of an icosahedral virion with a sphere we can write:

$$\frac{4}{3}\pi r^3 = 1182 \cdot L^{1.52}$$

We can thus solve for the radius of the virion:

$$r = (\frac{3}{4\pi} \cdot 1182 \cdot L^{1.52})^{1/3}$$

The size of an adenovirus virion ranges between 70 to 100 nm in diameter (Kennedy and Parks, 2009), whilst their genomes range from 25-48 kb (ICTV). Using the above equation and genome sizes, we would expect the diameter of adenovirus virions to vary between 67 and 93 nm (twice the predicted radius), which is consistent with the literature. Since the

genomes of *Mavericks* vary between 12.7 to 24.6 kb, the expected diameters of the virion are between 48 and 67 nm (like a small adenovirus).

References

Cui J, Schlub TE, Holmes EC. (2014). An allometric relationship between the genome length and virion volume of viruses. *Journal of Virology* 88: 6403—6410. doi: 10.1128/JVI.00362-14.

Kennedy MA, Parks RJ. (2009). Adenovirus virion stability and the viral genome: size matters. *Molecular Therapy* 17: 1664—1666. doi: 10.1038/mt.2009.202.