VIRAL DISCOVERY IN BLUEGILL SUNFISH (*LEPOMIS MACROCHIRUS*) AND GIANT GUITARFISH (*RHYNCHOBATUS DJIDDENSIS*) BY HISTOPATHOLOGY EVALUATION, METAGENOMIC ANALYSIS AND NEXT GENERATION SEQUENCING

by

JENNIFER ANNE DILL

(Under the Direction of Alvin Camus)

ABSTRACT

The rapid growth of aquaculture production and international trade in live fish has led to the emergence of many new diseases. The introduction of novel disease agents can result in significant economic losses, as well as threats to vulnerable wild fish populations. Losses are often exacerbated by a lack of agent identification, delay in the development of diagnostic tools and poor knowledge of host range and susceptibility. Examples in bluegill sunfish (*Lepomis macrochirus*) and the giant guitarfish (*Rhynchobatus djiddensis)* will be discussed here. Bluegill are popular freshwater game fish, native to eastern North America, living in shallow lakes, ponds, and slow moving waterways. Bluegill experiencing epizootics of proliferative lip and skin lesions, characterized by epidermal hyperplasia, papillomas, and rarely squamous cell carcinoma, were investigated in two isolated poopulations. Next generation genomic sequencing revealed partial DNA sequences of an endogenous retrovirus and the entire circular genome of a novel hepadnavirus. Giant Guitarfish, a rajiform elasmobranch listed as 'vulnerable' on the IUCN Red List, are found in the tropical Western Indian Ocean.

Proliferative skin lesions were observed on the ventrum and caudal fin of a juvenile male quarantined at a public aquarium following international shipment. Histologically, lesions consisted of papillomatous epidermal hyperplasia with myriad large, amphophilic, intranuclear inclusions. Deep sequencing and metagenomic analysis produced the complete genomes of two novel DNA viruses, a typical polyomavirus and a second unclassified virus with a 20 kb genome tentatively named Colossomavirus. The goals of this research were to: 1) describe the various lesions in fish and associated viral agents by light and electron microscopy; 2) characterize the agents using molecular techniques and investigate their evolutionary phylogeny; 3) develop methodologies to rapidly detect each agent; and 4) determine the presence of these agents in available contact animals. Accomplishment of these objectives has provided data on pathologic changes associated with four novel viruses in fish, molecular and phylogenetic characterizations of the agents and diagnostic protocols for their detection.

INDEX WORDS: aquarium; bluegill; *Lepomis macrochirus*; retrovirus; hepadnavirus; elasmobranch; guitarfish; *Rhynchobatus djiddensis;* polyomavirus; colossomavirus; next generation sequencing; PCR; *in situ* hybridization

VIRAL DISCOVERY IN BLUEGILL SUNFISH (*LEPOMIS MACROCHIRUS*) AND GIANT GUITARFISH (*RHYNCHOBATUS DJIDDENSIS*) BY HISTOPATHOLOGY EVALUATION, METAGENOMIC ANALYSIS AND NEXT GENERATION SEQUENCING

by

JENNIFER ANNE DILL

BS, Roger Williams University, 2007

DVM, University of Florida, 2011

A Dissertation Submitted to the Graduate Faculty of The University of Georgia in Partial Fulfillment of the Requirements for the Degree

DOCTOR OF PHILOSOPHY

ATHENS, GEORGIA

2016

VIRAL DISCOVERY IN BLUEGILL SUNFISH (*LEPOMIS MACROCHIRUS*) AND GIANT GUITARFISH (*RHYNCHOBATUS DJIDDENSIS*) BY HISTOPATHOLOGY EVALUATION, METAGENOMIC ANALYSIS AND NEXT GENERATION SEQUENCING

by

JENNIFER ANNE DILL

Major Professor: Alvin Camus
Committee: Susan Williams
Terry Fei Fan Ng

Electronic Version Approved:

Suzanne Barbour
Dean of the Graduate School
The University of Georgia
May 2016

DEDICATION

This dissertation is dedicated to my parents and my husband.

ACKNOWLEDGEMENTS

Many thanks to my PhD committee members for their guidance and constructive feedback, Dr. Buffy Howerth for her guidance with *in situ* hybridization techniques, John Leary for his help with qPCR assay development and interpretation, and Dr. Joel Cline and the staff at the Georgia aquarium for providing valuable help in collecting samples.

TABLE OF CONTENTS

Page

Chapter 1

INTRODUCTION

In this research the pathological, genetic, and phylogenic characterization of four novel, emerging viruses will be described. The concept of an 'emerging' virus is an arbitrary term that can be defined as a previously unknown virus that is newly recognized as an animal pathogen, or as a previously described virus in which the incidence or severity of induced disease has increased in relation to previous records (Knipe et al. 2013; Tompkins et al. 2015; Allison 2010). Multiple factors can alter or enhance emergence, but unfortunately, intensive surveillance and research aimed at characterizing and comprehending novel agents are usually only undertaken retrospectively and when human or companion or food animal mortalities are excessive. Since 2000, over half the agents with sufficient evidence of emergence, have been isolated from fish and other aquatic species, and many of those were microparasites, predominantly viruses (Tompkins et al. 2015). Higher profile examples in aquatic species include viral hemorrhagic septicemia (VHSV) in the Great Lakes fish, megalocytivirus (iridovirus) infections in Asian aquaculture and the aquarium trade, fibropapilloma-associated turtle herpesvirus (FPTHV) in sea turtles, and morbillivirus induced mortalities in seals and dolphins (Whittington et al. 2010; Meyers et al. 1995; Skall et al. 2005; Guardo et al. 2005; Valenti et al. 2011).

Commercial aquaculture, as well as trade in live fish and fish products, are global industries with immense estimated worth. The value of farmed food fish production was

estimated at $137.7 billion in 2012 and represents one of the world's fastest growing food producing sectors. Trade in ornamental fish, both cultured and wild caught, was valued at approximately $278 million in 2005 and is a major source of income in many local markets (Bostock et al. 2010; Livengood et al. 2007; FAO 2012; Whittington et al. 2007; Rosa et al. 2002; Kautsky et al. 1997; Folke et al. 1992; Whitmarsh et al. 2006).

In contrast to other animal production sectors, there is enormous species diversity, less regulation and disease surveillance, and resistance by producers, toward a national fish health plan is the US. Still, a national aquatic animal health plan has been proposed as there is a need to protect domestic resources and to parallel regulations by foreign governments effecting imports and exports (Beveridge et al. 1997, NAAHTF, 2008; Caffey et al. 2000). The Federal agencies with the primary responsibility for aquatic animal health are the U.S. Department of Agriculture (USDA), the U.S. Department of Commerce (DOC), and the U.S. Department of the Interior (DOI) (NAAHTF, 2008). Together they have developed a National Aquatic Animal Health Task Force to develop and implement a National Aquatic Animal Health Plan (NAAHP) that should provide principles and guidelines to U.S. Federal Agencies with jurisdiction over aquatic animal health including the Animal and Plant Health Inspection Service (APHIS), the National Oceanic and Atmospheric Administration Fisheries (NOAA), and the United States Fish and Wildlife Service (FWS) (NAAHTF, 2008). This plan calls for cooperation between industry, regional organizations, like state, local, and Tribal governments, and all other stakeholders. Goals of the NAAHP recommendations are to facilitate the legal and safe movement of aquatic animals, protect and improve the quality of farmed animals, ensure the availability of diagnostics, training programs, and certification services, and to minimize the impacts of disease.

This research was conducted with such goals in mind. To reinforce aquaculture as a viable business activity, to protect cultured and wild resources, and to ensure the future of commercial aquaculture and the aquarium trade, there is a need to identify, diagnose and assess the impacts of disease agents affecting both wild populations and aquarium collections. The translocation of millions of pounds of freshwater and marine fish poses a significant threat of disease introduction (Bostock et al. 2010; Livengood et al. 2007; Tompkins et al. 2015; Whittington et al. 2010; Meyers et al. 1995; Skall et al. 2005). Potential impacts are compounded by less funding for directed research, but an interest in aquatic virology has been stimulated in recent years by continued worldwide increases in aquaculture and the discovery of novel agents that have the potential to expose deep evolutionary mysteries (Bostock et al. 2010; Yutin et al. 2014; Koonin et al. 2015; Moniruzzaman et al. 2014). This includes the identification and characterization of apparently non-pathogenic viruses that could serve as ancient phylogenic resources, as well as surrogate models for understanding aspects of viral ecology including modes of transmission, host diversity and virulence mechanisms as they relate to other closely-related pathogenic viruses.

References

Allison, A. 2010. Genetic Mechanisms of Virus Evolution and Emergence: Recombination, Reassortment, Overprinting and Mutation. Diss. University of Georgia, 2010.

Beveridge MC, Phillips MJ, Macintosh DJ. 1997. Aquaculture and the environment: the supply of and demand for environmental goods and services by Asian aquaculture and the implications for sustainability. Aquaculture Research 28:797-807.

Bostock J, McAndrew B, Richards R, Jauncey K, Telfer T, Lorenzen K, Little D, Ross L, Handisyde N, Gatward I, Corner R. 2010. Aquaculture: global status and trends. Philos Trans R Soc Lond B Biol Sci 365:2897-2912.

Caffey RH, Kazmierczak Jr RF, Avault JW. 2000. Developing consensus indicators of sustainability for Southeastern United States aquaculture. LSU AgCenter, Department of Agricultural Economics & Agribusiness Working Draft Bulletin.

FAO, Food and Agriculture Organization of the United Nations. FAO Yearbooks 1996 to 2005 and 2012. Fishery Statistics, Commodities.

Folke C, Kautsky N. 1992. Aquaculture with its environment: prospects for sustainability. Ocean Coast Manag 17:5-24.

Guardo GD, Marruchella G, Agrimi U, Kennedy S. 2005. Morbillivirus infections in aquatic mammals: a brief overview. J Vet Med A 52:88-93.

Kautsky N, Berg H, Folke C, Larsson J, Troell M. 1997. Ecological footprint for assessment of resource use and development limitations in shrimp and tilapia aquaculture. Aquaculture Research 28:753-766.

Knipe DM, Howley PM (ed). 2013. Fields Virology, 6th ed. Wolters Kluwer Health/Lippincott Williams & Wilkins, Philadelphia, PA.

Koonin EV, Krupovic M, Yutin N. 2015. Evolution of double-stranded DNA viruses of eukaryotes: from bacteriophages to transposons to giant viruses. Ann N Y Acad Sci 1341:10-24.

Koonin EV, Senkevich TG, Dolja VV. 2006. The ancient virus world and evolution of cells. Biol Direct 1:29.

Livengood EJ, Chapman FA. 2007. The ornamental fish trade: An introduction with perspectives for responsible aquarium fish ownership. University of Florida IFAS Extension.

Meyers TR, Winton JR. 1995. Viral hemorrhagic septicemia virus in North America. Annual Review of Fish Diseases 5:3-24.

Moniruzzaman M, LeCleir GR, Brown CM, Gobler CJ, Bidle KD, Wilson WH, Wilhelm SW. 2014. Genome of brown tide virus (AaV), the little giant of the Megaviridae, elucidates NCLDV genome expansion and host–virus coevolution. Virology 466:60-70.

Rosa IL, Oliveira TP, Osório FM, Moraes LE, Castro AL, Barros GM, Alves RR. 2011. Fisheries and trade of seahorses in Brazil: historical perspective, current trends, and future directions. Biodivers Conserv 20:1951-1971.

Skall HF, Olesen NJ, Mellergaard S. 2005. Viral haemorrhagic septicaemia virus in marine fish and its implications for fish farming–a review. J Fish Dis 28:509-529.

Tompkins DM, Carver S, Jones ME, Krkošek M, Skerratt LF. 2015. Emerging infectious diseases of wildlife: a critical perspective. Trends Parasitol 31:149-159.

Valenti WC, Kimpara JM, de L Preto B. 2011. Measuring aquaculture sustainability. World Aquaculture 42:26.

Whitmarsh DJ, Cook EJ, Black KD. 2006. Searching for sustainability in aquaculture: an investigation into the economic prospects for an integrated salmon–mussel production system. Mar Policy 30:293-298.

Whittington RJ, Becker JA, Dennis MM. 2010. Iridovirus infections in finfish–critical review with emphasis on ranaviruses. J Fish Dis 33:95-122.

Whittington RJ, Chong R. 2007. Global trade in ornamental fish from an Australian perspective: the case for revised import risk analysis and management strategies. Prev Vet Med 81:92-116.

Yutin N, Wolf YI, Koonin EV. 2014. Origin of giant viruses from smaller DNA viruses not from a fourth domain of cellular life. Virology 466:38-52.

Chapter 2

LITERATURE REVIEW

Virus Classification

Living organisms are classified into three domains, the Archaea, Bacteria and Eukarya. Viruses are notably excluded (Woese et al. 1990; Nasir et al. 2014). This is intriguing, since viruses do possess certain characteristics of life and are by far the most abundant source of nucleic acid diversity on Earth. They can also be killed, become extinct and evolve by various mechanisms (Suttle 2007; Holmes 2011; Villarreal 2005). The virosphere, the portion of the Earth in which viruses occur or which is affected by viruses, is inclusive of every environment and the repertoire of viral genes is great. Every life form is undoubtedly infected with viruses, usually multiple (Suttle 2007; Holmes 2011; Villarreal 2005; Bandea 2009, Koonin et al. 2013). Although for a longtime it was assumed Archaeoviruses were all DNA viruses, possibly due to high temperature RNA instability, it is becoming increasingly recognized that all domains contain a variety of viruses (Nasir et al. 2014; Forterre et al. 2013; Bolduc et al. 2012) (Figure 2.1). Collectively, these sources advocate that viruses are everywhere, part of everything and are evolving at a very fast pace.

One definition of a virus is an infectious, obligate intracellular parasite comprising genetic material surrounded by a protein coat and/or an envelope derived from a host cell membrane (Racaniello 2014). Note that this definition makes no reference to the molecular

identity of the virus. This is likely because, while eukaryotes, bacteria, and archaea utilize double stranded DNA to pass on their genetic codes, various classes of virus can be encoded by all forms of nucleic acid types, including both single-stranded and double-stranded DNA and RNA. A classification scheme based on the nucleic acid type found in virions and the mechanisms of transcription and replication, as well as how that structure influences mRNA synthesis, was first instituted by Baltimore in 1971 (Baltimore 1971) (Figure 2.2). Briefly, RNA and DNA viral genomes can be classified as either double-stranded (ds) or single-stranded (ss), in addition to being either segmented or non-segmented. Genome polarity [negative-sense (-), positive-sense (+)] may be applied to the ssDNA viruses, but is most commonly used in reference to delineating between ssRNA viruses whose genomes are in message-sense (+), and are thus directly infectious, and those genomes that are complementary to message (-) and therefore must first transcribe their mRNAs from the (-) genome by a virion-associated RNA polymerase (Knipe 2013; Joklik et al 1980; Voyles 1993; Flint et al. 2000; Racaniello 2014).

The original Baltimore scheme recognized six groups, but now with the inclusion of the gapped genome of the hepadnaviruses, there are seven. Group I contains double stranded DNA viruses, including the adenoviruses, herpesviruses, poxviruses, and polyomaviruses. Group I viruses primarily use the host cell biosynthetic systems for expression and replication and, all but poxviruses, replicate within the host cell nucleus (Knipe 2013; Joklik et al 1980; Voyles 1993). Group II contains positive (+) strand, or sense, single stranded DNA viruses, the parvoviruses are a well-known example. These viruses often rely entirely on host replicative mechanisms and often need the S or synthesis phase of the cell cycle to replicate (Knipe 2013; Joklik et al 1980; Voyles 1993).

Group III double stranded RNA viruses, like reoviruses, use their negative (-), or antisense, strand as a templet for synthesis of viral mRNA by a polymerase located in the capsid. Single stranded RNA viruses are divided into Groups IV, V and VI. Group IV viruses have a (+) strand RNA that serves directly as the source of information for viral synthesis. The genomes of picornaviruses and togaviruses are composed of messenger RNA (Knipe 2013; Joklik et al 1980; Voyles 1993). Group V viruses are (-) strand, single stand RNA viruses that include the orthomyxoviruses and rhabdoviruses and are encoded by a RNA dependent RNA polymerase in their viral genome. Group VI contains (+) strand, single stranded RNA viruses with a reverse transcription requirement and utilize a DNA intermediate (Knipe 2013; Joklik et al 1980; Voyles 1993). Replication of these viruses, the retroviruses, requires conversion of viral RNA genomes by reverse transcription to DNA and then integration of a so-called provirus into host genome. The newly identified group, Group VII, contains the hepadnavirues. These viruses have a partially double-stranded DNA genome, which is covalently linked to the viral reverse transcriptase (Knipe 2013; Joklik et al 1980; Voyles 1993). Hepadnaviruses are the only DNA viruses of animals known to replicate this way.

The viruses in this research represent a broad range of genomic configurations, including both single and double stranded RNA and DNA genomes. For the research conduct here, the proposed name, taxonomic status, and genomic organization, along with the host(s) from which each virus was obtained, were as follows: bluegill retrovirus, *Retroviridae*, ssRNA-RT (Group VI), bluegill (*Lepomis macrochirus*); bluegill hepadnavirus, *Hepadnaviridae*, dsDNA-RT (Group VII), bluegill (*L. macrochirus*); guitarfish polyomavirus, *Polyomaviridae*, dsDNA (Group I), giant guitarfish (*Rhynchobatus djiddensis*)*;* guitarfish colossomavirus, tentative new viral family *Colossomaviridae*, dsDNA (Group I), giant guitarfish (*R. djiddensis*).

Viral Evolution

Virus origin is still a highly debated topic because no theory has been proven. The three major theories that dominate discussions of the origin of viruses include whether: 1) viruses have a pre-cellular origin and could have contributed to the fundamental architecture of the first cells; 2) viruses evolved after the first cellular organisms as "escaped genes" that acquired capsid proteins and the ability to replicate autonomously; and 3) viruses are regressed copies of cellular species that have shed those genes whose functions are provided by the host (Villarreal 2005; Bandea 2009; Knipe 2013; Nasir et al. 2014; Koonin 2006) (Figure 2.3).

Phylogenetic and sequence analyses do not support a cellular origin for any DNA viruses, or the theory that viruses originated from bacterial or other unicellular genomes (Koonin et al. 2006; Koonin et al 2014.; Yutin et al. 2014). While elucidation of a common viral ancestor remains intangible, phylogenetic data does strongly imply that many viruses have evolved from well-established viral lineages (Koonin 2014 et al.; Yutin et al. 2014). However, several conceivable independent viral origins have been proposed, and a cascade of influential processes including a combination of horizontal gene transfer, vertical descent, genome reduction, genome expansion, and processes of architectural fusion and fission resulting in rearrangement (Koonin et al. 2013; Knipe 2013; Villarreal 2005; Wang et al 2007; Wolf et al. 2013; Holmes 2011).

Additional hurdles to establishing the age and origins of most extant viruses include highly variable mutation rates and the lack of a true fossil record. A distinguishing characteristic of viruses is the tendency for high genomic variability within populations. It is now understood that virus populations are not composed of a single member with a defined nucleic acid sequence, but rather, a dynamic distribution of nonidentical members called quasispecies

(Villarreal 2005; Knipe 2013). A key point is that the genome sequences of viruses cluster around an average sequence, but every genome is probably different from that consensus and it may be necessary to separately detect individual variants within a population. Quasispecies populations, with related but nonidentical viruses, are the result of rapid and error prone replication.

Irrespective of a single origin theory, it is clear that viruses are subject to similar forces that shape the evolution of other species. Viral populations can exhibit high grade genetic variation by a variety of mechanism that can steer evolution and include mutation, recombination and reassortment (Knipe 2013; Joklik et al. 1980; Voyles 1993; Flint et al. 2000; Racaniello 2014). Although each mechanisms may operate individually, they do not necessarily function exclusively. As mutation, in its simplest form, involves a single nucleotide change, it is the genetic mechanism that is most commonly observed in all viruses. However, mutation rates for RNA viruses far exceed those observed in DNA viruses. Owing to the lack of a 3′ to 5′ exonuclease activity, RNA-dependent RNA polymerases and RNA-dependent DNA polymerases (i.e., reverse transcriptases) cannot remove misincorporated nucleotides once they are inserted into a growing nucleic acid strand (Steinhauer et al. 1992; Knipe 2013; Voyles 1993). In contrast, DNA-dependent RNA or DNA polymerases (i.e., enzymes found in DNA viruses, prokaryotes, and eukaryotes) have a 3′ to 5′ proofreading-repair activity and thus can remove such misincorporated bases (Abbotts et al. 1985; Abbotts et al. 1987; Knipe 2013; Voyles 1993).

Recombination involves the exchange of nucleotide sequences between two different RNA or DNA molecules. It has been proposed that recombination may occur in viruses by both a breakage-and-rejoining mechanism and by a copy-choice or template-switching mechanism (Lai 1992; Nagy et al. 1997; Bujarski 2008). The breakage-and-rejoining mechanism is the

predominant form observed in DNA viruses (Block et al. 1985; Worobey et al. 1999; Esposito et al., 2006), while the second is widely accepted as the primary mechanism of recombination in RNA viruses (Lai 1992; Kim et al. 2001). Although recombination can occur in all types of RNA viruses, it has been documented most often in those that have ss (+) genomes, such as coronaviruses, flaviviruses and caliciviruses (Jackwood et al. 2010; Twiddy et al. 2003; Forrester et al. 2008).

Finally, reassortment is defined as the exchange of complete RNA or DNA segments between two (or more) viruses co-infecting the same cell. Reassortment, as a prerequisite, is limited to viruses which have a segmented genome. Since genome segmentation is relatively uncommon among ss (+) RNA, ss DNA and ds DNA viruses, reassortment as a major evolutionary mechanism is confined primarily to segmented viruses that have either a ds RNA or ss (-) RNA genome (Webster et al. 1992; Mertens 1999). These dynamic and ongoing processes result in abundant diversity and constant evolutional change in the viral world, as well as highlight the complexities of understanding viral evolution. Consequently, the origins and age of most extant viruses remains elusive.

Our understanding of viral origins can be improved by studying viral biodiversity, by focusing on environments and potential hosts that to date have been poorly sampled, and by plunging further into phylogenetic analysis and using such results to track ancient evolutionary history (Holmes 2011; Gilbert et al. 2010; Koonin et al. 2013; Koonin et al. 2006). While evolutionary time scales for other domains have historically been based on fossil records, viruses don't leave fossils (Knipe 2013; Voyles 1993; Racaniello 2014). In the past, characterizations of viruses were based mainly on phenotypical traits or even by the species they infect or the diseases they cause, but now various routine diagnostic techniques are available to better

differentiate known viruses (Knipe 2013; Joklik et al. 1980; Voyles 1993; Flint et al. 2000; Racaniello 2014). Transmission and scanning electron microscopy have been and remain useful techniques for describing morphologic features of viruses, such as shape, size and surface structures, e.g. classification of retroviruses as type a, b or c particles using descriptive means (Kurth et al. 2010; Dudley 2010; Knipe 2013). Culture methods, including plague assays, are used to isolate viruses, demonstrate cytotoxicity and in quantification and viral titer determination, but are not always a viable option as some viruses cannot be grown in c and many viruses do not produce cytopathic effects (Knipe 2013; Voyles 1993; Racaniello 2014). In particular, the study of many fish viruses has been limited by a lack of appropriate cell lines for a given fish species or virus (Knüsel et al. 2007; Ariel et al 2009; Imajoh et al. 2007). For viruses that produce unique antigens capable of provoking a host antibody response, a variety of methods can be employed, including antibody based serologic tests, such as ELISA and immunohistochemistry (Evensen et al. 1994; Jessie et al. 2004; Pikarsky et al. 2004; Al-Hussinee et al. 2011). While these methods are useful in many species and form the basis of numerous commercial tests, application in fish can again be limited by a lack of suitable reagents. Increasingly, newer molecular methods, using PCR and sequence data, often faster and more sensitive, are being employed to develop the field of virology, particularly when classical methods, such as culture, fail or are unavailable (Ng et al. 2014; Cruz et al. 2013; Ng et al. 2012; Delwart 2007; Delwart 2013). Among these emerging technologies, deep sequencing methods and metagenomic analysis have allowed for the interpretation of massive data inputs and the discovery of numerous new viral genomes (Ng et al. 2015a; Ng et al 2015b; Ng et al 2015c; Edwards 2005). Using typical desktop computers, online databases like BLAST and bioinformatics software such as Geneious, nucleotide sequences and whole viral genomes can be

described and quickly compared, markedly enhancing the rate of viral discovery. However the success of such technologies is dependent on strong financial patronage, the continued development of infrastructure and analytical tools to handle the enormous datasets generated, and skilled people to interpret them.

Culture-independent approaches, including the use of metagenomic and high-throughput methods, focus on the same criteria used in the Baltimore classification system and identify and classify the virus molecularly by its genetic code. This growing availability of genetic data, with the combined advancement of computational power, has provided the study of virology and paleovirology with new vigor, by utilizing both nucleotide sequences, as well as protein sequences and structures (Ho et al. 1995; Takezaki et al. 1995; Seshadri et al. 2007; Moniruzzaman et al. 2014; Edwards 2005; Katzourakis et al. 2010). The global viral metagenome is still largely uncharacterized and while many viruses are typified by high substitution rates, as discussed above, preventing reconstruction of their long-term evolutionary history, molecular clock techniques, using protein and DNA sequences, have become a valuable component of phylogenetic analysis (Ho et al. 1995; Takezaki et al. 1995; Wang et al. 2007). DNA is the dominant form used, but protein sequences can still be utilized to study deep divergences (Moniruzzaman et al. 2014; Edwards 2005; Villarreal 2005). Molecular clocks can estimate evolutionary sequence divergence, allowing for the high rates of sequence change, and emerging information on the genomic sequences and architecture of newly discovered viruses have the potential to redefine our understanding of virus function and evolution (Wang et al. 2007; Takezaki et al. 1995; Nasir et al. 2015; Koonin et al. 2006).

Classification of viruses based on genome characteristics, including nucleic acid type and replication strategy, along with morphology of the virus particle, have been successful because

these properties are generally maintained during virus evolution (Bandea 2009; Holmes 2011; Nasir et al. 2015; Villarreal 2005; Bamford et al. 2005). Even if viruses share little sequence identity, they can share protein structural and biochemical properties indicative of a common ancient origin (Nasir et al. 2015; Holmes 201; Bamford et al. 2005). Because molecular structure is relatively robust, protein domains and RNA secondary structure are typically less prone to the effects of mutation, making both evolutionarily more conserved (Nasir 2015; Holmes 2011; Bamford et al. 2005; Caetano-Anollés et al. 2012). An example of deep structural similarity found in seemingly diverse viral lineages is the highly conserved jelly-roll capsid present in dsDNA and dsRNA viruses, ssRNA+ viruses, and some DNA phages (Caetano-Anollés et al. 2012; Holmes 2011). Analysis of conserved protein structure continues to provide significant information on viral evolution, but exact chronological details of evolutionary dynamics, including when, where and how viruses emerged remains debatable (Knipe 2013; Murzin et al. 1995; Caetano-Anollés et al. 2012; Bamford et al. 2005; Kim et al. 2012; Abroi et al. 2011).

Another component to the conundrum of ancient viral evolution involves the study of viruses that have left footprints of their evolution in the genomes of their hosts, referred to as endogenous viral elements (EVEs) (Knipe 2013; Weiss 2006; Kurth et al. 2010; Dudley et al. 2010). This is the study of paleovirology. EVEs are permanently integrated into the genome of germ line cells, accumulate over time and are passed vertically through generations in a Mendelian pattern (Knipe 2013; Weiss 2006; Holmes 2011). Retroviruses account for the major portion of known EVEs, as host genome integration is an obligate step in their replication strategy (Kurth et al. 2010; Dudley et al. 2010; Herniou et al. 1998). However, while less common, examples of EVEs derived from viruses using all known replication strategies can be found, including RNA viruses (*Reoviridae*, *Flaviviridae*, *Orthomyxoviridae*, *Bunyaviridae*,

Bornaviridae, *Filoviridae*, *Rhabdoviridae*) and DNA viruses (*Parvoviridae*, *Circoviridae*, *Hepadnaviridae*) (Horie et al. 2010; Holmes 2011; Gilbert et al. 2014; Gilbert et al. 2010; Kurth et al. 2010; Dudley et al. 2010). Although many EVEs are highly mutated or fragmented and nonfunctional, tracing their presence can reveal an extensive history of genome invasion by some retroviruses dating back at least 93 million years and even fragmented EVEs can be used to track ancestry lines or locate ancient verses recent species divisions (Katzourakis et al. 2010; Horie et al. 2010; Holmes 2011; Gilbert et al. 2014; Gilbert et al. 2010; Dudley et al. 2010; Holmes 2011) (Figure 2.4).

Lastly, there is no single shared gene that unifies all viral groups, like the 16S and 18S small subunit rRNA sequences found in bacteria and eukaryotic cells, respectively (Yutin et al. 2011; Koonin et al. 2014; Moniruzzaman et al. 2014). One explanation revisits the theories of viral evolution and furthers the idea that viruses evolved from established viral lineages. While there is no single phylogeny linking all types of viruses, a polythetic history has been suggested, where viruses are derived from multiple viral origins, possibly several origins for one individual virus (Holmes 2011; Yutin et al 2011; Koonin et al. 2014; Moniruzzaman et al. 2014). This lack of phylogenic resolution is not the same as complete absence of common ancestry, but most likely reflects the extreme levels of sequence divergence discussed above. While not a single gene is shared by all viruses, a set of viral hallmark genes that cover the entire diversity of genome strategies have been recent proposed (Koonin et al. 2013). These genes encode essential viral functions, like the capsid protein of icosahedral viruses, DNA and RNA polymerases, distinct helicases involved in genome replication, integrases that catalyze insertion of viral DNA into host genomes, and others (Koonin et al 2013; Yutin et al 2011). This existence of various genes that are fundamental to virus replication and structure that are shared by a variety of

viruses, but are missing from cellular genomes, suggests the existence of an ancient virus world and offers an opportunity for a data-driven exploration of the deep roots of viruses (Koonin 2006).

Aquatic Viruses

A brief review of the evolution of aquatic species should supplement viral evolutionary research. The Cambrian explosion of species occurred nearly 545 million year ago and provided an immense increase in species evolution that has led to all modern life forms (Marshall 2006; Valentine et al. 1999; Koonin et al. 2013). This is recognizable by fossil records with the abrupt appearance of numerous skeletal forms like mollusks and echinoderms (Marshall 2006; Valentine et al. 1999). Such urochordates or sea urchins, and protostomes that include Mollusca, are known to support the replication of various virus particles and types (Munn 2006; Villarreal 2005).

The earliest vertebrates were jawless fish, similar to living hagfish, which appeared between 500 and 600 million years ago (Wilkin et al. 2012; Smith et al. 2013) (Figure 2.5). They had a cranium, but no vertebral column and were all extinct by the end of the Devonian period. Modern jawless fish, such as the lampreys and hagfish, are not direct descendants of the Class Agnatha, but are instead distant cousins of the cartilaginous fish (Munn 2006; Villarreal 2005; Kumar et al. 1998; Klapenbach 2012; Wilkin et al. 2012; Smith et al. 2013). The cartilaginous fish, or Chondrichthyes, include the elasmobranchs (sharks, skates, rays and sawfish) and the holocephalans (chimeras). They are believed to have diverged from ancestral fish-like species nearly 500 million years ago (mya) (Klimley 2013; Martin 2001). The chondrichthyans have skeletons composed of cartilage, sometimes mineralized, but do not possess true bone, and lack

swim bladders. The boney fish, or Osteichthyes, first arose about 400 million years ago and diverged into two groups, one that evolved into the modern ray-finned fish, the Actnopterygii, and the lobe-finned fish, or Sarcopterygii, which includes the lungfish and coelocanths (Villarreal 2005; Kumar et al. 1998; Klapenbach 2012; Wilkin et al 2012; Smith et al 2013). These fleshy finned fish later gave rise to amphibians and ultimately all tetrapods (Wilkin et al 2012).

With the diversification of aquatic species there also occurred a diversification of the viruses that infect them. It is estimated that now there are over 10^{30} total viruses in the ocean, many of which are phages infecting marine microbes, but the quantity of viruses far exceeds the abundance of bacteria and archaea (Suttle 2007). While great advances in marine virology have been made in recent years, there is still a lot to learn and discover. Arguably, recent discoveries of viruses from marine systems, such as the giant mimiviruses of amoebae and algae, with their unique large dsDNA genomes, have provided knowledge that is being used to reformat the existing theories of the origins and phylogenic histories of viruses (Moniruzzaman et al. 2014). Mimiviruses highlight our profound ignorance of the virosphere. The huge genomes of these nucleocytoplasmic large DNA viruses (NCLDV), shatter the definition of "filterable agent" because virions do not pass through bacterial filters as particles. They are bigger than multiple bacteria and archaea, and also possess a novel, diverse gene content (Moniruzzaman et al. 2014; Koonin et al. 2012; Koonin et al. 2013). Members of the proposed supergroup NCLDV, or "Megavirales", share a number of key gene sequences and structural features. Analysis of these shared characteristics reveals apparent evolutionary relationships, not only amongst this supergroup, but between giant and smaller viruses, as well as components found in the Polinton group of dsDNA transposons (Yutin et al. 2014; Koonin et al. 2015). With evidence building,

such findings propose clearer, sturdier models for the primordial origins of DNA-based viruses that infect eukaryotes and link and reclassify distinct viral families and groups (Koonin et al. 2013; Suttle 2005).

Oceans cover 70% of the Earth's surface and are composed of water, sediments, microorganisms, invertebrates and vertebrates so it is reasonable to hypothesize that every type of marine organism is host to at least one type of virus (Suttle 2007; Munn 2006; Villarreal 2005) (Figure 2.6). Viruses are not always pathogenic, and have a wide range of beneficial effects, including the structuring of microbial communities (Suttle 2007; Moniruzzaman et al. 2014). Examples of marine viruses that infect aquatic microorganisms are numerous. Phycodnaviruses effect the bloom dynamics of the algae *Emiliania huxleyi*, Ostreid herpesvirus 1 is responsible for annual summer mortalities of juvenile bivalve molluscs, and one of the most important shrimp diseases, white-spot syndrome, caused by an enveloped dsDNA virus, recently named as Whispovirus (Munn 2006; Villarreal 2005; Suttle 2007). Figures 2.6 provides an overview of viruses that infect marine organisms including bacteria and archaea (Munn 2006).

Surprisingly, relatively few viral infections have been documented in cartilaginous fish and most descriptions are based entirely on morphological features with no supporting molecular data (Leibovitz et al. 1985; McAllister et al. 1993; Terrell 2004; Bowman et al. 2008; Garner 2013, Camus et al. 2016). In contrast, representatives of most virus groups found in mammals are known to infet modern bony fish, including recent discoveries of a picornavirus and a polyomavirus (Villarreal 2005; Barbknecht; Peretti et al. 2015). A summary of viral families with representatives occurring in fish, amphibians and reptiles, as of 2005, can be found in Figure 2.7.

A conservative sampling bias of human and profitable agricultural animals, known as "ascertainment bias," could be truncating our knowledge of viral evolution and likely one reason very little is known about viruses in Chondrichthyes (Martin 2001; Nasir et al. 2015). As previously discussed, many aspects of viral origin and evolution remain unknown, but as descendants of early vertebrates that diverged during the Cambrian and Silurian periods, contemporary elasmobranches and teleosts could hold the key to understanding viral emergence and evolution among early vertebrates.

Proving Causation

The Henle-Koch Postulates, known simply as Koch's Postulates, were developed to demonstrate causal relationships between recognized disease entities and potential etiological agents. In brief, the postulates dictate that: 1) an organism must be regularly associated with a disease and its characteristic lesions, 2) the organism must be isolated from a diseased host and grown in culture, 3) the disease must be reproduced when a pure culture of the organism is introduced into a healthy susceptible host, and 4) the same organism must be reisolated from the same experimentally infected host (Henle 1938; Koch 1884; Koch 1982). The postulates have aided the study of infectious disease by demanding a secure scientific foundation for cause and effect relationships between microbes and their hosts, but despite their importance, have severe limitations. For example, Koch himself could not satisfactorily fulfil his own postulates when trying to establish the causes of leprosy and cholera (Evans 1976; Evans 1977; Hanson 1988; Fredericks et al. 1996; Rivers 1937).

The shortcomings of Koch's postulates become even more obvious when viral diseases are considered, probably because viruses had not been discovered when the postulates were

formulated (Evans 1976; Evans 1977; Hanson 1988; Fredericks et al. 1996; Rivers 1937). Contrary to the first postulate, many viruses do not cause illness in all infected individuals and infections with the same virus may lead to markedly different diseases in the same species, while different viruses can cause diseases with similar or identical clinical signs (Axthelm et al. 2004; Chamberlain 1985; Del Piero et al. 2001; Johne et al. 2007). More significantly, the remaining postulates cannot be fulfilled for viruses that cannot be replicated in cell culture or for which a suitable animal model has not been identified. In addition, the postulates are not applicable to viral pathogens that produce an asymptomatic carrier state. Neither do they consider the biological spectrum of a disease, the role of epidemiological factors on outcome, effects of multiple infections, and instances where a single disease has multiple causes under different conditions (Hament et al. 1999; Evans 1977; Fredericks et al. 1996; Schnitzer 1979).

Due to the absence of *in vitro* systems for the isolation and purification of many viral agents, including those described in this dissertation, it became evident that additional methods were needed to prove causal relationships. Advances in technology and the application of nucleic acid-based testing and sequencing made it possible for a set of broad based molecular postulates to be developed in 1996 that were applied in this project (Fredericks et al. 1996). The updated tenets of Koch's postulates for the 21st century, as suggested by Fredricks and Relman, are listed below (Fredericks et al. 1996).

1) A nucleic acid sequence belonging to a putative pathogen should be present in most cases of an infectious disease. Microbial nucleic acids should be found preferentially in those organs or gross anatomic sites known to be diseased, and not in those organs that lack pathologic changes.

2) Fewer, or no, copy numbers of pathogen-associated nucleic acid sequences should occur in hosts or tissues without disease.

3) With resolution of disease, the copy number of pathogen-associated nucleic acid sequences should decrease or become undetectable, while with clinical relapse, the opposite should occur.
4) When sequence detection predates disease, or sequence copy number correlates with severity of disease or pathology, the sequence-disease association is more likely to be a causal relationship.
5) The nature of the microorganism inferred from the available sequence should be consistent with the known biological characteristics of that group of organisms.
6) Tissue-sequence correlates should be sought at the cellular level. Efforts should be made to demonstrate specific *in situ* hybridization of microbial sequence to areas of tissue pathology and to visible microorganisms or to areas where microorganisms are presumed to be located.
7) Sequence-based forms of evidence for microbial causation should be reproducible.

While there is a great need to develop *in vitro* protocols for the culture of novel fish viruses, which may lead to the fulfilment of Koch's postulates, this pathway is not feasible at this time. As a result, the molecular postulates developed by Fredericks and Relman were used to establish disease causation in this study. Nucleic acid presence, quantity, and disease correlation, along with tissue-sequence correlates in the form of *in situ* hybridization, were used in this project as well as in other disease examples (Negro et al. 1992; Nocton et al. 1884; Rowley et al. 1994; Stoler et al. 1986).

References

Abbotts J, Loeb LA. 1985. DNA polymerase α and models for proofreading. Nucleic Acids Res 13:261-274.

Abbotts J, Loeb LA. 1985. On the fidelity of DNA replication: use of synthetic oligonucleotide-initiated reactions. Biochim Biophys Acta 824:58-65.

Abroi A, Gough J. 2011. Are viruses a source of new protein folds for organisms?–Virosphere structure space and evolution. Bioessays 33:626-635.

Al-Hussinee L, Lord S, Stevenson RM, Casey RN, Groocock GH, Britt KL, Kohler KH, Wooster GA, Getchell RG, Bowser PR, Lumsden JS. 2011. Immunohistochemistry and pathology of multiple Great Lakes fish from mortality events associated with viral hemorrhagic septicemia virus type IVb. Dis Aquat Organ 93:117-127.

Allison, A. 2010. Genetic Mechanisms of Virus Evolution and Emergence: Recombination, Reassortment, Overprinting and Mutation. Diss. University of Georgia, 2010.

Ariel E, Skall HF, Olesen NJ. 2009. Susceptibility testing of fish cell lines for virus isolation. Aquac 298:125-130.

Axthelm MK, Koralnik IJ, Dang X, Wüthrich C, Rohne D, Stillman IE, Letvin NL. 2004. Meningoencephalitis and demyelination are pathologic manifestations of primary polyomavirus infection in immunosuppressed rhesus monkeys. J Neuropathol Exp Neurol 63:750-758.

Baltimore D. 1971. Expression of animal virus genomes. Bacteriol Rev 35:235.

Bamford DH, Grimes JM, Stuart DI. 2005. What does structure tell us about virus evolution?. Curr Opin Struct Biol 15:655-663.

Bandea CI. 2009. The origin and evolution of viruses as molecular organisms. Nature Precedings. hdl:10101/npre.2009.3886.1.

Barbknecht M, Sepsenwol S, Leis E, Tuttle-Lau M, Gaikowski M, Knowles NJ, Lasee B, Hoffman MA. 2014. Characterization of a new picornavirus isolated from the freshwater fish Lepomis macrochirus. J Gen Virol 95:601-613.

Beveridge MC, Phillips MJ, Macintosh DJ. 1997. Aquaculture and the environment: the supply of and demand for environmental goods and services by Asian aquaculture and the implications for sustainability. Aquaculture Research 28:797-807.

Block W, Upton C, McFadden G. 1985. Tumorigenic poxviruses: genomic organization of malignant rabbit virus, a recombinant between Shope fibroma virus and myxoma virus. Virology 140:113-124.

Bolduc B, Shaughnessy DP, Wolf YI, Koonin EV, Roberto FF, Young M. 2012. Identification of novel positive-strand RNA viruses by metagenomic analysis of archaea-dominated Yellowstone hot springs. J Virol 86:5562-5573.

Bostock J, McAndrew B, Richards R, Jauncey K, Telfer T, Lorenzen K, Little D, Ross L, Handisyde N, Gatward I, Corner R. 2010. Aquaculture: global status and trends. Philos Trans R Soc Lond B Biol Sci 365:2897-2912.

Bowman M, Ramer J, Proudfoot J, Stringer E, Garner M, Trupkiewicz J, Giray C. 2008. A novel adenovirus in a collection of wild-caught dusky smooth-hounds (*Mustelus canis*). In: Proceedings AAZV ARAV Joint Conference, Los Angeles, CA.

Bowser PR, Casey JW. 1993. Retroviruses of fish. Annual Review of Fish Diseases 3:209-224.

Caetano-Anollés G, Nasir A. 2012. Benefits of using molecular structure and abundance in phylogenomic analysis. Front Genet 3:172.

Caffey RH, Kazmierczak Jr RF, Avault JW. 2000. Developing consensus indicators of sustainability for Southeastern United States aquaculture. LSU AgCenter, Department of Agricultural Economics & Agribusiness Working Draft Bulletin.

Camus A, Dill J, McDermott A, Camus M, Fan NT. 2015. Virus-associated papillomatous skin lesions in a giant guitarfish Rhynchobatus djiddensis: a case report. Dis Aquat Organ 117:253-258.

Chamberlain RW. 1958. Vector Relationships of Arthropod-Borne Encephalitides in North America. Public Health Reports (1896-1970).

Cruz FN, Giannitti F, Li L, Woods LW, Del Valle L, Delwart E, Pesavento PA. 2013. Novel polyomavirus associated with brain tumors in free-ranging raccoons, western United States. Emerg Infect Dis 19:77.

Del Piero F, Wilkins PA, Dubovi EJ, Biolatti B, Cantile C. 2001. Clinical, pathologic, immunohistochemical, and virologic findings of eastern equine encephalomyelitis in two horses. Vet Pathol 38(4):451-456.

Delwart EL. 2013. A roadmap to the human virome. PLoS pathogens 9: e1003146. doi:10.1371/journal.ppat.1003146

Delwart EL. 2007. Viral metagenomics. Rev Med Virol 17:115-131.

Domingo E, Escarmis C, Sevilla N, Moya A, Elena SF, Quer J, Novella IS, Holland JJ. Basic concepts in RNA virus evolution. FASEB J. 10:859-864.

Dudley J (ed). 2010. Retroviruses and insights into cancer. Springer Science & Business Media, Austin, TX.

Earth History a New Approach. Online: http://www.earthhistory.org.uk/transitional-fossils/fish-to-amphibian

Edwards RA, Rohwer F. 2005. Viral metagenomics. Nat Rev Microbiol 3:504-510

Esposito JJ, Sammons SA, Frace AM, Osborne JD, Olsen-Rasmussen M, Zhang M, Govil D, Damon IK, Kline R, Laker M, Li Y. 2006. Genome sequence diversity and clues to the evolution of variola (smallpox) virus. Science 313:807-812.

Evans AS. 1976. Causation and disease: the Henle-Koch postulates revisited. Yale J Biol Med 49:175.

Evans AS. 1991. Causation and disease: effect of technology on postulates of causation. Yale J Biol Med 64:513.

Evans, A. S. 1977. Limitation of Koch's postulates. *Lancet* ii:1277–1278. (Letter.)

Evensen Ø, Meier W, Wahli T, Olesen NJ, Vestergård Jørgensen PE, Håstein T. 1994. Comparison of immunohistochemistry and virus cultivation for detection of viral haemorrhagic septicaemia virus in experimentally infected rainbow trout Oncorhynchus mykiss. Dis Aquat Organ 20:101-109.

FAO, Food and Agriculture Organization of the United Nations. FAO Yearbooks 1996 to 2005 and 2012. Fishery Statistics, Commodities.

Flint SJ, Enquist LW, Racaniello VR, Skalka AM, Barnum DR, de Evaluación E. 2000. Principles of Virology: Molecular Biology, Pathogenesis and. ASM Press, Washington DC.

Folke C, Kautsky N. 1992. Aquaculture with its environment: prospects for sustainability. Ocean Coast Manag 17:5-24.

Forrester NL, Moss SR, Turner SL, Schirrmeier H, Gould EA. 2008. Recombination in rabbit haemorrhagic disease virus: possible impact on evolution and epidemiology. Virology 376:390-396.

Forterre P, Prangishvili D. 2013. The major role of viruses in cellular evolution: facts and hypotheses. Curr Opin Virol 3:558-565.

Fredericks DN, Relman DA. 1996. Sequence-based identification of microbial pathogens: a reconsideration of Koch's postulates. Clin Microbiol Rev 19:18-33.

Garner MM. 2013. A retrospective study of disease in elasmobranchs. Vet Pathol 50:377-389.

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

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

Guardo GD, Marruchella G, Agrimi U, Kennedy S. 2005. Morbillivirus infections in aquatic mammals: a brief overview. J Vet Med A 52:88-93.

Hament JM, Kimpen JL, Fleer A, Wolfs TF. 1999. Respiratory viral infection predisposing for bacterial disease: a concise review. FEMS Immunol Med Microbiol 26:189-195.

Hanson RP. 1988. Koch is dead. J Wildl Dis 24:193-200.

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

Hillier LW, Miller W, Birney E, Warren W, Hardison RC, Ponting CP, Bork P, Burt DW, Groenen MA, Delany ME, Dodgson JB. 2004. Sequence and comparative analysis of the chicken genome provide unique perspectives on vertebrate evolution. Nature 432:695-716.

Ho SY, Duchêne S. 2014. Molecular-clock methods for estimating evolutionary rates and timescales. Mol Ecol 23:5947-5965.

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

Holmes EC. 2011. What does virus evolution tell us about virus origins?. J Virol 85:5247-5251.

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

Imajoh M, Ikawa T, Oshima SI. 2007. Characterization of a new fibroblast cell line from a tail fin of red sea bream, Pagrus major, and phylogenetic relationships of a recent RSIV isolate in Japan. Virus Res 126:45-52.

Jackwood MW, Boynton TO, Hilt DA, McKinley ET, Kissinger JC, Paterson AH, Robertson J, Lemke C, McCall AW, Williams SM, Jackwood JW. 2010. Emergence of a group 3 coronavirus through recombination. Virology 98:98-108.

Jessie K, Fong MY, Devi S, Lam SK, Wong KT. 2004. Localization of dengue virus in naturally infected human tissues, by immunohistochemistry and in situ hybridization. J Infect Dis 189:1411-1418.

Johne R, Müller H. 2007. Polyomaviruses of birds: etiologic agents of inflammatory diseases in a tumor virus family. J Virol 81:11554-11559.

Joklik WK, Phil D (ed). 1980. Principles of animal virology. Appleton-Century-Crofts, New York, NY.

Katzourakis A, Gifford RJ. 2010.Endogenous viral elements in animal genomes. PLoS Genet 11:e1001191.

Kautsky N, Berg H, Folke C, Larsson J, Troell M. 1997. Ecological footprint for assessment of resource use and development limitations in shrimp and tilapia aquaculture. Aquaculture Research 28:753-766.

Kim KM, Caetano-Anollés G. 2010. The evolutionary history of protein fold families and proteomes confirms that the archaeal ancestor is more ancient than the ancestors of other superkingdoms. BMC Evol Biol 12:1.

Kim MJ, Kao C. 2001. Factors regulating template switch in vitro by viral RNA-dependent RNA polymerases: Implications for RNA–RNA recombination. Proc Natl Acad Sci U S A 98:4972-4977.

Klapenbach, L. The Basics of Vertebrate Evolution. Available Online: http://animals.about.com/od/evolution/a/vertebrateevolu.htm

Klimley AP. 2013. The biology of sharks and rays. University of Chicago Press, Chicago, IL.

Knipe DM, Howley PM (ed). 2013. Fields Virology, 6th ed. Wolters Kluwer Health/Lippincott Williams & Wilkins, Philadelphia, PA.

Knüsel R, Bergmann SM, Einer-Jensen K, Casey J, Segner H, Wahli T. 2007. Virus isolation vs RT-PCR: which method is more successful in detecting VHSV and IHNV in fish tissue sampled under field conditions? J Fish Dis 30:559-568

Koch, R. 1884. Die Aetiologie der Tuberculose. Mitt. Kaiser. Gesundh.

Koch, R. 1982. Ueber bakteriologische Forschung. *In* Verh. X. Int. Med. Congr. Berlin, 1890:35.

Koonin EV, Dolja VV. 2013. A virocentric perspective on the evolution of life. Curr Opin Virol 3:546-557.

Koonin EV, Dolja VV. 2014. Virus world as an evolutionary network of viruses and capsidless selfish elements. Microbiol Mol Biol Rev 78:278-303.

Koonin EV, Senkevich TG, Dolja VV. 2006. The ancient Virus World and evolution of cells. Biol Direct 1:29.

Kumar S, Hedges SB. 1998. A molecular timescale for vertebrate evolution. Nature 392:917-920.

Kurth R, Bannert N (eds). 2010. Retroviruses: molecular biology, genomics and pathogenesis. Caister Academic Press, Norfolk, UK.

Lai MM. 1992. RNA recombination in animal and plant viruses. Microbiol Rev 56:61-79.

Leibovitz L, Lebouitz SS. 1985. A viral dermatitis of the smooth dogfish, Mustelus canis (Mitchill). J Fish Dis 8:273-279.

Livengood EJ, Chapman FA. 2007. The ornamental fish trade: An introduction with perspectives for responsible aquarium fish ownership. University of Florida IFAS Extension.

Marshall CR. 2006. Explaining the Cambrian "explosion" of animals. Annu Rev Earth Planet Sci 34:355-84.

Martin A. 2001. The phylogenetic placement of Chondrichthyes: inferences from analysis of multiple genes and implications for comparative studies. Genetica 111:349-357.

McAllister PE, Stoskopf MK. 1993. Shark viruses, p 780−782. *In* Stoskopf MK (ed) Fish medicine. WB Saunders, Philadelphia, PA.

Mertens PP. 1999. Orbiviruses and coltiviruses–general features. Encyclopedia of Virology: 1043-1061.

Meyers TR, Winton JR. 1995. Viral hemorrhagic septicemia virus in North America. Annual Review of Fish Diseases 5:3-24.

Moniruzzaman M, LeCleir GR, Brown CM, Gobler CJ, Bidle KD, Wilson WH, Wilhelm SW. 2014. Genome of brown tide virus (AaV), the little giant of the Megaviridae, elucidates NCLDV genome expansion and host–virus coevolution. Virology 466:60-70.

Munn CB. 2006. Viruses as pathogens of marine organisms—from bacteria to whales. J Mar Biol Assoc U.K. 86:453-467.

Murzin AG, Brenner SE, Hubbard T, Chothia C. 1995. SCOP: a structural classification of proteins database for the investigation of sequences and structures. J Mol Biol 247:536-540.

Nagy PD, Simon AE.1997. New insights into the mechanisms of RNA recombination. Virology 235:1-9.

Nasir A, Forterre P, Kim KM, Caetano-Anollés G. 2015. The distribution and impact of viral lineages in domains of life. Recent Discoveries in Evolutionary and Genomic Microbiology 4:26

Nasir A, Sun FJ, Kim KM, Caetano-Anollés G. 2015. Untangling the origin of viruses and their impact on cellular evolution. Ann N Y Acad Sci 1341:61-74.

National Aquatic Animal Health Task Force. "National aquatic animal health plan for the United States." (2008): 147.

Negro F, Pacchioni D, Shimizu Y, Miller RH, Bussolati G, Purcell RH, Bonino F. 1992. Detection of intrahepatic replication of hepatitis C virus RNA by in situ hybridization and comparison with histopathology. Proc Natl Acad Sci USA 89:2247–2251.

Ng TF, Chen LF, Zhou Y, Shapiro B, Stiller M, Heintzman PD, Varsani A, Kondov NO, Wong W, Deng X, Andrews TD. 2014. Preservation of viral genomes in 700-y-old caribou feces from a subarctic ice patch. Proc Natl Acad Sci 111:16842-16847.

Ng TF, Kondov NO, Deng X, Van Eenennaam A, Neibergs HL, Delwart E. 2015. A metagenomics and case-control study to identify viruses associated with bovine respiratory disease. J Virol 89:5340-5349.

Ng TF, Miller MA, Kondov NO, Dodd EM, Batac F, Manzer M, Ives S, Saliki JT, Deng X, Delwart E. 2015. Oral papillomatosis caused by *enhydra lutris* papillomavirus 1 (ELPV-1) in southern sea otters (*enhydra lutris nereis*) in California, USA. J Wildl Dis 51: 446-453.

Ng TF, Wellehan JF, Coleman JK, Kondov NO, Deng X, Waltzek TB, Reuter G, Knowles NJ, Delwart E. 2015. A tortoise-infecting picornavirus expands the host range of the family Picornaviridae. Arch Virol 160:1319-1323.

Nocton JJ, Dressler F, Rutledge BJ, Rys PN, Persing DH, Steere AC. 1994. Detection of Borrelia burgdorferi DNA by polymerase chain reaction in synovial fluid from patients with Lyme arthritis. N Engl J Med 330:229–234.

Peretti A, FitzGerald PC, Bliskovsky V, Pastrana DV, Buck CB. 2015. Genome sequence of a fish-associated polyomavirus, black sea bass (*Centropristis striata*) polyomavirus 1. Genome Announc 3:e01476-14.

Pikarsky E, Ronen A, Abramowitz J, Levavi-Sivan B, Hutoran M, Shapira Y, Steinitz M, Perelberg A, Soffer D, Kotler M. 2004. Pathogenesis of acute viral disease induced in fish by carp interstitial nephritis and gill necrosis virus. J Virol 78:9544-9551.

Racaniello, Vincent. Virology – Biology W3310/4310. Columbia University. Online Course. Spring 2014.

Rivers TM. 1937. Viruses and Koch's postulates. J Bacteriol 33:1.

Rosa IL, Oliveira TP, Osório FM, Moraes LE, Castro AL, Barros GM, Alves RR. 2011. Fisheries and trade of seahorses in Brazil: historical perspective, current trends, and future directions. Biodivers Conserv 20:1951-1971.

Rowley AH, Wolinsky SM, Relman DA, Sambol SP, Sullivan JA, Terai M, Shulman ST. 1994. Search for highly conserved viral and bacterial nucleic acid sequences corresponding to an etiologic agent of Kawasaki disease. Pediatr Res 36:567–571.

Schnitzer TJ, Gonczol E. 1979. Phenotypic mixing between murine oncoviruses and murine cytomegalovirus. J Gen Virol 43:691-695.

Seshadri R, Kravitz SA, Smarr L, Gilna P, Frazier M. 2007. CAMERA: a community resource for metagenomics. PLoS biology 5.

Skall HF, Olesen NJ, Mellergaard S. 2005. Viral haemorrhagic septicaemia virus in marine fish and its implications for fish farming–a review. J Fish Dis 28:509-529.

Smith JJ, Kuraku S, Holt C, Sauka-Spengler T, Jiang N, Campbell MS, Yandell MD, Manousaki T, Meyer A, Bloom OE, Morgan JR. 2013. Sequencing of the sea lamprey (Petromyzon marinus) genome provides insights into vertebrate evolution. Nat Genet 45:415-421.

Steinhauer DA, Domingo E, Holland JJ. 1992. Lack of evidence for proofreading mechanisms associated with an RNA virus polymerase. Gene 122:281-288.

Stoler MH, Broker TR. 1986. In situ hybridization detection of human papillomavirus DNAs and messenger RNAs in genital condylomas and a cervical carcinoma. Hum Pathol 17:1250–1258.

Suttle CA. 2007. Marine viruses—major players in the global ecosystem. Nat Rev Microbiol 5: 801-812.

Takezaki N, Rzhetsky A, Nei M. 1995. Phylogenetic test of the molecular clock and linearized trees. Mol Biol Evol 12:823-833.

Terrell SP. 2004. An introduction to viral, bacterial, and fungal diseases of elasmobranchs p 427−431 *In* Smith M, Warmolts D, Thoney D, Heuter R (eds) Elasmobranch husbandry manual:captive care of sharks, rays, and their relatives. Ohio Biological Survey, Columbus, OH.

Tompkins DM, Carver S, Jones ME, Krkošek M, Skerratt LF. 2015. Emerging infectious diseases of wildlife: a critical perspective. Trends Parasitol 31:149-159.

Twiddy SS, Holmes EC. 2003. The extent of homologous recombination in members of the genus Flavivirus. J Gen Virol 84:429-440.

Valenti WC, Kimpara JM, de L Preto B. 2011. Measuring aquaculture sustainability. World Aquaculture 42:26.

Valentine JW, Jablonski D, Erwin DH. 1999. Fossils, molecules and embryos: new perspectives on the Cambrian explosion. Development 126:851-859.

van Regenmortel MH, Mahy BW (ed). 2010. Desk encyclopedia of general virology. Academic Press San Diego, CA.

Villarreal LP. 2005. Viruses and the Evolution of Life. American Society of Microbiology Press, Washington DC.

Voyles, BA. 1993. The biology of viruses. Mosby, St. Louis, MO.

Wang M, Yafremava LS, Caetano-Anollés D, Mittenthal JE, Caetano-Anollés G. 2007. Reductive evolution of architectural repertoires in proteomes and the birth of the tripartite world. Genome Res 17:1572-1585.

Webster RG, Bean WJ, Gorman OT, Chambers TM, Kawaoka Y. 1992. Evolution and ecology of influenza A viruses. Microbiol Rev 56:152-179.

Weiss RA. 2006. The discovery of endogenous retroviruses. Retrovirology 3:1.

Whitmarsh DJ, Cook EJ, Black KD. 2006. Searching for sustainability in aquaculture: an investigation into the economic prospects for an integrated salmon–mussel production system. Mar Policy 30:293-298.

Whittington RJ, Becker JA, Dennis MM. 2010. Iridovirus infections in finfish–critical review with emphasis on ranaviruses. J Fish Dis 33:95-122.

Whittington RJ, Chong R. 2007. Global trade in ornamental fish from an Australian perspective: the case for revised import risk analysis and management strategies. Prev Vet Med 81:92-116.

Wilkin, Douglas. Vertebrate Evolution. Online: http://www.ck12.org/book/CK-12 BiologyConcepts/r17/section/12.5/ Vertebrate-Evolution/

Woese CR, Kandler O, Wheelis ML. 1990. Towards a natural system of organisms: proposal for the domains Archaea, Bacteria, and Eucarya. Proc Natl Acad Sci 87:4576-4579.

Wolf YI, Koonin EV. 2013. Genome reduction as the dominant mode of evolution. Bioessays 35:829-837.

Worobey M, Holmes EC. 1999. Evolutionary aspects of recombination in RNA viruses. J Gen Virol 80:2535-2543.

Yutin N, Wolf YI, Koonin EV. 2014. Origin of giant viruses from smaller DNA viruses not from a fourth domain of cellular life. Virology 466:38-52.

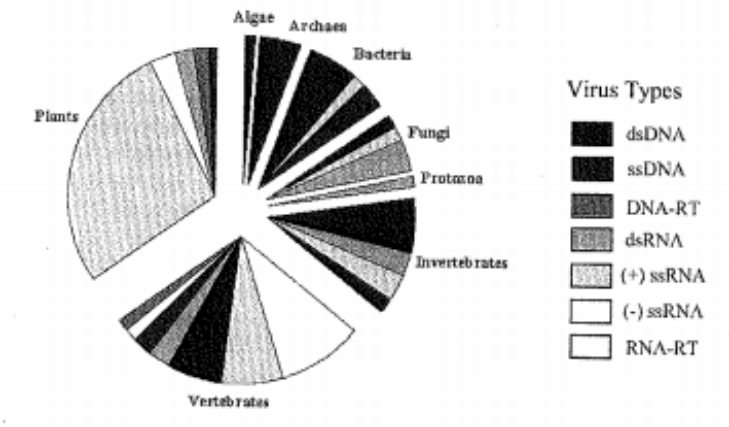


Figure 2.1 Host groups and the distribution of virus types found in each as known in 2005. From: Villarreal LP. 2005. Viruses and the Evolution of Life. American Society of Microbiology Press, Washington DC.

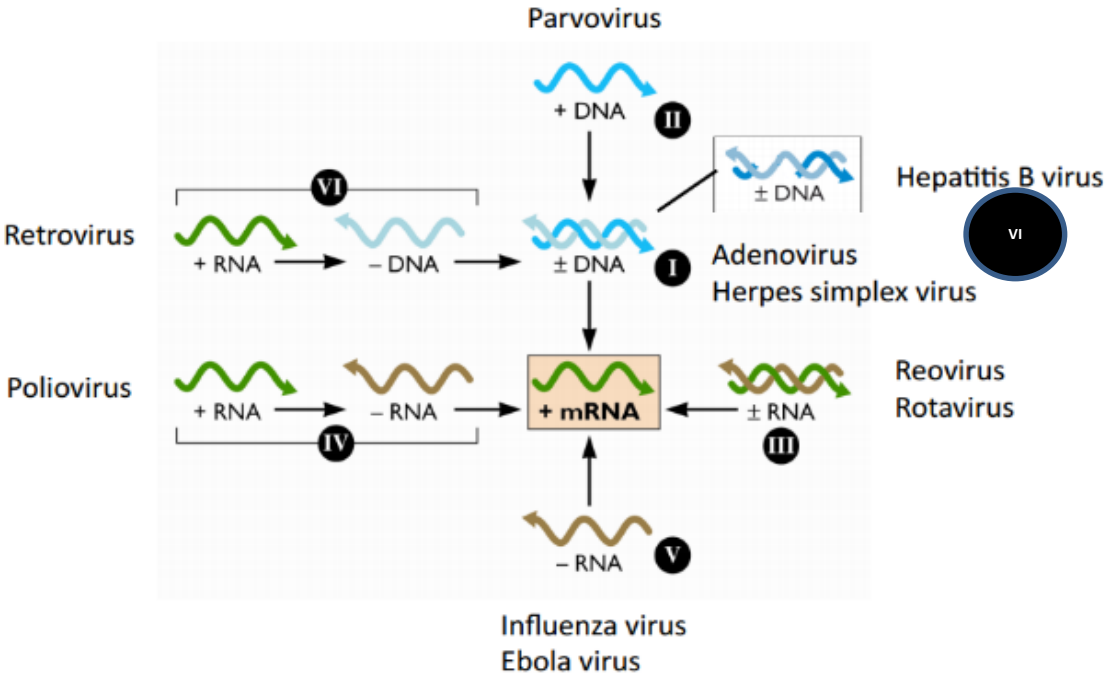


Figure 2.2 The Baltimore classification scheme. Original included I-VI, but now VII includes Hepadnaviruses. From: Racaniello, Vincent. Virology – Biology W3310/4310. Columbia University. Online Course. Spring 2014.

Concept	Principal message	References	Brief critique/comment
Cell degeneration model of virus origin	Viruses, at least complex ones, evolved as a result of degeneration of cells, perhaps, through a stage of intracellular parasites	[40, 43, 45, 50]	This route of virus evolution appears to be inconsistent with the results of viral comparative genomic, in particular, the prominence of genes without cellular counterparts in the conserved cores of viral genomes
Escaped-genes model of virus origin	Viruses evolved from within cells, through autonomization of the appropriate genes, e.g., those coding for polymerases	[40, 43, 45, 55]	Similarly, this model lacks support from virus genome comparison
Origin of viruses from a primordial gene pool	Viruses are direct descendants of primordial genetic elements	[40, 43, 87]	Generally, this appears to be the most plausible path for the origin of viruses. However, non-trivial conceptual development is required, given that viruses are intracellular parasites and, technically, could not precede cells during evolution
An ancient lineage of viruses spanning the three domains of cellular life	The presence of JRC in a variety of groups of DNA viruses is taken as evidence of the existence of an ancient lineage of viruses infecting all three domains of cellular life	[13–15]	This concept capitalizes on a truly remarkable observation of the near ubiquity of JRC in viruses. However, inferring an ancient lineage of viruses on the basis of the conservation of a single protein smacks of essentialism and does little to explain the trajectories of most other virus-specific and virus hallmark genes. Besides, this concept does not specify the cellular context in which the ancient virus lineage might have emerged
Three DNA viruses to replicate genomes of RNA cells	The hypothesis postulates that at least three major lineages of RNA viruses emerged by the escaped-genes route from RNA-based progenitors of archaea, bacteria and eukaryotes. These ancient RNA viruses are thought to have given rise to three independent lineages of DNA viruses that imparted DNA replication onto their cellular hosts	[49, 55]	This concept is based on important general notions of the ancient origin of viruses and their major role in evolution of cells. However, the specific model of Forterre appears to be critically flawed as it stems from a model of cellular evolution that appears not to be defendable (see text)

Figure 2.3 Major concepts in viral evolution. From: Koonin EV, Senkevich TG, Dolja VV. 2006. The ancient Virus World and evolution of cells. Biol Direct 1:29.

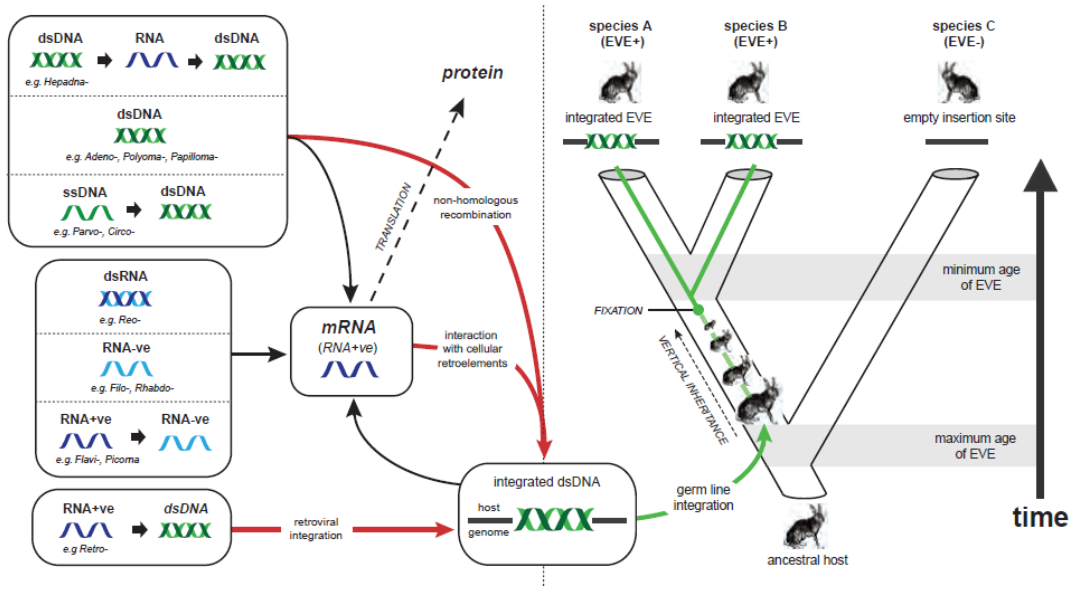


Figure 2.4 How endogenous virus elements are generated and can be used to estimate the age of viral families. From: Katzourakis A, Gifford RJ. 2010.Endogenous viral elements in animal genomes. PLoS Genet 11:e1001191 and Holmes. 2013. Virus Evolution, p 286 – 467. *In* Knipe DM, Howley PM (ed), Fields Virology, 6th ed. Wolters Kluwer Health/Lippincott Williams & Wilkins, Philadelphia, PA.

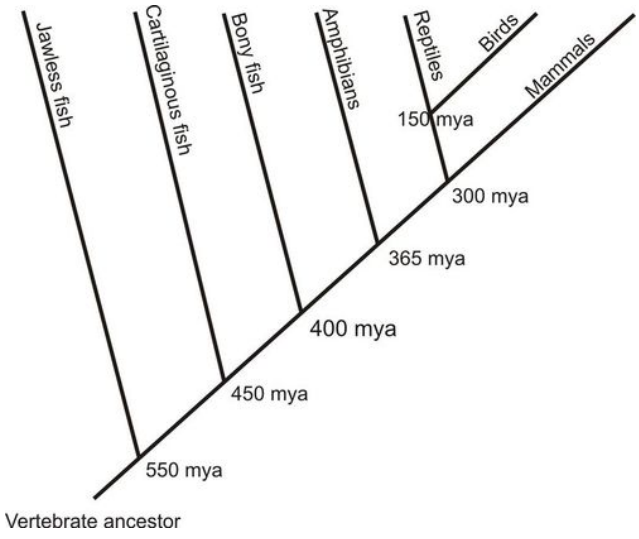


Figure 2.5 This phylogenetic tree gives an overview of vertebrate evolution. The earliest vertebrates were jawless fish that lived between 500 and 600 million years ago. As more data become available, new ideas about vertebrate evolution emerge. From: Wilkin, Douglas et al. Vertebrate Evolution. Online: http://www.ck12.org/book/CK-12 BiologyConcepts/r17/section/ 12.5/Vertebrate-Evolution/

Virus family	Morphology	Size (nm)*	Host
	Double-stranded DNA viruses		
Baculoviridae	Enveloped rods, some with tails	200–450 × 100–400	Crustacea
Corticoviridae, Tectiviridae	Icosahedral with spikes	60–75	Bacteria
Herpesviridae	Pleomorphic, icosahedral, enveloped	150–200	Molluscs, fish, mammals, turtles
Iridoviridae	Round, icosahedral	190–200	Molluscs, fish
Lipothrixviridae	Thick rod with lipid coat	40 × 400	Archaea
Mimiviridae	Icosahedral with microtubule-like projections	400	Protozoa (?)
Myoviridae	Polygonal head (icosahedral) with contractile tail (helical)	50–110 (head)	Bacteria
Nimaviridae	Enveloped, ovoid with tail-like appendage	120 × 275	Crustacea
Papovaviridae	Round, icosahedral	40–50	Molluscs
Phycodnaviridae	Icosahedral	130–200	Algae
Podoviridae, Siphoviridae	Icosahedral with noncontractile tail	60 (head)	Bacteria
	Single stranded DNA viruses		
Microviridae	Icosahedral with spikes	25–27	Bacteria
Parvoviridae	Round, icosahedral	20	Crustacea
	Double stranded RNA viruses		
Birnaviridae	Round, icosahedral	60	Molluscs, fish
Cystoviridae	Icosahedral with lipid coat	60–75	Bacteria
Reoviridae	Icosahedral, some with spikes	50–80	Crustacea (?), molluscs, fish
Totiviridae	Round, icosahedral	30–45	Protozoa
	Single stranded RNA viruses		
Bunyaviridae	Round, enveloped	80–120	Crustacea (?)
Caliciviridae	Round, icosahedral	35–40	Fish, mammals
Coronaviridae	Rod-shaped with projections	200 × 42	Crustacea
Dicistroviridae	Round, icosahedral	30	Crustacea
Leviviridae	Round, icosahedral	26	Bacteria
Marnaviridae	Round, icosahedral	25	Algae
Nodaviridae	Round, icosahedral	30	Fish
Orthomyxoviridae	Round, with spikes	80–120	Fish
Paramyxoviridae	Various, mainly enveloped, filamentous	60–300 × 1000	Mammals
Picornaviridae	Round, icosahedral	27–30	Algae, crustacea (?), thraustochytrids,
Rhabdoviridae	Bullet-shaped with projections	45–100 × 100–430	Fish

*, For rod-shaped viruses, dimensions are shown as diameter × length.

Figure 2.6 Viruses that infect marine organisms have been identified as belonging to a range of virus groups. From: Munn CB. 2006. Viruses as pathogens of marine organisms—from bacteria to whales. J Mar Biol Assoc U.K. 86:453-467.

Virus family	Fishes		Amphibians		Reptiles			
	Sharks	Teleosts	Anurans	Salamanders	Lizards	Snakes	Turtles	Crocodiles
Orthomyxoviridae	−	+	−	−	−	−	−	−
Paramyxoviridae	−	+	−	−	+	+	+	−
Rhabdoviridae	−	+	−	−	+	−	−	−
Bunyaviridae	−	−	−	−	−	−	+[a]	−
Retroviridae	−	+	+	−	−	+	+	−
Coronaviridae	−	+	−	−	−	−	−	−
Calciviridae	−	+	+	−	−	+	−	−
Togaviridae	−	+	−	−	+[a]	+[a]	+[a]	−
Picornaviridae	−	+	−	−	−	+	−	−
Nodaviridae	−	+	−	−	−	−	−	−
Flaviviridae	−	−	−	−	+[a]	+[a]	+[a]	−
Reoviridae	−	+	−	−	+	+	−	−
Birnaviridae	−	+	−	−	−	−	−	−

[a] Arthropod-borne viruses termed "arboviruses."

A.

Virus family	Fishes		Amphibians		Reptiles			
	Sharks	Teleosts	Anurans	Salamanders	Lizards	Snakes	Turtles	Crocodiles
Parvoviridae	−	−	−	−	+	+	−	−
Iridoviridae	−	+	+	+	+	+	+	−
Poxviridae	−	−	+	−	+	−	−	+
Herpesviridae	+	+	+	−	+	+	+	−
Adenoviridae	−	+	+	−	+	+	−	+
Polyomaviridae	−	+	+	−	+	−	+	−

B.

Figure 2.7 A: RNA viruses that infect poikilothermic vertebrates – fishes, amphibians and reptiles. **B**: DNA viruses that infect poikilothermic vertebrates – fishes, amphibians and reptiles. From: Villarreal LP. 2005. Viruses and the Evolution of Life. American Society of Microbiology Press, Washington DC.

Chapter 3

A NEW CLADE OF FISH VIRUSES REVEALS THE EVOLUTION AND RECOMBINATION OF DOUBLE STRANDED DNA VIRUSES IN EARLY VERTEBRATES

Dill JA, Camus AC, Koda S, Subramaniam K, Waltzek T, Wen CM, Koonin EM, Pipas JM, Buck CB, and Ng TFF

To be submitted to Proceedings of the National Academy of Sciences USA

Abstract

Currently there are eight families of double stranded DNA viruses known to infect vertebrates (dsDNA; Baltimore system Group I). The majority of viral evolutionary history is deduced from the molecular phylogeny of viruses found in extant hosts and the increased study of viral biodiversity focusing on environments and potential hosts that have been poorly sampled could reformat the existing theories of the viral origins. An investigation of four recently discovered viruses in fish, the oldest group of living vertebrates, with cross-disciplinary techniques, including culture, transmission electron microscopy, metagenomic analysis, phylogenetic analysis, protein structural modeling, histopathology and *in situ* hybridization revealed the existence of a previously unknown clade of dsDNA viruses, with the tentative proposed name of colossomaviruses. The four genomes exhibit high flexibility in gene organization, but are united by ultrastructure, a complete circular dsDNA genome, a conserved helicase and string of homologous open reading frames (ORFs) that encode structural genes. The colossomaviruses appear to have arisen through recombinant chimerization events involving primitive adenoviruses and polyomaviruses. This surprising new family of viruses ties together a model that reveals distant evolutionary relationship among adenoviruses, papillomaviruses, polyomaviruses and parvoviruses.

Introduction

Viruses are the most abundant source of nucleic acid diversity on Earth and infect all domains of life (Holmes 2013; Tompkins et al. 2015; Nasir et al. 2015; Suttle 2007). While eukaryotes, bacteria, and archaea utilize double stranded (ds) DNA to pass their genetic codes, various viral types encode all forms of nucleic acid forms, including single-stranded (ss) and ds DNA and RNA. Based on the nucleic acid type found in virions and the mechanisms of transcription and replication, viruses are classified into seven types in the Baltimore system (Baltimore 1971). Currently there are eight families of double stranded DNA viruses known to infect vertebrates (dsDNA; Baltimore system Group I). These include super-clades, the *Herpesvirales* and the giant Nucleo-Cytoplasmic Large DNA viruses (NCLDV), as well as three ungrouped viral families, the *Adenoviridae*, *Papillomaviridae* and *Polyomaviridae*. The evolution of the *Herpesvirales* and NCLDV independently trace their origins prior to the emergence of chordates (Munn 2006; McGeoch et al. 2008; Colson et al. 2013;Arslan et al. 2011;Yutin et al. 2013; Andrarade 2014). For example, *Herpesvirales* include *Herpesviridae* and *Alloherpesviridae* that infect vertebrates, as well as *Malacoherpesviridae* that infect bivalve molluscs (Munn 2006; McGeoch et al. 2008). The NCLDV include viruses that infect both vertebrate (*Iridoviridae*, *Asfarvivirdae*, *and Chordopoxvirinae* of the *Poxviridae*) and non-vertebrate (*Mimiviridae*, *Phycodnaviridae*, and the insect *Entomopoxvirinae* of the *Poxviridae*, etc) eukaryotes (Arslan et al. 2011; Moniruzzaman et al. 2014). However, the origin and ancient evolutionary history of the three ungrouped dsDNA viral families, the *Adeno-Papillo-Polyomaviridae*, remains an important unresolved question in viral evolution (Moniruzzaman et al. 2014; Villarreal 2011; Koonin 2009; Koonin et al. 2013; Koonin et al. 2015; Nasir et al. 2015; Yutin et al. 2014).

Viruses do not leave a fossil record. Their evolutionary history is usually deduced from the molecular phylogeny of viruses found in extant hosts, except in the occasional specimens where viruses have been exceptionally preserved in extremely cold or stable environments (Moniruzzaman et al. 2014; Koonin et al. 2015; Nasir et al. 2015; Ng et al. 2014). Collectively, fish represent the oldest group of living vertebrates, with a transitional fossil record dating to the mid-Cambrian/Ordovician periods 600-500 million years ago (Martin 2001, Braasch et al. 2016). Our group has recently reported the first polyomaviruses discovered in fish, suggesting polyomaviruses may have a more ancient vertebrate origin (Peretti et al. 2015; Dill et al. 2016; Buck et al. 2016). In this respect, fish may represent fertile ground for detection of undescribed viral lineages that could reshape our understanding of dsDNA virus evolution and fill in key gaps allowing us to construct a more detailed model of these vertebrate viruses. This investigation revealed the existence of a previously unknown family of viruses, tentatively named colossomavirus that contain uniting features (genome organization, genome size, and virion structure) reminiscent of adenoviruses, papillomaviruses, and polyomaviruses. Using cross-disciplinary techniques, including histopathology, transmission electron microcopy (TEM), sequence-independent metagenomics, and *in situ* hybridization, we investigated the biology of a novel colossomaviruses causing proliferative skin lesions in a giant guitarfish (Camus et al. 2015) and compared features of its genomic composition and organization to that of related viruses in three other fish species. This new family of viruses ties together a model in which colossomavirus genomes appear to have arisen through recombinant chimerization events involving primitive adenoviruses and polyomaviruses, suggesting portions of the genomes of colossomaviruses, adenoviruses and the “-oma” families descended from common viral ancestors.

Materials and Methods

Virus evaluation

A male giant guitarfish (*Rhynchobatus djiddensis*) was identified with erythematous lesions of various sizes distributed over its ventral surface (Figure 3.1A). Histopathological examination of biopsy specimens revealed proliferations of the epidermis with intranuclear inclusions (Figure 3.1B). Microscopic and TEM results were suggestive of an unclassified viral agent (Camus et al. 2015). In order to circumvent the absence of known viral sequences in elasmobranchs, a sequence-independent metagenomic approach was performed to identify any underlying viruses. Deep sequencing of a skin biopsy from the guitarfish revealed the presence of a previously unknown polyomavirus, the sequence of which we have recently reported (accession NC_026244) (Buck et al. 2016; Dill et al. 2016), and a much larger number of reads representing the complete genome of an additional circular DNA virus. The virus is tentatively named guitarfish colossomavirus (GFCV). GfCv was detected by PCR, qPCR and Sanger-confirmed in the lesion, and was not detected in other unaffected elasmobranch tissue. Furthermore, using a GfCv-specific probe targeting the primase region, a strong positive *in situ* hybridization signals localized the virus to the nucleus of affected cells (Figure 3.1C). Cell culture was not attempted due to a lack of suitable cell lines.

In addition, a circular dsDNA genome was isolated from a red discus cichlid (*Symphysodon discus*) at the University of Florida using next-generation sequencing. The virus is tentatively named red discus cichlid colossomavirus (RdCV). Another circular dsDNA virus, tentatively named marbled eel colyomavirus (MeCV; previously called AMPyV), was discovered in an autogenous cell line from marbled eels (*Anguilla marmorata*) in Taiwan (Wen et al. 2015). The genomic and virion properties of the three new viruses, GfCV, RdCV, and MeCV, were

compared and analyzed. Analysis also included a previously reported related taxa, the Japanese eel endothelial cells-infecting virus (JEECV) (GenBank AB543063), which infects Japanese eels (*Anguilla japonica*) (Mizutani et al. 2011; Okazaki et al. 2015).

Metagenomic sequencing and NGS analyses.

Unbiased metagenomic sequencing of the viral particles was performed according to previously described protocols (Ng et al. 2015; Ng et al. 2013; Ng et al. 2012; Victoria et al. 2008), consisting of the following steps: filtration of tissue homogenate to enrich viral particles, depletion of host nucleic acid in filtrate using nucleases, unbiased sequence-independent amplification using random priming. Specifically, nucleic acids from nuclease-resistant viral particles were extracted using the QIAquick viral RNA column purification system (Quiagen). Reverse transcription was performed using a 28-base oligonucleotide whose 3′ end consisted of eight random nucleotides (primer N1_8N, CCTTGAAGGCGGACTGTGAGNNNNNNNN). Second strand was synthesized using Klenow fragment DNA polymerase (New England BioLabs). The resulting double-stranded cDNA and DNA were then PCR amplified using AmpliTaq Gold DNA polymerase and a 20-base primer (primer N1, CCTTGAAGGCGGACTGTGAG). A duel-indexed sequencing library was prepared using the Nextera XT DNA Sample Prep Kit (Illumina, San Diego, CA), and after pooling, the final library was sequence using the MiSeq sequencing system with 2 × 250 bp paired-end sequencing reagents (Illumina MiSeq Reagents V2, 500 cycles).

An in-house analysis pipeline running on a 32-node Linux cluster was used to process the data (University of California San Francisco). A total of 8,119,238 million reads were generated and analyzed as previous described (Ng et al. 2013; Ng et al. 2012). Adaptor and primer sequences were trimmed using VecScreen (McGinnis et al. 2004), while duplicate reads and

low-sequencing-quality tails were removed using a Phred quality score of 10 as the threshold. The cleaned reads were de novo assembled using an in-house sequence assembler employing an ensemble strategy (Deng et al. 2015) consisting of SOAPdenovo2, ABySS, meta-Velvet, and CAP3. The assembled sequences were compared with an in-house viral proteome database using BLASTx. Once the viral contigs are identified, they were further assembled iteratively to obtain the complete circular genome of GfCv.

Molecular screening of guitarfish

Skin lesions were sampled at the time of original diagnosis, 10 weeks later, and following their resolution after 25 weeks. GfCv specific primers were designed from the genome, targeting major genes, for molecular screening and probe construction for *in situ* hybridization (Table 3.1). Total nucleic acids were extracted from tissue using QIAamp Viral RNA Mini Kit (Qiagen). PCR was performed using One Taq DNA Polymerase kits (New England Biolabs, Ipswich, MA), with a touch-down thermocycling condition previously described (Ng et al. 2013). The PCR cycle consisted of an initial denaturation at 94°C for 30 s, followed by 45 cycles of 94°C for 30 s, 58°C for 30 s with a -2°C touchdown each cycle, and 68°C for 120 s, with a final elongation at 68 °C for 5 min. The PCR products were analyzed using 1.5% agarose gel electrophoresis and the resulting amplicons were verified by Sanger sequencing.

Quantitative (q)PCR was used to assess the presence and quantity of viral DNA. Primers were designed from the LO7 gene (Table 3.1). In order to obtain amplicon standards, the primer set was used in a standard PCR reaction with DNA extracted from the guitarfish. The DNA was run on a 2% agarose gel, purified (Qiaquick Gel Extraction Kit) and quantitated (NanoDrop 2000, Thermo Fisher). DNA was adjusted to 1 ng/µl. Ten-fold dilutions of this stock were made

in water for qPCR standard curve generation. Preliminary analysis indicated that the 10^{-1} through 10^{-8} dilutions (10^{-1} -10^{-8} ng) would cover the dynamic (linear) range of the assay ($R^2 \geq 0.95$). qPCRs for the tissue samples and standards were performed on a Bio-Rad IQ5 iCycler using iQ5 system software for analysis. One µl of extracted DNA was added to each 25 µl reaction mix containing iQ SYBR Green Supermix (Bio-Rad) and 100 nmol each of the indicated primers. A 2-step cycling program was used as follows: an initial 95°C for 3 min followed by 35 cycles of 95°C for 10 seconds and 60°C for 30 seconds. Initial screening of all samples was done twice using one PCR well/sample. Final assessment of viral DNA presence was made on samples run in triplicate.

Fluorescent *in situ* hybridization

In situ hybridization assays using a digoxigenin-labeled probe and a biotin-streptavidin method were adapted from previously described protocols (Dawson et al. 2001; Tate et al. 2013). Primers specific for GfCv, designed from the genome, were used to generate a digoxigenin-labeled PCR probe. Briefly, 3-um sections of formalin-fixed paraffin-embedded skin were deparaffinized and rehydrated. Tissue proteases were digested in Ready-to-Use Proteinase K (DAKO) for 15 minutes. Slides were placed in a BioRad Frame-Seal Incubation chamber and denatured with 100% formamaide (Sigma) at 105° C for 5 minutes. The probe was diluted in molecular grade water and previously prepared hybridization solution, applied to slides, denatured at 105°C for 5 minutes and then left to hybridize overnight in a 37°C humidified oven. On day 2, slides were washed with 5X sodium chloride-sodium citrate buffer (SSC) at room temp followed by 2X SSC incubated at 37°C. Sections were blocked with universal blocking buffer (BioGenex). Mouse anti-digoxigenin antibodies (ROCHE) diluted 1:500 with Antibody Diluent (DAKO) were applied to the sections for 60 minutes and detected by serial application of

goat anti-mouse biotinylated immunoglobulins (Biogenex), streptavidin alkaline phosphatase (Biogenex) and napthol fast red substrate (DAKO) and mounted with aqueous mounting adhesive. All steps were carried out at room temperature unless otherwise noted. For fluorescent in situ hybridization normal goat serum (Rockland) was used as a block, Streptavidin-Alexa Fluor® 532 as a conjugate and ProLong Gold with DAPI as a mountant (Molecular Probes).

Structural modeling

As a first step to obtain models for colossomavirus helicase, the sequence was analyzed using the PsiPred website to determine whether reliable templates from solved structures could be identified. Small regions of guitarfish colossomavirus helicase (525-700aa) and parvovirus NS1 (roughly 200-400aa) share homology with ATPase domains of SF3 helicases, including papillomavirus E1, AAV2 rep40 and SV40 large T antigen. Based on fold domain recognition, the guitar fish colossomavirus helicase seems more distant from SV40 large T, but both AAV2 rep 40 and papillomavirus E1 scored almost equally well. The models were then generated using papillomavirus E1 as the template. Next, a surface conservation map was generated using multiple alignment of 500 homologous proteins including papillomavirus E1 helicases, polyomavirus LTs and parvovirus NS1 proteins. A conservation score was calculated for each alignment residue, and then mapped onto the solved structure of HPV18 E1.

Sequence comparisons and phylogenetic analysis

Coding sequences of representative helicase genes and late open reading frames (LOs) were downloaded from GenBank that were of sufficient length to conduct phylogenetic analyses. Amino acid sequence alignment of data sets were inferred using multiple cycles of the MUSCLE algorithm (Edgar 2004). Based on these alignments, maximum likelihood (ML) phylogenetic

trees were estimated. Finally, pairwise sequence similarities were calculated using the translated amino acid sequences with the Sequence Demarcation Tool (Muhire et al. 2014).

Results

Characterization of colossomaviruses

Using metagenomic sequencing, the 21,527 bp complete genome of of a dsDNA viurs, named giant guitarfish colossomavirus (GfCv) was identified from a skin lesion with intranuclear inclusions in the giant guitarfish (*Rhynchobatus djiddensis*) (Figure 3.2). GfCv nucleic acid was detected by conventional and real time PCR and Sanger-confirmed in the lesion, and, using a GfCv-specific probe, strong positive *in situ* hybridization signals localized GfCv DNA to intranuclear inclusions of affected cells (Figure 3.1). Intranuclear hexagonal, average 75 nm virions were identified with TEM (Camus et al. 2015) (Table 3.2, Figure 3.2).

RdCV was identified from a skin lesion in a red discus cichlid (*Symphysodon discus*) using next-generation sequencing. The complete 19,275 bp genome was used for further comparisons. There was no evidence of intranuclear inclusions on histopathology and cell culture was unsuccessful.

MeCV was isolated from marbled eels (*Anguilla marmorata*) with hemorrhage and congestion throughout the body and gills, using a cell line established from the pectoral fin of marbled eels (Wen et al. 2015). MeCV nucleic acids were PCR confirmed from cultured cells, and the full MeCV genome was obtained by analyzing transcriptome data from the cell line. Reads assembled to MeCV did not constitute any poly-A sequences, suggesting MeCV, a DNA virus, was co-sequenced during the transcriptome experiment. The complete genome was reanalyzed and a final genome of 16,930 bp was obtained. Icosahedral ~75 nm virions were identified on electron microscopy of the cell culture (Wen et al. 2015).

JEECV causes viral endothelial cell necrosis of eel (VECNE) disease in Japanese eels (*Anguilla japonica*) (Mizutani et al. 2011; Okazaki et al. 2015). Seventy-five nm virions were identified on TEM and nucleic acids have been confirmed in both Japanese eel autogenous cell culture, as well as fish tissue from natural habitats (Mizutani et al. 2011; Okazaki et al. 2015). While the colossomavirus infections have primarily been identified in external epithelial cells (skin, GfCV and RdCV; gill, MeCV and JEECV), the viruses, particularly in eels, have been reported to proliferate in diverse cell types (Wen et al. 201; Mizutani et al. 2011; Okazaki et al. 2015).

Sequencing, cell culture, in situ hybridization, and TEM results, along with previous findings (Wen et al. 201; Mizutani et al. 2011; Okazaki et al. 2015; Camus et al. 2015), are evidence that these four colossomaviruses are fish-infecting viruses that share similar morphologic properties and genome size (Table 3.2).

Genome modularity of colossomaviruses

The complete novel genomes of GfCV, RdCV and MeCV were analyzed together with the previously published JEECV genome. These four circular genomes range from 16,930 to 21,527 base pairs (Table 3.2), distinguishing them from other dsDNA viral families, including papillomaviruses (7.0 – 8.6 kb; circular), polyomaviruses (3.9 – 7.3 kb; circular), and adenovirus (26 – 45 kb; linear). These genomes contain a GC content of 44%-49%, and share no significant nucleotide homology to any known viral sequences (BLASTn).

Each of the colossomaviruses encode two cassettes of bi-directionally transcribed genes, the non-structural early open reading frames (EO) EO1-5 and structural late open reading frames (LO) LO1-8 (Fig 3.3, Table 3.3). As a non-structural gene, EO1 likely functions as helicase and

is the most conserved gene among colossomaviruses, with inter-taxa protein identities between 22-28% (Table 3.4). The EO1 genes share distant protein homology (<35% by local alignment/ BLASTp) to other helicase genes including the E1 proteins of papillomaviruses, LT proteins of polyomaviruses, and NS proteins of parvoviruses. Within the EO1 gene of all colossomaviruses, like other DNA viruses, the conserved helicase domain superfamily 3 of DNA viruses was recognized (SF3 helicase; Prosite accession PS51206). Additionally, in the EO1 helicase of JEECV and MeCV, a DNAJ domain, a unique hallmark of polyomaviruses, was identified.

EO2-5 genes showed the most plasticity in genome arrangement among the different colossomaviruses (Figure 3.3) and vary by strand position, direction and number of open reading frames (ORFs). In MeCV, GfCV and RdCV, they are on the opposite strands to the EO1, while in JEECV, they are on the same strand. EO2-4 remained as separate ORFs in MeCV, RdCV, and JEECV, but are concatenated into a single ORF in GfCV. In GfCV, this EO2-5 encodes a >300 kD protein with an N-terminal domain showing a high degree of similarity to the catalytic subunit of archael-eukaryotic DNA primases (AEPs) (Figure 3.4). Although similarity between the colossomavirus AEP domain and an AEP domain-containing protein of baculoviruses, late expression factor 1 (LEF-1), could be detected in Blast searches restricted to viruses, neither LEF-1 nor any other virally-encoded proteins appeared among the top few thousand hits in unrestricted searches. This GfCv primase contains an LXCXE motif that is found in all polyomavirus large T antigens and bind directly to the retinoblastoma family of tumor suppressor proteins including pRb.

All viruses share a 7kb tandem array of homologous LO4 -LO8 genes. LO1-8 are a cassette of viral genes that are usually transcribed antisense to the EO1 helicase gene (Figure 3.3). LO1-LO3 are only detected in the eel viruses, JEECV and MeCV, but LO2-3 are fused in

MeCv, and remain as individual ORFs in JEECV. The functions of LO1-LO3 are unknown, as they share no recognizable nucleotide or protein identities to any entry in GenBank. Instead of LO1-3, a small ORF with similarity to cellular Su(var)3-9, Enhancer-of-zeste, Trithorax (SET) proteins, is found in GfCV and RdCV (Figure 3.3). This GfCV SET gene shares recognizable homologs to histone-lysine N-methyltransferase proteins from a wide range of metazoans. To our knowledge, no other viruses have incorporated a SET gene in their genome.

It was suspected that the LO arrays encode for structural proteins that make up the viral capsid and distant homology searches to link these to capsid proteins of adenoviruses. LO4 contains a coiled-coil structure similar to that of adenovirus protein IX (pIX) and may act as a cement joining the hexon major capsid proteins together to make up the facets of the icosahedral virion. LO5 shares amino acid homology with adenovirus penton protein and therefore, might have structural function resembling penton protein. LO6 resembles that of adenovirus protein Mu (pX) and protein VI (pVI), so it may carry a membrane-destabilizing peptide important for infectious entry. LO7 shares distant protein homology with the adenovirus major capsid protein (MCP) hexon. Adenovirus MCP is composed of two separate domains that make up a pair of 8-stranded jellyroll folds, while polyomaviruses and papillomaviruses MCPs (L1 and VP1 correspondingly) contain only a single beta-jellyroll (Kim et al. 2012; Holmes 2011). The colossomavirus LO7 gene is half the size of the adenovirus hexon and only contains a single beta-jellyroll. Therefore, colossomavirus LO7 shares protein homology with adenoviruses, but shares structural homology with papilloma and polyomaviruses. Finally, LO8 shares distant protein homology with adenovirus cysteine protease and therefore likely facilitates proteolytic cleavage of cytokines during capsid maturation.

Structural protein modeling

Computer-based structural models detected striking overall resemblance of GfCv helicase with those from parvoviruses and papillomaviruses (Figure 3.5). The Walker A motif associated with phosphate binding, and the Walker B motif, were detected, with conserved residues identified in all viruses. The agreement between the models and the template, as well as the presence of conserved Walker motifs within the guitarfish helicase protein, indicates that GfCv helicase contains a functional ATPase domain. Moreover, the surface conservation analysis maps showed that surface residues involved in the core functions of the helicase (hexamerization, ATP binding and hydrolysis) are highly conserved among the three groups of viruses (Figure 3.6A). In contrast, residues mediating E2 binding, which is specific for papillomavirus E1, are much less conserved (Figure 3.6B). Altogether, these analyses support the conclusion that GfCv encodes a DNA helicase homolog related to those of parvoviruses and polyomaviruses and that these families of proteins share a common functioning DNA helicase.

Phylogenetic analysis

Phylogenetic analysis based on this helicase domain revealed the GfCV and RdCV constitutes a novel, family-level clade with equal distance to members of the *Papillomaviridae* and *Polyomaviridae* (Figure 3.7). The helicase domain of JEECV and MeCV, on the other hand, groups within polyomaviriuses. Concordantly, the DNAJ domain, a unique hallmark of polyomaviruses (Sheng et al. 1997), is recognized in JEECV and MeCV.

Discussion

A paucity of shared genes between all viral groups effectively rules out viral origin from a common viral ancestor. However, the discovery of giant or NCLDV viruses and analysis of their evolutionary relationships has put us closer to having a coherent scenario for the evolution of most dsDNA virus groups found in eukaryotes (Yutin et al. 2014; Koonin et al. 2015: Moniruzzaman et al. 2014). The proposed supergroup, "*Megavirales*," which includes adenoviruses and NCLDV viruses of amoebae and algae, have very large genomes and share a number of key sequence and structural features (Koonin 2009; Koonin et al. 2013; Koonin et al. 2014). Evidence of approximately 50 shared ancestral genes, include those encoding essential structural and non-structural viral functions, such as capsid proteins of the icosahedral viruses, DNA and RNA polymerases, distinct helicases, and integrases provide strong evidence that these giant viruses share common ancestry (Yutin et al. 2014; Koonin et al. 2015; Koonin et al. 2013; Iyer et al. 2001; Iyer et al. 2006; Koonin et al. 2001).

Small DNA viruses, including the dsDNA *Polyomaviridae* and *Papillomaviridae* ("oma" viruses), as well as the ssDNA *Parvoviridae* are united by the conserved fold of an ancestral helicase protein (Koonin 2009; Koonin et al. 2013; Koonin et al. 2014). Based on conservation among their helicases, including LT and E1 proteins, it has been suggested that these viruses share a separate origin from the larger dsDNA viruses of eukaryotes (Koonin et al. 2015; Krupovic 2013). Until now, there has been no evidence to link ancestry between these two groups of large and small DNA viruses. The discovery of these novel, highly divergent DNA viruses in lower vertebrates may be central to developing an evolutionary model that accounts for interrelationships of all dsDNA viruses.

Together, GfCv, RdCv, JEECV and MeCv represent a new clade of dsDNA viruses. Additionally, GfCv represents a prototype virus of a distinct DNA virus family tentatively named *Colossomaviridae*. Genomes are highly novel and support their grouping in an isolated family separate from the *Polyomaviridae*. First, the genome sizes are strikingly large compared to other dsDNA viruses with a circular genome. The 21,527 kp genome of GfCv is much larger than the 7-8k bp genomes of mammalian and avian papillomas and the 4-5k bp genomes of polyomaviruses. In contrast, it is smaller than all *Megavirales* genomes and the linear, 26 – 45 kb genomes of adenoviruses. The only viruses that share similar genome size and gene organization are the related viruses discussed in this paper, including the 19,275 bp, 16,930 bp and the 15,131 bp genomes of RdCV, MeCV, and JEECV, respectively (Wen et al. 201; Mizutani et al. 2011; Okazaki et al. 2015). Similarly, GfCV, RdCV, MeCV, and JEECV all produce viral particles of ~ 75 nm, with icosahedral symmetry and a hexagonal face (Wen et al. 201; Mizutani et al. 2011; Okazaki et al. 2015; Camus et al. 2015) (Figure 3.2, Table 3.2). Together these morphologies are distinct from the knobby 45-60 nm virions of polyomaviruses and papillomaviruses, the 80 -100 nm virions of adenoviruses, and bear no resemblance the 120 nm virions of herpesviruses (Cheville 1994).

Colossomavirus genome organization and its gene products are distinct from established DNA viral families (Figure 3.3). The existence of various genes fundamental to virus replication and structure that are shared by a variety of virus groups, but are missing from cellular genomes, suggests these features were inherited from a common viral ancestor. Identified structural and non-structural genes suggest ancient distant recombinant chimerization events involving both primitive adenoviruses and polyomaviruses.

The most significant protein conservation was in the helicase gene, where protein homology was found to be related to parvoviruses, papillomaviruses, and polyomaviruses, signatures preserved in the architectural repertoire (Kim et al. 2012, Bamford et al. 2005). Even though the helicase displayed only limited protein sequence homology, structural modeling of the GfCv showed marked resemblance with other helicases (Figure 3.5, 3.6). A unifying structure among GfCv helicase, parvovirus N1 and papillomavirus E1, is strongly suggestive of the ancient common ancestry of these helicase genes. The model postulates the existence an ancient "proto-oma" that supplied the helicase to colossomaviruses and other small DNA viruses (Figure 3.8). However, JEECV and marbled eel virus appear to encode a captured genuine polyomavirus LT in place of the GfCv "proto-oma" provided helicase. This is also supported by the presence of the classic DNAJ domain within both eel virus LTs (Figure 3.7).

The four colossomaviruses also share a 7kb tandem array of homologous open reading frames designated late genes LO4-8. Some of these ORFs have been predicted to represent structural elements with integral functions responsible for icosahedral virion construction (Krupovič et al. 2010). These structural proteins of the colossomaviruses share distant homologies with those of adenoviruses leading us to hypotheize the existence of an ancient proto-adenoviruses lineage that contributed to the structure of colossomavirus. The model postulates the existence an ancient "proto-adeno" that is the origin of the structural (capsid) genes of the colossamaviruses (Figure 3.8).

This report describes four prototype dsDNA viruses that infect fish, including the first lesion associated virus of elasmobranchs to be fully characterized by morphologic features, as well as cross disciplinary techniques including molecular techniques, phylogenetic analysis, and protein structural modeling. These novel viruses provide clues as to how this previously

unknown clade, tentatively named colossomavirus, evolved and suggest distant recombinant chimerization events involving primitive adenoviruses and polyomaviruses that could have shaped the evolution of multiple dsDNA viral lineages. The timing of the apparent inter-familial recombination events that gave rise to the chimeric colossomavirus lineages is unclear. However, the very high degree of divergence between both primary nucleic acid sequence and the helicase domains, suggests that the recombination events occurred in the very ancient past.

References

Arslan D, Legendre M, Seltzer V, Abergel C, Claverie JM. 2011. Distant Mimivirus relative with a larger genome highlights the fundamental features of Megaviridae. Proc Natl Acad Sci 108:17486-91.

Baltimore D. 1971. Expression of animal virus genomes. Bacteriol Rev 35:235

Bamford DH, Grimes JM, Stuart DI. 2005. What does structure tell us about virus evolution? Curr Opin Struct Biol 15:655-663.

Braasch I, Gehrke AR, Smith JJ, Kawasaki K, Manousaki T, Pasquier J, Amores A, Desvignes T, Batzel P, Catchen J, Berlin AM. 2016. The spotted gar genome illuminates vertebrate evolution and facilitates human-teleost comparisons. Nat Genet 48:427-437.

Buck CB, Van Doorslaer K, Peretti A, Geoghegan EM, Tisza MJ, An P, Katz JP, Pipas JM, McBride AA, Camus AC, McDermott AJ, Dill JA, Delwart E, Ng TFF, Farkas K, Varsani A. 2016. The Ancient Evolutionary History of Polyomaviruses. Plos Pathog 12:e1005574.

Camus A, Dill J, McDermott A, Camus M, Fan NT. 2016. Virus-associated papillomatous skin lesions in a giant guitarfish Rhynchobatus djiddensis: a case report. Dis Aquat Organ 117:253-258.

Cheville NF. 1994. Cytopathology of viral disease p 490-615. In Ultrastructural pathology, an introduction to interpretation. Iowa State University Press, Ames, Iowa.

Colson P, De Lamballerie X, Yutin N, Asgari S, Bigot Y, Bideshi DK, Cheng XW, Federici BA, Van Etten JL, Koonin EV, La Scola B. 2013. "Megavirales", a proposed new order for eukaryotic nucleocytoplasmic large DNA viruses. Arch Virol 158:2517-2521.

Dawson JE, Paddock CD, Warner CK, Greer PW, Bartlett JH, Ewing SA, Munderloh UG, Zaki SR. 2001. Tissue diagnosis of Ehrlichia chaffeensis in patients with fatal ehrlichiosis by use of immunohistochemistry, *in situ* hybridization, and polymerase chain reaction. Am J Trop Med Hyg 65:603-609.

Deng X, Naccache SN, Ng T, Federman S, Li L, Chiu CY, Delwart EL. 2015. An ensemble strategy that significantly improves de novo assembly of microbial genomes from metagenomic next-generation sequencing data. Nucleic Acids Res 43:e46.
GfPV Genome Announc

Grogan ED, Lund KM, Greenfest-Allen E. 2012. The origins and relationships of early chondrichthyans p 3-30. In Carrier JC, Musick JA, Heithaus MR (ed), Biology of sharks and their relatives. CRC press; Boca Raton, FL.

Holmes EC. 2013. Virus Evolution, p 286 – 313. In Knipe DM, Howley PM (ed), Fields Virology, 6th ed. Wolters Kluwer Health/Lippincott Williams & Wilkins, Philadelphia, PA.

Holmes, EC. 2001. What does virus evolution tell us about virus origins? J Virol 85: 5247-5251.

Iyer LM, Aravind L, Koonin EV. 2001. Common origin of four diverse families of large eukaryotic DNA viruses. J Virol 75:11720-11734.

Iyer LM, Balaji S, Koonin EV, Aravind L. 2006. Evolutionary genomics of nucleo-cytoplasmic large DNA viruses. Virus Res 117:156-184.

Kim KM, Caetano-Anollés G. 2012. The evolutionary history of protein fold families and proteomes confirms that the archaeal ancestor is more ancient than the ancestors of other superkingdoms. BMC Evol Biol 12:1.

Koonin EV, Dolja VV. 2013. A virocentric perspective on the evolution of life. Curr Opin Virol 3:546-557.

Koonin EV, Dolja VV. 2014. Virus world as an evolutionary network of viruses and capsidless selfish elements. Microbiol Mol Biol Rev 78:278-303.

Koonin EV, Krupovic M, Yutin N. 2015. Evolution of double-stranded DNA viruses of eukaryotes: from bacteriophages to transposons to giant viruses. Ann N Y Acad Sci 1341:10-24.

Koonin EV, Yutin N. 2001. Origin and evolution of eukaryotic large nucleo-cytoplasmic DNA viruses. Intervirology 53:284-292.

Koonin, EV. 2009. On the origin of cells and viruses: primordial virus world scenario. Ann N Y Acad Sci 1178:47-64.

Krupovič M, Bamford DH. 2010. Order to the viral universe. J Virol 84:12476-12479.

Krupovic M. 2013. Networks of evolutionary interactions underlying the polyphyletic origin of ssDNA viruses. Curr Opin Virol 3:578-586.

Martin A. 2001. The phylogenetic placement of Chondrichthyes: inferences from analysis of multiple genes and implications for comparative studies. Genetica 111:349-357.

Martin A. 2012. The phylogenetic placement of Chondrichthyes: inferences from analysis of multiple genes and implications for comparative studies. Genetica 111:349-357.

McGeoch DJ, Davison AJ, Dolan A, Gatherer D, Sevilla-Reyes EE. 2008. Molecular evolution of the Herpesvirales. Origin and evolution of viruses 23:447-475.

McGinnis S, Madden TL. 2004. BLAST: at the core of a powerful and diverse set of sequence analysis tools. Nucleic Acids Res 32: W20-5.

Mizutani T, Sayama Y, Nakanishi A, Ochiai H, Sakai K, Wakabayashi K, Tanaka N, Miura E, Oba M, Kurane I, Saijo M. 2011. Novel DNA virus isolated from samples showing endothelial cell necrosis in the Japanese eel, Anguilla japonica. Virology 412:179-187.

Moniruzzaman M, LeCleir GR, Brown CM, Gobler CJ, Bidle KD, Wilson WH, Wilhelm SW. 2014. Genome of brown tide virus (AaV), the little giant of the Megaviridae, elucidates NCLDV genome expansion and host–virus coevolution. Virology 466:60-70.

Munn CB. 2006. Viruses as pathogens of marine organisms—from bacteria to whales. J Mar Biol Assoc U.K. 86:453-467.

Nasir A, Forterre P, Kim KM, Caetano-Anollés G. 2015. The distribution and impact of viral lineages in domains of life. Recent Discoveries in Evolutionary and Genomic Microbiology 4:26.

Nasir A, Sun FJ, Kim KM, Caetano-Anollés G. 2015. Untangling the origin of viruses and their impact on cellular evolution. Ann N Y Acad Sci 1341:61-74.

Ng TF, Chen LF, Zhou Y, Shapiro B, Stiller M, Heintzman PD, Varsani A, Kondov NO, Wong W, Deng X, Andrews TD. 2014. Preservation of viral genomes in 700-y-old caribou feces from a subarctic ice patch. Proc Natl Acad Sci 111:16842-16847.

Ng TF, Driscoll C, Carlos MP, Prioleau A, Schmieder R, Dwivedi B, Wong J, Cha Y, Head S, Breitbart M, Delwart E. 2013. Distinct lineage of vesiculovirus from big brown bats, United States. Emerg Infect Dis 19:1978-80.

Ng TF, Kondov NO, Deng X, Van Eenennaam A, Neibergs HL, Delwart E. 2015. A metagenomics and case-control study to identify viruses associated with bovine respiratory disease. J Virol 89:5340-5349.

Ng TF, Marine R, Wang C, Simmonds P, Kapusinszky B, Bodhidatta L, Oderinde BS, Wommack KE, Delwart E. 2012. High variety of known and new RNA and DNA viruses of diverse origins in untreated sewage. J Virol 86:12161-75.

Okazaki S, Manabe H, Omatsu T, Tsuchiaka S, Yamamoto T, Chow S, Shibuno T, Watanabe K, Ono S, Kuwada H, Mizutani T. 2015. Detection of Japanese eel endothelial cells-infecting virus (JEECV) in the Japanese eel Anguilla japonica (Temminck & Schlegel), living in natural habitats. J Fish Dis 38:849-852.

Peretti A, FitzGerald PC, Bliskovsky V, Pastrana DV, Buck CB. 2015. Genome sequence of a fish-associated polyomavirus, black sea bass (Centropristis striata) polyomavirus 1. Genome Announc 3:e01476-14.

Sheng QI, Denis DE, Ratnofsky MA, Roberts TM, DeCaprio JA, Schaffhausen BR. 1997. The DnaJ domain of polyomavirus large T antigen is required to regulate Rb family tumor suppressor function. J Virol 71:9410-9416.

Suttle CA. 2007. Marine viruses—major players in the global ecosystem. Nat Rev Microbiol 5:801-812.

Tate CM, Howerth EW, Mead DG, Dugan VG, Luttrell MP, Sahora AI, Munderloh UG, Davidson WR, Yabsley MJ. 2013. *Anaplasma odocoilei* sp. nov.(family Anaplasmataceae) from white-tailed deer (*Odocoileus virginianus*). Ticks Tick Borne Dis 4:110-9.

Tompkins DM, Carver S, Jones ME, Krkošek M, Skerratt LF. 2015. Emerging infectious diseases of wildlife: a critical perspective. Trends Parasitol 31:149-159.

Vellinga J, van den Wollenberg DJ, van der Heijdt S, Rabelink MJ, Hoeben RC. 2005. The coiled-coil domain of the adenovirus type 5 protein IX is dispensable for capsid incorporation and thermostability. J Virol 79:3206-3210.

Victoria JG, Kapoor A, Dupuis K, Schnurr DP, Delwart EL. 2008. Rapid identification of known and new RNA viruses from animal tissues. PLoS Pathog 4:e1000163.
Villarreal, LP. 2005. Viruses and the Evolution of LIfe. ASM Press, Washington D.C.

Wang M, Yafremava LS, Caetano-Anollés D, Mittenthal JE, Caetano-Anollés G. 2007. Reductive evolution of architectural repertoires in proteomes and the birth of the tripartite world. Genome Res 17:1572-1585.

Wen CM, Chen MM, Wang CS, Liu PC, Nan FH. 2015. Isolation of a novel polyomavirus, related to Japanese eel endothelial cell-infecting virus, from marbled eels, Anguilla marmorata (Quoy & Gaimard). J Fish Dis: doi:10.1111/jfd.12423.

Yutin N, Colson P, Raoult D, Koonin EV. 2013. Mimiviridae: clusters of orthologous genes, reconstruction of gene repertoire evolution and proposed expansion of the giant virus family. Virol J 10:1.

Yutin N, Wolf YI, Koonin EV. 2014. Origin of giant viruses from smaller DNA viruses not from a fourth domain of cellular life. Virology 466:38-52.

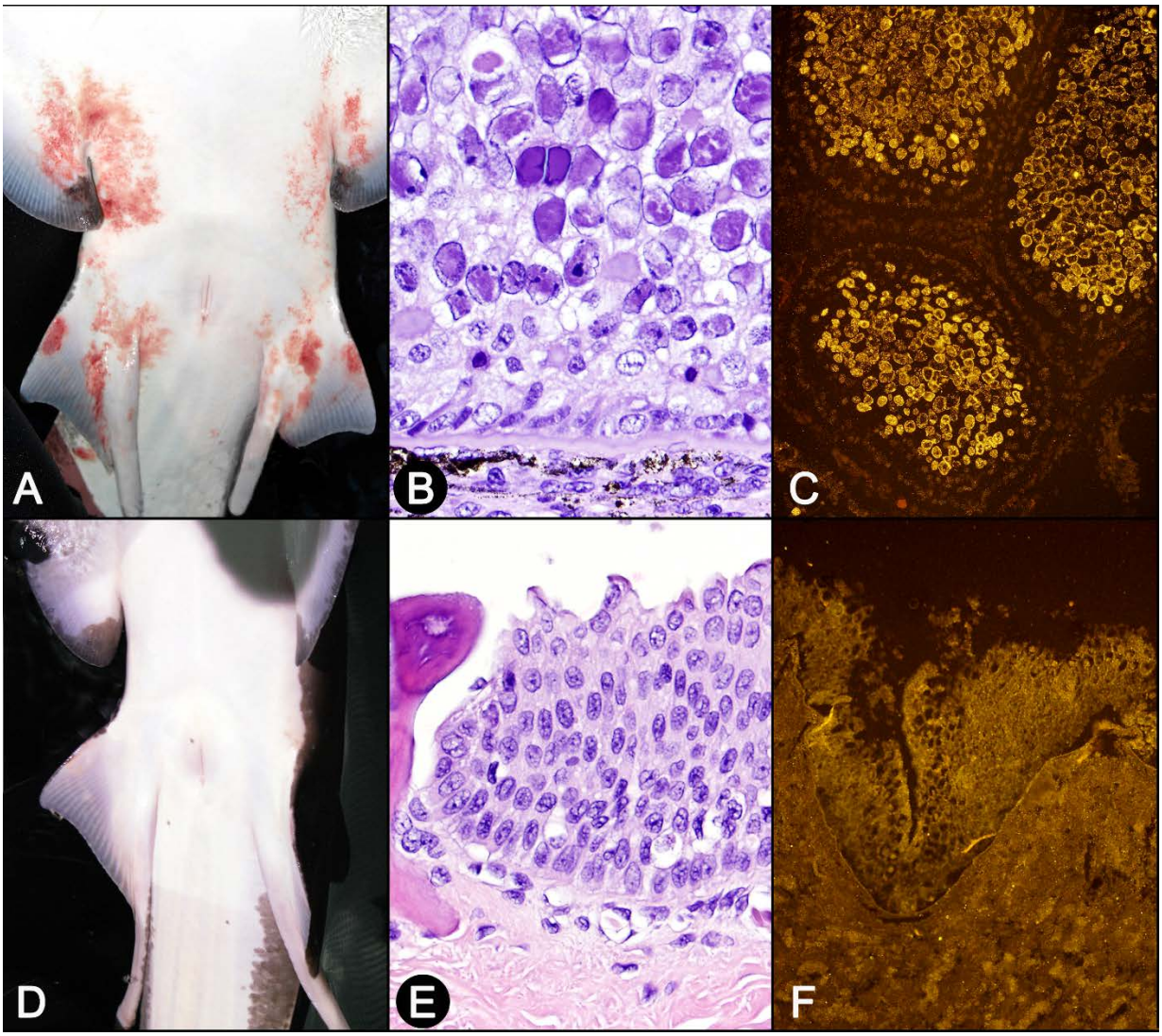


Figure 3.1 Giant guitarfish skin lesion pathology and in situ hybridization findings of active and (A-C) resolved lesions (D-F). A) Gross lesions on ventrum are raised, pink and filiform, involving the pectoral fins, pelvic fins and claspers. B) Photomicrograph of the lesioned skin. The epithelium is hyperplastic and disorganized with widespread cytokaryomegaly. Many nuclei contained hyaline, amphophilic inclusions that fill the nucleus and marginate chromatin. Saliar epithelial cells are unaffected. C) Strong positive fluorescent *in situ* hybridization signals localized to the nucleus of affected epithelial cells. There is an absence of hybridization signal in basilar cells and connective tissue. E) Ventral skin following lesion resolution is white and smooth. F) Photomicrograph of normal skin characterized by organized layers of uniformly sized epithelial, free of inclusion bodies, resting on a basement membrane. A dermal denticle is visible on the left. F) Fluorescent *in situ* hybridization produced no signal in normal skin.

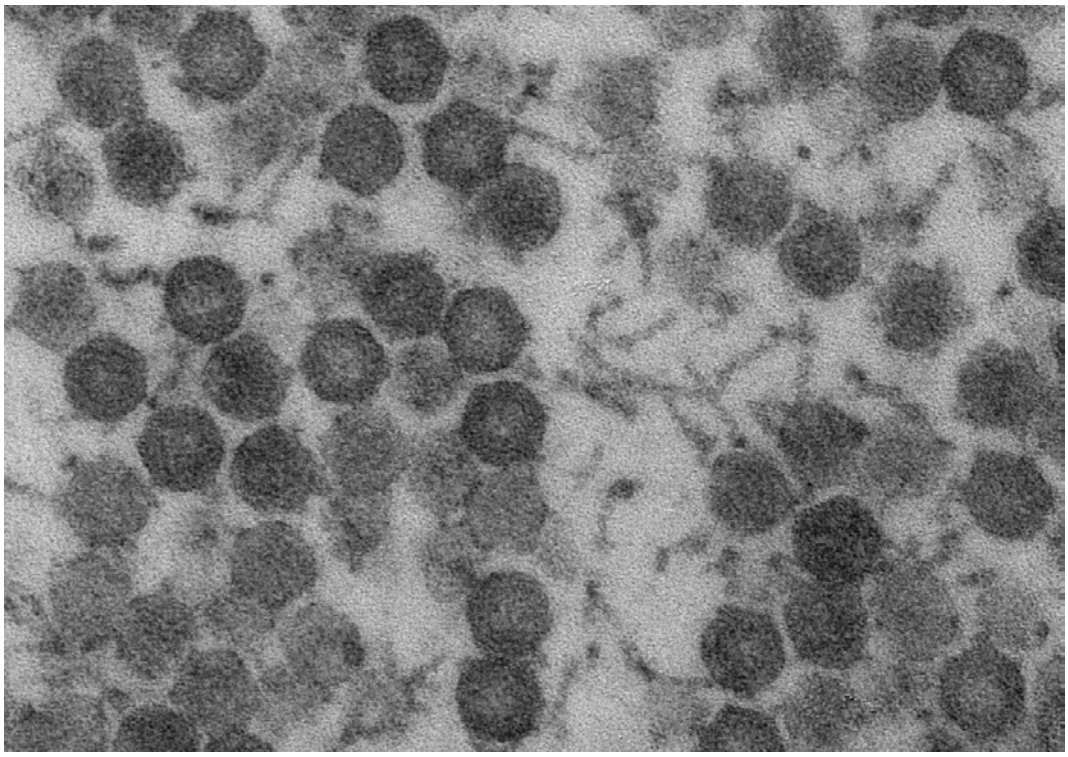


Figure 3.2 Transmission electron microscopy of guitarfish colossomavirus. Hexogonal viral particles ~70 -75nm from the guitarfish skin lesions. Virions were in arrays limited to the nucleus, with no evidence of budding.

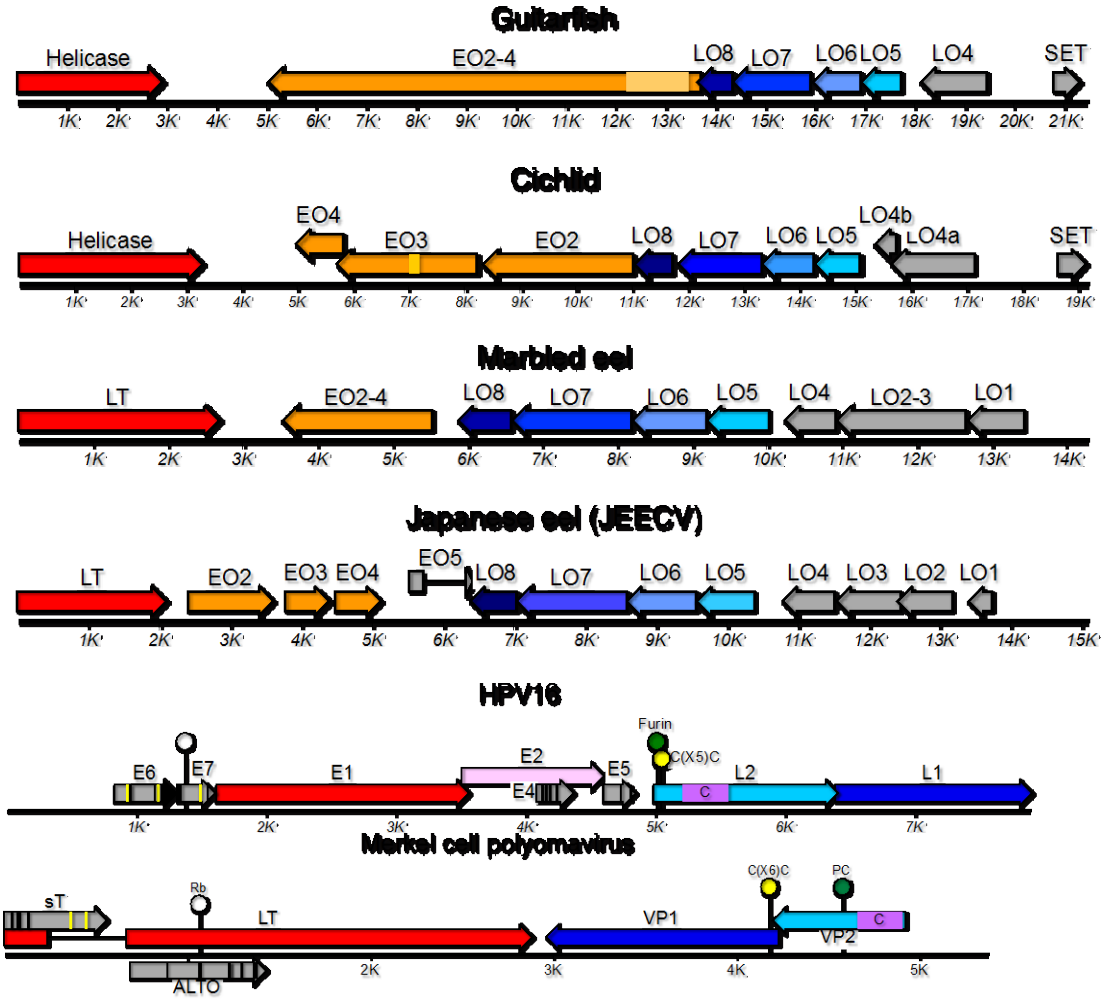


Figure 3.3 Genome organization of the colossomaviruses and a papillomavirus (HPV16) and polyomavirus (Merkel cell) for comparison. Open reading frames encoding the helicase (helicase, LT and E1), non-structural early ORFs (EO2-4) and structural late ORFs (LO1-8) are indicated by colors. Circular genomes are linearized.

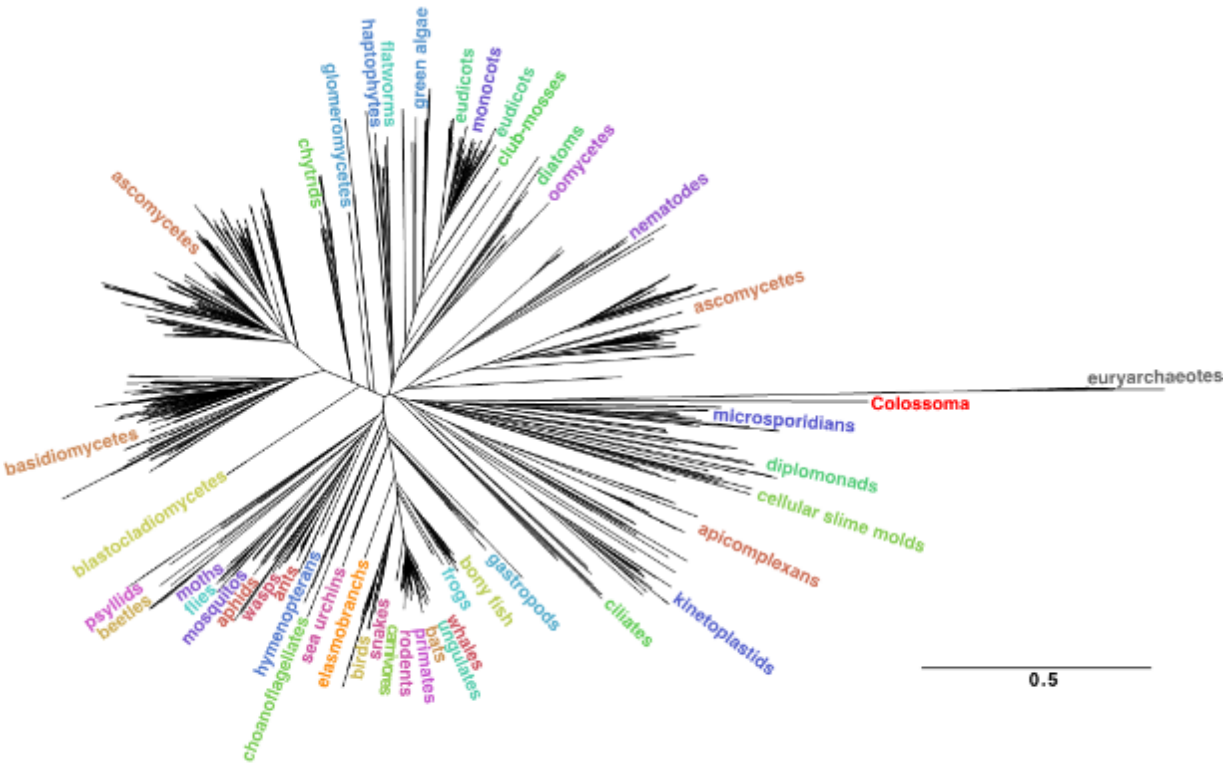
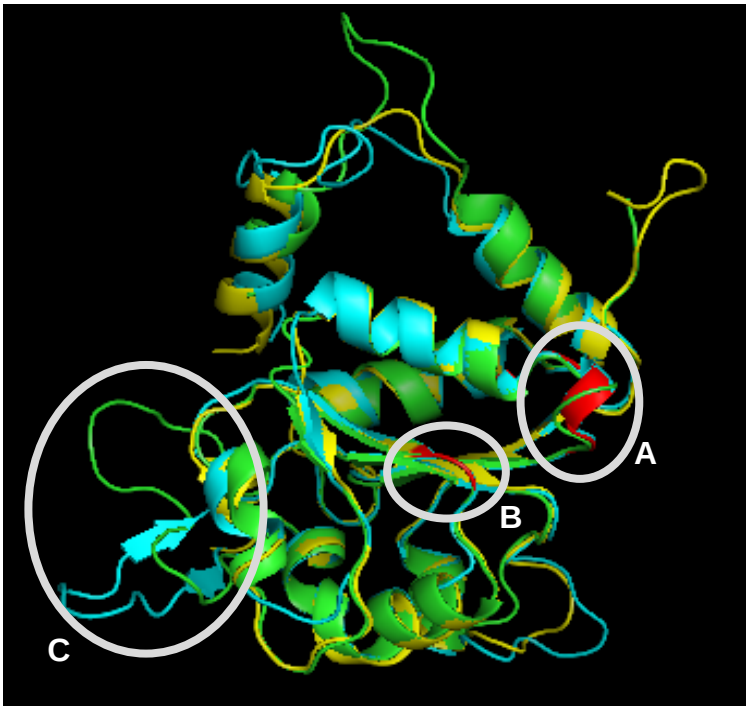


Figure 3.4 A standard BLASTP search was used to find the top 20,000 protein sequences most similar to the guitarfish colossomvirus conserved primase catalytic domain. A multiple alignment was performed and used to construct a phylogenetic tree. The locations of select taxa are indicated in colored text.



A. **Walker A motif**

Guitar Fish	**EQKKCGKS**
Parvo NS1	**GPSNTGKS**
Papilloma E1	**GPANTGKS**

B. Walker B motif

(Two negatively charged residues following a stretch of bulky hydrophobic residues)

Guitar Fish	**NVIIED**
Parvo NS1	**LLLWEE**
Papilloma E1	**VAMLDD**

C. Potential structural variations

Guitar Fish_loop
Parvo NS1_ antiparallel beta strands
Papilloma E1_loop (invisible in solved structure)

Figure 3.5 Protein model of guitarfish colossomavirus (GfCv) A) Protein model of GfCv helicase compared and superimposed with homologous region in parvovirus NS1 protein and papillomavirus E1 protein. Walker A and B motifs are highlighted in red.

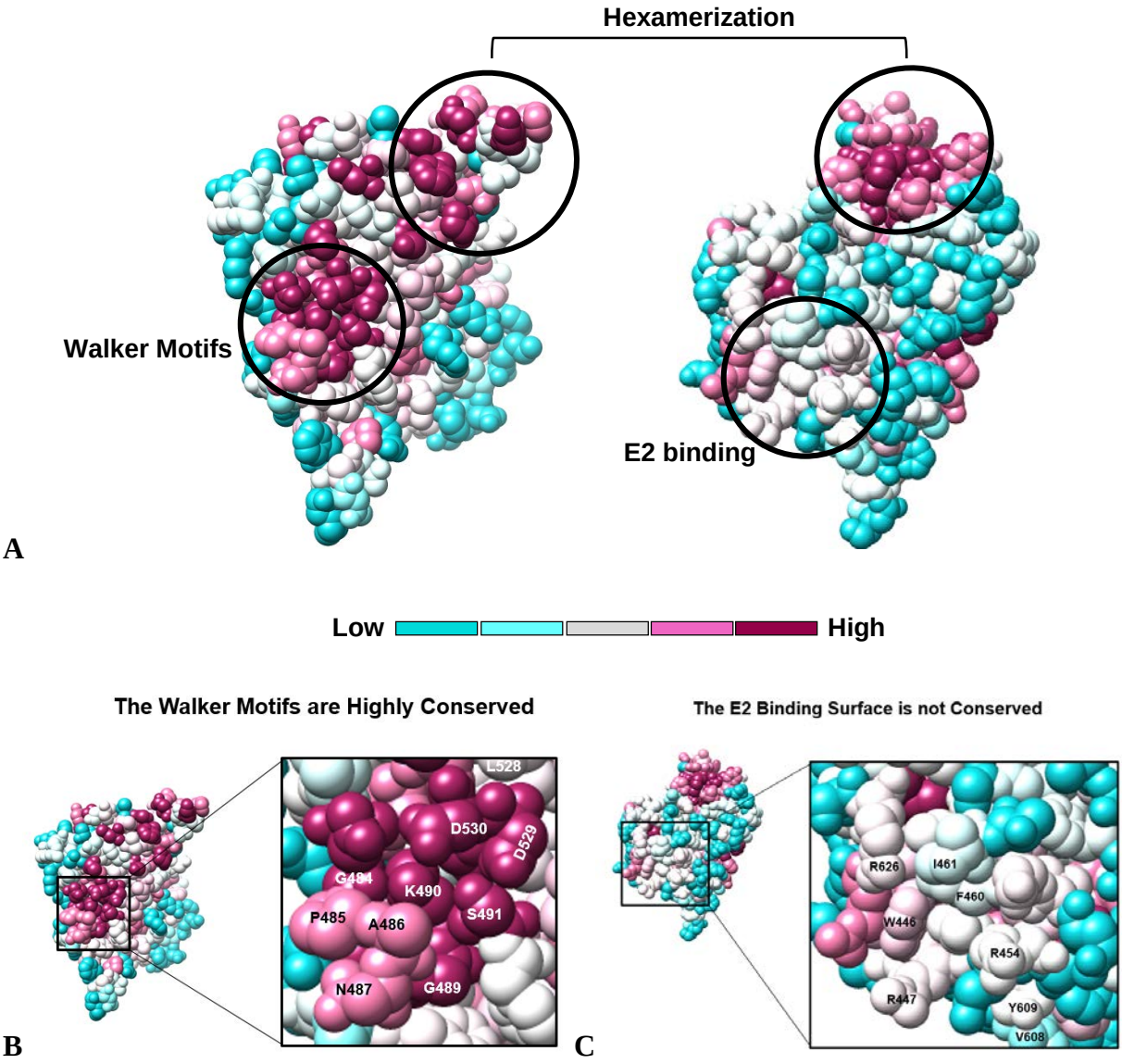


Figure 3.6 Surface conservation analysis of guitarfish colossomavirus (GfCv). A) Surface conservation analysis of GfCv helicase and 500 others. B) Surface residues involved in the core functions of the helicase, like this walker motif are highly conserved (magenta). C) Residues mediating E2 binding, which is specific for papillomavirus E1, are much less conserved (white).

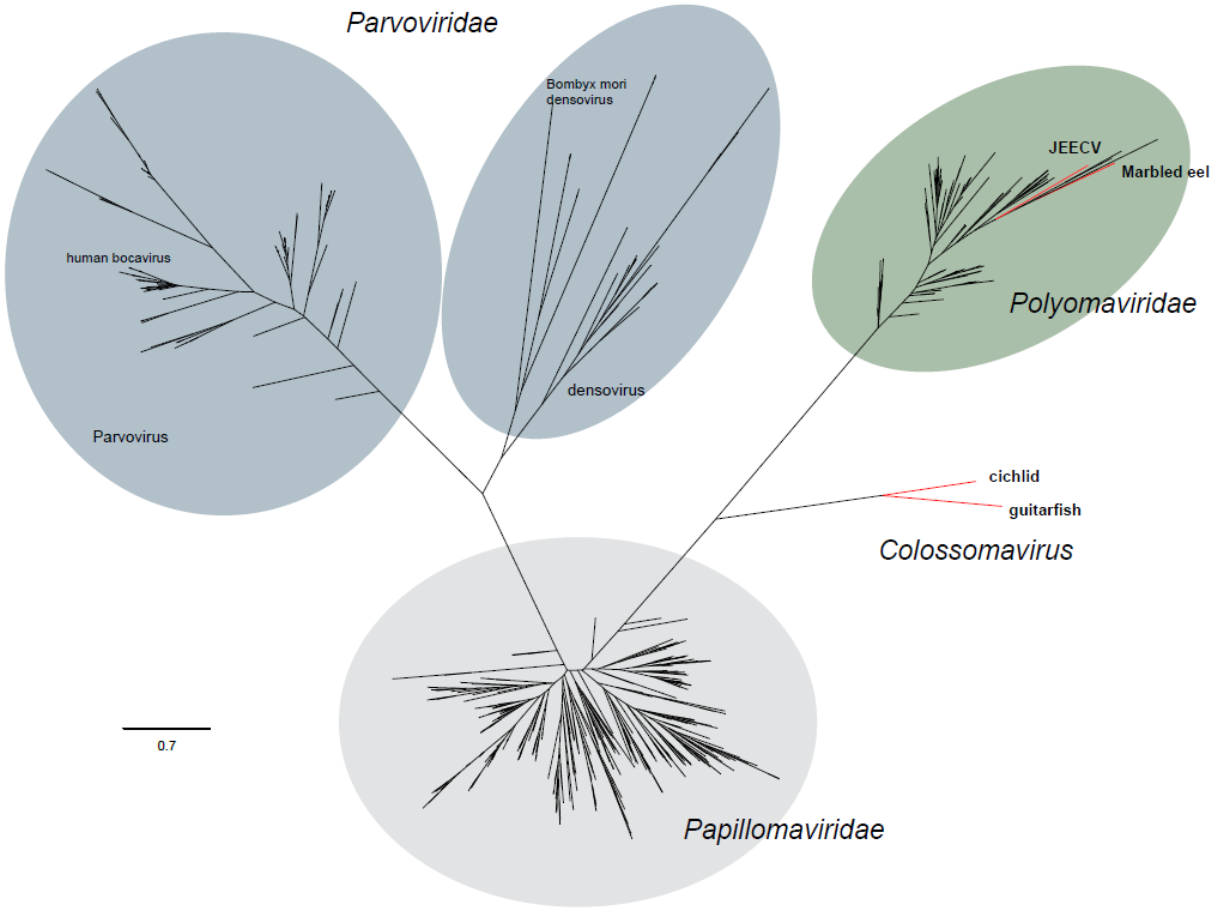


Figure 3.7 Maximum likelihood phylogeny of the helicase gene. Phylogenetic analysis based on this helicase domain revealed giant guitarfish colossomavirus (GfCV) and red discus cichlid colossomavirus (RdCV), constitutes a novel, family-level clade with equal distance to *Papillomaviridae* and *Polyomaviridae*. The helicase domain of marbled eel colossomavirus (MeCV) and JEECV, on the other hand, groups within polyomaviriuses.

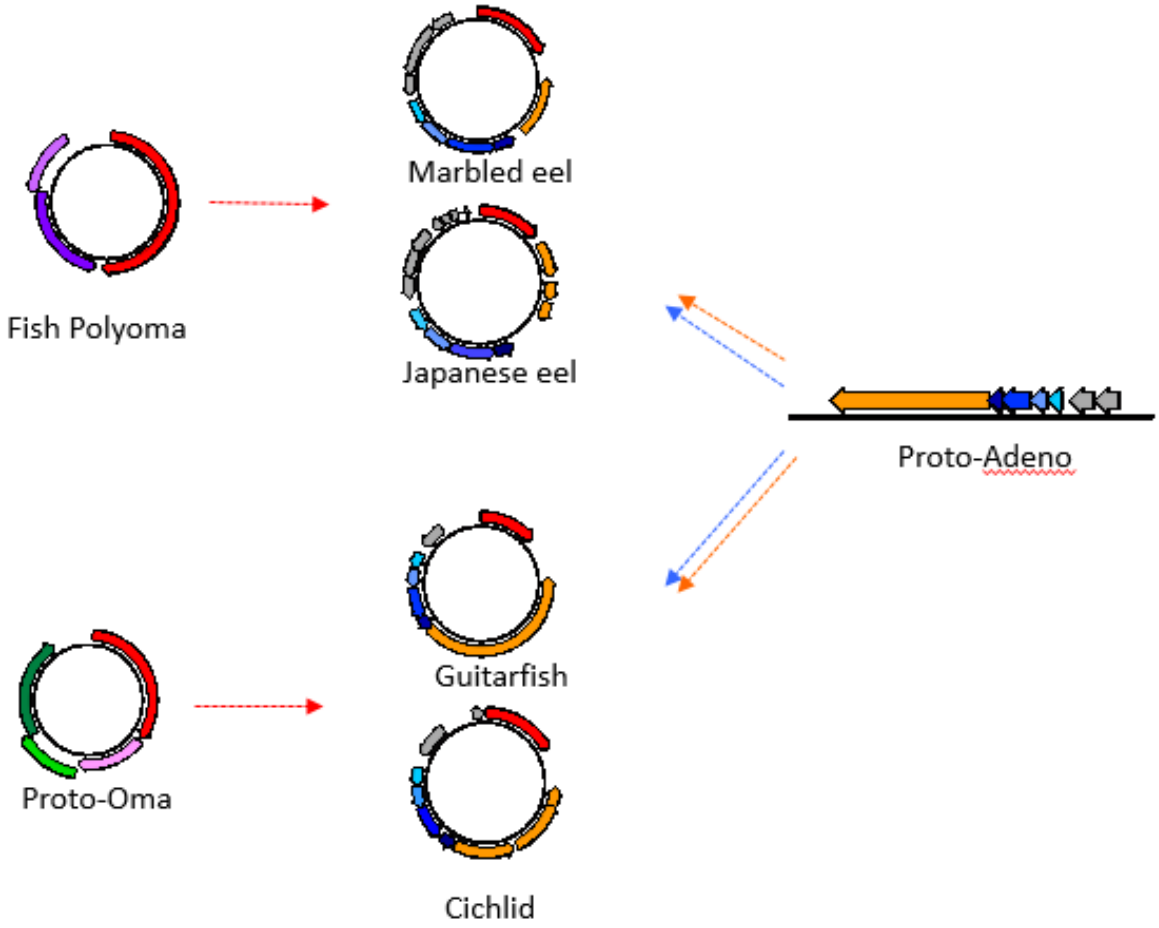


Figure 3.8 A speculative evolutionary model of colossomaviruses that accommodates current available information. Solid lines indicate direct evolutionary transition, dashed lines indicate inter-familial chimerization.

Table 3.1 Targeted gene, primer sequences and product size for guitarfish colossomaviurs (GfCv).

Guitarfish Colossomavirus Primers

Name	Sequence	Amplicon (bp)	Corresponding Gene
AF	TCACTCACAGCTCCAAATGC	390	primase
AR	TCCGTACCTGCCACACACTA		
BF	TGCTGTCAGAGGTGAAGGTG	322	helicase
BR	ACCATTCCCCTTCCTAATGG		
CF	CCAGAGGAAGATGGTGCAAT	352	LO7
CR	CCTCCCTGGAATCGTCTGTA		
DF	GGTACAGGCAGGACGACAAT	340	primase
DR	CTCGCTTATAATGCCGAAGC		

Table 3.2 Genome characteristics of colossomaviruses. Genomes are united by a complete circular dsDNA genome, ultrastructure, a conserved helicase and string of homologous ORFs that encode non-structural genes

Virus	Genome Size	GC Content %	Host (Scientific Name)	Genome Sequencing (platform)	Isolated in Cell Culture	TEM Morphology & Size
GfCV	21,527	44	Giant guitarfish (*Rhynchobatus djiddensis*)	Metagenome of tissue (MiSeq)	No (Attempted)	hexogonal 75 nm
RdCv	19,275	44	Red discus cichlid (*Symphysodon discus*)	Metagenome of tissue (MiSeq)	No (Attempted)	hexagonal 60-70nm
JEECV	15.131	48	Japanese eels (*Anguilla japonica*)	Metagenome of culture isolate (454; previous report)	Yes (Japanese eel cells)	hexagonal 75nm
MeCV	16,930	49	Taiwanese marbled eels (*Anguilla marmorata*)	Transcriptome of culture isolate (Miseq)	Yes (Marbled eel cells)	hexogonal 70-80 nm

Table 3.3 Gene predication and annotations for guitarfish colossomavirus (GfCv).

	Closest Homolog	**Amino Acid Identities**
Early Genes		
SET	**various SET domain protein**	**34%**
Helicase	**E1 [various papillomaviruses]**	**25-28%**
Late Genes		
LO4	**no homolog**	**n/a**
LO5	**LO5 protein [Japanese eel endothelial cells-infecting virus]**	**28%**
LO6	**LO6 protein [Japanese eel endothelial cells-infecting virus]**	**23%**
LO7	**LO7 protein [Japanese eel endothelial cells-infecting virus]**	**30%**
LO8	**LO8 protein [Japanese eel endothelial cells-infecting virus]**	**28%**
DNA primase	**DNA primase [Brugia malayi]**	**24%**

Table 3.4 Percent amino acid identity of giant guitarfish colossomavirus (GfCV), the red discus cichlid colossomavirus (RdCV), and the marbled eel colyomavirus (MeCV; previously called AMPyV), and Japanese eel endothelial cells-infecting virus (JEECV) predicted genes compared and partitioned by open reading frame.

			Helicase			LO5				LO6		
Virus	**size (aa)**	**% Amino Acid Identity**			**size (aa)**	**% Amino Acid Identity**			**size (aa)**	**% Amino Acid Identity**		
		RdCV	MeCV	JEECV		RdCV	MeCV	JEECV		RdCV	MeCV	JEECV
GuitarFish CV	1102	27%	19%	22%	265	24%	24%	29%	318	20%	22%	22%
Red discus cichlid CV	973		23%	20%	271		31%	24%	311		27%	22%
Marbled Eel CV	895			28%	278			38%	329			34%
Japanese Eel CV	699				273				327			

		LO7				LO8		
Virus	**size (aa)**	**% Amino Acid Identity**			**size (aa)**	**% Amino Acid Identity**		
		RdCV	MeCV	JEECV		RdCV	MeCV	JEECV
GuitarFish CV	1102	28%	31%	29%	235	25%	27%	31%
Red discus cichlid CV	973		26%	27%	228		34%	32%
Marbled Eel CV	895			37%	237			37%
Japanese Eel CV	699				216			

Chapter 4

COMPLETE SEQUENCE OF THE SMALLEST POLYOMAVIRUS GENOME, GIANT GUITARFISH (*RHYNCHOBATUS DJIDDENSIS*) POLYOMAVIRUS 1

Dill JA, Ng TFF and Camus AC

To be submitted to ASM Genome Announcements

Abstract

Polyomaviruses are known to infect mammals and birds. Deep sequencing and metagenomic analysis identified the first polyomavirus from a cartilaginous fish, the giant guitarfish (*Rhynchobatus djiddensis*). Giant guitarfish polyomavirus 1 (GfPyV1) has typical polyomavirus genome organization, but is the smallest polyomavirus genome (3.96 kb) described to date.

Polyomaviruses have been found in a range of avian and mammalian species. Although some persist asymptomatically, other polyomavirus species cause diseases ranging from urinary tract hemorrhage to neoplasia (Essbauer et al. 2001; DeCaprio et al. 2013;Baron et al. 2013; Voyles 1993; Guerin et al. 2000). Historically, taxonomic classification included three genera, the *Orthopolyomavirus* and *Wukipolyomavirus* from mammals and *Avipolyomavirus* from birds (Flint et al. 2000; Johne et al. 2011). However, a recent taxonomy proposal delineated four new genera, designated *Alpha-*, *Beta-*, *Gamma-* and *Delta-* polyomavirus (Calvignac-Spencer et al. 2016). Recently, black sea bass-associated polyomavirus 1 (BassPyV1, GenBank accession number KP071318), the first polyomavirus associated with a bony fish (*Centropristis striata*), was described (Peretti et al. 2015). Here, a complete polyomavirus genome is reported from a cartilaginous fish, the giant guitarfish (*Rhynchobatus djiddensis*), a batoid elasmobranch (Order Rajiformes). The presence of proliferative skin lesions, characterized microscopically by large intranuclear inclusions containing 75 nm icosahedral viral particles, initiated an investigation of the causative agent (Camus et al. 2015). To circumvent the lack of known viral genetic information in elasmobranchs, a sequence-independent metagenomic approach was performed to identify viral sequences within the lesions (Ng et al. 2011; Schuurman et al. 1990).

A complete, circular, double-stranded, 3,962 bp DNA genome was characterized. This virus, giant guitarfish polyomavirus 1 (GfPyV1), has characteristic polyomavirus arrangement of major open reading frames, including LT, VP1 and VP2. Although transmission electron microscopy failed to identify polyomavirus-like particles in tissue and virus isolation was not attempted due to lack of compatible cell lines, the presence of GfPyV1 LT and VP1 nucleic acids in skin lesions were confirmed using nested PCR and Sanger sequencing.

The genome size of 3.96 kb makes GfPyV1 the smallest described polyomavirus, compared with other genomes of 4.7 to 7.4 kb (DeCaprio et al. 2013; Schuurman et al. 1990; Stevens et al. 2013). While the GfPyV1 genome showed typical polyomavirus organization, its nucleotide sequence is highly divergent from other polyomaviruses. The predicted 1,794 bp large T (LT) protein is encoded by a single open reading frame, in contrast to the spliced LT genes of other polyomaviruses. BLAST searches revealed roughly 30% identity to a variety of mammalian and avian polyomavirus LT proteins. The LT from GfPyV1contains predicted DnaJ, Ori-binding, and helicase domains typical of polyomaviruses (An et al. 2012). A possible small T antigen-like ORF encoding a 75 amino-acid-long protein was also predicted in the GfPyV1 early region, but a BLAST search revealed no sequence identity to any proteins in GenBank.

The predicted major capsid protein (VP1) contains 277 amino acids, smaller than all known VP1 proteins (DeCaprio et al. 2013). It shares roughly 25% identity with various polyomavirus VP1 coat proteins by BLASTp search. At 500 amino acids, the predicted minor capsid protein (VP2) is longer than typical VP2 proteins (DeCaprio et al. 2013). The VP2 encodes a possible N-terminal myristoylation signal.

Comparing the two fish polyomaviruses using Sequence Declamation Tool (SDT) v.1.0 (Muhire et al. 2014), BassPyV1 and GfPyV1 share 19.3%, 26%, 27.8%, and 22.9% protein

identity in the viral genes LT, ST, VP1 and VP2, respectively. Although GfPyV1 DNA was present in associated tissues, preliminary data suggests that it was not the cause of the skin lesions.

Nucleotide sequence accession number. The complete genomic sequence of guitarfish polyomavirus 1 was deposited in GenBank under the accession number NC_026244.1/ KP264963.1.

Acknowledgements: We thank Eric Delwart and Beatrix Kapusinszky at the University of California, San Francisco and the Blood Systems Research Institute for assistance with sequencing. We thank Christopher B. Buck for his sequence analysis support.

References

An P, Sáenz Robles MT, Pipas JM. 2012. Large T antigens of polyomaviruses: amazing molecular machines. Ann Rev Microbiol 66:213-236.

Baron HR, Howe L, Varsani A, Doneley RJ. 2013. Disease screening of three breeding populations of adult exhibition budgerigars (*Melopsittacus undulatus*) in New Zealand reveals a high prevalence of a novel polyomavirus and avian malaria infection. Avian Dis 58:111-117.

Calvignac-Spencer S, Feltkamp MC, Daugherty MD, Moens U, Ramqvist T, Johne R, Ehlers B. 2016. A taxonomy update for the family Polyomaviridae. Arch Virol. DOI 10.1007/s00705-016-2794-y.

Camus A, Dill J, McDermott A, Camus M, Fan NT. 2016. Virus-associated papillomatous skin lesions in a giant guitarfish *Rhynchobatus djiddensis*: a case report. Dis Aquat Organ 117:253-258.

DeCaprio JA, Imperiale MJ, Major EO. 2013. Polyomaviruses, p 1633 – 1661. *In* Knipe DM, Howley PM (ed), Fields Virology, 6th ed. Wolters Kluwer Health/Lippincott Williams & Wilkins, Philadelphia, PA.

Essbauer S, Ahne W. 2001. Viruses of lower vertebrates. J Vet Med B 48: pp.403-475.

Flint SJ, Enquist LW, Racaniello VR, Skalka AM, Barnum DR, de Evaluación E. 2000. Principles of virology: molecular biology, pathogenesis and control. ASM Press, Washington DC.

Guerin JL, Gelfi J, Dubois L, Vuillaume A, Boucraut-Baralon C, Pingret JL. 2000. A novel polyomavirus (goose hemorrhagic polyomavirus) is the agent of hemorrhagic nephritis enteritis of geese. J Virol 74:4523-9.

Johne R, Buck CB, Allander T, Atwood WJ, Garcea RL, Imperiale MJ, Major EO, Ramqvist T, Norkin LC. 2011. Taxonomical developments in the family Polyomaviridae. Arch Virol 156:1627-34.

Muhire BM, Varsani A, Martin DP. 2014. SDT: a virus classification tool based on pairwise sequence alignment and identity calculation. PLoS One 9:e108277.

Ng TF, Driscoll C, Carlos MP, Prioleau A, Schmieder R, Dwivedi B, Wong J, Cha Y, Head S, Breitbart M, Delwart E. 2013. Distinct lineage of vesiculovirus from big brown bats, United States. Emerg Infect Dis 19:1978-80.

Ng TF, Wheeler E, Greig D, Waltzek TB, Gulland F, Breitbart M. 2011. Metagenomic identification of a novel anellovirus in Pacific harbor seal (*Phoca vitulina richardsii*) lung samples and its detection in samples from multiple years. J Gen Virol 92:1318-23.

Peretti A, FitzGerald PC, Bliskovsky V, Pastrana DV, Buck CB. 2015. Genome sequence of a fish-associated polyomavirus, black sea bass (*Centropristis striata*) polyomavirus 1. Genome Announc 3:e01476-14.

Schuurman R, Sol C, Van Der Noordaa J. 1990. The complete nucleotide sequence of bovine polyomavirus. J Gen Virol 71:1723-35.

Stevens H, Bertelsen MF, Sijmons S, Van Ranst M, Maes P. 2013. Characterization of a novel polyomavirus isolated from a fibroma on the trunk of an African elephant (*Loxodonta africana*). PloS one 8:e77884.

Voyles BA. 1993. The biology of viruses. Mosby, St. Louis, Missouri.

Appendix

Additional information regarding guitarfish polyomavirus 1 is included in the appendix as “The Ancient Evolutionary History of Polyomaviruses” published in Plos Pathogens by Buck et al. including myself and committee members. This paper used GfPv1 alongside other novel polyomavirus genomes to investigate and provide a theoretical framework for understanding the deep evolutionary history of the *Polyomaviridae* family. In depth bioinformatics, gene analysis, structural modeling and phylogenetic approaches were able to highlight potential pitfalls of the previously accepted taxonomic system, estimate evolutionary rates including last common ancestors, and suggest recombination and chimera events that have shaped the lengthy evolution of polyomaviruses. *In situ* hybridization for GfPv1 was also summarized in this paper.

Chapter 5

DISTINCT VIRAL LINEAGES FROM FISH AND AMPHIBIANS REVEAL THE COMPLEX EVOLUTIONARY HISTORY OF HEPADNAVIRUSES

Dill JA, Camus AC, Leary JH, Giallonardo FD, Holmes EC, Ng TFF
To be submitted to Journal of Virology

Abstract

Hepadnaviruses (HBVs) are the only animal viruses that replicate their DNA by reverse transcription of an RNA intermediate. Next generation sequencing and metagenomic analysis of papillomas from the lips and skin of bluegill (*Lepomis macrochirus*) revealed a novel exogenous hepadnavirus representing a second prototype fish hepadnavirus. In addition, *in silico* analyses of the whole-genome shotgun (wgs) and transcriptome Shotgun assembly (TSA) databases revealed novel homologs of hepatitis B viruses (HBVs) in another fish, the African cichlid (*Ophthalmotilapia ventralis*) and an amphibian, the Tibetan frog (*Nanorana parkeri*). Residues in the core proteins, designated motif I, II and III, were highly conserved in all vertebrate HBVs, likely to maintain proper formation of capsid monomer, dimer and inter-subunit interactions. Surface proteins in all vertebrate HBVs contain similar predicted membrane topology, characterized by the three transmembrane domains, even though pairwise identities are very low (<40%). However, none of the fish and amphibian viruses contained an X protein homolog common to the mammalian orthohepadnaviruses. Most striking was that the bluegill hepadnavirus (BGHBV), the African cichlid hepadnavirus (ACHBV), and the previously described white sucker hepadnavirus (WSHBV) did not form a fish-specific monophyletic group. Notably, BGHBV was more closely related to the mammalian hepadnaviruses, indicating that cross-species transmission events have played a major role in viral evolution. Evidence of cross-species transmission was also observed with TFHBV. Hence, these data indicate that the evolutionary history of the hepadnaviruses is more complex than previously realized and combines both virus-host co-divergence over millions of years and host species jumping.

Importance

Hepadnaviruses are responsible for significant disease in humans (hepatitis B virus) and have been reported from a diverse range of vertebrates as both exogenous and endogenous viruses. We report the full length genome of a novel hepadnavirus from a fish and the first hepadnavirus genome from an amphibian. The novel fish hepadnavirus, sampled from bluegill, was more closely related to mammalian hepadnaviruses than to other fish viruses. This phylogenetic pattern reveals that although hepadnaviruses have likely been associated with vertebrates for hundreds of millions of years, they have also been characterized by species jumping across wide phylogenetic distances.

Key Words: Hepatitis B Virus, *Hepadnaviridae*, fish hepadnavirus, amphibian hepdnavirus, evolution, phylogeny

Introduction

The *Hepadnaviridae* are characterized by extremely small (3-3.3kbp), partially double-stranded DNA (dsDNA) genomes. Viral particles are spherical, with a diameter of approximately 42 nm, each containing a single copy of the genome covalently linked to the viral reverse transcriptase (RT) that provides DNA polymerase activity (Knipe 2013, Voyles 1993; Flint et al. 2000). The hepadnaviruses are unique among animal viruses in that they replicate their DNA by reverse transcription of an RNA intermediate and comprise the only Group VII animal virus (dsDNA-RT virus) of the Baltimore system, which classifies viruses according to their genome composition and method of replication (Baltimore 1971; Knipe 2013).

At present, the *Hepadnaviridae* are subdivided into two genera (Orito et al 1989; Suh et al. 2013; ven Hemert et al 2011): the genus *Orthohepadnavirus* that infects mammals, including humans, and the genus *Avihepadnavirus* that infects birds (Knipe 2013; Drexler et al. 2013; Saif 2008; Siddiqui et al. 1981; Kodama 1985; Summers et al. 1978; Prassolov et al 2003). Within both genera, the circular viral genomes exhibit multiple overlapping open reading frames (ORF), comprising the polymerase, pre C/C, and pre S/S ORFs that encode the viral polymerase (P), core (C), and surface (S) proteins, respectively. In the *Orthohepadnavirus* genus, a fourth ORF encodes protein X. Despite these similar genome organizations, nucleotide sequence identity between hepadnavirus genera is limited, with the exception of some highly conserved functional domains (Gilbert 2014; Suh 2014).

Human hepatitis B virus (HBV) affects more than one third of the human population and infections have the potential to cause both severe chronic liver disease and hepatocellular carcinoma (WHO, Knipe 2013, Voyles 1993; Flint et al. 2000). Interestingly, chronic infection by woodchuck hepatitis B virus (WHBV) can result in similar pathologic changes in that species

(Kodama 1985; Summers et al. 1978). Liver pathology is less commonly induced by avihepadnaviruses, although duck hepatitis B virus (DHBV) can cause liver necrosis (Saif 2008). The first hepadnavirus from a bony fish, the white sucker (*Catostomus commersonii*), class Actinopterygii, was described in 2015, although no disease association was observed. To-date, no other exogenous reptilian or amphibian hepadnaviruses have been described.

In addition to exogenous hepadnaviruses, a number of endogenous sequences (eHBV), in the form of endogenous viral elements (EVEs), have been identified in animal genomes. Hepadnaviral EVEs have been documented in turtles, crocodiles, snakes, and birds (Cui et al. 2014; Gilbert et al. 2010; Gilbert et al. 2014; Robertson et al. 2002), although no mammalian, amphibian or fish endogenous hepadnaviruses have yet been detected. The presence of EVEs has helped provide a time-scale of hepadnavirus evolution, particularly as some of the endogenization events may have occurred as early as 200 million years ago (Suh 2014). Hence, although there is clear evidence for some cross-species transmission (Starkman et al. 2003), current data suggests that hepadnavirus evolution largely follows a pattern of virus-host co-divergence that extends to at least the origin of the ray-finned fishes.

To better understand the host range and evolution of the hepadnaviruses in vertebrates, particularly the extent of virus-host co-divergence, we investigated new fish and amphibian (exogenous) hepadnaviral homologs that are highly divergent from the hepadnaviruses previously described in mammals and birds. These include the second fish hepadnavirus, from bluegill sunfish (*Lepomis macrochirus*), the first amphibian hepadnavirus from a Tibetan frog (*Nanorana parkeri*), and analysis of a hepadnavirus-like sequence from Lake Tanganyika African cichlid fish (*Ophthalmotilapia ventralis*).

Materials and Methods

Sample collection

Five bluegill sunfish from a mixed species exhibit were submitted by a public aquarium to the Aquatic Pathology Service at the College of Veterinary Medicine, University of Georgia, in 2009 as part of an investigation into an epizootic of papillomas involving the lips and skin of this species. Complete necropsies were performed and samples of major organs and lesions were fixed in 10% neutral buffered formalin. Tissues were processed routinely, sectioned at 5 μm, and stained with hematoxylin and eosin for histologic evaluation. Portions of lip and skin lesions were collected separately and archived in a -80°C freezer.

In 2014, similar proliferative lesions were observed on bluegill by a private pond owner in Waleska, Georgia and five fish were submitted for diagnostic evaluation on April 14, 2014. Additional submissions of 13 bluegill, seven with proliferative lip lesions and five without, were made on July 7, 2014. Eight bluegill, four with lesions and four without, and one largemouth bass (*Micropterus salmoides*) were submitted on July 4, 2015. Additional samples used in the study included five bluegill, two redbreast sunfish (*Lepomis auritus*) and two redear sunfish (*Lepomis microlophus*), submitted by a commercial fish hatchery in Hawkinsville, Georgia on January 16, 2015. Four bluegill and one green sunfish (*Lepomis cyanellus*) were also received from local anglers in the Athens, Georgia area September 1, 2015. All fish were processed for histopathology as described above. In addition to lip and skin lesions, pooled samples of liver, spleen and kidney were frozen at -80°C, as well as gonadal tissue from some fish.

Fin clip samples from two *O. ventralis* cichlids were provided by a local hobbyist and archived at -80°C.

Viral metagenomics and bioinformatics analysis of next-generation sequencing (NGS) data

Histopathological examination and transmission electron microscopy (TEM) did not reveal a causative agent for the lesions (data not shown). Metagenomic sequencing was performed according to previously described protocols to further investigate a potential underlying viral etiology (Ng et al. 2015; Ng et al. 2012; Victoria et al. 2008). In brief, a tissue homogenate was centrifuged through a 0.22 µm filter to enrich viral particles by size, then treated with nucleases to deplete host nucleic acids. Nucleic acids from nuclease-resistant viral particles were extracted using the QIAquick viral RNA column purification system, followed by sequence-independent amplification using random priming. First strand synthesis (for both DNA and RNA) was performed using a 28-base oligonucleotide whose 3′ end consisted of eight random nucleotides (primer N1_8N, CCTTGAAGGCGGACTGTGAGNNNNNNNN) using superscript III reverse transcriptase (Invitrogen) (Ng et al. 2015; Ng et al. 2012; Victoria et al. 2008). A second strand was synthesized using Klenow fragment DNA polymerase (New England BioLabs). The resulting double-stranded cDNA and DNA were then PCR amplified using AmpliTaq Gold DNA polymerase and a 20-base primer (primer N1, CCTTGAAGGCGGACTGTGAG). A duel-indexed sequencing library was then prepared using the Nextera XT DNA Sample Prep Kit (Illumina, San Diego, CA). After pooling, the final library was sequenced using the MiSeq sequencing system, with 2 × 250 bp paired-end sequencing reagents (Illumina MiSeq Reagents V2, 500 cycles).

A total of 11 million reads were generated and analyzed as previous described (Ng et al. 2012). An in-house analysis pipeline running on a 32-node Linux cluster was used to process the data (University of California, San Francisco). Adaptor and primer sequences were trimmed using VecScreen (McGinnis et al. 2004), while duplicate reads and low-sequencing-quality tails

were removed using a Phred quality score of 10 as the threshold. The cleaned reads were *de novo* assembled using an in-house sequence assembler employing an ensemble strategy (Deng et al. 2015) that consists of SOAPdenovo2, ABySS, meta-Velvet, and CAP3. The assembled sequence was compared with an in-house viral protein sequence database using BLASTx. Viral contigs were further inspected manually using Geneious (version R6; Biomatters, Auckland, New Zealand).

Complete genome sequencing

PCR was performed using primers BGHBV-CirF 5- CAACGCCAACAGCATTTTTA-3 and BGHBV-CirR 5- TAATATCGGTCGAGACTGCG-3, which anchored in the polymerase and core ORFs to obtain the last 1% of the genome, bridging the intergenic region. The resulting 373-bp amplicons were sequenced using Sanger methods to confirm the circularity of the genome.

Molecular screening

Tissues from 40 bluegill, three related *Lepomis* species, and one largemouth bass were extracted using Qiagen DNA extraction kits. Screening for BGHBV was accomplished by traditional PCR, targeting the polymerase with primer sets BGHBV-PolF 5-TGTGGACAAAAATCCACGAA-3 and BGHBV-PolR 5-CGTAAAGCACCTATGGGCAT-3 (Table 5.1) using a previously described touch down protocol (Ng et al. 2013). Additional primers targeting the polymerase, capsid and core proteins were also designed and verified (Table 5.2).

Quantitative (q)PCR was used to assess the presence of viral DNA from the selected tissues as indicated (Table 5.1). Primers were designed from the polymerase gene to yield a 110 bp amplicon (PolQpcrF and PolNestR, Table 5.2). The primer set was used in a standard PCR

reaction with DNA extracted from bluegill GAI-2 (referred to as the positive control). The DNA was run on a 2% agarose gel, purified (Qiaquick Gel Extraction Kit) and quantitated (NanoDrop 2000, Thermo Fisher). DNA was adjusted to 1 ng/µl. Ten-fold dilutions of this stock were made in water for qPCR standard curve generation. Preliminary analysis indicated that the 10^{-1} through 10^{-8} dilutions (10^{-1} -10^{-8} ng) would cover the dynamic (linear) range of the assay ($R^2 \geq 0.95$). qPCR was performed on a Bio-Rad IQ5 iCycler using iQ5 system software for analysis. One µl of extracted DNA was added to each 25 µl reaction mix containing iQ SYBR Green Supermix (Bio-Rad) and 100 nmol each of the indicated primers. A 2-step cycling program was used as follows: an initial 95° C for 3 min followed by 35 cycles of 95° C for 10 seconds and 60° C for 30 seconds. Initial screening of all samples was performed twice using one PCR well/sample. Final assessment of viral DNA presence was made on samples run in triplicate.

Endpoint PCR were performed to test the cichlids for ACHBV. Fin biopsies from two *O. ventralis* cichlids were extracted using spin columns as described above. Tissue DNA was screened for the presence of cichlid hepadnavirus DNA using primers specific to the cichlid hepadnavirus polymerase sequence (ACHBV-PolF and ACHBV-PolR , Table 5.2). PCR for Cytochrome b was used as a positive control to verify extraction and PCR methods (Primer OVCytBF, OVCytBR, Table 5.2) (Morita et al. 2014).

***In silico* screening of public sequence data**

The core, polymerase, and surface protein sequences from BGHBV were used as queries in a BLAST analysis against the GenBank whole-genome shotgun (wgs) and transcriptome Shotgun assembly (TSA) databases, employing an e-value of 10 e^{-4}, to detect hepadnavirus homologs in amphibians and fish. The resulting sequences were then re-analyzed by reverse-BLAST, ORF predication, sequence comparison and alignment, as well as bioinformatics

analysis to validate the initial assembly. Other orthohepadnavirus and avihepadnavirus proteins used as queries detected identical sequences as that from BGHBV (data not shown).

Sequence comparisons and phylogenetic analysis

Coding sequences of representative hepadnavirus C, P, and S genes were downloaded from GenBank and combined with those of BGHBV and TFHBV. To be as broad as possible, the background GenBank data set included both exogenous *Avihepadnavirus*, *Orthohepadnavirus*, and white sucker hepatitis B virus (WSHBV) sequences, as well as available avian and reptilian (crocodilian) endogenous (e) hepadnavirus sequences that were of sufficient length to conduct phylogenetic analyses, although sequence availability differed by gene. A full list of the sequences utilized are available (Table 5.3).

Amino acid sequence alignment of the core, polymerase, and surface data sets were inferred using multiple cycles of the MUSCLE algorithm (Edgar 2004). Because the highly divergent nature of some sequences could compromise phylogenetic accuracy, alignment gaps and ambiguously aligned sequences were removed using the Gblocks program with relatively relaxed settings (i.e. allowing smaller final blocks and less strict flanking regions) (Talavera & Castresana 2007). This resulted in final multiple sequence alignments lengths of (i) P = 35 taxa, 272 amino acids; C = 34 taxa, 110 amino acids; S = 24 taxa, 187 amino acids. Based on these alignments, maximum likelihood (ML) phylogenetic trees were estimated using PhyML (Guindon et al. 2010), employing the LG+Γ model of amino acid substitution and 1000 bootstrap replicates. Finally, pairwise sequence similarities were calculated using the translated amino acid sequences with the Sequence Demarcation Tool (Muhire et al. 2014) (Table 5.4).

Core (capsid) protein modeling

The structure of capsid dimer and homo hexamer was based on the published structure of the HuHBV virion (PDB: 3J2V and 5E0I, respectively) (Klumpp et al. 2015, Yu et al. 2013). Protein modeling and color manipulation were performed using PyMOL software (http://www.pymol.org, version 1.8.0.0). No protein crystal has been resolved in any fish or amphibian HBV, but all share conserved residues with HuHBV.

Membrane protein prediction

The PreS/Surface gene encodes for three envelope proteins: L, M, and S. Since L is the largest and contains the sequence and the membrane configuration of M and S (Brass 2004), we focused on this protein. Transmembrane prediction was performed on known (HuHBV and DHBV) and putative (BGHBV, WSHBV and TFHBV) L protein sequences using Hidden Markov models in TMHMM (Krough et al. 2001; Moller et al. 2001). The results were compared against an established transmembrane model (Brass 2004) to predict membrane topology. Alternative start codon positions for the envelope protein (L) were detected in TFHBV, resulting in two potential sizes (361 and 490 amino acids), so both protein sequences were analyzed.

Data availability

Sequences were deposited into GenBank under accession KX058433-5.

Results

Viral metagenomics of a divergent hepadnavirus in bony fish

Histological examination of the lips and skin of 40 bluegill revealed typical, well-differentiated papillomas in 20 fish, suggesting a possible viral etiology. A sequence-independent metagenomic approach was performed to identify viral sequences within the lesions

and internal organs to circumvent the lack of known viral genetic information in teleosts and the paucity of cell lines for culture. Accordingly, a novel virus, denoted bluegill hepadnavirus (BGHBV; GenBank KX058433), was identified in the next generation sequence data (Figure 5.1). Over 5,000 NGS reads covered 99% of the genome with more than 10X coverage. The remaining sequence, as well as the circular nature of the viral genome, was confirmed by PCR and Sanger sequencing, using primers anchored in the polymerase and core ORFs that spanned the entire noncoding region (Figure 5.2).

Molecular screening of BGHBV in *Lepomis* species

Forty-five *Lepomis* spp. fish, including 40 bluegill, two redear sunfish, two redbreast sunfish and one green sunfish, and one largemouth bass were screened by endpoint and real time PCR to investigate whether BGHBV was endogenized in the *Lepomis* spp. genomes. At least one tissue from all 46 fish was screened by both techniques and 12 samples were selected from the total survey for qPCR replicate analysis, which ranged from 0 fg to the highest concentration of 146 fg in the skin of one fish (Table 5.1). Among the four *Lepomis* spp. examined, BGHBV was only identified by PCR in 6/40 bluegill and, with the exception of one archived lip sample, five positive fish came from a single pond. Fish from two additional locations were all PCR negative. Although the quantity of viral DNA in each tissue varied, BGHBV nucleic acid was identified in 3/18 grossly visible lip lesions, 3/22 non-lesioned lip samples, 3/12 pooled organ samples and 3/7 skin samples (Table 5.1). This prevalence data did not indicate that BGHBV was associated with the lip lesion even though it was initially discovered from a diseased individual. Taken together with the circular nature of the BGHBV genome as indirect evidence against an endogenization, these results indicate that BGHBV is not derived from the germline of the bluegill.

Characterization of a prototype amphibian hepadnavirus

In silico screening of GenBank for novel hepadnaviruses identified hepadnavirus-like sequences from the whole genome sequence data of the Tibetan frog (Sun et al 2015) that shared 33% protein sequence similarity with the polymerase protein of BGHBV. The initial contig (GenBank accession number JYOU01126907) was analyzed using SOAPdenovo assembler in the original report (Sun et al 2015), resulting in a linear sequence of 3,137 bp. Our subsequent analysis using the original 4.4B Illumina Hiseq reads confirmed the circular nature of the sequence by identifying overlapping read coverage at both sequence termini (Figure 5.2), resulting in a complete genome sequence of length 3,138 bp (Tibetan frog hepadnavirus, TFHBV; GenBank KX058435). The Tibetan frog data set contained 13 whole genome sequencing (DNA) runs, of which only a small portion (<0.003%) of the total reads were hepadnaviral (Table 5.5). All runs were performed on a single muscle sample (Sun et al 2015), so all data sets contained TFHBV sequences. No other hepadnavirus-like sequences were identified in other amphibian whole genome or transcriptome assembled data sets at the time of analysis.

Identification of hepadnavirus in cichlids

A hepadnavirus-like sequence was also identified from the transcriptome data set of the African cichlid *Ophthalmotilapia ventralis* (GenBank accession number JL559376) (Baldo et al. 2011; Hahn et al. 2015).. Notably, this is the only hepadnaviral sequence in the entire 454 transcriptome, comprising the polymerase polyprotein (Figure 5.1). Using the original reads, our analysis obtained a final sequence of 2,485 bp (KX058434). This is clearly a partial sequence in which a circular genome could not be obtained with the available data. To further investigate if this African cichlid hepadnavirus-like sequence (ACHBV) was endogenized into the host

cellular genome, we examined cellular DNA from skin samples of two *O. ventralis* using end point PCR. PCRs targeting the hepadnavirus-like sequence was negative in both, while the positive control PCR targeting the cytochrome b gene of *O. ventralis*, was validated, indicating that ACHBV was not incorporated in the cellular genome in the samples investigated.

Genome organization of fish and amphibians hepadnaviruses

The bluegill hepadnavirus (BGHBV) and the Tibetan frog hepadnavirus (TFHBV) are complete circular genomes with 3,260 bp and 3,138 bp, respectively (Figure 5.1). The complete circular genomes of BGHBV and TFHBV have typical hepadnaviral organization, comprising three overlapping reading frames that encode the core, polymerase and surface proteins (Figure 5.1). Interestingly, an X protein homolog was not detected in the fish and amphibian hepadnaviruses in this study, and this protein is known to be absent in *Avihepadnavirus*. Consequently, these data confirmed that the X protein is a distinctive feature of the mammal-infecting orthohepadnaviruses (Kew 2011; van Hemert et al. 2011).

The hepadnavirus core gene encodes phosphoproteins that are assembled into subviral capsids. The core polyproteins are 181 amino acid (aa) (BGHBV) and 266 aa (TFHBV) in length, but were not detected in the partial genome of ACHBV. The two fish hepadnaviruses, BGHBV and the recently described WSHBV, encode some of the shortest core proteins among known hepadnaviruses (Table 5.4). The BGHBV core protein shares 24% aa identity with WSHBV, 37% aa identity with TFHBV, and 32 to 44% aa identity with avian and mammalian hepadnaviruses. Similarly, TFHBV shares 31% aa identity with WSHBV, 37% aa identity with BGHBV, and 24 to 36% aa identity with avian and mammalian hepadnaviruses. The C-terminals of BGHBV and TFHBV both contain an arginine-rich domain, a hallmark of hepadnavirus core

proteins, which contains a signal for nuclear transport required for pregenome encapsidation (Yeh et al. 1990; Nassal 1992).

The polymerase gene encodes the viral DNA polymerase, the sole enzyme produced by hepadnaviruses. This gene, with lengths of 781 aa (BGHBV), 744 aa (TFHBV), and 828 aa (ACHBV), covers over half the hepadnavirus genome and its open reading frame overlaps with that of the core and surface proteins. The proteins from these three viruses share 23-42% aa sequence identity among themselves and other hepadnaviruses (Table 5.4). The newly identified amphibian and fish hepadnavirus polymerase genes contain several conserved domains homologous to known avi- and ortho- hepadnaviruses, including the viral DNA polymerase C (pfam00336) and N (PSSM-ID 249709) termini and the reverse transcriptase LTR (PSSM-ID 238825) (Figure 5.3). Mammalian orthohepadnaviruses contain an expanded reverse transcriptase domain with more than 40 additional amino acids. Strikingly, such an expansion was also observed in BGHBV, but not in other fish (WSHBV and ACHBV), amphibian (TFHBV), or avian hepadnaviruses, thereby supporting the phylogenetic analysis that shows BGHBV shares common ancestry with the mammalian hepadnaviruses (see below). In contrast, an expansion of the viral DNA polymerase N-terminal domain was only observed in mammalian orthohepadnaviruses (Figure 5.3).

In known ortho- and avi- hepadnaviruses, the surface polyprotein gene encodes three integral transmembrane envelope glycoproteins; S, M and L. The surface polyproteins were 328 aa and 443 aa in length for BGHBV and TFHBV, respectively, but were not detected in the partial ACHBV genome (Figure 5.1). The amphibian TFHBV contains the largest PreS/S gene of all known hepadnaviruses, encoding a 490 aa protein. However, an alternative start codon was also detected which will produce a shorter, 361 aa protein. The surface proteins from BGHBV

and TFHBV share 34 % aa identity between themselves and 28-39 % aa identity to other hepadnaviruses (Table 5.4).

Conserved core motifs in vertebrate HBVs

Characterization of the prototype fish and amphibian hepadnaviruses allowed us to identify family-wise conserved domains in the core protein. Besides the arginine-rich domain, several conserved motifs were identified among all avian, mammalian, fish and amphibian hepadnaviruses. These include Core motif I, LPXD(F/Y)FPXXXXX(V/L), Core motif II, WXHXX(S/C)(L/I)X(W/F)G, and Core motif III, WXXTPXXYRXXXAPX(I/L) (Figure 5.4). Although Core motif I is close to the N terminus, while Motif II and III are close to the C terminus, all three motifs are in close proximity with each other when the capsid dimers are assembled in the protein model (Figure 5.4B - D). In a typical HBV, two monomers associate to give a compact dimer in which the two a-helical hairpins form a four-helix bundle (Figure 5.4B) (Wynne et al. 1999). Residues at the antigenic sites located near the major immunodominant tips of the four-helix bundle, and residues that made up the four helix bundles are not conserved among vertebrate HBVs.

In exogenous HBVs from the four classes of vertebrate, the three motifs (I, II, and III) contain a total of 15 fully conserved residues. Three additional residues, including the start codon, Asp-4, and His-47, are also conserved, but not included in these motifs (positions as of HuHBV). By visualizing the core protein using the well-established HuHBV model, all three motifs are located at the base of the capsid monomer (Figure 5.4B), containing hydrophobic residues essential for folding of the capsid monomer (Wynne et al. 1999). In HuHBV and other orthohepadnaviruses, the Cys-61 residues of two capsid monomers form a disulfide bond to each other at the dimer interface (Yamada et al. 2008); the same Cys(C) disulfide bonds are also

found in WSHBV and BGHBV. Instead of Cys, Thr(T) is found in avihepadnaviruses, and His(H) is found in TFHBV in the homologous position. One possible explanation for the lack of Cys residue conservation among all vertebrate HBVs is that it is not essential for dimer or capsid formation as evident by mutagenesis studies of this residue (Nassal et al. 1992).

Motif III is likely also important for the interactions between capsid subunits in five-fold and two-fold axes as it contains the proline-rich loop (128–136 in HuHBV) essential for such interactions (Wynne et al. 1999). In particular, Tyr-132 and Pro-129 are both conserved in all vertebrate HBVs investigated. In crystallized protein, Tyr-132 is fully buried in the capsid which is important for proper capsid folding, while Pro-129 is important in inter-subunit packaging (Wynne et al. 1999).

Membrane proteins

The TMHMM analysis predicted that the fish hepadnaviruses (BGHBV and WSHBV) have a membrane protein folding and topology similar to the known model for orthohepadnaviruses and avihepadnaviruses (Figure 5.5) (Bruss 2004). The C-terminal region is hydrophobic and is most likely embedded in host membranes. Two additional hydrophobic domains were detected, forming a hairpin structure with a cytosolic loop (Eble et al. 1986). Alternative start codon positions were detected for TFHBV PreS/S ORF, resulting in putative envelop proteins L of 490 and 361 amino acids. While analysis of the smaller L protein of TFHBV suggested that it might fold in agreement with the other HBVs, the longer L protein was predicted to have an additional transmembrane domain near the N-terminus, potentially forming an extra loop in the ER lumen and exposing the N-terminus in the cytosol (Figure 5.5B). Start codon usage and putative membrane folding predictions clearly need to be experimentally confirmed.

Evolutionary analysis of hepadnaviruses

Phylogenetic analysis of ACHBV, BGHBV and TFHBV, along with representative exogenous and endogenous hepadnaviruses from mammals, birds, reptiles and fish (WSHBV), was performed to determine their relationships and evolutionary history. Although the (Gblocks cleansed) sequence alignments of the polymerase, core and surface genes are necessarily short, they are consistent in clearly showing that the three fish hepadnaviruses do not form a monophyletic group (Figure 5.6). While ACHBV and WSHBV fell in divergent phylogenetic positions, both exhibiting very long branches, BGHBV is clearly more closely related to mammalian viruses of the genus *Orthohepadnavirus*, a relationship supported by a high level of bootstrap support (93-100%). Importantly, although the location of the root of these phylogenies is uncertain, no rooting position would force the fish viruses to be monophyletic. In contrast, the amphibian TFHBV sequence was most closely related to the endogenous hepadnaviruses sampled from crocodillians in the P and the C genes, with 73% and 80% bootstrap support, respectively. The sequences of the surface genes of these endogenous viruses were unavailable for comparison.

Discussion

Relatively little is known about the host range and evolutionary history of the *Hepadnaviridae*. Until recently, the only described exogenous hepadnaviruses were from mammals and birds, comprising approximately twenty ortho- and avi- hepadnavirus genomes from humans, non-human primates, rodents, bats, and birds. The study describing a hepadnavirus in the white sucker fish (Hahn et al. 2015), and our discovery of the second fish hepadnavirus and the first amphibian hepadnavirus are evidence that hepadnaviruses have a broader host range

than previously appreciated. Indeed, the analysis of these new genomes, as well as previously described exogenous (HBV) and endogenous hepadnavirus (eHBV) (Gilbert et al. 2014; Suh et al 2014) sequences, indicates that the *Hepadnaviridae* have been able to infect all five major groups of vertebrates, namely mammals, birds, reptiles, amphibians and fishes (Table 5.6; Figure 5.6).

Although hepadnaviral EVEs have been described in birds and reptiles (Table 5.6), we found no evidence that the bluegill and cichlid viruses were incorporated into the fish germline or caused lesions. In particular, the confirmation of a circular genome and the presence of the virus in some, but not all, bluegill provides strong evidence against BGHBV being endogenous. Similarly, the absence of ACHBV in the genomes of the *O. ventralis* cichlids examined suggests that it is not an EVE. In addition, sequence analysis of TFHBV revealed no insertion site linking the viral genome to that of the host, and a complete circular genome was identified, again suggesting it constitutes an exogenous virus. Unfortunately, a lack of tissue specimens precluded verification of the presence or absence of the virus in additional frogs.

The genome organization of the fish and amphibian hepadnaviruses is similar to that of orthohepadnaviruses and avihepadnaviruses although, with the exception of the highly conserved functional domains (Gilbert 2014; Suh 2014), the sequence identities between these virus groups is very low (Table 5.4). The polymerase in BGHBV, TFHBV and ACHBV also contained conserved domains, including the viral DNA polymerase C and N termini and the reverse transcriptase LTR. Perhaps of most note was that the expanded reverse transcriptase domain detected in BGHBV, but not in the other fish (WSHBV, ACHBV) or amphibian (TFHBV) viruses, is concordant with the phylogenetic analysis showing that BGHBV shares common ancestry with the mammalian hepadnaviruses.

Our analysis of core (capsid) and surface (membrane) proteins revealed features that unify all vertebrate viruses, identifying conserved core protein residues or membrane protein topography that play an important role of hepadnavirus infection and evolution. First, BGHBV and TFHBV contain an arginine-rich domain at the C-terminals, a hallmark of hepadnavirus core proteins (Figure 5.4) (Yeh et al. 1990, Nassal et al. 1992, Glebe et al. 2007). Second, 18 residues in the core proteins were fully conserved among examined vertebrate HBVs. Core motif I, II, and III account for 15 of those conserved residues. The majority of the conserved residues are located at the hydrophobic core of the capsid, while residues at the antigenic tips, as well as the four-helix bundle, are not conserved at all. Based on protein models, the core motifs are conserved, probably because they play key roles in the formation of capsid monomer, dimer, and in the inter-subunit interactions (Wynne et al. 1999). Structural constraints to maintain proper capsid formation seems to be a key force in hepadnaviral core capsid evolution. Since Phe-23 and Trp-102 in motif I and II are important for the interaction with a replication inhibitor drug (Klumpp et al. 2015), further analysis of conserved residues could be worthy of investigation as anti-viral targets.

The new fish and amphibian hepadnaviruses contain three hydrophobic domains similar to ortho- and avihepadnavirus. Therefore, it appears that all vertebrate hepadnaviruses share membrane protein topology similar to those of orthohepadnavirus (Figure 5.5). The second half of the surface protein contains more conserved residues than the N terminus, probably due to conserved transmembrane residues.

The X gene encodes a soluble cytoplasmic X protein that is required for efficient infection by orthohepadnaviruses *in vivo* (Zoulim et al. 1994). Although it is suspected to be involved in the generation of tumors in chronic hepadnaviral infections in humans and

woodchucks, its exact role in the viral replication cycle is not known (Feitelson et al. 2007; Fourel et al. 1990; Fourel et al. 1994; Hansen et al. 1993; Sung et al. 2012; Wen at al. 2008). The presence or absence of an X protein represents a major genomic difference between the ortho- and avihepadnaviruses. Notably, an X protein homolog was not identified in the fish or amphibian viral genomes, providing further evidence that the X protein is a distinctive feature of orthohepadnaviruses in mammalian hosts (Kew et al. 2011; venHemert et al. 2011), and that it evolved by overprinting in these taxa only (Suh et al. 2013; van Hemert et al. 2011). This is consistent with the absence of a detectable X gene in lower vertebrate species, including fish, amphibians and birds.

In addition to increasing our understanding of their genome structure and host range, the data presented here sheds important new light on hepadnavirus evolution. ACHBV falls deep in all the phylogenetic trees and has a common ancestry with the white sucker virus in the polymerase gene tree (albeit with low bootstrap support), observations that are compatible with the long-term co-divergence of hepadnaviruses with their vertebrate hosts over time-scales spanning hundreds of millions of years. However, our data also provide compelling evidence for cross-species transmission. First, although the data is tentative, the single amphibian virus (TFHBV) is clearly most closely related to the eHBVs from crocodilians, whereas strict virus-host co-divergence should place TFHBV as the sister-group to viruses from reptiles, birds and mammals. Far more dramatic, however, was the observation that the bluegill virus (BGHBV) formed a strongly supported monophyletic group with the mammalian orthohepadnaviruses in all three gene trees (Figure 5.6). While this is consistent with the shared presence of an expanded reverse transcriptase domain among these taxa, BGHBV differs from the orthohepadnaviruses in that it lacks both the expansion in the polymerase N-terminal domain and the X protein.

That BGHBV falls as the sister-group to the mammalian hepadnaviruses suggests a far more complex evolutionary history than that of strict virus-host co-divergence, such that multiple species jumps need to be involved. Indeed, it is striking that the fish hepadnaviruses do not form a monophyletic group. While the precise history of these species jumps is difficult to determine, the most parsimonious scenario is that fish harbor an extensive diversity of hepadnaviruses, evident in the long branches leading to ACHBV and WSHBV, and that one of these lineages, represented by BGHBV, jumped to terrestrial vertebrates giving rise to the mammalian orthohepadnaviruses that circulate today. If so, this would be one of the few cases in which viruses have jumped such a wide taxonomic distance. Alternatively, it is possible that BGHBV represents a successful spill-back lineage from terrestrial vertebrates to fish, although this again requires a species jump across a substantial phylogenetic distance. Which of these, or other, evolutionary scenarios is correct will require a far greater sampling of vertebrate hepadnaviruses.

This study shows that fish carry a remarkable diversity of hepadnaviruses, one of which forms a sister-group to mammalian hepadnaviruses. Although the evolution of this important group of viruses is uncertain, a clear prediction from the current study is that there are many more vertebrate hepadnaviruses to be discovered, particularly in species where there has been little active surveillance to date. For example, the observation of a hepadnavirus in a frog suggests there could be additional undiscovered hepadnaviruses with unknown significance to the health of amphibian populations. Increased viral surveillance is especially important as amphibian populations continue to decline as a result of infectious disease and habitat loss (56-60). Finally, the increasing detection of hepadnaviruses in fish (BGHBV and WSHBV) clearly warrants additional investigations to further elucidate their host range and potential pathogenic effects.

Funding

ECH is funded by an NHMRC Australia Fellowship (AF30).

Acknowledgements

We thank Eric Delwart and Beatrix Kapusinszky at the University of California, San Francisco and the Blood Systems Research Institute for assistance with sequencing.

References

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

Baldo L, Santos ME, Salzburger W. 2011. Comparative transcriptomics of Eastern African cichlid fishes shows signs of positive selection and a large contribution of untranslated regions to genetic diversity. Genome Biol Evol 3:443-455.

Bruss V. 2004 Envelopment of the hepatitis B virus nucleocapsid. Virus Res 106:199-209.

Cui J, Zhao W, Huang Z, Jarvis ED, Gilbert MT, Walker PJ, Holmes EC, Zhang G. 2014. Low frequency of paleoviral infiltration across the avian phylogeny. Genome Biol 15:539.

Deng X, Naccache SN, Ng T, Federman S, Li L, Chiu CY, Delwart EL. 2015. An ensemble strategy that significantly improves de novo assembly of microbial genomes from metagenomic next-generation sequencing data. Nucleic Acids Res 43:e46.

Drexler JF, Geipel A, König A, Corman VM, van Riel D, Leijten LM, Bremer CM, Rasche A, Cottontail VM, Maganga GD, Schlegel M. 2013. Bats carry pathogenic hepadnaviruses antigenically related to hepatitis B virus and capable of infecting human hepatocytes. Proc Natl Acad Sci U S A 110:16151-16156.

Eble BE, Lingappa VR, Ganem D. 1986. Hepatitis B surface antigen: an unusual secreted protein initially synthesized as a transmembrane polypeptide. Mol Cell Biol 6:1454-63.

Edgar RC. 2004. MUSCLE: multiple sequence alignment with high accuracy and high throughput. Nucleic Acids Res 32:1792-1797.

Fauquet CM, Mayo MA, Maniloff J, Desselberger U, Ball LA, editors. 2005. *Hepadnaviridae,* p. 373–384 *In* Virus taxonomy: VIIIth report of the International Committee on Taxonomy of Viruses. Academic Press, San Diego, CA.

Feitelson MA, Lee J. 2007. Hepatitis B virus integration, fragile sites, and hepatocarcinogenesis. Cancer Lett 252:157-170.

Flint SJ, Enquist LW, Racaniello VR, Skalka AM, Barnum DR, de Evaluación E. 2000. Principles of virology: molecular biology, pathogenesis and control. ASM Press, Washington DC.

Fourel G, Couturier J, Wei Y, Apiou F, Tiollais P, Buendia MA.1994. Evidence for long-range oncogene activation by hepadnavirus insertion. EMBO J 13:2526.

Fourel G, Trepo C, Bougueleret L, Henglein B, Ponzetto A, Tiollais P, Buendia MA. 1990. Frequent activation of N-myc genes by hepadnavirus insertion in woodchuck liver tumours. Nature 347: 294 - 298 (Letter).

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

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

Glebe D, Urban S. 2007. Viral and cellular determinants involved in hepadnaviral entry. World J Gastroenterol 13:22.

Guindon S, Dufayard JF, Lefort V, Anisimova M, Hordijk W, Gascuel O. 2010. New algorithms and methods to estimate maximum-likelihood phylogenies: assessing the performance of PhyML 3.0. Syst Biol 59:307-21.

Hahn CM, Iwanowicz LR, Cornman RS, Conway CM, Winton JR, Blazer VS. 2015. Characterization of a Novel Hepadnavirus in the White Sucker (*Catostomus commersonii*) from the Great Lakes Region of the United States. J Virol 89:11801-11.

Hansen LJ, Tennant BC, Seeger CH, Ganem D. 1993. Differential activation of myc gene family members in hepatic carcinogenesis by closely related hepatitis B viruses. Mol Cell Biol 13:659-67.

Joklik, WK, Phi D. 1980. Principles of animal virology. Appleton-Century-Crofts, New York, NY.

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

Kew MC. 2011. Hepatitis B virus x protein in the pathogenesis of hepatitis B virus-induced hepatocellular carcinoma. J Gastroenterol Hepatol 26:144-52.

Kodama KA, Ogasawara NA, Yoshikawa HI, Murakami SE. 1985. Nucleotide sequence of a cloned woodchuck hepatitis virus genome: evolutional relationship between hepadnaviruses. J Virol 56:978-86.

Krogh A, Larsson B, Von Heijne G, Sonnhammer EL. 2001. Predicting transmembrane protein topology with a hidden Markov model: application to complete genomes. J Mol Biol 305:567-80.

McGinnis S, Madden TL. 2004. BLAST: at the core of a powerful and diverse set of sequence analysis tools. Nucleic Acids Res 32: W20-5.

Möller S, Croning MD, Apweiler R. 2001. Evaluation of methods for the prediction of membrane spanning regions. Bioinformatics 17:646-53.

Morita M, Awata S, Yorifuji M, Ota K, Kohda M, Ochi H. 2014. Bower-building behaviour is associated with increased sperm longevity in Tanganyikan cichlids. J Evol Biol 27:2629-43.

Muhire BM, Varsani A, Martin DP. 2014. SDT: a virus classification tool based on pairwise sequence alignment and identity calculation. PLoS One 9:e108277.

Nassal M. 1992. The arginine-rich domain of the hepatitis B virus core protein is required for pregenome encapsidation and productive viral positive-strand DNA synthesis but not for virus assembly. J Virol 66:4107-16.

Ng TF, Driscoll C, Carlos MP, Prioleau A, Schmieder R, Dwivedi B, Wong J, Cha Y, Head S, Breitbart M, Delwart E. 2013. Distinct lineage of vesiculovirus from big brown bats, United States. Emerg Infect Dis 19:1978-80.

Ng TF, Kondov NO, Deng X, Van Eenennaam A, Neibergs HL, Delwart E. 2015. A metagenomics and case-control study to identify viruses associated with bovine respiratory disease. J Virol 89:5340-9.

Ng TF, Marine R, Wang C, Simmonds P, Kapusinszky B, Bodhidatta L, Oderinde BS, Wommack KE, Delwart E. 2012. High variety of known and new RNA and DNA viruses of diverse origins in untreated sewage. J Virol 86:12161-75.

Orito E, Mizokami M, Ina Y, Moriyama EN, Kameshima N, Yamamoto M, Gojobori T. 1989. Host-independent evolution and a genetic classification of the hepadnavirus family based on nucleotide sequences. Proc Natl Acad Sci 86:7059-62.

Prassolov A, Hohenberg H, Kalinina T, Schneider C, Cova L, Krone O, Frölich K, Will H, Sirma H. 2003. New hepatitis B virus of cranes that has an unexpected broad host range. J Virol 77:1964-76.

Robertson BH, Margolis HS. 2002. Primate hepatitis B viruses–genetic diversity, geography and evolution. Rev Med Virol 12:133-41.

Saif Y. 2008. Viral diseases, p 405 -448. *In* Diseases of Poultry, Vol. 12. Blackwell Publishing, Ames, IA.

Seeger C, Zoulim F, Mason WS. 2013. Hepadnaviruses, p 2185 – 2221. *In* Knipe DM, Howley PM (ed), Fields Virology, 6th ed. Wolters Kluwer Health/Lippincott Williams & Wilkins, Philadelphia, PA.

Siddiqui AL, Marion PL, Robinson WS. 1981. Ground squirrel hepatitis virus DNA: molecular cloning and comparison with hepatitis B virus DNA. J Virol 38:393-397.

Starkman SE, MacDonald DM, Lewis JC, Holmes EC, Simmonds P. 2003. Geographic and species association of hepatitis B virus genotypes in non-human primates. Virology 314:381-393.

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

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:e1004559.

Summers J, Smolec JM, Snyder R. 1978. A virus similar to human hepatitis B virus associated with hepatitis and hepatoma in woodchucks. Proc Natl Acad Sci 75: 4533-4537.

Sun YB, Xiong ZJ, Xiang XY, Liu SP, Zhou WW, Tu XL, Zhong L, Wang L, Wu DD, Zhang BL, Zhu CL. 2015. Whole-genome sequence of the Tibetan frog *Nanorana parkeri* and the comparative evolution of tetrapod genomes. Proc Natl Acad Sci U S A 112:E1257-E1262.

Sung WK, Zheng H, Li S, Chen R, Liu X, Li Y, Lee NP, Lee WH, Ariyaratne PN, Tennakoon C, Mulawadi FH. 2012. Genome-wide survey of recurrent HBV integration in hepatocellular carcinoma. Nat Genet 44:765-769.

Talavera G, Castresana J. 2007. Improvement of phylogenies after removing divergent and ambiguously aligned blocks from protein sequence alignments. Syst Biol 56:564-577.

van Hemert FJ, van de Klundert MA, Lukashov VV, Kootstra NA, Berkhout B, Zaaijer HL. 2011. Protein X of hepatitis B virus: origin and structure similarity with the central domain of DNA glycosylase. PloS one 6:e23392.

Victoria JG, Kapoor A, Dupuis K, Schnurr DP, Delwart EL. 2008. Rapid identification of known and new RNA viruses from animal tissues. PLoS Pathog 4:e1000163.

Voyles BA. 1993. The biology of viruses. Mosby, St. Louis, Missouri.

Wen Y, Golubkov VS, Strongin AY, Jiang W, Reed JC. 2008. Interaction of hepatitis B viral oncoprotein with cellular target HBXIP dysregulates centrosome dynamics and mitotic spindle formation. J Biol Chem 283:2793-2803.

Yeh CT, Liaw YF, Ou JH. 1990. The arginine-rich domain of hepatitis B virus precore and core proteins contains a signal for nuclear transport. J Virol 64:6141-6147.

Zoulim F, Saputelli J, Seeger C. 1994. Woodchuck hepatitis virus X protein is required for viral infection in vivo. J Virol 68:2026-2030.

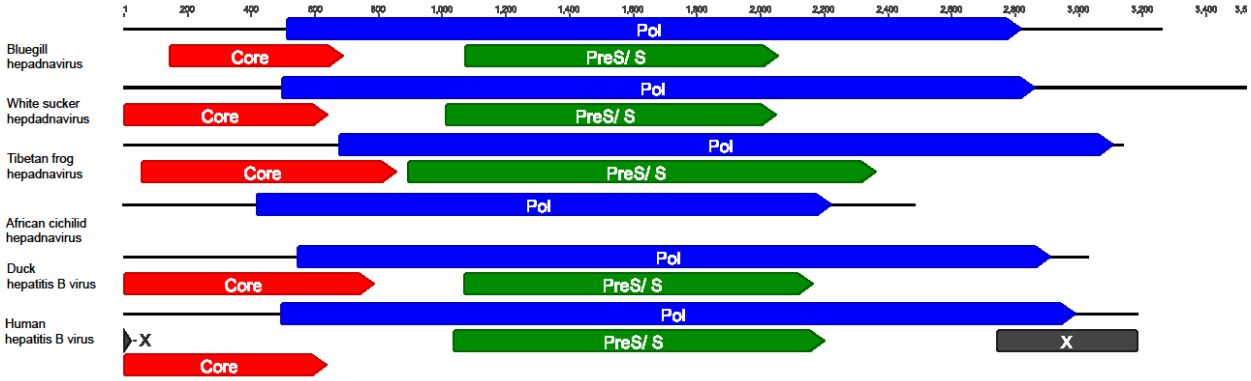


Figure 5.1 Genome organization of the hepadnaviruses. Open reading frames encoding the polymerase (Pol), core, surface (PreS/S), and X proteins are indicated by colors. Circular genomes are linearized, with the exception of the partial sequence of the African cichlid hepadnavirus (ACHBV).

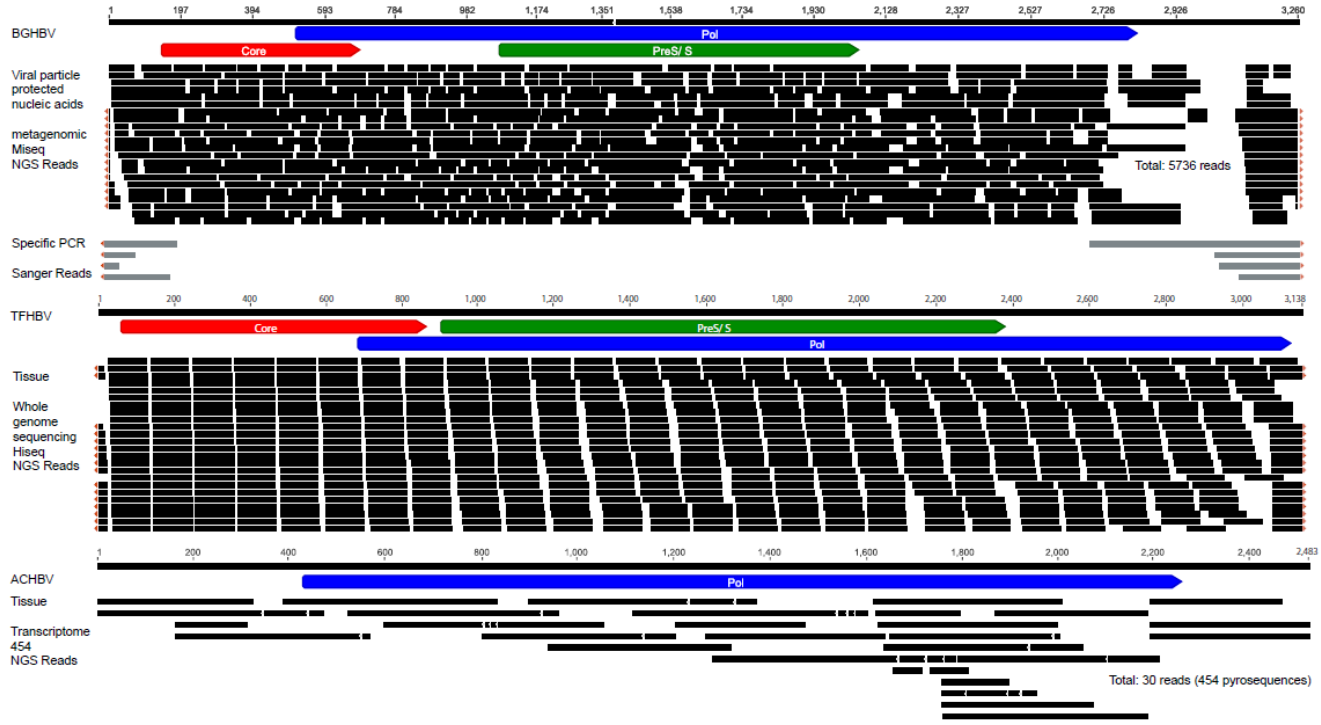


Figure 5.2 Coverage map for the bluegill hepadnavirus (BGHBV), Tibetan frog hepadnavirus (TFHBV) and African cichlid hepadnavirus-like sequence (ACHBV). Circular genomes of BGHBV and TFHBV are linearized, and sequence coverage over 15 reads are collapsed for display purpose. The overlapping sequences confirming the circular nature of the genomes are annotated with small orange triangles.

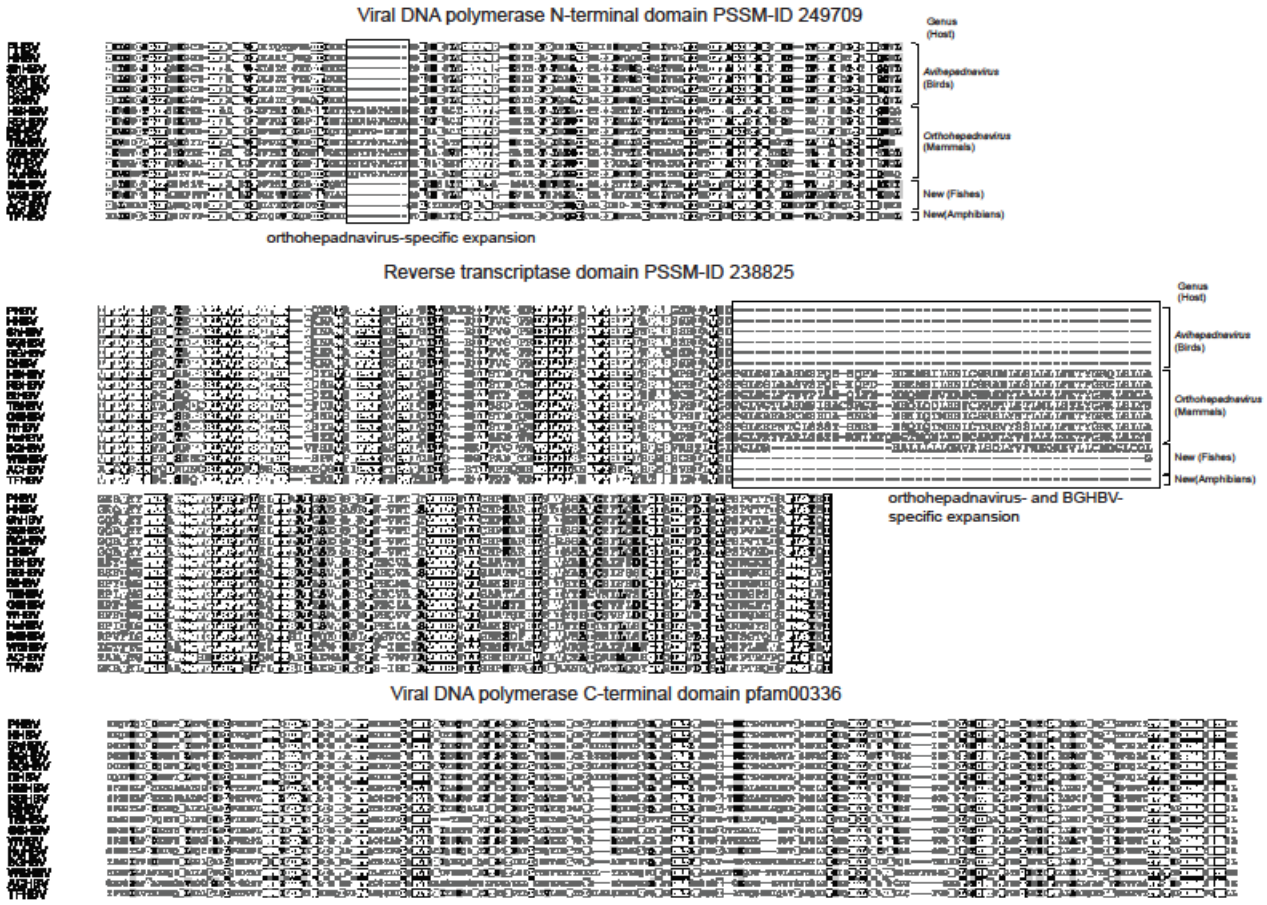


Figure 5.3 Conserved motifs in the polymerase protein of mammal, avian, fish and amphibian hepadnaviruses. An expanded reverse transcriptase domain is evident in mammalian orthohepadnaviruses and BGHBV. The expansion of the viral DNA polymerase N-terminal domain was only observed in mammalian orthohepadnaviruses. Bluegill HBV (BGHBV), white sucker HBV (WSHBV), Tibetan frog HBV (TFHBV), parrot HBV (PHBV), heron HBV (HHBV), sheldgoose HBV (ShHBV), snow goose HBV (SGHBV), Ross's goose HBV (RGHBV), duck HBV (DHBV), horseshoe bat HBV (HBHBV), roundleaf bat HBV (RBHBV), bat HBV (BtHBV), tent-making bat HBV (TBHBV), ground squirrel HBV (GSHBV), woodchuck HBV (WHBV), and human HBV (HuHBV).

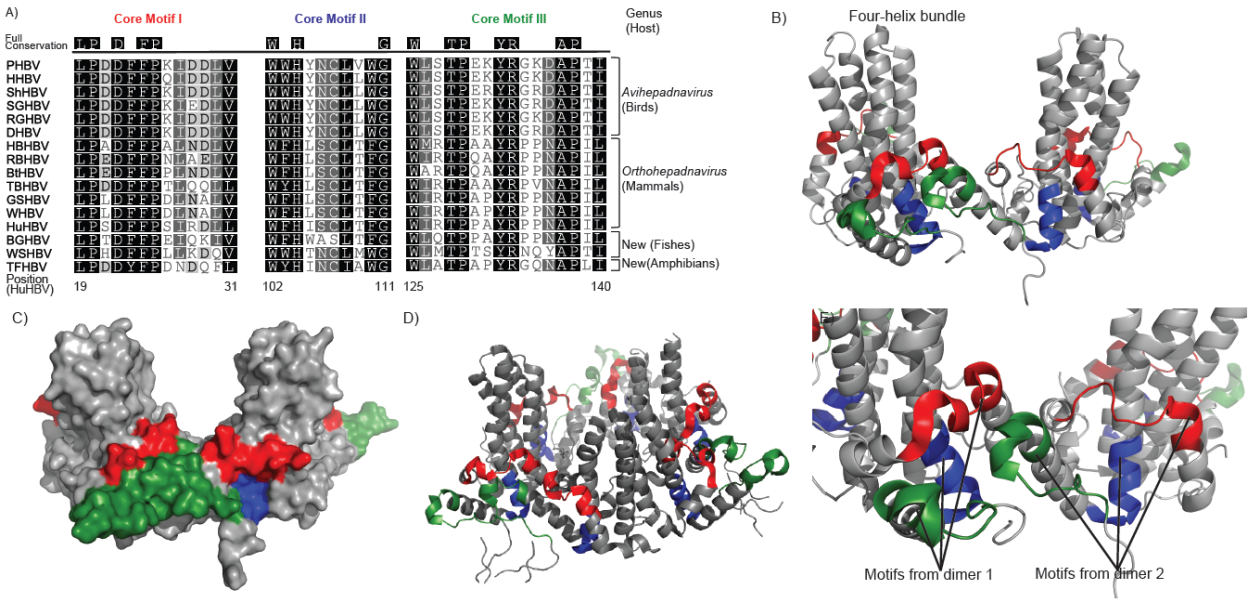


Figure 5.4 Conserved motifs in the core protein of mammal, avian, fish and amphibian hepadnaviruses. A) Amino acid sequence alignment of the three conserved motif in the core proteins. Position is indicated for HuHBV protein (43) B) Motif I, II and III (Red, Blue and Green respectively) in the capsid protein dimer using HuHBV as model. C) Motif locations in the surface representation of the capsid dimer. D) and E) Homo hexamer representation showing the proximity of the motifs between subunits. Acronyms are indicated in FIG. 3. Accession numbers of included HBV protein sequences are listed in Table 1.

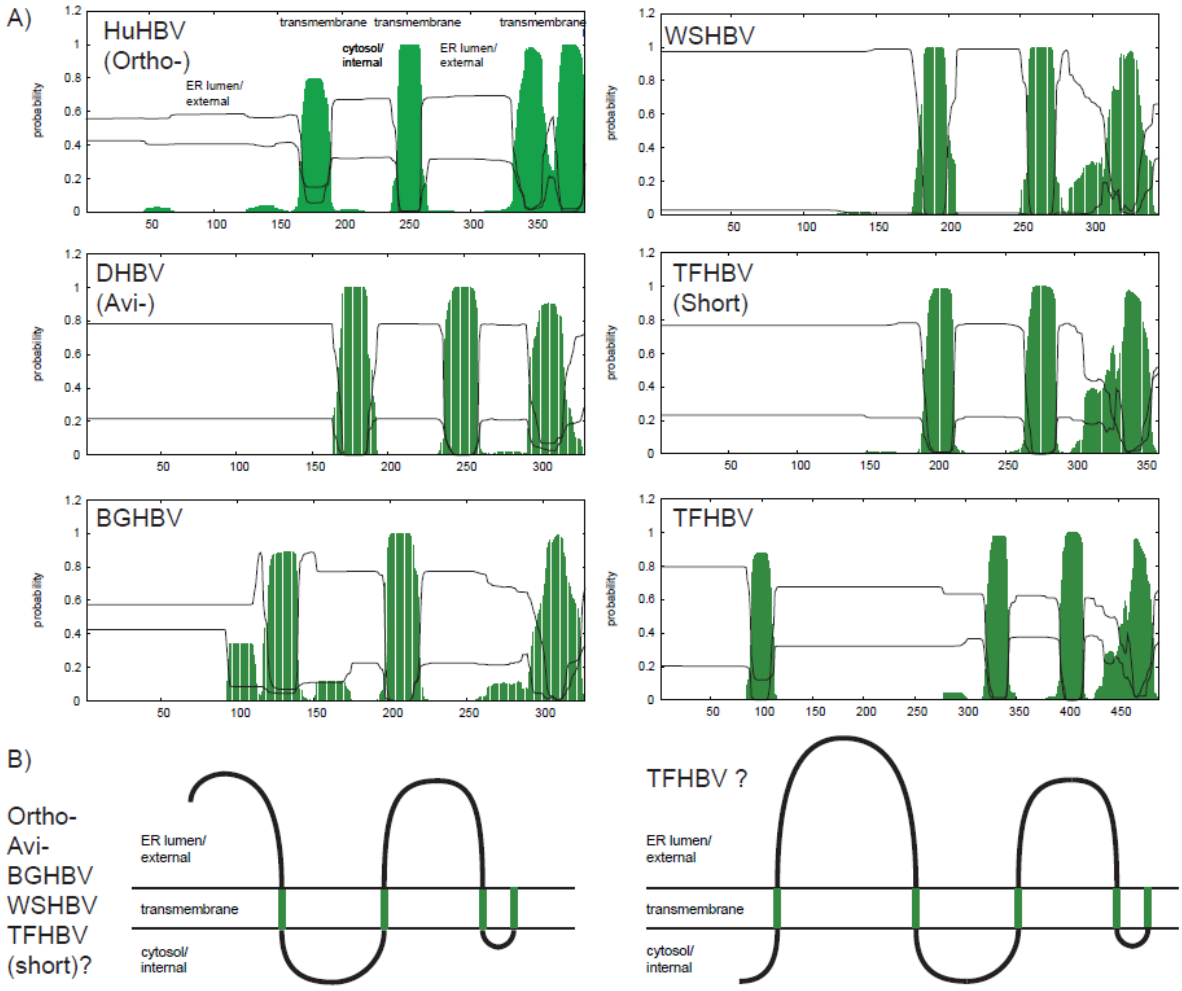


Figure 5.5 Membrane protein analysis. A) TMHMM analysis for the fish hepadnaviruses, bluegill HBV (BGHBV) and white sucker HBV (WSHBV), as well as Tibetan frog HBV (TFHBV) with known models for orthohepadnaviruses (human HBV, HuHBV) and avihepadnaviruses (duck HBV, DHBV) (35). B) Membrane protein folding and topology compared to the established model of orthohepadnavirus (46). Since alternative start codon positions resulting in different length envelope proteins were detected for TFHBV, the analysis was performed on both.

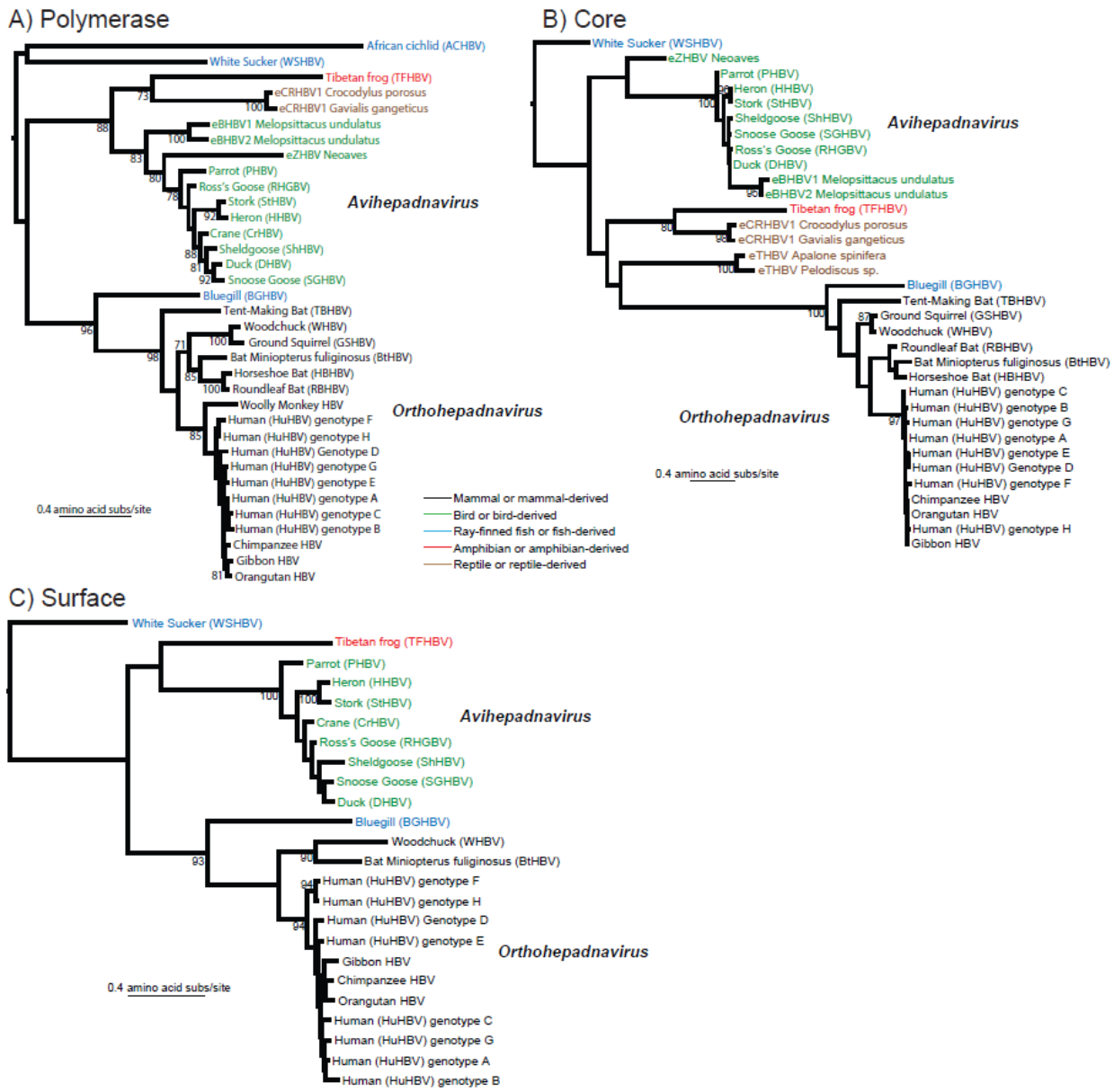


Figure 5.6 Maximum likelihood phylogenetic trees of the (A) Polymerase, (B) Core, and (C) Surface genes of exogenous and endogenous (e) vertebrate hepadnaviruses. Viruses are color-coded to reflect their host group of origin. All trees are drawn to a scale of amino acid substitutions per site (scale bars shown) and rooted on the fish (WSHBV and where available ACHBV) sequences as (i) these are the most divergent and (ii) this rooting position maximises the extent of virus-host co-divergence. Bootstrap support values >70% are shown for relevant nodes.

Table 5.1 Molecular screening of BGHBV in bluegill and related *Lepomis* species. Lip papilloma and skin lesion in the individual animal is indicated. Triplicate qPCR analysis values in femtograms (standard deviation). na – sample not available for evaluation. + lesion or nucleic acid present. – lesion or nucleic acid not identified in tissue. *denotes initial bluegill lip sample used for next generation sequencing and metagenomic analysis.

	Fish	Species	Lip Papilloma	Skin Lesion	Hepadnavirus PCR Positive		
					Lip	Skin	Liver
Aquarium 9-9-09							
1	GAI-1	*L. macrochirus*	+	-	-	na	na
2	GAI-2	*L. macrochirus*	+	-	+ 13.4 (4)	na	na
3	GAI-3	*L. macrochirus*	+	-	-	na	na
4	GAI-4	*L. macrochirus*	+	-	-	na	na
5	GAI-5	*L. macrochirus*	+	-	- 0.00187 (0.0009)	na	na
Waleska, GA 4-14-14							
6	WA-A1	*L. macrochirus*	-	+	nr	na	na
7	WA-A2	*L. macrochirus*	+	-	-	na	na
8	WA-A3	*L. macrochirus*	-	-	nr	na	na
9	WA-A4	*L. macrochirus*	+	-	-	na	na
10	WA-A5	*L. macrochirus*	-	+	-	-	na
Waleska, GA 7-7-14							
11	WA-B1	*L. macrochirus*	-	-	- 0.0003 (0.00005)	na	- 0 (0)
12	WA-B2	*L. macrochirus*	-	-	-	na	-
13	WA-B3	*L. macrochirus*	-	-	+	na	na
14	WA-B4	*L. macrochirus*	-	-	nr	-	na
15	WA-B5	*L. macrochirus*	-	-	+ 5.81 (1.42)	+ 111 (10)	+ 0.179 (0.0475)
16	WA-B6	*L. macrochirus*	+	-	-	na	na
17	WA-B7	*L. macrochirus*	+	-	-	na	na
18	WA-B8	*L. macrochirus*	-	-	-	na	na
19	WA-B9	*L. macrochirus*	+	-	+	na	na
20	WA-B10	*L. macrochirus*	+	-	-	na	-
21	WA-B11	*L. macrochirus*	+	-	-	na	na
22	WA-B12	*L. macrochirus*	+	-	+	+ 146 (12.9)	+ 0.0076 (0.00024)
23	WA-B13	*L. macrochirus*	+	-	-	na	na
Waleska GA 7-4-15							

24	WA-C1	*L. macrochirus*	+	-	-	na 0.0012 (0.001)	na
25	WA-C2	*L. macrochirus*	-	-	-	na	na
26	WA-C3	*L. macrochirus*	+	-	-	-	-
27	WA-C4	*L. macrochirus*	+	-	-	na	na
28	WA-C5	*L. macrochirus*	+	-	-	na	-
29	WA-C6	*L. macrochirus*	-	-	-	na	na
30	WA-C7	*L. macrochirus*	-	-	+	+	+
31	WA-C8	*L. macrochirus*	-	-	-	na	-
32	WA-C9	*M.. salmoides*	-	-	-	na	-
Hawkinsville, GA 1-16-15							
33	OW-1	*L. macrochirus*	-	-	- 0 (0)	na	- 0.012 (0.0046)
34	OW-2	*L. macrochirus*	-	-	-	na	na
35	OW-3	*L. macrochirus*	-	-	-	-	na
36	OW-4	*L. macrochirus*	-	-	-	na	na
37	OW-5	*L. macrochirus*	-	-	-	na	na
38	OW-6	*L. microlophus*	-	-	-	-	-
39	OW-7	*L. microlophus*	-	-	-	na	-
40	OW-8	*L. auritus*	-	-	-	-	-
41	OW-9	*L. auritus*	-	-	-	na	-
Athens, GA 9-1-15							
42	SC-1	*L. macrochirus*	-	-	-	na	na
43	SC-2	*L. macrochirus*	-	-	na	na	-
44	SC-3	*L. macrochirus*	-	-	-	na	na
45	SC-4	*L. macrochirus*	-	-	-	na	na
46	SC-5	*M. cyanellus*	-	-	-	na	-

Table 5.2 Targeted gene, primer sequences and product size for bluegill hepadnavirus (BGHBV) and African cichlid hepadnavirus-like sequence (ACHBV). *denotes primers used to assess the presence of BGHBV in fish surveys via end-point PCR. ** denotes primers designed for real time PCR.

Bluegill Hepadnavirus Primers					
Gene	Primer		Primer		Product Size
Core	CoreF	GACCAAATTGACTCGGCTGT	CoreR	ATTTGGTCCACCAGCCATAA	327
Polyemerase & Capsid	BGHBV-PolF*	TGTGGACAAAAATCCACGAA	BGHBV-PolR*	ATGCCCATAGGTGCTTTACG	387
Polyemerase & Capsid	PolNestF	CACCACACTTGCCAACAAAC	PolNestR	TGCTCCCAGAACACGTACAG	287
Circle	BGHBV-CirF	CAACGCCAACAGCATTTTTA	BGHBV-CirR	CGCAGTCTCGACCGATATTA	301
Polymerase	PolQpcrF**	CCTGGCTCTGTTCGTCATACT	PolNestR**	TGCTCCCAGAACACGTACAG	110

African cichlids Hepadnavirus Primers					
Gene	Primer		Primer		Product Size
Polymerase	ACHBV-PolF	TGGGCATTCAACACAAAAGA	ACHBV-PolR	GCGTGCATGACCTCTGAGTA	302
CytochromeB	OVCytBF	TGACGCACTTGTTGACCTTC	OBCytBR	GGAGAACGTAGCCCACAAAA	300

Table 5.3 GenBank Accession number for the sequences used in the phylogenetic analysis. "-" denotes not applicable.

Virus	C	P	S
ACHBV	-	This study	-
BGHBV	This study	This study	This study
TFHBV	This study	This study	This study
WSHBV	AKT95193	AKT95195	AKT95194
eCRHBV1 Crocodylus porosus	From Suh et al. 2014	Suh et al.	-
eCRHBV1 Gavialis gangeticus	From Suh et al. 2014	From Suh et al. 2014	-
eBHBV1 Melopsittacus undulatus	From Suh et al. 2014	From Suh et al. 2014	-
eBHBV2 Melopsittacus undulatus	From Suh et al. 2014	From Suh et al. 2014	-
eTHBV Apalone spinifera	From Suh et al. 2014	-	-
eTHBV Pelodiscus sp.	From Suh et al. 2014	-	-
eZHBV Neoaves	From Suh et al. 2014	From Suh et al. 2014	-
PHBV	YP_004956862	YP_004956864	YP_004956865
RHGBV	AAR89922	YP_024968	YP_024969
StHBV	AJ251934	CAC80820	AJ251934
HHBV	NP_040997	NP_040998	NP_040999
CrHBV	-	CAD29588	CAD29589
ShHBV	YP_024973	YP_024974	YP_024975
DHBV	ADP55743	1803562C	NP_039824
SGHBV	YP_031693	AAD21995	YP_031696
TBHBV	YP_009046002	KC790381	-
WHBV	NP_671816	AAA19183	AAA19182
GSHBV	NP_040993	NP_040994	-
BtHBV	YP_007678002	YP_007677999	YP_007678000
HBHBV	YP_009045998	KC790377	-
RBHBV	YP_009045994	YP_009045991	-
Woolly Monkey HBV	-	AAO74855	-
HuHBV Genotype A	BAD91278	CCK33754	Q4R1S6
HuHBV Genotype B	BAO96185	BAO96176	BAK32999
HuHBV Genotype C	BAU25817	BAO96196	BAQ95566
HuHBV Genotype D	CCH63726	ABC87304	BAJ51643
HuHBV Genotype E	BAD91272	CCK33758	CCK33757
HuHBV Genotype F	CCK33700	CCK86729	CCK33685
HuHBV Genotype G	BAM05705	CCK86644	BAD91285
HuHBV Genotype H	BAF49207	BAN75948	BAN75949
Chimpanzee HBV	P12901	P12900	P12911
Gibbon HBV	P89951	P87744	AAG01444
Orangutan HBV	AAF33123	AAF33121	AAF33124

Suh A, Weber CC, Kehlmaier C, Braun EL, Green RE, et al. (2014) Early Mesozoic Coexistence of Amniotes and Hepadnaviridae. PLoS Genet 10(12): e1004559. doi:10.1371/journal.pgen.1004559

Table 5.4 Percent amino acid identity of bluegill HBV (BGHBV), African cichlid hepadnavirus-like sequence (ACHBV), and Tibetan frog HBV (TFHBV) compared to hepadnaviruses partitioned by open reading frame. Size and GenBank accession numbers of the predicted and published hepadnavirus core, polymerase and surface proteins are also included. White sucker HBV (WSHBV), stork HBV (STHBV), heron HBV (HHBV), parrot HBV (PHBV), duck HBV (DHBV), crane hepatitis B virus (CHBV), snow goose HBV (SGHBV), woodchuck HBV (WHBV), bat HBV (BtHBV), ground squirrel HBV (GSHBV), and human HBV (HuHBV).

	Polymerase Protein					Surface Protein				Core Protein			
Virus	GenBank accession #	size (aa)	% Amino Acid Identity			GenBank accession #	size (aa)	% AA Identity		GenBank accession #	size (aa)	% AA Identity	
			BGHBV	TFHBV	ACHBV			BGHBV	TFHBV			BGHBV	TFHBV
BGHBV	KX058433	781	-	35%	30%	KX058433	328	-	34%	KX058433	181		37%
ACHBV	KX058434	828	30%	25%	-	-	-	-	-	-	-	-	-
WSHBV	AKT95195	789	35%	34%	30%	AKT95194.2	346	39%	31%	AKT95193	213	24%	31%
TFHBV	KX058435	744	35%	-	25%	KX058435	443	34%	-	KX058435	266	37%	-
StHBV	CAC80820	790	36%	36%	27%	AJ251934	337	35%	37%	AJ251934	305	33%	33%
HHBV	NP_040998	788	35%	35%	23%	NP_040999	335	36%	34%	NP_040997	305	32%	35%
PHBV	YP_004956864	795	31%	38%	33%	YP_004956865	375	34%	31%	YP_004956862	305	35%	29%
DHBV	NP_039822	788	35%	36%	25%	NP_039824	330	33%	36%	ADP55743	262	33%	33%
CrHBV	CAD29588	785	35%	37%	25%	CAD29589	327	32%	35%	-	-	-	-
SGHBV	YP_031695	787	36%	37%	25%	YP_031696	329	36%	36%	YP_031693	305	32%	31%
WHBV	NP_671813	884	41%	30%	35%	NP_671814	431	37%	33%	NP_671816	188	42%	32%
BtHBV	YP_007677999	853	40%	33%	29%	YP_007678000	399	35%	28%	YP_007678002	217	43%	24%
GSHBV	NP_040994	881	41%	30%	29%	NP_040995	282	39%	39%	NP_040993	217	44%	35%
HuHBV	NP_647604	843	42%	31%	28%	YP_355333	400	37%	32%	YP_355335	212	43%	36%

Table 5.5 The whole genome sequence data of the Tibetan frog with initial contig (GenBank accession number JYOU01126907) (38). The Tibetan frog data set contained 13 whole genome sequencing (DNA) runs.

Project	Accession	Total Read	TFHBV Reads	% TFHBV Reads
Nanorana parkeri Genome sequencing	SRX514761	374,982,742	4497	0.001199%
Nanorana parkeri Genome sequencing	SRX514762	309,742,412	6200	0.002002%
Nanorana parkeri Genome sequencing	SRX514763	313,197,132	7213	0.002303%
Nanorana parkeri Genome sequencing	SRX514764	389,751,764	1928	0.000495%
Nanorana parkeri Genome sequencing	SRX514765	433,440,774	1694	0.000391%
Nanorana parkeri Genome sequencing	SRX514766	366,201,646	156	0.000043%
Nanorana parkeri Genome sequencing	SRX514767	354,666,118	101	0.000028%
Nanorana parkeri Genome sequencing	SRX514768	273,217,750	322	0.000118%
Nanorana parkeri Genome sequencing	SRX514769	321,698,530	745	0.000232%
Nanorana parkeri Genome sequencing	SRX514770	336,644,766	2128	0.000632%
Nanorana parkeri Genome sequencing	SRX514771	337,281,318	1454	0.000431%
Nanorana parkeri Genome sequencing	SRX514772	302,142,366	4287	0.001419%
Nanorana parkeri Genome sequencing	SRX514773	267,159,900	3657	0.001369%

Table 5.6 Current knowledge of hepanaviral host range and viral life cycle. Endogenous hepadnavirus is absent in mammalian genomes (6). Avian and reptile endogenous hepadnavirus have been described previously (14, 17-19). The fish WSHBV was described in a concurrent study (16).

Host	Exogenous	Exclusively endogenous
Mammal	*Orthohepadanvirus*	n/a
Avian	*Avihepadanvirus*	Avian hepadnavirus EVE
Reptile	n/a	Reptilian hepadnavirus EVE
Fish	WSHBV, BGHBV and ACHBV (this study)	n/a
Amphibian	TFHBV (this study)	n/a

Chapter 6

EPIZOOTIC PAPILLOMATOSIS IN THE BLUEGILL SUNFISH *LEPOMIS MACROCHIRUS*

Dill JA, Williams SM, Ng TFF, Camus AC

To be submitted to Veterinary Pathology or Journal of Fish Disease

Abstract

Skin tumors, particularly papillomas of the lips, have been recognized in fish for over a century. Although the etiologies of epizootic neoplasia in fish are varied, viruses, primarily herpesviruses and retroviruses, have been implicated in some skin neoplasms. This study sought to identify a viral agent as the cause of well-differentiated papillomas of the lips, epidermal hyperplasia of the trunk and fins, and rare squamous cell carcinomas in bluegill (*Lepomis macrochirus*). Histopathological examination provided no insight as to a cause and viral particles were not observed with transmission electron microscopy. Next generation sequencing and metagenomic analysis of lesioned skin revealed partial sequences of retroviral envelope and polymerase genes. Further analysis of sequence data demonstrated continuity of the envelope sequence with the host genome. PCR primers designed against the envelop gene subsequently amplified the sequence in 100% of skin samples from 20 bluegill with lesions and 20 without, as well as in normal tissues from three additional *Lepomis* species, including redear (*L. microlophus*), redbreast (*L. auritus)* and green (*L. cyanellus*) sunfish. PCR failed to amplify the envelope sequence in tissue from a largemouth bass (*Micropterus salmoides*), or in non-centrarchid fish, avian and mammalian samples. Findings suggest an endogenous retroviral element has been integrated into the germlines of multiple *Lepomis* species, throughout the 14.6 million years of their divergence, and is likely not related to tumor formation.

Key words: Bluegill, *Lepomis macrochirus*, papillomas, retrovirus, endogenous viral elements

Introduction

Bluegill (*Lepomis macrochirus*) are familiar centrarchid fish indigenous to static and slow moving bodies of water in eastern North America from Quebec to Mexico. Growing to a maximum length of 40 cm, bluegill are a popular gamefish produced by private, state and federal fish hatcheries for stocking purposes on public and private lands (Manooch and Raver 1991). A high incidence of epithelial hyperplasia and papillomas, with rare progression to squamous cell carcinoma, on the skin, fins and lips of bluegill displayed at a public aquarium and from a private pond in Georgia prompted investigation of a suspected viral etiology.

Orocutaneous neoplasms are common in fish and have been recognized for over a century (Mawdesley-Thomas 1975). As seen in these bluegill, most are described histologically as benign epidermal hyperplasias and papillomas (Roberts 2001). Squamous cell carcinomas in fish are rare (Groff 2004). In commonly affected species, such as the brown bullhead (*Ameiurus melas*), there is evidence to suggest a chemical etiology, while virus or virus-like particles have been demonstrated in many other cases (Grizzle et al. 1984; Coffee et al. 2013; Pinkney et al. 2014). Several tumor associated virus particles have been tentatively identified, based on morphologic features only, using transmission electron microscopy (TEM). These include herpesviruses, retroviruses, adenoviruses, and others (Anders & Yoshimizu 1994; Coffee et al. 2013). While virus isolation has been impeded by a lack of fish cell culture lines, advanced molecular techniques, such as next generation sequencing and *in situ* hybridization methods, have made it increasingly possible to establish more definitive relationships between viral agents and neoplasms in fish.

Oncogenicity has been clearly demonstrated for herpesviruses, notably Salmonid herpesvirus 2 (SalHV-2) and Cyprinid herpesvirus 1 (CyHv-1), which induce papillomas in masu

salmon (*Onchorhynchus masou*) and common carp (*Cyprinus carpio*). These viruses can be transmitted by cohabitation, waterborne challenge, and injection of cell culture filtrates. Both induce high mortalities in juvenile fish and a relatively high proportion of survivors develop tumors months later (Kimura et al. 1981, Sano et al 1985). In salmonids, tumors are exophytic and most develop around the mouth. In contrast, lesions in carp are more common on body surfaces, appearing as slightly raised mucoid plaques (Plumb 2011; Roberts 2001).

The pathology of retroviral associated tumors and hyperplasias in fish, both epithelial and mesenchymal, have been reviewed by Coffee et al. 2013. While some associations are well supported by sequence data and transmission trials, others are based entirely on TEM and the detection of reverse transcriptase activity. The genus *Epsilonretrovirus* contains the complex retroviruses that infect fish (Knipe 2013; Flint et al 2000; Voyles et al. 1993; Kurth et al. 2010). Examples of tumor associated viruses, partially or completely sequenced, include walleye dermal sarcoma virus, walleye epidermal hyperplasia viruses 1 and 2, and perch discrete epidermal hyperplasia viruses 1 and 2 (Bowser et al. 1993; Holzschu et al. 1995; Coffee et al. 2013).

Potentially confounding investigations of suspected tumorigenic retroviruses are the presence of fragmented, partial or complete retroviral genomes, representing previous, often ancient, integrations into host germlines. Retroviruses account for the majority of known endogenous viral elements (EVEs), as host genome integration is essential to their replication (Herniou et al. 1998; Kurth et al. 2010; Dudley et al. 2011). Transmitted vertically, endogenous retroviruses (ERVs) are common in birds and mammals and may comprise 10% of all mammalian genomes. Most are incapable of producing infectious virus, becoming increasingly defective over time, and cause no negative effects. However, functional gene products implicated in cellular proliferation are sometimes transcribed and some may provide genetic diversity to

novel pathogenic retroviruses. Replication and reintegration of recently integrated functional retroviruses may cause neoplasia by insertional mutagenesis of proto-oncogenes or tumor suppressor genes (Weiss 2006; Kurth et al. 2010; Flint et al. 2000; Dudley et al. 2011; Brown et al. 2014). In fish, the first fully sequenced ERV was discovered in zebrafish. The 11.2 kb provirus contains intact open reading frames (ORFs) for the gag, pol, env and LTR sequences (Shen & Steiner 2004).

Materials and Methods

Sample collection

Five bluegill from a mixed species aquarium exhibit were submitted to the Aquatic Pathology Service at the College of Veterinary Medicine, University of Georgia, in 2009, as part of an investigation into an epizootic of orocutaneous papillomas in this species. The origin of these fish could not be traced. Complete necropsies were performed and samples of lesions and major organs were fixed in 10% neutral buffered formalin. Tissues were processed routinely, sectioned at 5 μm, and stained with hematoxylin and eosin (H&E) for histologic evaluation. Portions of lip and skin lesions were collected separately and archived in a -80°C freezer.

Similar lesions were observed on bluegill by a private pond owner in Waleska, Georgia and five fish were submitted for evaluation April 14, 2014. The approximately 1 acre pond was surrounded by a small woodland, but received runoff from adjacent pastureland. Thirteen bluegill, seven with proliferative lip lesions and five without, were submitted July 7, 2014. Eight bluegill, four with lesions and four without, and one largemouth bass (*Micropterus salmoides*) were submitted July 4, 2015. Included in the study were a number of fish from other sources, all lacking lesions. On January 16, 2015, five bluegill, two redbreast sunfish (*Lepomis auritus*) and

two redear sunfish (*Lepomis microlophus*) were received from a commercial hatchery in Hawkinsville, Georgia. Four bluegill and one green sunfish (*Lepomis cyanellus*) were received from local anglers in the Athens, Georgia area September 1, 2015. All fish were processed for histopathology as described above. In addition to lip and skin, pooled samples of liver, spleen and kidney were frozen at -80°C, as well as gonadal tissue from some fish.

Portions of spleen from a goliath grouper (*Epinephelus itajara*), liver from an African penguin (*Spheniscus demersus*), gonad from a female koi (*Cyprinus carpio*), fin from a rummy-nose tetra (*Hemigrammus rhodostomus*) and a cheek swab from Dalmatian dog (*Canis lupus*) were also collected and frozen.

Electron microscopy

Approximately 2 mm cubes of neoplastic tissue from multiple fish were fixed immediately in a cold glutaraldehyde based fixative modified by the addition of picric acid (Karnovsky 1965; McDowell & Trump 1976) and processed routinely for TEM (Bozzola 1992). Ultrathin sections were cut on a Reichert Ultracut S ultramicrotome (Leica, Inc., Deerfield, IL), stained with lead citrate and examined on a JEM-1210 transmission electron microscope (JEOL USA, Inc., Peabody, MA).

Next-generation sequencing and metagenomic analysis

Next generation sequencing (NGS) was performed on tissue from eight bluegill using previously described protocols (Victoria et al. 2008; Ng et al. 2012; Ng et al. 2015). Samples included two lip lesions and pooled skin lesions from archived aquarium fish tissue, one individual and one pooled sample of lip lesions from pond fish, and lip tissue from an unaffected hatchery fish. In brief, a tissue homogenate was centrifuged through a 0.22 µm filter, to enrich for viral particles, then treated with nucleases to deplete host nucleic acids, followed by

sequence-independent amplification using random priming. Nucleic acids from any nuclease-resistant viral particles were extracted using the Qiagen QIA quick viral RNA column purification system. Reverse transcription was performed using a 28-base oligonucleotide whose 3′ end consisted of eight random nucleotides (primer N1_8N, CCTTGAAGGCGGACTGTGAGNNNNNNNN). A second strand was synthesized using Klenow fragment DNA polymerase (New England BioLabs). The resulting double-stranded cDNA and DNA were then PCR amplified using AmpliTaq Gold DNA polymerase and a 20-base primer (primer N1, CCTTGAAGGCGGACTGTGAG). A duel-indexed sequencing library was then prepared using the Nextera XT DNA Sample Prep Kit (Illumina, San Diego, CA). After pooling, the final library was sequence using the MiSeq sequencing system with 2 × 250 bp paired-end sequencing reagents (Illumina MiSeq Reagents V2, 500 cycles).

A total of 870,000 reads were generated and analyzed as previously described (Ng et al. 2012). An in-house analysis pipeline running on a 32-node Linux cluster was used to process the data (University of California, San Francisco). Adaptor and primer sequences were trimmed using VecScreen (McGinnis et al. 2004), while duplicate reads and low-sequencing-quality tails were removed using a Phred quality score of 10 as the threshold. The cleaned reads were *de novo* assembled using an in-house sequence assembler employing an ensemble strategy (Deng et al. 2015) consisting of SOAPdenovo2, ABySS, meta-Velvet, and CAP3. The assembled sequence was compared with an in-house viral protein sequence database using BLASTx. Viral contigs were further inspected manually using Geneious (version R6; Biomatters, Auckland, New Zealand).

Sequence comparisons and phylogenetic analysis

Based on the next generation sequence results, coding sequences of representative retroviral envelop (env) and polymerase (pol) genes were downloaded from GenBank and compared to the bluegill dataset. To be as broad as possible, the background GenBank data set included both exogenous and endogenous sequences that were of sufficient length to conduct phylogenetic analyses, although sequence availability differed by gene, and included zebrafish endogenous retrovirus (AY075045), avian leukosis virus (NC001408), equine foamy virus (NP054716), human immunodeficiency virus type 1 (NC001802), human T-cell leukemia virus type 1 (NC001436), mouse mammary tumor virus (NC001503), porcine endogenous retrovirus (CAC82505), Atlantic salmon swim bladder sarcoma virus (NC007654), walleye dermal sarcoma virus (NP045937), walleye epidermal hyperplasia type 1 (AAD30048), walleye epidermal hyperplasia type 2 (AAC59311) and snakehead retrovirus (NC001724).

Molecular screening

Tissues samples were extracted using Qiagen QIAmp Viral RNA MiniKit and Qiagen DNeasy Blood and Tissue Kits. Screening for the bluegill env and pol sequences was accomplished by end point PCR using a previously described touch down protocol (Ng et al. 2013). Initial targeting of a 160 bp env amplicon was performed with primer sets BF-EnvF-5-CCAATGATAGATGCCCTGCT -3 and BR-EnvR-5- CCAATGATAGATGCCCTGCT-3, while a 100 bp pol amplicon was amplified with primer sets CF-PolF 5-TGCCAGCATCTGTAGAAGACA-3 and CR-PolR 5- CATGTGAAGTTTCCATGTGCT-3. The DNA was electrophoresed on a 2% agarose gel, purified (Qiaquick Gel Extraction Kit) and quantitated (NanoDrop 2000, Thermo Fisher). Detected viral nucleic acids were confirmed by

Sanger sequencing (MCLAB, San Francisco, CA) and the new sequence information was used to design additional primers, with the goal of lengthening and bridging the original NGS fragments.

Samples were initially processed with a Qiagen OneStep RT-PCR kit, but a two-step process, using SuperScript® III Reverse Transcriptase (RT) to obtain cDNA and then One Taq DNA polymerase kit (New England Biolabs), was also undertaken. The inclusion of a sample extracted by the Qiagen QIAmp Viral RNA Mini Kit that did not undergo the 1st RT step was used to determine the type of nucleic acid present in the retroviral sequences. An extraction using the RNeasy Mini Kit with DNase digestion steps, followed by RNA purification was similarly used to verify the type of nucleic acid present in the sequences. Degenerate "generic" retroviral primers used to identify highly conserved genome regions, the reverse transcriptase and protease, in all seven genera of retroviruses (Hernious et al. 1998, Burmesiter et al. 2001) were also attempted, unsuccessfully, on the bluegill.

Results

Gross and histopathologic findings

The verrucous skin masses varied in size and shape, were soft, fleshy, and pale pink to dark gray, with pedunculated to broad based attachment to underlying tissue (Figure 6.1). Lip tumors, interpreted as papillomas, were composed of exophytic, villiform proliferations of well-differentiated squamous epithelial cells. Thin fibrovascular cores supported orderly, 15-20 cell thick, epithelial layers, with scattered goblet cells, resting upon intact basement membranes (Figure 6.1). Larger, more reddened masses were present on the skin and caudal fin of one bluegill (Figure 6.1) and operculum of a second. In contrast to the benign papillomas, epithelial cells in superficial areas of these masses were moderately disorganized and formed branching

and anastomosing cords and trabeculae that infiltrated the dermis and underlying skeletal muscle. Epithelial cells in these squamous cell carcinomas had variably distinct cell borders and exhibited moderate anisocytosis and anisokaryosis. Nuclei in papillomas were finely stippled and euchromatic, while the carcinomas had vacuolated nuclei, with marginated, hyperbasophilic chromatin, and a single prominent nucleolus. In both tumor types, mitoses were rare and necrotic cells and lymphocytes were scattered throughout.

Electron microscopy

Viral particles were not observed in papillomas from five bluegill (Figure 6.2).

Viral metagenomics of endogenous retrovirus gene sequences

To circumvent the lack of known viral genetic information in teleosts and paucity of available cell lines, a sequence-independent metagenomic approach was performed in an attempt to identify viral genetic material within neoplasms. Accordingly, portions of retroviral env and pol gene sequences (to be deposited in GenBank) were identified in next generation sequence data (Table 6.1, Figure 6.3). Primers were chosen based on deep sequencing genomic results to target the envelope and polymerase.

Molecular screening in *Lepomis* species

At least one tissue from all 45 *Lepomis* spp. fish was screened and found positive for the env gene sequence, regardless of their origin and independent of lesion presence or absence (Table 6.2). The env sequence was not amplified in the other fish, avian or mammalian samples tested. Longer sequences of envelope were obtained by PCR and Sanger methods, but the env and pol genes could not be linked and additional genes could not be identified. The positive bands on agarose gels for the extraction samples run with the two-step process without the essential RT step and the loss of PCR bands once DNases were employed, strongly suggests that

the env and pol sequences were composed of DNA, not RNA. Additional analysis of the next generation sequence data revealed the env sequence to be directly linked to the host genome in two fish and in close proximity to host DNA in the remaining four samples (Figure 6.3).

Sequence comparisons and phylogenetic analysis

Phylogenetic relationships of the predicted env and pol gene products were compared with nucleotide sequences in GenBank using BLAST searches. The translated envelope sequence had 53% identity to the envelope of Atlantic salmon swim bladder sarcoma virus (YP_443923.1) and 61% identity to the predicated envelope of an endogenous retrovirus in *Austrofundulus limnaeus* (XP_013886217.1). No similarities were identified for the nucleotide or translated pol sequence. Neighbor-joining phylogenetic trees were estimated using Geneious software and published exogenous and endogenous retrovirus sequences. The closest branches for the env and pol sequences were Atlantic salmon swim bladder sarcoma virus and zebra fish endogenous retrovirus, respectively.

Discussion

Although relatively common, a cause has not been established for many epizootics of orocutaneous neoplasia in fish. In addition to viral causes, environmental contaminants and the cumulative effects of trauma have all been advanced as possibly associated, potentially multifactorial, influences involved in tumor development (Smith et al. 1989, Groff 2004, Pinkney et al. 2014). Despite negative electron microscopy findings, a metagenomic approach was undertaken in an attempt to elucidate a viral agent as the cause of orocutaneous neoplasms in these two isolated bluegill populations. Although NGS did not reveal a complete genome, retroviral env and pol gene sequences were discovered in 100% of lesioned and non-lesioned

skin samples from bluegill, pooled organ samples from non-lesioned bluegill and non-lesioned skin from three additional *Lepomis* species using endpoint PCR. In addition, elimination of env and pol bands from gels using DNase clean up kits on reverse transcriptase (RT) preparations indicates a host DNA genomic, rather than exogenous RNA virus, origin. This conclusion was further supported by NGS data demonstrating integration of the env gene sequence into the host genome. These findings are most consistent with an endogenous viral element (EVE) permanently integrated into the germlines of multiple *Lepomis* spp. Failure to link the env and pol genes suggests the ERV is fragmented and likely nonfunctional.

Endogenous retroviruses have been identified in almost all vertebrate genomes. While most are defective, some do remain intact (Shen & Steiner 2004). In fish, fragments derived from endogenous retroviral elements have been identified, but they exhibit extensive mutations and deletions and are unlikely to generate functional proteins (Herniou et al. 1998). The presence of the EVE in the four *Lepomis* spp. tested, and its absence in the largemouth bass, indicates a long association with the lepomids. *Micropterus* and *Lepomis* are sister-taxa that diverged an estimated 24.81 mya (Near et al. 2004). The first divergence within the genus *Lepomis*, which separates *L. macrochirus* and *L. cyanellus* from *L. microlophus* and *L. auritus* is estimated to have occurred 14.64 mya (Near et al. 2004), suggesting the ERV has been evolving with the genomes of these species for at least that amount of time.

The cause of neoplasia in these bluegill remains undetermined. While NGS data largely rules out the possibility of a viral etiology, the potential effects of anthropogenic contaminants remain a consideration. Historical information and the samples provided on both fish groups were limited. In the aquarium population, tumors persisted for over three years under constant environmental conditions. Although the pond owner reported the presence of tumors year round,

samples were only obtained in spring and summer months. No fish additions had been made to the pond in over 15 years. No chemical analysis was performed, but the pond was subject to agricultural runoff, suggesting xenobiotic exposures could be involved.

References

Anders K, Yoshimizu M. 1994. Role of viruses in the induction of skin tumors and tumor-like proliferations of fish. Dis Aquat Organ 19:215-232.

Bowser PR, Casey JW. 1993. Retroviruses of fish. Annual Review of Fish Diseases 3:209-224.

Bozzola JJ. 1992. Electron Microscopy: Principles and Techniques for Biologists. Jones and Bartlett, Sudbury, MA.

Burmeister T, Schwartz S, Thiel E. 2001. A PCR primer system for detecting oncoretroviruses based on conserved DNA sequence motifs of animal retroviruses and its application to human leukaemias and lymphomas. J Gen Virol 82:2205-2013.

Coffee LL, Casey JW, Bowser PR. 2013. Pathology of tumors in fish associated with retroviruses: A review. Vet Pathol 50:390-403.

Deng X, Naccache SN, Ng T, Federman S, Li L, Chiu CY, Delwart EL. 2015. An ensemble strategy that significantly improves de novo assembly of microbial genomes from metagenomic next-generation sequencing data. Nucleic Acids Res 43:e46.

Dudley J (ed). 2011. Retroviruses and Insights into Cancer. Springer Science & Business Media, Austin, TX.

Flint SJ, Enquist LW, Racaniello VR, Skalka AM, Barnum DR, de Evaluación E. 2000. Principles of Virology: Molecular Biology, Pathogenesis and. ASM Press, Washington DC.

Grizzle JM, Melius P, Strength DR. 1984. Papillomas on fish exposed to chlorinated wastewater effluent. J Natl Cancer Inst 73:1133-1142.

Groff JM. 2004. Neoplasia in fishes. Vet Clin Exot Anim 7:705-756.

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

Holzschu DL, Martineau D, Fodor SK, Vogt VM, Bowser PR, Casey JW. 1995. Nucleotide sequence and protein analysis of a complex piscine retrovirus, walleye dermal sarcoma virus. J Virol 69:5320-5331.

Karnovsky, M.J. 1965. A formaldehyde-glutaraldehyde fixative of high osmolarity for use in electron microscopy. J Cell Biol 27: 137A.

Kimura T, Yoshimizu M, Tanaka M. 1981. Studies on a new virus (OMV) from *Onchorhynchus masou*–II. Onchogenic nature. Fish Pathol 15:149-153.

Knipe DM, Howley PM (ed). 2013. Fields Virology, 6th ed. Wolters Kluwer Health/Lippincott Williams & Wilkins, Philadelphia, PA.

Kurth R, Bannert N (eds). 2010. Retroviruses: molecular biology, genomics and pathogenesis. Caister Academic Press, Norfolk, UK.

Manooch CS, Raver D. 1991. In: Fishes of the Southeastern United States. North Carolina State Museum of Natural History, Raleigh, NC, pp 70-71.

Mawdsley-Thomas LE. 1975. Neoplasia in fish, p 805-870. *In* Ribelin WE, Migaki G (eds), The Pathology of Fishes. University of Wisconsin Press, Madison,WI.

McDowell EM, Trump BF. 1976. Histologic fixatives suitable for diagnostic light and electron microscopy. Arch Pathol Lab Med 100: 405.

McGinnis S, Madden TL. 2004. BLAST: at the core of a powerful and diverse set of sequence analysis tools. Nucleic Acids Res 32: W20-25.

Near TJ, Bolnick DI, Wainwright PC. 2004. Investigating phylogenetic relationships of sunfishes and black basses (Actinopterygii: Centrarchidae) using DNA sequences from mitochondrial and nuclear genes. Mol Phylogenet Evol 32:344-357.

Ng TF, Marine R, Wang C, Simmonds P, Kapusinszky B, Bodhidatta L, Oderinde BS, Wommack KE, Delwart E. 2012. High variety of known and new RNA and DNA viruses of diverse origins in untreated sewage. J Virol 86:12161-12175.

Ng TF, Driscoll C, Carlos MP, Prioleau A, Schmieder R, Dwivedi B, Wong J, Cha Y, Head S, Breitbart M, Delwart E. 2013. Distinct lineage of vesiculovirus from big brown bats, United States. Emerg Infect Dis 19:1978-1980.

Ng TF, Kondov NO, Deng X, Van Eenennaam A, Neibergs HL, Delwart E. 2015. A metagenomics and case-control study to identify viruses associated with bovine respiratory disease. J Virol 89:5340-5349.

Pinkney AE, Harshbarger JC, Rutter MA. 2014. Temporal and spatial patterns in tumor prevalence in brown bullhead *Ameiurus nebulosus* (Lesueur) in the tidal Potomac River watershed (USA). J Fish Dis 37:863-876.

Plumb JA. 2011. Trout and salmon viruses, p 147-217. *In* Plumb JA, Hanson LA (eds) Health Maintenance and Principal Microbial Diseases of Cultured Fishes, 3rd ed. Iowa State University Press, Ames, IA.

Roberts RJ. 2012. The virology of teleosts, p 186-291. *In* Roberts RJ (ed) Fish pathology, 4th ed. WB Saunders, Edinburgh, Scotland.

Sano T, Fukuda H, Furukawa M. 1985. *Herpesvirus cyprini*: Biological and oncogenic properties. Fish Pathol 20:381-388.

Shen CH, Steiner LA. 2004. Genome structure and thymic expression of an endogenous retrovirus in zebrafish. J Virol 78:899-911.

Smith IR, Ferguson HW, Hayes MA. 1989. Histopathology and prevalence of epidermal papillomas epidemic in brown bullhead, *Ictalurus nebulosus* (Lesueur), and white sucker, *Catostomus commersoni*, (Lacépède), populations from Ontaria, Canada. J Fish Dis 12:373-388.

Victoria JG, Kapoor A, Dupuis K, Schnurr DP, Delwart EL. 2008. Rapid identification of known and new RNA viruses from animal tissues. PLoS Pathog 4:e1000163.

Voyles, BA. 1993. The Biology of Viruses. Mosby, St. Louis, MO.

Weiss RA. 2006. The discovery of endogenous retroviruses. Retrovirology 3:67. doi:10.1186/1742-4690-3-67

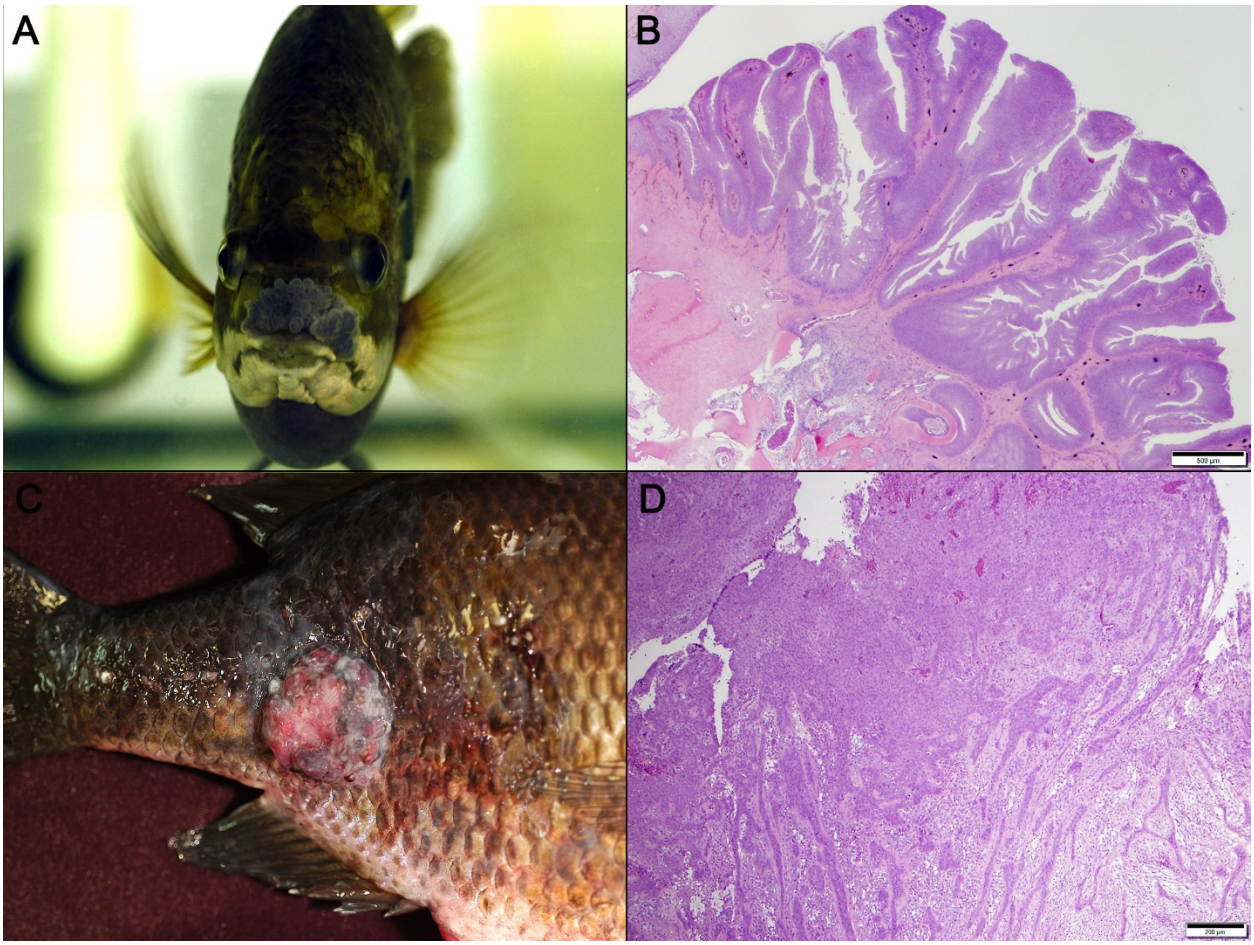


Figure 6.1 Gross and photomicrographic images of bluegill (*Lepomis macrochirus*) skin neoplasms. A) Gross verrucous papilloma on upper and lower lips. B) Micrograph of typical benign papilloma, with exophytic, branching, well-differentiated epithelial fronds supported by scant fibrovascular stroma. H&E stain, Bar = 500 µm. C) Gross image of large squamous cell carcinoma on trunk. D) Micrograph of squamous cell carcinoma with thick dysplastic epidermis and extensive invasion of the dermis by neoplastic epithelial cords and trabeculae. H&E stain, Bar = 200 µm.

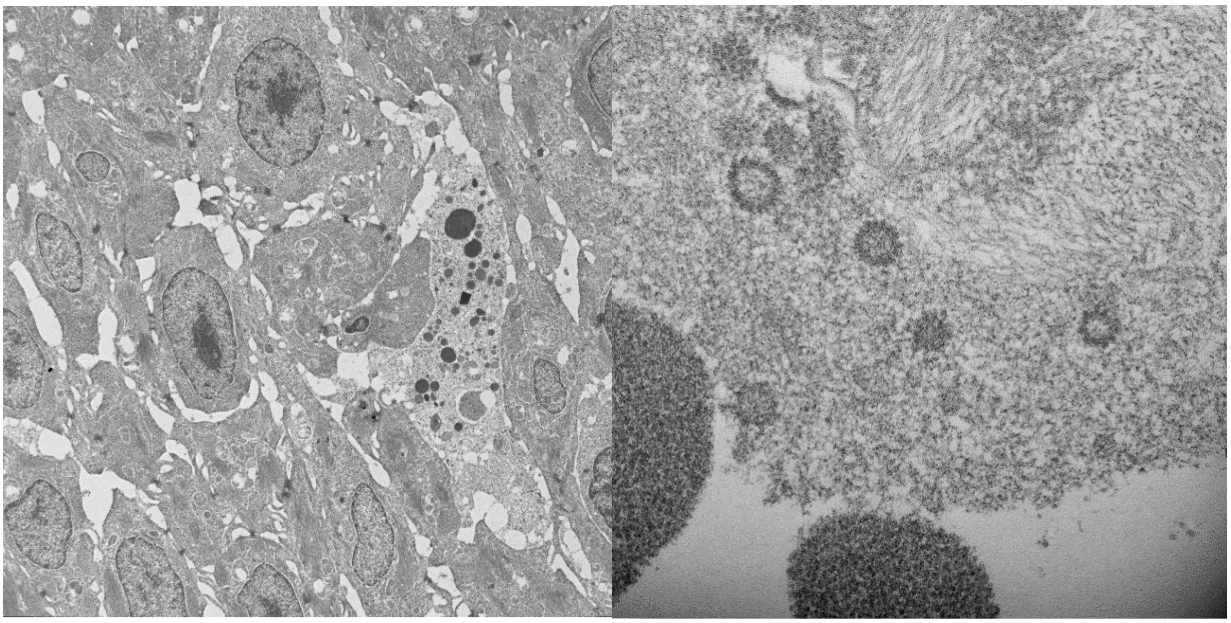


Figure 6.2 Transmission electron micrographs of lip papillomas from bluegill (*Lepomis macrochirus*). In rare epithelial cell nuclei, structural elements showed size and ultrastructural characteristics that resembled type-c retrovirus virions, but were confirmed nonviral.

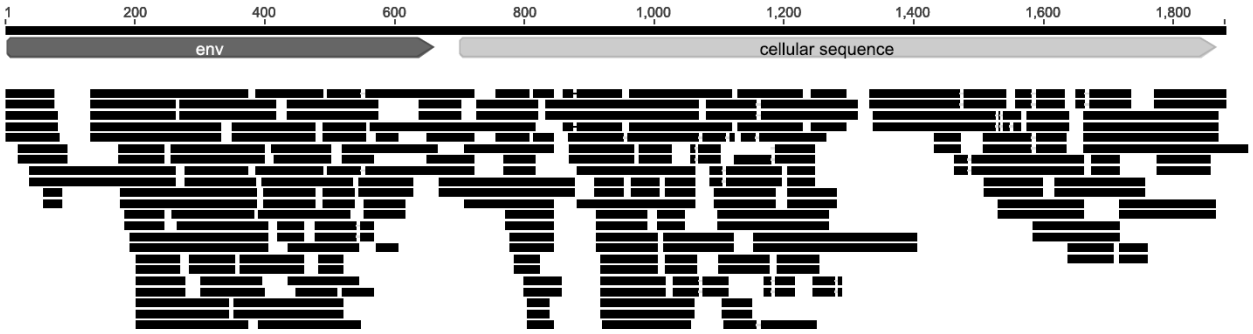
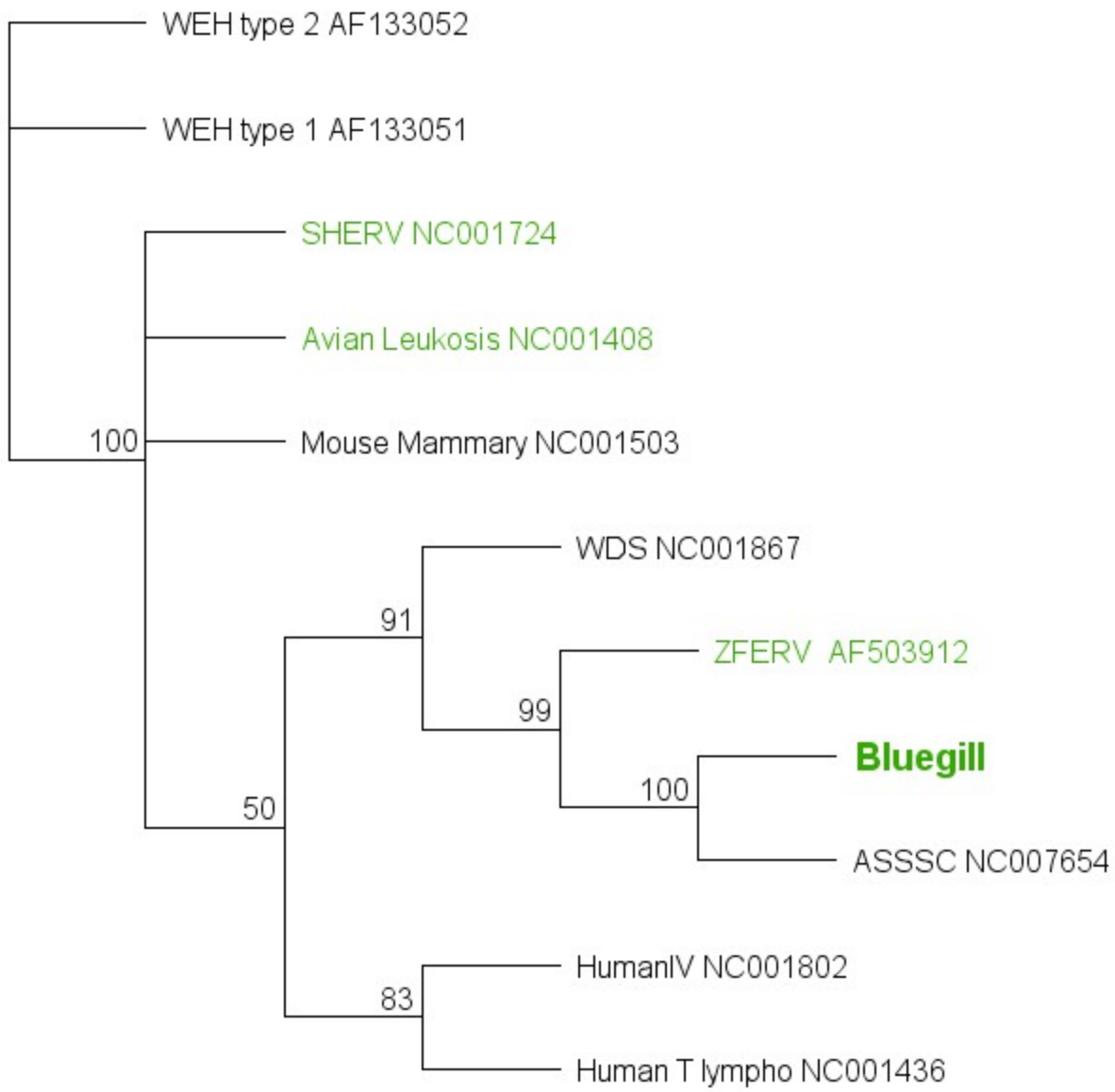
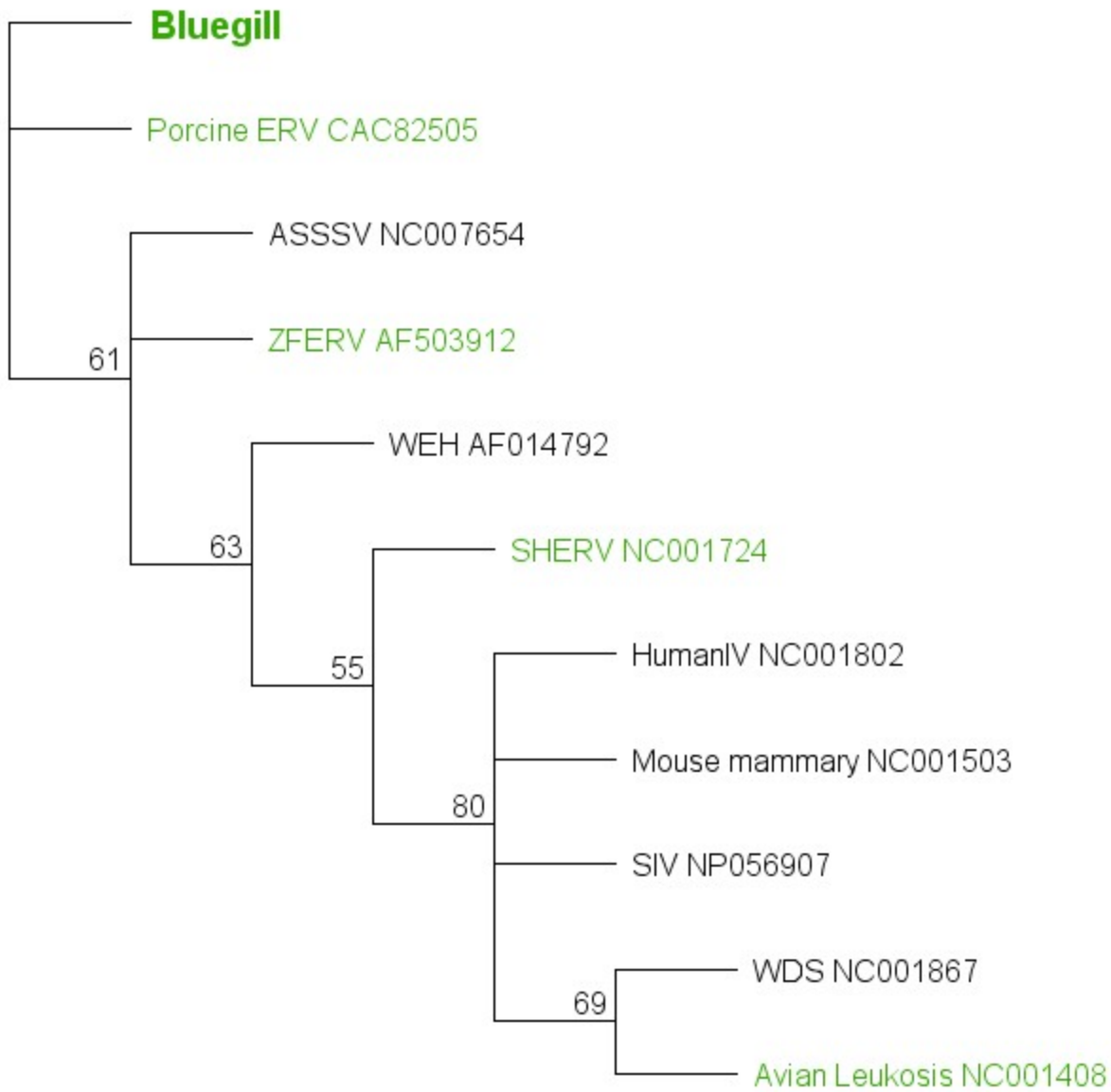


Figure 6.3 Coverage map for the bluegill envelope retrovirus sequence. Sequence coverage over 15 reads are collapsed for display purpose. This molecular data also supports that the bluegill envelope sequence (env) is endogenous as there is no break between viral genes and host (cellular sequence).


WEH type 2 AF133052
WEH type 1 AF133051
SHERV NC001724
Avian Leukosis NC001408
100
Mouse Mammary NC001503
WDS NC001867
91
ZFERV AF503912
99
Bluegill
100
50
ASSSC NC007654
HumanIV NC001802
83
Human T lympho NC001436

a)



b)

Figure 6.4 Neighbor-joining phylogenetic trees of the a) envelope and b) pol gene of exogenous and endogenous (green) vertebrate retroviruses. Zebra fish endogenous retrovirus (ZFERV, AY075045), avian leukosis virus (ALV, NC001408), human immunodeficiency virus type 1 (HIV, NC001802), human T-cell leukemia virus type 1 (NC001436), mouse mammary tumor virus (MMTV, NC001503), simian immunodeficiency virus (SIV, NP056907) porcine endogenous retrovirus (PERV, CAC82505), Atlantic salmon swim bladder sarcoma virus (ASSSC, NC007654), walleye dermal sarcoma virus (WDS, NP045937), walleye epidermal hyperplasia type 1 (WEH1, AAD30048), walleye epidermal hyperplasia type 2 (WEH2, AAC59311) and snakehead retrovirus (SHERV, NC001724).

Table 6.1 After analysis of the viral hits from the 870,000 NGS reads the bluegill retroviral sequence top match hits by BLASTx. The sequence hits to the polymerase and envelope gene of fish retroviruses.

>@s24146_BG_hit_to_Atlantic_salmon_swim_bladder_sarcoma_virus_Env_pos_9000
CACTACAACGTTCAGAAACTAGGAAATTGGACCCAAAGCGGTTTCGAGGCCATTCA
TGACCAACTTGCTGCCACCTCTCT
CATGGCGTTCCAGAACCGAATAGCCATAGATATGGTA

>@s35021_BG_hit_to_sacroma_Walleye_epidermal_hyperplasia_virus_1_Pol
GAGCATAAGACAGTATCCATTAAAAGAAGAAGCACAAGAAGGAATAAAACCAGTA
ATAGAAGATTTGCTTAAAGCAGGAG
TAATAATGAAATGTGAAGACTCCCCTTGTAATATTCCAATCTTCTGCGAATGTCCAA
CCATCTATTCAC

>@s57378_BG_hit_to_Walleye_epidermal_hyperplasia_virus_1_pol_4400
GCCAATGATAGATGCCCTGCTGGTAAAAGGAGTGTTGAAAGAAACGACGAGCTCGT
GTAATACGCCGATATTTCCGATAA
AAAAAGCAGGAAGAGAGGAGTATAGGATGATA

>@s61699_BG_hit_to_Sacroma_Pol_6000
GGATTTCTCACAGCAGGAAATCAGCCAATCAAACATGAAGAAGGAATGAAGGAACT
GGCAGAGGCCTTGCTCGTTCACAG
TGAAGTCGCAGTTGTTAAGTGCAAAGGACACGA

>@s68847_BG_hit_to_Atlantic_salmon_swim_bladder_sarcoma_virus_Env_8500
CTCCTACAAGTAGTGCCAGCATCTGTAGAAGACAGTTGTGCTATTGACCTGATGAAT
AATACCAATCCTAAGAAAAGCTG
TCAGAAATGGGATTCGGTGTTCCCTGTTGTTGCAGCAGACAAAGAGAAACCTTTATT
CTCTAAAAGAGTAGCACATGGAA
ACTTCACATGCATAAAT

Table 6.2 Molecular screening of retrovirus envelope in bluegill and related *Lepomis* species. Lip papilloma and skin lesion in the individual animal is indicated. na – sample not available for evaluation. * - fish used for deep sequencing.

	Fish	Species	Lip Papilloma	Skin Lesion	PCR Positive			
					Lip	Liver	Skin	Gonad
Aquarium 9-9-09								
1	GAI-1*	*L. macrochirus*	+	-	+	na	na	na
2	GAI-2*	*L. macrochirus*	+	-	+	na	na	na
3	GAI-3*	*L. macrochirus*	+	-	na	na	na	na
4	GAI-4*	*L. macrochirus*	+	-	+	na	na	na
5	GAI-5	*L. macrochirus*	+	-	+	na	na	na
Waleska, GA 4-14-14								
6	WA-A1	*L. macrochirus*	-	+	+	na	na	na
7	WA-A2	*L. macrochirus*	+	-	+	na	na	na
8	WA-A3*	*L. macrochirus*	-	-	na	na	na	na
9	WA-A4*	*L. macrochirus*	+	-	+	na	na	na
10	WA-A5*	*L. macrochirus*	-	+	+	na	+	na
Waleska, GA 7-7-14								
11	WA-B1	*L. macrochirus*	-	-	+	+	na	na
12	WA-B2	*L. macrochirus*	-	-	+	na	na	na
13	WA-B3	*L. macrochirus*	-	-	+	+	na	na
14	WA-B4	*L. macrochirus*	-	-	+	na	na	na
15	WA-B5	*L. macrochirus*	-	-	+	na	na	na
16	WA-B6	*L. macrochirus*	+	-	+	na	na	na
17	WA-B7	*L. macrochirus*	+	-	+	na	na	na
18	WA-B8	*L. macrochirus*	-	-	+	na	na	na
19	WA-B9	*L. macrochirus*	+	-	+	na	na	na
20	WA-B10	*L. macrochirus*	+	-	+	+	+	na
21	WA-B11*	*L. macrochirus*	+	-	+	na	na	na
22	WA-B12	*L. macrochirus*	+	-	+	+	+	na
23	WA-B13	*L. macrochirus*	+	-	+	na	na	na
Waleska GA 7-4-15								
24	WA-C1	*L. macrochirus*	+	-	+	na	na	na
25	WA-C2	*L. macrochirus*	-	-	+	na	na	na
26	WA-C3	*L. macrochirus*	+	-	+	+	na	na

27	WA-C4	*L. macrochirus*	+	-	+	na	na	na
28	WA-C5	*L. macrochirus*	+	-	+	na	na	na
29	WA-C6	*L. macrochirus*	-	-	+	na	na	na
30	WA-C7	*L. macrochirus*	-	-	+	+	+	+
31	WA-C8	*L. macrochirus*	-	-	+	na	na	na
32	WA-C9	*M. salmoides*	-	-	-	na	na	na
Hawkinsville, GA 1-16-15								
33	OW-1*	*L. macrochirus*	-	-	+	+	+	na
34	OW-2	*L. macrochirus*	-	-	+	+	na	na
35	OW-3	*L. macrochirus*	-	-	+	+	na	+
36	OW-4	*L. macrochirus*	-	-	+	na	+	+
37	OW-5	*L. macrochirus*	-	-	+	na	na	na
38	OW-6	*L. microlophus*	-	-	+	+	+	+
39	OW-7	*L. microlophus*	-	-	+	+	+	+
40	OW-8	*L. auritus*	-	-	+	+	+	+
41	OW-9	*L. auritus*	-	-	+	+	+	+
Athens, GA 9-1-15								
42	SC-1	*L. macrochirus*	-	-	+	na	na	na
43	SC-2	*L. macrochirus*	-	-	+	+	na	na
44	SC-3	*L. macrochirus*	-	-	+	na	na	na
45	SC-4	*L. macrochirus*	-	-	+	na	na	na
46	SC-5	*M. cyanellus*	-	-	+	+	na	na

Chapter 7

PERSPECTIVES/CONCLUSION

In this research the pathological, genetic, and phylogenic characterization of four novel, emerging viruses was described. The potential role of these viral agents in the induction of proliferative skin lesions in giant guitarfish (*Rhynchobatus djiddensis)* and bluegill (*Lepomis macrochirus)* was investigated by designing and using molecular and histopathologic methods. To ensure the future of commercial aquaculture, wild fish populations, and the aquarium trade, there is an increased need to rapidly identify and assess the impacts of emerging disease agents. This includes the identification and characterization of apparently non-pathogenic viruses that could serve as ancient phylogenetic resources, as well as surrogate models for understanding aspects of viral ecology, including modes of transmission, host diversity and virulence mechanisms as they relate to other closely-related pathogenic viruses. This research was conducted with such goals in mind. The study was consistent with the history of fish disease research and serves the dual purpose of fostering collaboration with practicing veterinarians, aquariums and research institutions such as the Centers for Disease Control and Prevention (CDC) and the National Institutes of Health (NIH). Identification of these viral agents adds significantly to the body of knowledge concerning oncogenic viruses in fish and will potentially limit their spread. The characterization of these viruses has helped to shape our knowledge of viral evolution and highlights the need for further sampling of lower vertebrates.

APPENDICES

THE ANCIENT EVOLUTIONARY HISTORY OF POLYOMAVIRUSES

Buck CB, Van Doorslaer K, Peretti A, Geoghegan EM, Tisza MJ, An P, Katz JP, Pipas JM, McBride AA, Camus AC, McDermott AJ, Dill JA, Delwart E, Ng TFF, Farkas K, Austin C, Kraberger S, Davison W, Pastrana DV, Varsani A. 2016. The ancient evolutionary history of polyomaviruses. PLoS Pathog 12:e1005574.

PLOS | PATHOGENS

RESEARCH ARTICLE

The Ancient Evolutionary History of Polyomaviruses

Christopher B. Buck[1]*, Koenraad Van Doorslaer[2], Alberto Peretti[1], Eileen M. Geoghegan[1], Michael J. Tisza[1], Ping An[3], Joshua P. Katz[3], James M. Pipas[3], Alison A. McBride[2], Alvin C. Camus[4], Alexa J. McDermott[5], Jennifer A. Dill[4], Eric Delwart[6,7], Terry F. F. Ng[6,7¤], Kata Farkas[8], Charlotte Austin[8], Simona Kraberger[8], William Davison[8], Diana V. Pastrana[1], Arvind Varsani[8,9,10]

1 Lab of Cellular Oncology, NCI, NIH, Bethesda, Maryland, United States of America, **2** Lab of Viral Diseases, NIAID, NIH, Bethesda, Maryland, United States of America, **3** Department of Biological Sciences, University of Pittsburgh, Pittsburgh, Pennsylvania, United States of America, **4** Department of Pathology, University of Georgia, Athens, Georgia, United States of America, **5** Animal Health Department, Georgia Aquarium, Inc., Atlanta, Georgia, United States of America, **6** Blood Systems Research Institute, San Francisco, California, United States of America, **7** Department of Laboratory Medicine, University of California, San Francisco, San Francisco, California, United States of America, **8** School of Biological Sciences, University of Canterbury, Christchurch, New Zealand, **9** Structural Biology Research Unit, Department of Clinical Laboratory Sciences, University of Cape Town, Cape Town, South Africa, **10** Department of Plant Pathology and Emerging Pathogens Institute, University of Florida, Gainesville, Florida, United States of America

¤ Current address: DVD, NCIRD, Centers for Disease Control, Atlanta, Georgia, United States of America
* buckc@mail.nih.gov

Citation: Buck CB, Van Doorslaer K, Peretti A, Geoghegan EM, Tisza MJ, An P, et al. (2016) The Ancient Evolutionary History of Polyomaviruses. PLoS Pathog 12(4): e1005574. doi:10.1371/journal.ppat.1005574

Editor: Denise A. Galloway, Fred Hutchinson Cancer Research Center, UNITED STATES

Received: November 18, 2015

Accepted: March 23, 2016

Published: April 19, 2016

Data Availability Statement: All relevant data are within the paper and its Supporting Information files. In addition to providing all data needed to replicate the work, as a convenience we have posted compiled sequence datasets at the following internet address http://home.ccr.cancer.gov/Lco/PyVE.asp. Although the posted data are not strictly required for understanding or reproducing the current work, the availability of compiled data should make it much faster and easier for colleagues to replicate our work.

Funding: This work was funded in part by the National Institutes of Health Intramural Research Program, with support from the National Cancer

Abstract

Polyomaviruses are a family of DNA tumor viruses that are known to infect mammals and birds. To investigate the deeper evolutionary history of the family, we used a combination of viral metagenomics, bioinformatics, and structural modeling approaches to identify and characterize polyomavirus sequences associated with fish and arthropods. Analyses drawing upon the divergent new sequences indicate that polyomaviruses have been gradually co-evolving with their animal hosts for at least half a billion years. Phylogenetic analyses of individual polyomavirus genes suggest that some modern polyomavirus species arose after ancient recombination events involving distantly related polyomavirus lineages. The improved evolutionary model provides a useful platform for developing a more accurate taxonomic classification system for the viral family *Polyomaviridae*.

Author Summary

Polyomaviruses are a family of DNA-based viruses that are known to infect various terrestrial vertebrates, including humans. In this report, we describe our discovery of highly divergent polyomaviruses associated with various marine fish. Searches of public deep sequencing databases unexpectedly revealed the existence of polyomavirus-like sequences in scorpion and spider datasets. Our analysis of these new sequences suggests that polyomaviruses have slowly co-evolved with individual host animal lineages through an

Institute Center for Cancer Research. AAM and KVD were funded by the Intramural Research Program of the National Institute of Allergy and Infectious Disease. Trematomus pennellii were collected in the Antarctic under the 2011/08R animal ethics permit and the field work was supported by a grant (K057) awarded to WD from Antarctica New Zealand. The Trematomus pennelli molecular work was supported by personal funds of AV. The structural analyses for the large T antigens by PA, JPK and JMP were supported by an R21 grant AI109339 awarded to JMP by National Institute of Allergy and Infectious Diseases. AP is supported by a grant from the Italian Foundation for Cancer Research (FIRC). Aside from the contribution of personal funds by AV, other funders had no role in study design, data collection and analysis, decision to publish, or preparation of the manuscript.

Competing Interests: AJM is an employee of Georgia Aquarium, Inc., a 501(c)3 not-for-profit organization. This affiliation does not alter our adherence to all PLOS Pathogens policies on sharing data and materials.

established mechanism known as intrahost divergence. The proposed model is similar to the mechanisms through with other DNA viruses, such as papillomaviruses, are thought to have evolved. Our analysis also suggests that distantly related polyomaviruses sometimes recombine to produce new chimeric lineages. We propose a possible taxonomic scheme that can account for these inferred ancient recombination events.

Introduction

Murine polyomavirus (MPyV) was discovered in the mid-1950s as a filterable infectious agent that could induce salivary tumors in experimentally exposed mice [2, 3]. It was quickly established that the virus is potently carcinogenic, causing many different types of tumors (Greek *poly* + *oma*) in various experimental systems. When the first primate polyomavirus, simian vacuolating virus 40 (SV40), was discovered as an abundant contaminant in early poliovirus vaccines that had already been administered to millions of individuals, it posed significant cause for alarm (reviewed in [4]). The ensuing rush to study the molecular biology of polyomaviruses provided a great wealth of insights into basic cell biology and the fundamental mechanisms of tumorigenesis (reviewed in [5]).

There is no conclusive evidence for productive transmission of SV40 among humans and it does not appear that the virus caused discernible disease in poliovirus vaccine recipients (reviewed in [6]). However, SV40 is closely related to human JC and BK polyomaviruses (JCV and BKV), both of which cause disease in immunosuppressed patients. JCV was discovered in a patient (initials JC) who was suffering from a lethal brain disease called progressive multifocal leukoencephalopathy (PML)[7]. BKV is rarely found in the brain, but causes serious kidney damage in up to 10% of kidney transplant recipients [8]. Conflicting reports suggest possible associations between JCV and BKV and additional human diseases, including prostate, colorectal, and kidney cancers [5, 9]. A more recently discovered human polyomavirus, Merkel cell polyomavirus (MCV), plays a key causal role in the development of a rare form of skin cancer, Merkel cell carcinoma [10]. Other recently discovered human polyomaviruses have been associated with a variety of disease states, ranging from thymic and lymphoid cancers to non-malignant skin dysplasias and vascular myopathy [11–14]. Efforts to discover additional human and animal polyomaviruses, and the conclusive establishment of further links to disease states, will undoubtedly remain highly active research areas for the foreseeable future.

It has been difficult to achieve consensus on the development of systems for taxonomic classification of polyomaviruses. This is regrettable, in the sense that the availability of a robust classification scheme could help guide researchers and clinicians toward an understanding of where to expect biological similarities and differences among established and newly discovered polyomavirus species. A key barrier to the development of a consensus taxonomic scheme has been the lack of a clear model for the evolutionary history of polyomaviruses. Approaches to this question have been limited by the fact that known polyomavirus species are derived from a restricted subset of terrestrial vertebrates. In this study, we report our discovery of polyomaviruses in several species of fish. Searches of shotgun genomics datasets also revealed previously unknown polyomavirus-like sequences in a surprisingly wide variety of additional animals, including insects and arachnids. We make use of these new, highly divergent polyomavirus sequences to develop an evolutionary model that might account for the interrelationships of extant polyomavirus species.

Results

Acquisition of divergent polyomavirus sequences

In an effort to obtain more divergent polyomaviruses to use as reference points for understanding polyomavirus evolution, we sampled a variety of fish species. We have recently published a brief announcement describing the sequence of a polyomavirus found in samples of a perciform fish, black sea bass (*Centropristis striata*)[15]. In the current report, we present our discovery of another polyomavirus species found in a different perciform fish, the sharp-spined notothen (*Trematomus pennellii*) from McMurdo Sound (Ross Sea, Antarctica). The predicted genetic organization of these viruses is shown in Fig 1.

We also report a previously unknown polyomavirus species found in a giant guitarfish (*Rhynchobatus djiddensis*) suffering from papillomatous skin lesions. Guitarfish are members of the subclass Elasmobranchii, which includes sharks and rays. Elasmobranchs and bony vertebrates are thought to have diverged during the Cambrian period, about half a billion years ago [16]. Although the guitarfish polyomavirus encodes the characteristic polyomavirus arrangement of major open reading frames (Fig 1), its 3,962 bp genome is substantially smaller than the 4,697 bp genome of bovine polyomavirus 1, which had previously been the smallest known member of the family (see S1 File). To confirm that the virus directly infected the giant guitarfish (as opposed to an unknown environmental source), we performed in situ hybridization using a probe targeting the VP1 ORF. Hybridization signal was observed in small numbers of cells in resolving skin lesions (S1 Fig), confirming that the virus directly infects guitarfish.

We have recently reported the sequences of three polyomavirus species found in supermarket ground beef [17, 18]. In a follow-up effort using the same methods, we sampled supermarket ground turkey, American bison, and lamb. Although no polyomaviruses were found in the turkey or bison samples, a single previously unknown polyomavirus species was identified in the ground lamb (*Ovis aries*, sheep) meat sample. In light of recent scandals identifying traces of horse meat in supermarket ground beef products [19], the association of this virus with sheep should be considered tentative.

In GenBank keyword searches we noticed that a genomic DNA segment of a South African social spider (*Stegodyphus mimosarum*) had been annotated as having a patch of sequence similarity to polyomavirus LT (accession KK122585). The apparent endogenized "fossil" LT gene, which is integrated into a putative spider transcription elongation factor locus, was inferred to have one frameshift mutation and one nonsense mutation.

Polyomavirus protein sequences, including the novel fish polyomavirus LTs and a "resurrected" version of the social spider LT, were used to query translated nucleotide sequences in various NCBI databases. An additional fossil LT sequence was detected at a second locus in the social spider Whole Genome Shotgun (WGS) dataset. At least half a dozen fossil LT-like sequences could be detected in WGS entries for the common house spider (*Parasteatoda tepidariorum*). A short (170 bp) LT-like contig was identified in a third spider species, the Brazilian whiteknee tarantula (*Acanthoscurria geniculata*). Nearly a dozen transcripts with clear similarity to LT proteins were found in the Transcriptome Shotgun Assembly (TSA) datasets for two primitive insects, *Machillis hrabei* and *Meinertellus cundinamarcensis* (commonly called bristletails). More recently, some additional arthropod polyomavirus LT and VP1 transcripts have appeared in the TSA datasets for brown widow (*Latrodectus geometricus*) and cupboard spider (*Steatoda grossa*). Accession numbers for these newer sequences are listed in the "fragments" tab of S1 File.

Several polyomavirus-like sequences were also observed in TSA datasets for vertebrates, including a short VP1-like sequence in guineafowl (*Numida meleagris*), a short LT-like

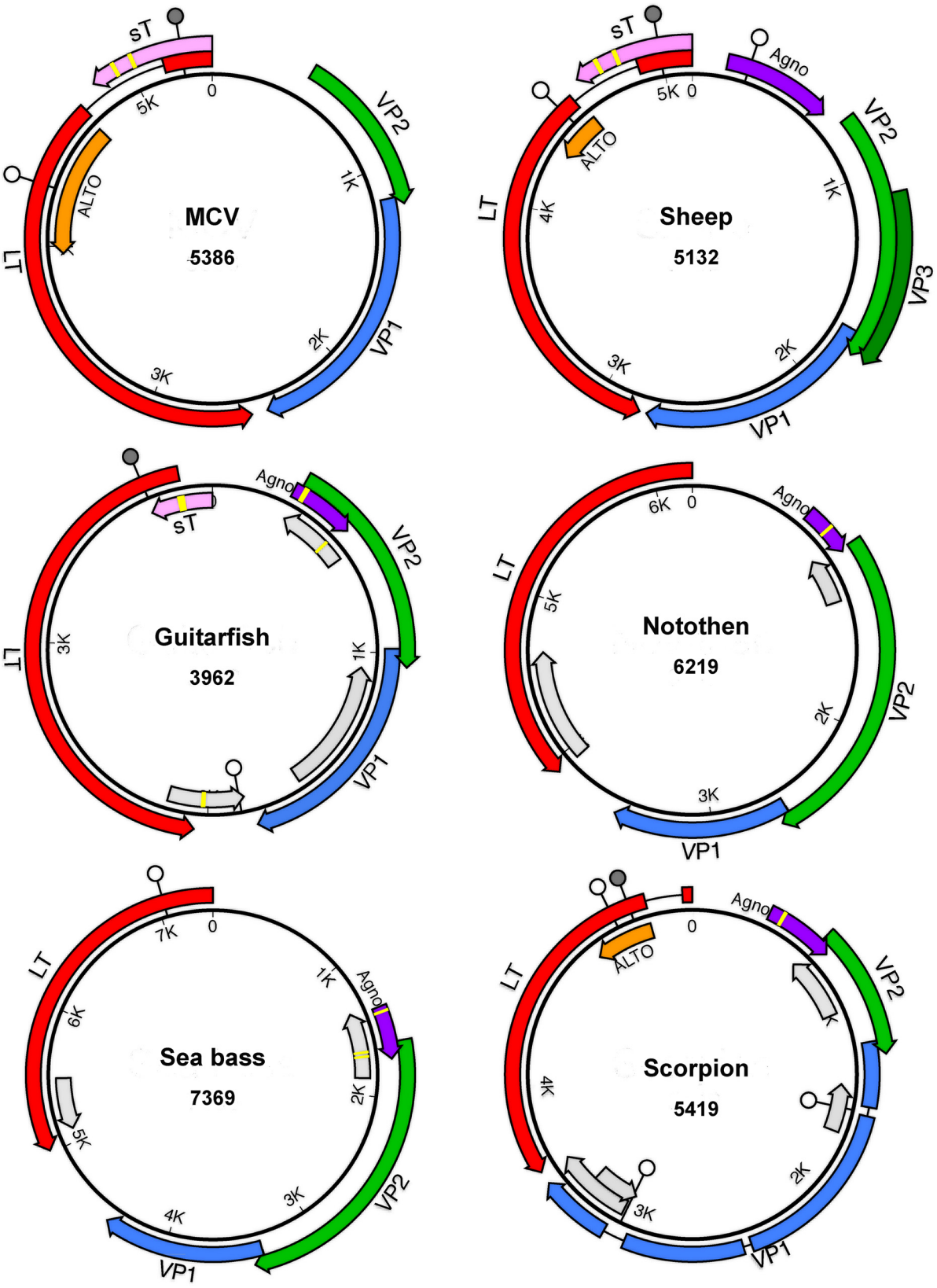

sT
5K
0
VP2
1K
ALTO
MCV
5386
LT
2K
3K
VP1
sT
Agno
5K 0
VP2
ALTO
1K
LT
4K
Sheep
5132
VP3
3K
2K
VP1
Agno
sT
VP2
3K
Guitarfish
3962
1K
LT
VP1
6K 0
Agno
LT
5K
Notothen
6219
2K
VP2
3K
VP1
7K
0
1K
Agno
LT
6K
Sea bass
7369
2K
5K
3K
4K
VP2
VP1
Agno
0
ALTO
VP2
LT
4K
Scorpion
5419
2K
3K
VP1

Fig 1. Predicted genetic organization of newly discovered polyomaviruses. Merkel cell polyomavirus (MCV) is shown as a well-studied reference species. The size of each genome (in basepairs) is listed below the species name. Large T antigen (LT) is indicated in red. Dark gray lollipops indicate the signature HPDKGG motif of the LT "DNAJ" domain (which appears to be missing from the sea bass and notothen polyomaviruses). White lollipops indicate LXCXE motifs, which are hypothetically involved in binding pRb and related tumor suppressor proteins. Each virus encodes a potential myristoylation signal that defines the N-terminus of the minor capsid protein VP2 (green). The VP2 of the supermarket sheep meat-associated virus encodes an internal MALXXΦ motif [1] that defines the N-terminus of a predicted VP3 minor capsid protein, while the other viruses do not. Predicted VP1 major capsid protein genes are shaded blue. ORFs found in the same general arrangement as previously described accessory proteins are also shown. These include small T antigen (sT, pink) Agnoprotein (purple), and the recently described ALTO (orange), which is overprinted in the LT +1 frame. Un-named ORFs of potential interest are shaded light gray. Yellow bars indicate hypothetical metal-binding motifs (CXCXXC or related sequences) observed in some of the predicted accessory proteins. Aside from MCV, for which expressed proteins have been experimentally confirmed, the predicted proteins are hypothetical and do not necessarily account for possible spliced transcripts.

doi:10.1371/journal.ppat.1005574.g001

fragment in Carolina anole lizard (*Anolis carolinensis*), and an apparently complete set of spliced LT, VP1, and VP2 transcripts in the TSA dataset for dark-eyed junco (*Junco hyemalis*).

The most important discovery in the WGS database was a single contig (AXZI01204118) that appears to represent a nearly complete polyomavirus genome associated with Baja California bark scorpion (*Centruroides exilicauda*). Extension of the contig using individual reads from the parent Sequence Read Archive (SRA) datasets revealed two variants (~92% identity) of a circular non-integrated polyomavirus-like sequence. It thus appears that the individual animal used for the genome sequencing project happened to be productively infected with a polyomavirus. Although the complete, apparently episomal sequences show the usual organization of polyomavirus genomes, with highly divergent homologs of the standard LT and VP2 proteins (Fig 1), BLAST alignments using the inferred VP1 protein do not yield any convincing hits (E values >0.5).

Structural modeling of divergent LT proteins

Computer-based modeling was used to investigate the possible structural conservation of the apparent LTs of the new fish and arthropod polyomaviruses. SV40 LT is divided into discrete structural domains that are thought to exist in a "beads on a string" configuration (reviewed in [20]). The structures of individual LT domains have been solved [21, 22]. The modeled structures of the scorpion and fish LT origin binding domain (OBD), zinc finger domain, and ATPase domain each show a good fit with the known SV40 structures (Fig 2). A conservation map for the DNAJ and Zn-ATPase domains is shown in S2 Fig. These results confirm that the fish- and scorpion-derived sequences represent *bona fide* polyomavirus LT proteins.

LT proteins typically carry an N-terminal domain with sequence and structural similarity to cellular DNAJ chaperone proteins. The domain is defined by a hallmark linear motif, HPDKGG. The guitarfish and scorpion viruses share this motif, and the N-terminal domains of their LT proteins can readily be modeled onto known DNAJ structures (Fig 2). In contrast, the predicted sea bass and notothen polyomavirus LT proteins lack HPDKGG motifs. The two viruses are unique among known polyomaviruses in their apparent lack of any sequences that can be modeled onto known DNAJ structures. The novel N-terminal domains of the two perciform fish LT proteins share only about 25% similarity to one another, show no clear similarity to any other known proteins or protein structures, and are predicted to be unstructured. A possible explanation could be that the LT DNAJ domain is a common ancestral feature that was lost during development of the perciform fish polyomavirus lineage.

Phylogenetic analysis of LT and VP1 proteins

A phylogenetic tree was constructed for the complete LT protein sequences of examples of all currently known polyomavirus species and sub-genomic fragmentary sequences available prior to November, 2015. The phylogenetic analyses also included putative LT protein sequences

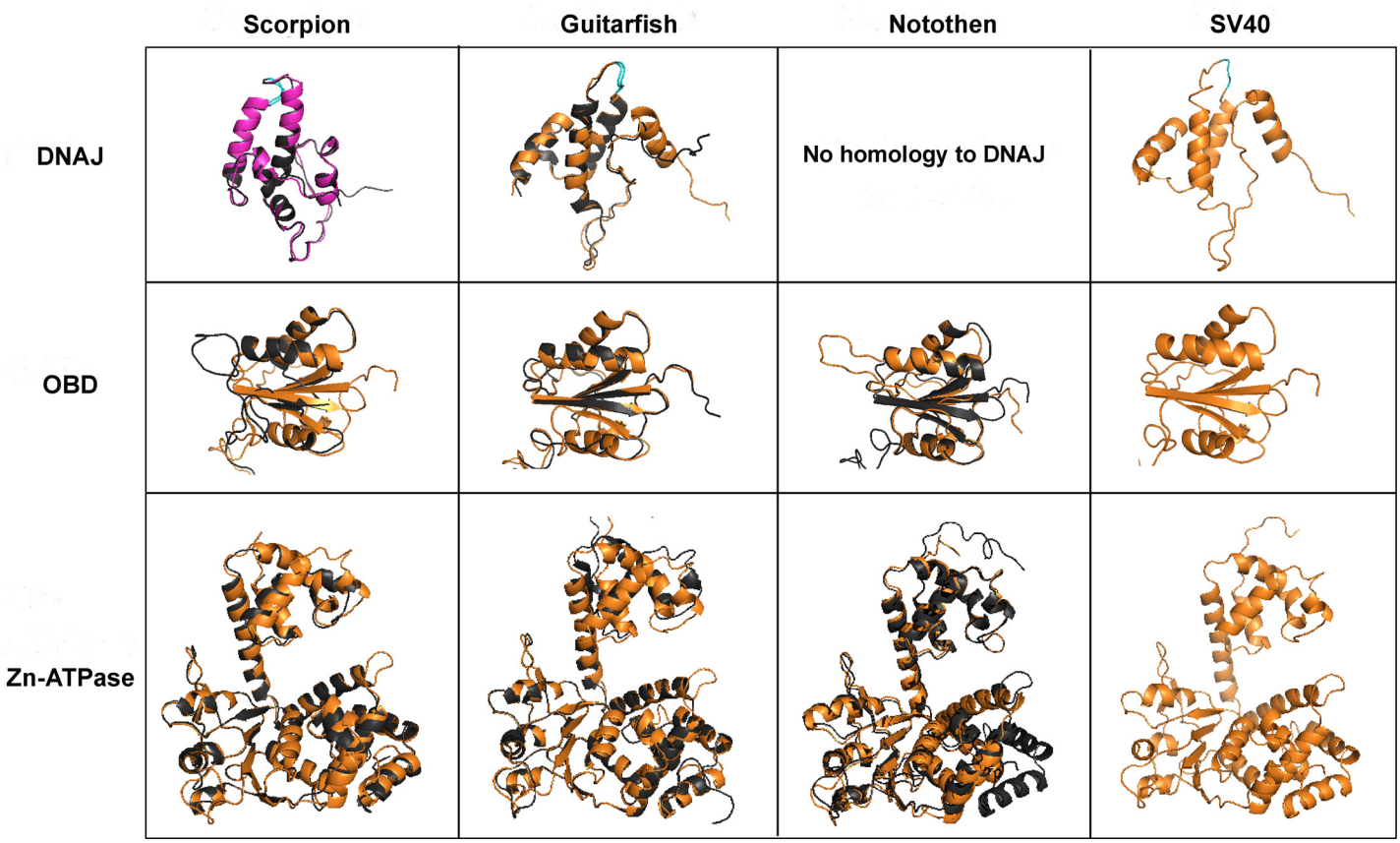


Fig 2. Structural modeling of LT proteins. The solved OBD-Zn-ATPase SV40 LT structure (PDB identifier 4GDF) was used as template for all OBD and Zn-ATPase domain models. The model of the guitarfish polyomavirus J domain was generated using the solved structure of the SV40 LT DNAJ domain (PDB identifier 1GH6) as template. For the DNAJ domain of scorpion polyomavirus LT, the best modeling template match is a *Thermus thermophilus* DNAJ protein (PDB identifier 4J7Z). The solved structure of the bacterial DNAJ is highlighted in magenta in the pairwise superimposition (top left). The LT proteins of the indicated polyomavirus species are shown in black. The known structures of SV40 LT domains are superimposed in gold. The conserved HPD motif of the DNAJ domain is positioned on the top and highlighted in cyan. The N-terminal domain of the notothen polyomavirus has no discernible structural similarity to known DNAJ structures.

doi:10.1371/journal.ppat.1005574.g002

found in a pair of viral species that cause carcinomatosis in an Australian marsupial, the western barred bandicoot (*Perameles bougainville*). The bandicoot viruses appear to have arisen after recombinant chimerization involving an unidentified polyomavirus and a member of a known group of marsupial-tropic papillomaviruses [23, 24]. The apparently chimeric viruses encode a polyomavirus LT-like gene on one strand and genes for papillomavirus-like L1 and L2 capsid proteins on the other strand.

Like the bandicoot viruses, a different apparently chimeric virus called Japanese eel endothelial cells-infecting virus (JEECV) encodes a protein with typical LT features, including an N-terminal DNAJ-like sequence domain [25]. A similar virus has recently been discovered in Taiwanese marbled eels [26]. Aside from the clear 2.1 kb LT gene, the remaining ~13 kb of the JEECV genome bears little similarity to sequences in GenBank. It thus appears that JEECV and the marbled eel virus arose through recombination between a bony fish-associated polyomavirus and a member of another DNA virus family that remains unidentified.

Phylogenetic analysis of LT proteins (Fig 3) shows distinct clades corresponding to fish- and arthropod-associated sequences, as well as the previously recognized mammalian Ortho and Almi clades [27, 28]. Avian and bandicoot LT sequences together occupy a distinct clade.

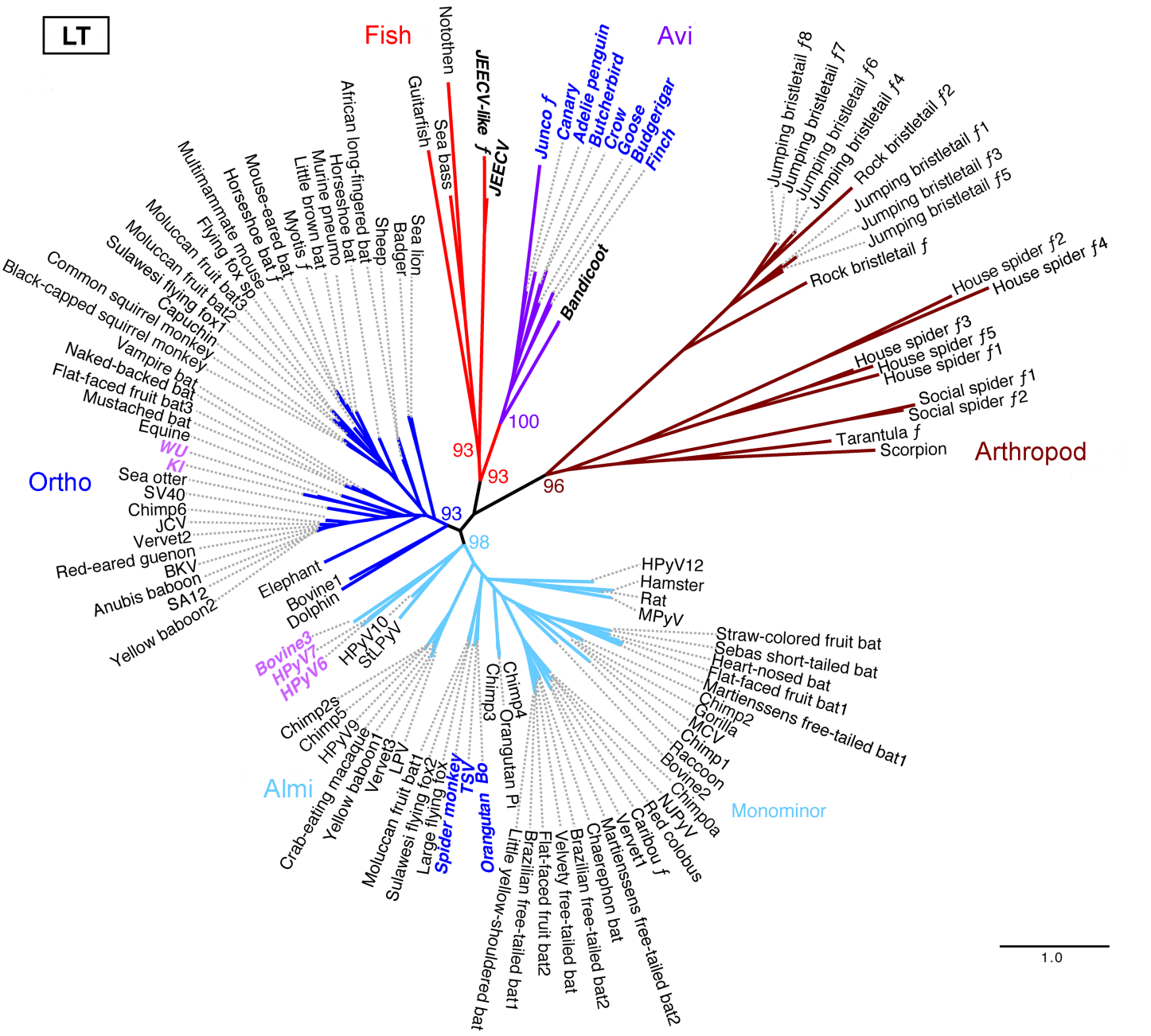


Fig 3. Midpoint-rooted phylogenetic tree for polyomavirus Large T antigen (LT) protein sequences. Species with different clade affiliations in VP1 analyses (Fig 4) are indicated in colored bold oblique text. The script *f* character indicates fragmentary (sub-genomic) sequences. A key to species nicknames, genetic characteristics, and accession numbers is provided in S1 File. Percent bootstrap values are indicated for selected nodes. A FigTree file containing detailed bootstrap values is provided as S2 File. Scale bar shows one substitution per site.

doi:10.1371/journal.ppat.1005574.g003

The appearance of the bandicoot virus LT protein sequences within this clade suggests that polyomaviruses with Avi-like early regions may infect modern marsupials. The avian and bandicoot LT proteins occupy a larger super-clade that loosely includes the newly identified fish-associated LT sequences. The LT protein sequences of the fish-associated polyomaviruses form a distinct clade that includes JEECV LT.

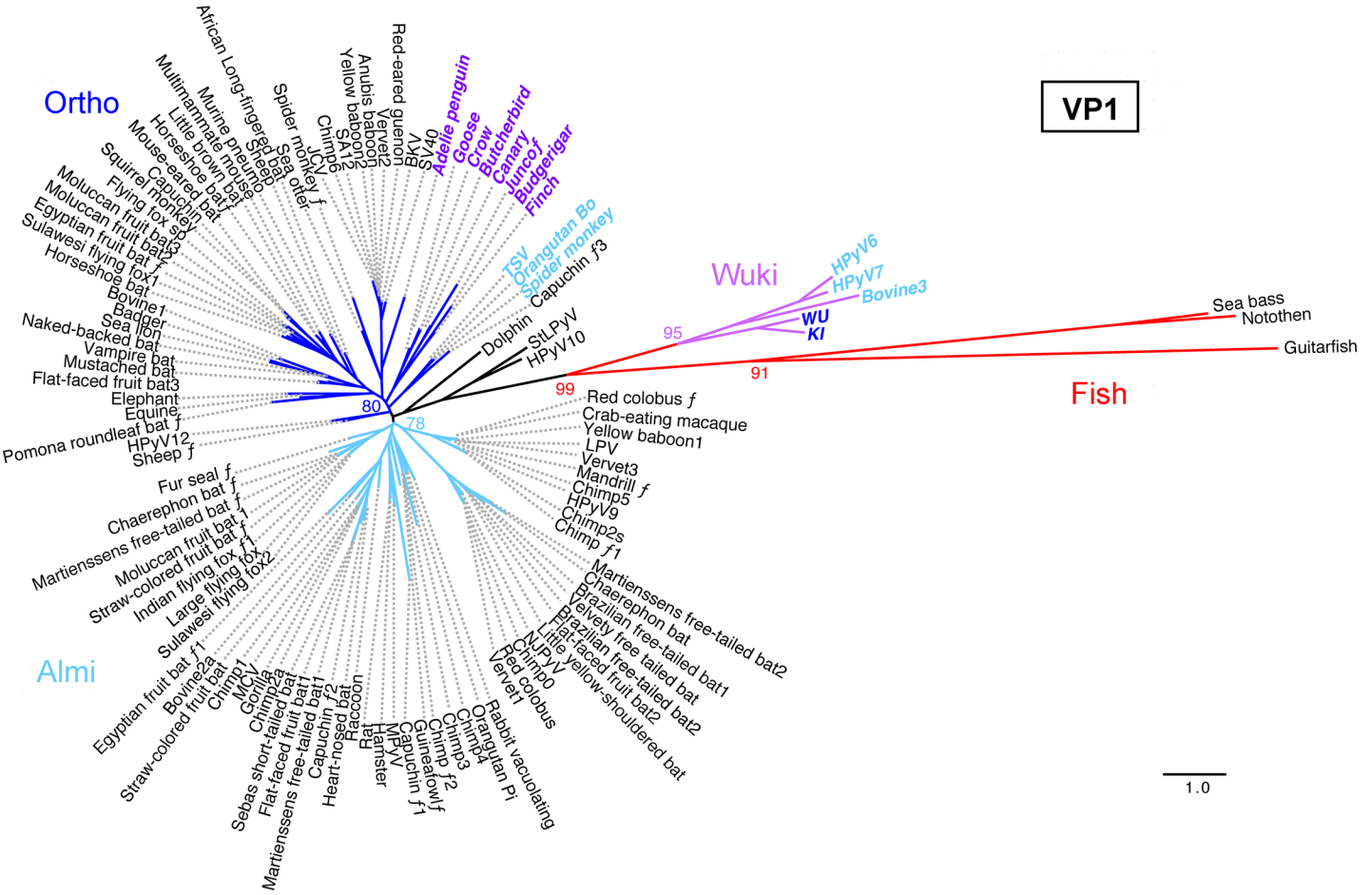


Fig 4. Midpoint-rooted phylogenetic tree for polyomavirus VP1 protein sequences. Species with different clade affiliations in LT analyses (Fig 3) are indicated in colored bold oblique text. The script *f* character indicates fragmentary (sub-genomic) sequences. Percent bootstrap values for selected nodes are indicated. A FigTree file containing detailed bootstrap values is provided as S3 File. Scale bar shows one substitution per site.

doi:10.1371/journal.ppat.1005574.g004

Phylogenetic analyses of VP1 protein sequences (Fig 4) reveal somewhat different patterns. In contrast to avian polyomavirus LT protein sequences, avian polyomavirus VP1 sequences are interspersed among mammalian Ortho VP1 sequences. Phylogenetic analyses of VP2 protein sequences (presented in FigTree format in S4 File) are concordant with the VP1 analysis in this regard.

Members of the previously recognized Wuki clade [29] encode VP1 protein sequences that occupy a highly divergent clade that distantly encompasses fish-associated VP1 sequences, while the early regions of Wuki species encode Ortho- or Almi-LT-like genes. Thus, relative to "classic" Ortho polyomaviruses the Avi clade shows a highly divergent early region while the Wuki clade shows a highly divergent late region. In Fig 5 we illustrate a recombination scheme that could account for this strangely mixed phylogeny.

Accessory ORFs

The carboxy-terminal halves of Avi and bandicoot small T antigens (sT) show no linear sequence similarity to the sT proteins of Ortho or Almi polyomavirus species. In particular, Avi-type sT proteins lack highly conserved cysteine motifs that have recently been shown to coordinate iron-sulfur clusters in mammalian sT proteins [30]. It is also noteworthy that a

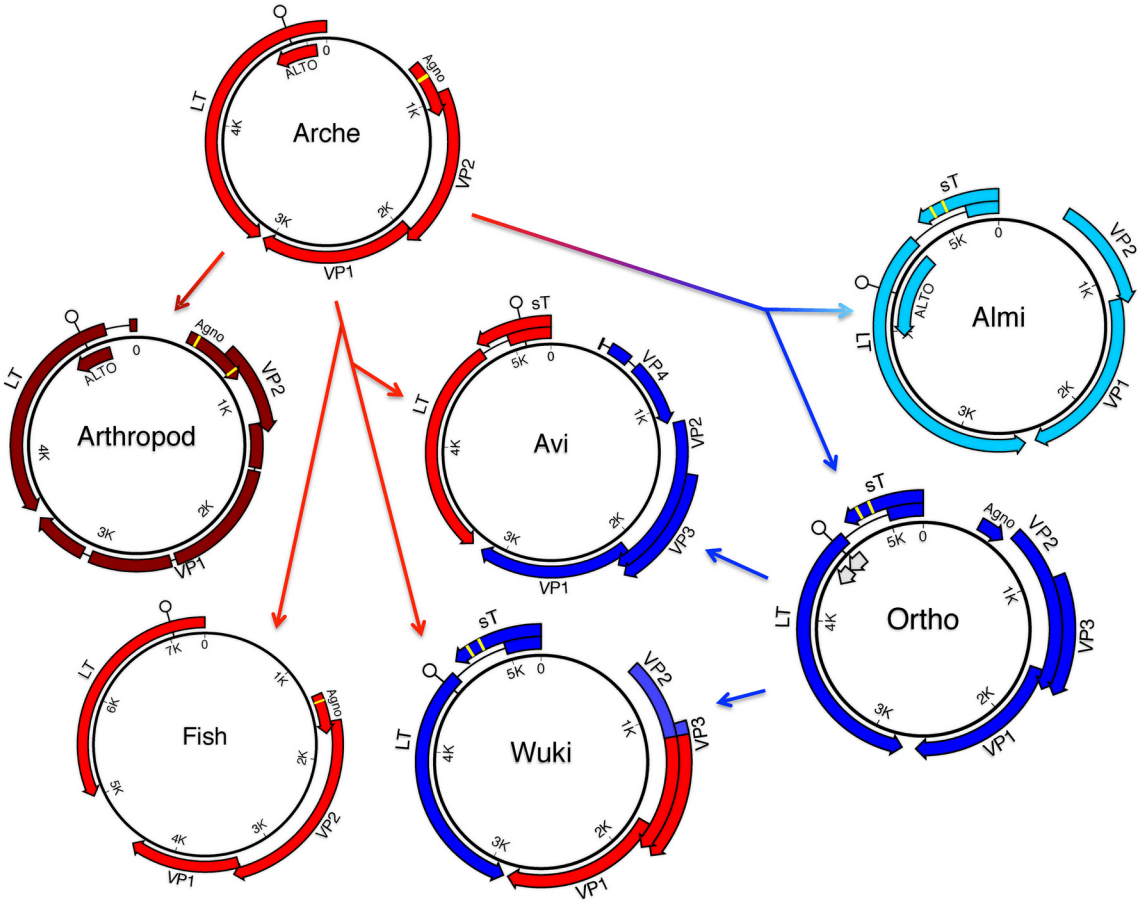


Fig 5. A hypothetical framework for ancient recombination events among major polyomavirus clades. The model attempts to reconcile observed incongruities between LT and VP1 phylogenetic trees shown in Figs 3 and 4. In the model, a hypothetical ancient polyomavirus, designated Arche, is inferred to have infected the last common ancestor of bilaterian animals. The ancient Arche lineage then gave rise to separate polyomavirus lineages found in arthropods and fish, as well as the mammalian Ortho/Almi lineages. The figure depicts Avi and Wuki clades arising after recombination events involving an unknown vertebrate-Arche lineage and Ortho-like species. The figure does not depict the inferred evolution of the HPyV6/7 clade, which appears to have arisen after a separate recombination event involving the late region of a hypothetical vertebrate-Arche lineage and the early region of a basal Almi-like species. The TSV lineage, which shows evidence of recombination between the Ortho and Almi lineages, is also omitted. White lollipops represent predicted pRb-binding motifs (LXCXE or related sequences). Yellow bars represent hypothetical metal-binding motifs (CXCXXC or related sequences). The absence of metal-binding motifs in Avi small T antigen (sT) proteins suggests a different evolutionary origin than the classic metal-binding Ortho/Almi sT. Possible ALTO-like ORFs predicted for some Ortho species are shaded gray.

doi:10.1371/journal.ppat.1005574.g005

conserved LXCXE motif (thought to be involved in interactions with the pRb family of tumor suppressor proteins and suppression of innate antiviral immunity [31]) is located on the shared sT/LT leader sequence in the Avi and bandicoot viruses, whereas the LXCXE motif is instead located in the second exon of LT in Ortho and Almi species. This suggests that Avi sT has a different evolutionary origin than Ortho/Almi sT. A possible explanation would be that Ortho/Almi sT arose after re-location of an ancestral cysteine motif-containing accessory gene into an N-terminal LT intron. For example, duplication of the scorpion polyomavirus ORF labeled "Agno" (Fig 1) into the LT intron could roughly reproduce an Ortho/Almi sT-like arrangement. A prediction of this idea would be that some of the hypothetical accessory ORFs of fish and arthropod polyomaviruses may be metal-binding proteins with Ortho/Almi sT-like functions, such as manipulation of cellular protein phosphatase 2A proteins [32].

ALTO is a recently discovered accessory gene that is "overprinted" in the +1 frame of the second exon of LT [27, 33, 34]. Although the function of ALTO is unknown, it shows sequence

similarity to the C-terminal transmembrane domain of the well-studied middle T antigen of MPyV. This suggests that ALTO might, like middle T, function by mimicking activated growth factor receptors (reviewed in [35]). In their initial report demonstrating the existence of MCV ALTO, Carter et al. suggested that the gene might have first arisen in the Almi (ALTO/middle T) lineage after its divergence from the Ortho lineage. However, Carter and colleagues also noted that the ATG codon thought to initiate the translation of MCV ALTO and a hydrophobic sequence near the C-terminus of ALTO are partially conserved in other polyomaviruses outside the defined Almi clade. Puzzlingly, many recently discovered non-Almi polyomaviruses appear to have ALTO-like ORFs with lengths similar to some of the shorter examples of recognized Almi-LT ALTOs (summarized in S1 File). For example, the two variants of the scorpion polyomavirus potentially encode 9 or 13 kD ATG-initiated proteins in the +1 frame of the second exon of their LT sequences (see Fig 1). Despite the fact that the new supermarket sheep meat-associated polyomavirus occupies the Ortho clade, the +1 frame of its LT second exon encodes a potential 10 kD ALTO-like protein. One conceivable explanation for these observations might be that ALTO-like ORFs are an ancient ancestral feature that has been lost in some polyomavirus lineages. A possible example of occult or remnant ALTO/MT-like genes might be found in the small clade of primate polyomaviruses that encompasses SV40. Members of this group of viruses encode a short Met-initiated ORF in the +1 frame of the second exon of LT and a separate short downstream LT +1 frame ORF with a splice acceptor near its 5' boundary (S3 Fig). It will be important to experimentally test the hypothesis that polyomavirus species outside the Almi-LT group express LT +1 frame ORFs as functional accessory proteins.

Virus-host co-divergence

Three previously established [36–38] virus-host co-evolutionary models are summarized in simplified cartoon form in Fig 6. In the strict co-divergence model, the rate at which viruses "speciate" from one another exactly matches the rate at which host animals speciate. A group of retroviruses known as foamy viruses are an example of a viral genus that may at least roughly follow this evolutionary model [39]. Many prior studies have established that the family *Polyomaviridae*, as a whole, does not conform to the strict co-divergence model [37, 40–42].

In the co-divergence with host switching model (middle panel of Fig 6), viruses and hosts generally co-diverge, but viruses are occasionally productively transmitted between distantly related host animals. In the example, such events are reflected in finding closely related viral sequences in bats and great apes (see light blue branches). Ebola and influenza viruses are familiar examples of viruses with clear evidence of occasional long-range host switching.

The first known polyomavirus of birds was discovered in diseased budgerigar fledglings (reviewed in [43]). Sequences >99% identical to the original budgerigar fledgling disease polyomavirus have subsequently been found in a surprisingly wide range of distantly related bird species [44–47]. Likewise, sequences nearly identical to goose hemorrhagic polyomavirus have been found in ducks [48, 49](accession JF304775). These prior findings are displayed as points close to the x-axis in Fig 7. Although the findings indicate that the host-switching model shown in the middle panel of Fig 6 might be applicable to some avian polyomaviruses, an important caveat is that all documented instances of inter-species Avi polyomavirus transmission have involved captive animals. It thus remains uncertain whether Avi polyomavirus host-switching occurs in the wild over longer timescales.

In contrast to Avi polyomaviruses, there are currently no examples of an individual polyomavirus species being found in more than one mammalian host (Fig 7). Most strikingly, there is no evidence of productive polyomavirus transmission between humans and any of the various polyomavirus-bearing animals we commonly live with or eat (i.e., budgerigars, canaries,

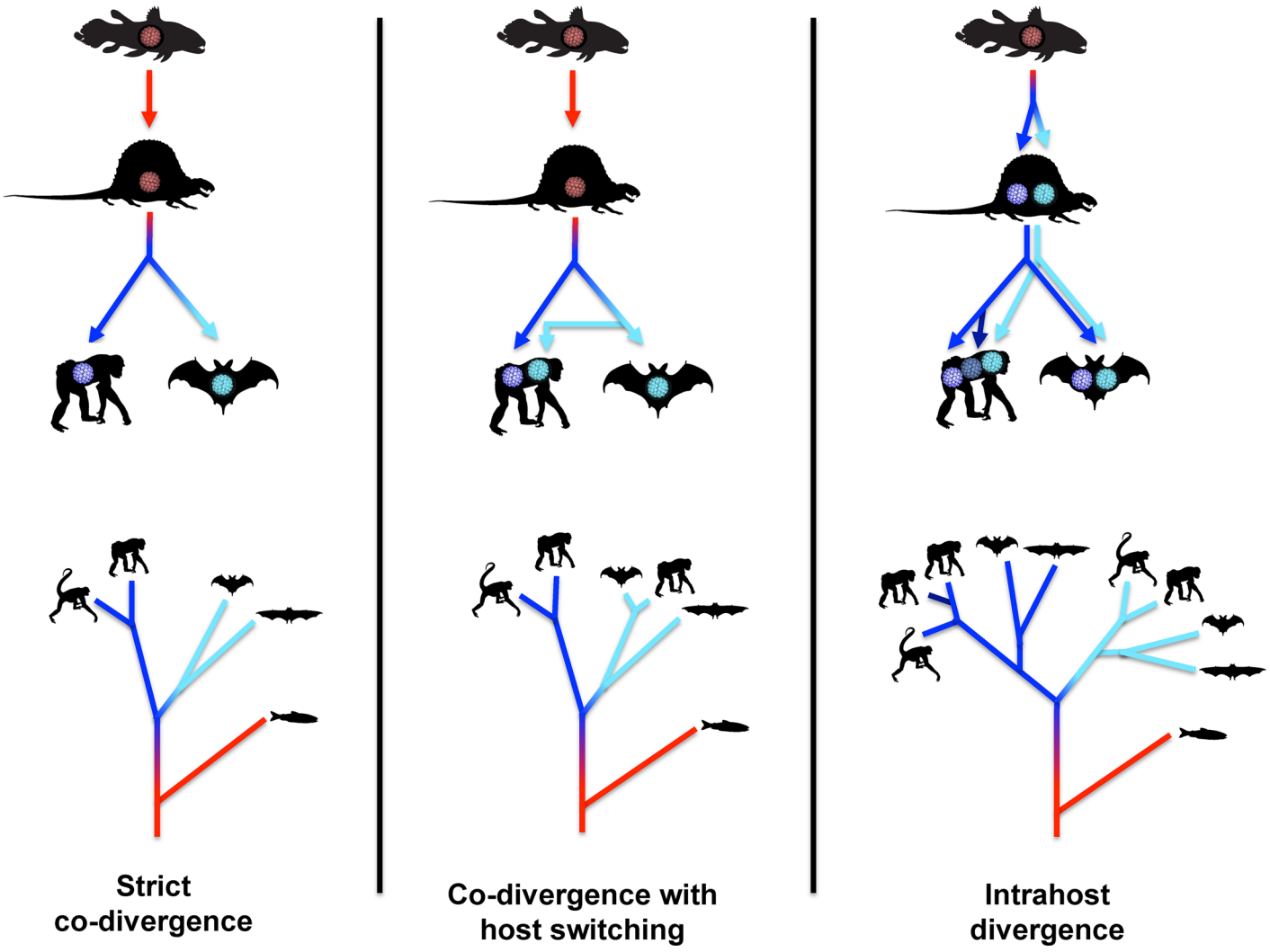


Fig 6. Standard virus/host co-divergence models. The top panels depict the evolution of polyomaviruses within animal lineages. Idealized cartoon trees in the bottom panels represent the expected polyomavirus phylogeny. The silhouettes in the bottom panels represent the animal type in which the polyomavirus at the branch tip would be found.

doi:10.1371/journal.ppat.1005574.g006

geese, ducks, mice, rats, hamsters, capuchins, horses, cattle, sheep, caribou, or sea bass). The fact that the Rhesus macaque polyomavirus SV40 seems not to have gained a detectable foothold in the human population despite extremely widespread human exposure is also noteworthy in this regard. These observations suggest that the host-switching model is not generally applicable to mammalian polyomaviruses.

In the intrahost divergence model (right-hand panel of Fig 6), viruses diverge from one another at a faster rate than host animal speciation. Ancient viral divergence events occurring within a single host animal lineage eventually give rise to separate viral clades that co-occupy a single animal species. The model does not invoke transmission of viruses between distantly related host animals, but could accommodate viral transmission between closely related animal species or subspecies. In the shown example, the dark blue and light blue lobes of the viral phylogenetic tree each internally resemble the phylogeny of host animals. More recent intrahost viral divergence events are reflected as distinct but closely related viral species found within a

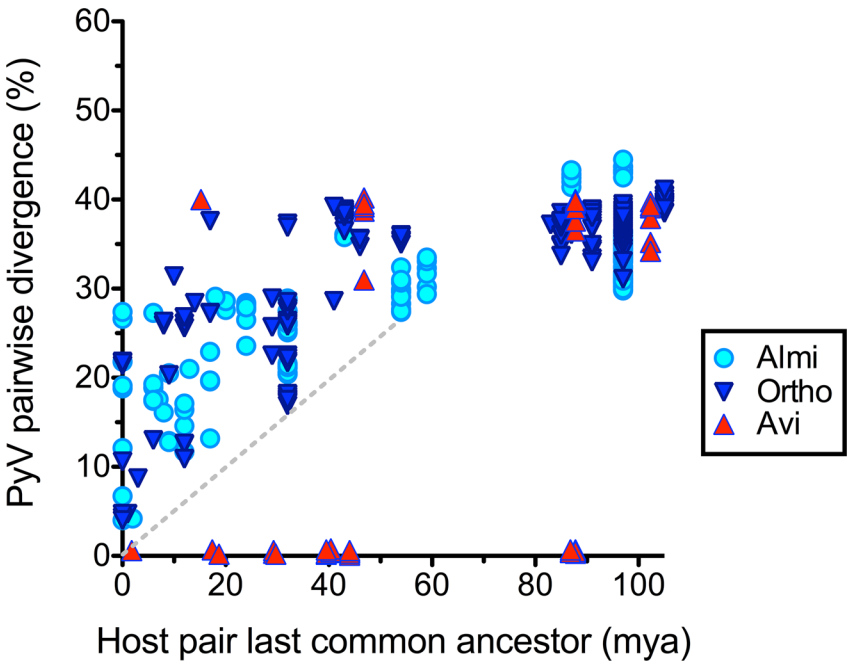


Fig 7. Virus-host co-divergence plot. SDT software was used to score individual pairs of polyomaviruses within various clades for percent divergence across the entire viral genome. The nucleotide divergence score was plotted against the estimated time (in millions of years ago, mya) of the last common ancestor of the host animals in which the polyomavirus pair was found. Apparent recent transmission of some Avi polyomaviruses between distantly related bird species is represented by points close to the x-axis. The absence of such points in the Almi and Ortho clades indicates a lack of evidence for recent transmission of polyomaviruses between distantly related mammal species. The arbitrary dashed reference line has a slope of about 0.5% polyomavirus divergence per million years after host divergence.

doi:10.1371/journal.ppat.1005574.g007

single host animal (see dark blue branches). Herpesviruses and some retrovirus genera are well-documented examples of this form of viral evolution [50, 51].

The phylogeny of mammalian polyomavirus species is qualitatively similar to the intrahost divergence model. In particular, the two Almi "Monominor" sub-clades (defined as species that encode a recognizable large ALTO but lack VP3 [1, 27]) recapitulate the expected topology of the intrahost divergence model (S2–S5 Files, S4 Fig).

The intra-host divergence model predicts that homologs of polyomavirus species that occupy currently depauperate lobes of the tree, such as the small clade that encompasses only WU and KI, may ultimately be found in other mammals. This prediction is consistent with the recent discovery of two WU/KI-like polyomavirus species associated with two European vole genera [52] (see Discussion).

Observed divergence analysis

A 2007 study by Carroll and colleagues showed that the VP1 nucleotide sequences of a panel MPyV strains found in feral mice collected in various locations in the United States all exactly matched the sequence of an MPyV laboratory isolate propagated in culture since 1953 [53]. Likewise, recent avian polyomavirus isolates are nearly identical to the isolate originally discovered in budgerigar fledglings in 1981 [54]. The concept that individual polyomavirus lineages may remain perfectly static over historical timescales is also consistent with the fact that BKV, JCV, MCV, and TSV strains with nearly or exactly identical nucleotide sequences have repeatedly been isolated from people residing on different continents [55, 56]. Historical sampling

thus does not appear to be a tractable approach to measuring polyomavirus nucleotide sequence divergence rates. We set out to instead compare the observed divergence of different polyomavirus species to the estimated time of divergence of the host animals in which they were found. The analysis rests on the starting assumption that productive transmission of polyomaviruses between different mammal genera is rare or non-existent (Fig 7, S3 Fig, S5 File).

In the intrahost divergence model, distantly related polyomaviruses found in closely related animals reflect ancient polyomavirus divergence events that occurred long prior to the divergence of the host animal pair. Under this scenario, data points in the top left quadrant of Fig 7 would give an artificially fast estimate of the rate of polyomavirus sequence divergence. Despite this caveat, it seems reasonable to assume that polyomavirus divergence events might sometimes happen to coincide with host animal speciation events. This would be reflected as the lowermost Ortho and Almi points in the scatter plot shown in Fig 7. The arbitrary dashed line in the figure connects polyomavirus pairs that hypothetically happened to diverge from one another at about the same time that the host animal pair diverged. The slope of the line is consistent with the idea that at least some Ortho and Almi polyomavirus pairs cumulatively diverged by roughly 0.5% per million years, at least during the first 60 million years after divergence. This crude estimate is consistent with a more sophisticated phylogenetics-based Bayesian rate estimate by Krumbholz et al. of about 0.8% per million years (8×10^{-9} nucleotide substitutions per site per year) for the protein-coding segments of Ortho polyomaviruses [57]. A separate observed-divergence analysis of LT and VP1 proteins suggests that the two genes have independently accumulated non-silent changes at comparable long-term rates (S5 Fig). This rough result is also consistent with the more sophisticated prior work of Krumbholz and colleagues.

We also performed additional computational analyses to further confirm the prior rate estimates of Krumbholz et al. These analyses focused on the phylogenetically tractable Monominor clade. A ParaFit analysis of the clade as a whole indicates that the null hypothesis that polyomaviruses evolved independently of their hosts can be rejected, with a p-value of 0.0258. Based on the assumption that the separate Monominor A and B sub-clades arose after an ancient intrahost divergence event that pre-dated the first placental mammals, we performed separate ParaFit analyses on each Monominor sub-clade. These analyses indicate an even more confident rejection of the null hypothesis, with p-values of 1×10^{-4} and 8×10^{-4} for the A and B sub-clades, respectively. A BEAST analysis of concatenated LT and VP1 genes for the Monominor clade confirms that codon positions 1 and 2 evolve at a long-term rate of about 5×10^{-9} substitutions per site per year (i.e., 0.5% per million years), while codon position 3 evolves at a rate of about 2×10^{-8} substitutions per site per year. A time-resolved phylogenetic tree of the entire Monominor clade based on host phylogeny is shown in S5 File.

Discussion

In this report, we propose a comprehensive theoretical framework for understanding the evolutionary history of the viral family *Polyomaviridae*. Our model suggests that the last common ancestor of arthropods and vertebrates harbored at least one polyomavirus. In the ensuing roughly half billion years, polyomaviruses appear to have accumulated genetic change at a remarkably slow cumulative long-term pace, in a pattern consistent with the intrahost divergence model diagrammed in Fig 6. Qualitative comparisons of phylogenetic trees suggest the occurrence of ancient recombination events involving distantly related polyomavirus species.

The intrahost divergence model also seems applicable to the evolution of papillomaviruses [58, 59]. A striking difference between polyomaviruses and papillomaviruses is the much greater number of known papillomavirus types. Our model could explain this difference simply

by postulating that papillomaviruses evolve (and therefore undergo intrahost divergence) at a slightly faster rate than polyomaviruses. This is consistent with the findings of Rector and colleagues, who used phylogenetic analyses to estimate that papillomaviruses diverge at an observed long-term rate of 2% per million years (i.e., slightly faster than our phylogenetics-based long-term rate estimates for polyomaviruses)[57, 60].

In the current classification system approved by the International Committee on Taxonomy of Viruses (ICTV), all members of the family *Polyomaviridae* belong to a single genus, *Polyomavirus*. We have previously contributed to a proposal that the family be divided into three genera to be officially named *Orthopolyomavirus*, *Avipolyomavirus*, and *Wukipolyomavirus* [29]. A recent case study [61] helped us to appreciate a potential pitfall of the previously proposed taxonomic system. Clinical colleagues approached us about a lung transplant recipient whose lung-wash samples showed strong immunohistochemical reactivity with an antibody known to detect BKV and JCV LT proteins. Puzzlingly, the samples were negative for BKV and JCV by PCR. Although WU and KI were initially discovered in human respiratory samples [62, 63], we reasoned that the observed immunohistochemical staining was unlikely to represent cross-detection of WU or KI, since they occupy a different proposed genus than BKV and JCV. Hypothesizing that the sample might instead contain an undiscovered human polyomavirus related to BKV and JCV, we applied virion purification, random-primed RCA and deep sequencing methods. The deep sequencing revealed high levels of WU and no other polyomaviruses. With hindsight, we realize that a taxonomic system highlighting the close phylogenetic relationship between the LT proteins of BKV/JCV and WU/KI would have served us by suggesting the less time-consuming approach of performing simple WU/KI-specific PCR on the lung wash sample. In short, our failure to appreciate the now-apparent problem of intergeneric polyomavirus chimeras resulted in wasted effort.

An established taxonomic approach to the problem of chimerization is to separately categorize each major gene product. The most familiar example is the classification of influenza virus hemagglutinin (H) and neuraminidase (N) genes (e.g., H1N1, H5N1, etc.). As an example of applying this type of approach to polyomaviruses, BKV could be described simply as an Ortho species, while WU could be described as an Ortho-LT/Wuki-VP1 species. Like the influenza virus classification system, this form of nomenclature could serve as a colloquial set of conventions operating as an adjunct to official ICTV classifications (which can only be applied to entire organisms, as opposed to individual gene segments).

Our proposed colloquial classification scheme is in conflict with a recent formal proposal currently being considered by the ICTV. The new ICTV proposal suggests classifying polyomaviruses into four official genera based solely on the phylogeny of LT proteins [64]. Although the proposal is appealingly simple, it suffers from the "chimera-blindness" described in the case study above. For example, the proposal fails to recognize that all seven members of proposed genus *Gammapolyomavirus* encode VP1 and VP2 proteins that are monophyletic with proposed genus *Betapolyomavirus* VP1 and VP2 proteins. We suggest that the slightly greater complexity of the colloquial "flu-style" classification system proposed in the current study is justified by its greater taxonomic accuracy. Since the new four-genus proposal would awkwardly preclude the use of the more accurate flu-style classification, we concur with Tao and colleagues' recent argument [41] in favor of preserving the existing ICTV standard, under which all polyomavirus species would officially remain in a single genus, *Polyomavirus*.

The intrahost divergence model predicts that multiple polyomaviruses with varying degrees of divergence will often be found within individual host animal species. Although we continue to favor the traditional cutoff of 81–84% identity across the entire viral genome for polyomavirus species distinctions, we note that this standard could be considered an arbitrary cutoff applied to a theoretically continuous variable. Since knowing the host animal species of origin

appears to be of paramount importance for understanding polyomavirus evolution, we suggest that, in the future, it would be useful for new polyomavirus species names to reference the host animal species in which they were found. A possible problem with this approach is that, in some cases, a newly discovered virus might theoretically represent environmental contamination (as opposed to productive infection of the sampled animal). Our model provides a rough "back-of-the-envelope" approach to this question. As a concrete example, two recently discovered polyomavirus species whose genome sequences differ by about 20% were found separately in common voles and bank voles [52]. These two host species are thought to have diverged about 10 million years ago [65]. The two rodent-associated polyomaviruses differ from their nearest previously known relatives, human polyomaviruses WU and KI, by about 50%. Primates and rodents diverged about 90 million years ago [66]. Given the rough consistency of the observed divergences of the two new viruses with the >0.5% per million year "rule of thumb" shown in Fig 7, there seems to be no affirmative reason to suspect that the putative vole viruses originated in a non-rodent host. As polyomavirus phylogenetic trees become better populated, such guesswork could become increasingly confident.

It will be interesting to learn whether any un-recombined examples of the hypothetical vertebrate Arche lineage infect modern mammals. Since the chimeric bandicoot papilloma/polyomaviruses appear to carry an Arche-LT, it seems possible that Australian marsupials would be a promising group of animals in which to search for un-recombined Arche polyomaviruses. Similarly, a small Almi-VP1-like contig from the TSA dataset for helmeted guineafowl (*Numida meleagris*) raises the possibility that some modern birds may harbor un-recombined examples of the Almi clade. It will also be important to search for additional polyomavirus species in wild *Mus musculus*, as well as other common laboratory animals, such as zebrafish, sea urchin, *Caenorhabditis elegans*, *Xenopus laevis*, and *Drosophila melanogaster*. In addition to providing experimentally tractable models for exploring polyomavirus/host interactions, discovering new polyomavirus species in any of these animal lineages would shed additional light on the seemingly languid evolution of this fascinating family of viruses.

Methods

Sequence acquisition

Complete genome sequences of known polyomavirus species, as well as sub-genomic polyomavirus fragments, were downloaded from GenBank. Final database searches and downloads for sequences included in the shown phylogenetic analyses were performed on August 5, 2015. When necessary, the circular genome map was rearranged to comply with the convention that the initiator ATG of the Large T antigen (LT) CDS comprises the 5' end of the antisense strand (see genome maps in Fig 1). In some instances, predicted splice sites or initiation codon annotations were altered based on alignments against other known polyomaviruses. MacVector 13 software was used to construct graphical maps.

Polyomavirus sequences that share <85% genome-wide pairwise nucleotide identity with other polyomaviruses are traditionally considered to be distinct viral species [29]. A few exceptions to the species cutoff rule were made for polyomavirus genomes with ≥85% identity that were isolated from different animal species. Examples of this exception include LPV/Vervet3 and Vervet2/Baboon2/SA12. Multiple representatives of each polyomavirus species were included in instances where different isolates with 85–95% identity could be found within the designated species. Each polyomavirus species was assigned a familiar nickname based on either a common name for the animal species with which it is associated or an established abbreviation (e.g., SV40, BKV, JCV).

As part of an ongoing *Trematomus* species physiology study in the Ross Sea, we sampled seven individual *Trematomus pennellii* (common name: sharp-spined notothen) caught using hook and line in McMurdo Sound during the summer field season of 2012–2013. *T. pennellii* are benthic nototheniid fish with a maximum body length of ~24 cm and are endemic to the Southern Ocean at a typical depth of ~ 1–100 meters. Their range can extend as far as ~700 meters [67]. Approximately 1 g of stomach, gills, liver and skin from the seven fish were grouped and each sample type was homogenized in 20 ml of SM buffer (0.1 M NaCl, 50 mM Tris/HCl–pH 7.4, 10 mM $MgSO_4$) using a mortar and pestle, as previously described by Varsani et al. [68, 69]. Extracted DNA was sequenced on an Illumina HiSeq 2000 sequencer at Macrogen Inc. (South Korea) and the paired-end reads *de novo* assembled using ABySS v1.5.2 [70] assembler (kmer = 64). In BLASTX [71] analyses, we identified a contig of ~6000 nt from the stomach sample that had similarity to polyomavirus LT. Based on this ~6000 nt *de novo* assembled sequence contig we designed abutting primers (PES-F: 5'-GTC GAC TTC TGT GCT GAC GTG ACT GAG-3'; PES-R: 5'-AGG TCC AGC CAT CTT CGG TGT ATC ACT T-3') to recover the complete circular DNA molecule encompassing the LT-like sequence. Using the abutting primer pair with KAPA Hifi Hotstart DNA polymerase (Kapa Biosystems, USA) we amplified the polyomavirus-like circular molecule using the following protocol: initial denaturation at 95°C for 3 min followed by 25 cycles at 98°C for 20 sec, 60°C for 15 sec, 72°C for 5min and a final extension at 72°C for 5min. We were able to recover the ~6 kb amplicon from the liver and the stomach samples and these were cloned into pJET1.2 plasmid (ThermoFisher, USA), and Sanger-sequenced by primer walking at Macrogen Inc. (Korea). The Sanger-sequences were assembled using DNAbaser v.4 (Heracle BioSoft S.R.L., Romania). The complete genome of sharp-spined notothen (*Trematomus pennellii*) polyomavirus 1 (6219 nt) was 100% identical in both the stomach and liver deep sequencing samples and has been deposited in GenBank (accession KP768176).

Giant guitarfish (*Rhynchobatus djiddensis*) polyomavirus 1 (GenBank accession KP264963) was detected using previously reported methods [72] in specimens from an aquarium animal suffering from proliferative skin lesions. The guitarfish polyomavirus was discovered alongside much higher levels of a member of a different DNA virus family. The sequence of the other virus, and details on the pathology of the guitarfish specimen, will be published in a separate report.

Previously reported methods were used to discover sheep (*Ovis aries*) meat-associated polyomavirus 1 (GenBank accession KP890267) in a sample of ground lamb meat purchased at a US supermarket [18].

Baja California bark scorpion (*Centruroides exilicauda*) polyomavirus 1 was initially identified in a TBLASTN [71] search of the NCBI Whole Genome Shotgun database (WGS) using the LT protein sequence of black sea bass polyomavirus as bait. A single contig, accession number AXZI01204118, was curated back to the original reads (Sequence Read Archive (SRA) accession number SRX476227). The back-curation revealed that small segments were missing from the ends of the original contig. The SRA dataset contained at least three distinct viral sequence variants. The two most abundant variants were compiled separately. The putative LT intron (where the original contig ends fell) was an apparent polymorphic hotspot. The extensive variation in this portion of the polyomavirus genome could explain why the contig assembly process failed at this particular point. No chimeric reads (potentially representing integration of the viral genome into the host animal's DNA) were detected, suggesting that both viral genomes were carried in an episomal form. Because current GenBank policies do not allow deposits of third-party sequence assemblies, the two scorpion polyomavirus sequences were instead deposited at EMBL (accession numbers LN846618 and LN846619).

Abbreviation and naming conventions

In the interest of clarity, this manuscript favors the use of host animal common names and avoids the extensive use of abbreviations. In our view, when abbreviations are necessary they should be short, easily inferred as representing the host animal species of origin, and, ideally, should serve as pronounceable "sigla" http://ictvonline.org/codeofvirusclassification_2012.asp. We suggest that newly coined abbreviations should use a condensation of a common name for the host animal and "PyV" for polyomavirus. Examples of pronounceable abbreviations might be ShePyV1 for supermarket sheep meat-associated polyomavirus 1 or ChimPyV1 for *Pan troglodytes verus* polyomavirus 1.

Possible accessory proteins were detected by analyzing genome sequences for ORFs of at least 25 codons. Small T antigen (sT) was defined as an ORF encoding an ATG-initiated protein of at least 10 kD near the 5' end of the LT gene. ALTO was defined as a >250 bp ATG-initiated ORF in the LT +1 frame located near the 5' end of the LT exon encoding the helicase domain. In nearly all cases, the ALTO ORF overlaps the segment of LT encoding the putative pRb-interaction motif LXCXE. Agno was defined as an ORF encoding a >10 kD protein initiated from an ATG codon located upstream of the inferred VP2 ORF.

An attempt was made to infer the LT-binding sites associated with the viral origin of replication. The "classic" Oris of SV40 and MPyV were defined as paired palindromic GRGGCY motifs adjacent to an A/T tract. Hypothetical Avi and fish Ori sequences were defined as paired palindromic YYTGSCA motifs adjacent to an A/T tract. A hypothetical arthropod Ori was defined as paired palindromic ATCACGYG motifs flanked on both sides by A/T tracts.

Structural modeling

The analyses of the Large T antigens (LTs) from scorpion, guitarfish and notothen polyomaviruses were performed using multiple bioinformatics tools from the psipred server, http://bioinf.cs.ucl.ac.uk/psipred/?disopred=1 [73]. In order to obtain models of high quality, the structural relationships between the novel LTs and previously solved protein structures were determined through fold recognition using pGenTHREADER and pDomTHREADER from the psipred server [74]. Matching structures with the highest scores were then selected as templates for predicting structures of the novel LTs. Models for DNAJ and OBD-Zn-ATPase were generated separately. All structures and models were visualized and compared using PyMOL (The PyMOL Molecular Graphics System, Version 1.2r3pre, Schrödinger, LLC).

MEME suite 4.10.0 http://meme.nbcr.net/meme/ [75] was used to facilitate the identification of possible palindromically arranged LT-binding motifs in candidate Ori regions. Inferred candidate motifs are indicated in the legend of Fig 5.

Phylogenetic analyses

Curated polyomavirus sequence sets used in this work are posted at http://home.ccr.cancer.gov/Lco/PyVE.asp. The site includes annotated genomes for examples of all currently known polyomavirus species and compiled protein sequences.

Initial exploratory phylogenetic analyses were performed using the Phylogeny.fr website http://phylogeny.lirmm.fr/ in "One Click" mode without Gblocks [76]. FigTree software v1.4.2 http://tree.bio.ed.ac.uk/software/figtree/ was used to display trees. Confirmatory analyses were performed by aligning sequences using MUSCLE [77] and manually editing the output. Maximum-likelihood phylogenetic trees (with approximate likelihood branch support, aLRT) were inferred using PHYML 3 [78] with LG+I+G as the best substitution model determined using ProtTest [79]. Branches with <80% aLRT branch support were collapsed. Confirmatory Bayesian phylogenetic analyses showed essentially identical tree topology. However, Bayesian

phylogenetic trees for VP1 proteins showed poor support values. The results are consistent with a pending ICTV proposal http://talk.ictvonline.org/files/proposals/animal_dna_viruses_and_retroviruses/m/animal_dna_under_consideration/5637.aspx. Because of their better bootstrap values, maximum-likelihood analyses were favored for the current study.

Nucleotide divergence calculations were performed for individual sequence pairs using Sequence Demarcation Tool (SDT) version 1.2 in MUSCLE mode [80, 81] http://web.cbio.uct.ac.za/~brejnev/. Pairwise calculations were performed on discrete clades, specifically: the separate "Monominor" A and B sub-clades, the Ortho-LT clade (excluding WU and KI), the "Blympho" clade (which houses B-lymphotropic polyomavirus (LPV) and HPyV9), and the two small clades that separately house TSV and Chimp3. For Avi polyomaviruses, sequences found in the "fragments" tab of S1 File were included in the analysis. The analysis was performed in January 2015 and does not include polyomavirus sequences made public after that time.

Estimates of the time to last common ancestor of animal species pairs were based on various references [82–87]. In most cases, the estimates were based primarily on sequence analyses, as opposed to fossil records. Estimates are consistent (to within 10%) with the "Expert Result" in Time Tree of Life http://www.timetree.org/ [65].

Test of polyomavirus and host co-speciation

To ensure maintenance of codon information, nucleotide sequences for the VP1 and Large T coding regions were translated into protein sequences. The translated proteins were aligned using Mafft (implementing the L-ins-I algorithm)[88]. Next, the aligned protein sequences were reverse translated into nucleotide sequences. Finally, the individual alignments were concatenated into a supermatrix.

To test for potential substitutional saturation [89, 90] the index of substitutional saturation statistic was calculated for the supermatrix (test implemented in DAMBE version 6.0.0 [91]). The results indicated that the observed saturation index of 0.5865 was smaller than the critical saturation index (Iss.c = 0.8023), suggesting that the sequences have experienced little substitutional saturation, thus conserving sufficient phylogenetic signal for phylogenetic reconstruction.

PartitionFinder v1.1.1 was used to select the best-fit partitioning schemes and partition-specific substitution models under the Bayesian information criterion (BIC) [92]. PartitionFinder suggested the use of 4 different partitions [(Large T codon position 1, VP1_CP1), (Large T_CP2, VP1_CP2), (VP1_CP3), and (Large T_CP3)]. All partitions were estimated to evolve under the General Time Reversible (GTR) model of nucleotide substitution with invariant sites (I) and Γ distributed rate variation among sites (GTR+I+G).

Parafit was used to formally test the hypothesis of coevolution between Monominor polyomaviruses and their associated hosts [93, 94]. The null hypothesis (H0) of the global test is that the evolution of polyomavirus species and the host animals in which they were found has been independent. The test, as implemented within the R package (APE) version 3.3 [95] requires two phylogenetic trees and the set of host-parasite association links. The host tree was constructed using phyloT (available from http://phylot.biobyte.de/). PhyloT uses NCBI taxonomy identification numbers to generate a phylogenetic tree. The obtained tree was manually edited to include branch lengths of unit length. MrBayes 3.2.6 [96, 97], as implemented within the CIPRES Science Gateway V. 3.3 [98], was used to estimate the Monominor phylogenetic tree. The selected GTR+I+G substitution model was implemented. The analysis was run using two independent chains for a total chain length of one million iterations, with a sampling frequency every 1,000th step. Following a 10% burn-in, the tree was summarized. The

GlobalParafit was estimated to be 3633.384, with a p-value = 0.0258 (based on 1,000 permutations), providing support in favor of co-speciation.

Estimation of the evolutionary rate of the Monominor clade

The supermatrix described in the previous section was used for this analysis.

The Bayesian analysis (Beast 1.8 [99]) as implemented within the CIPRES Science Gateway V. 3.3 [98], was performed using linked substitution rates for the first and second codon positions (CP_{12}), while allowing independent rates in CP_3. The uncorrelated lognormal relaxed molecular clock was used to accommodate rate variation among lineages. Monophyletic constraints were placed on the separate Monominor A and B clades. Based on the posterior distributions obtained for the host [84, 100], normal priors were imposed on specific nodes used to calibrate the evolutionary rates (S5 File). Three independent Markov Chain Monte Carlo (MCMC) analyses were run for 10 million generations each, with samples from the posterior drawn every 1,000 generations. The first 10% of each run was discarded prior to the construction of the posterior probability distributions of parameters. Each analysis was run sufficiently long that effective sample sizes for parameters were >400. The results from the three runs were combined to generate a maximum clade credibility tree and rate and divergence time summaries (S5 File).

Supporting Information

S1 Fig. In situ hybridization analysis of guitarfish polyomavirus in resolving skin lesions. A hybridization assay adapted from previously reported methods [101, 102] was used to stain sections of guitarfish skin lesions biopsied during the resolution of symptoms. Guitarfish polyomavirus VP1 probe hybridization signal (red) was observed in unidentified round cells. Arrows indicate selected positively-stained cells. The cells appear to have histiocytic or macrophage-like morphology. Free speckled brown/black patterns are attributable to melanin. Scale bar represents 20 μm.
(TIF)

S2 Fig. Conservation maps for LT DNAJ and Zn-ATPase domains. The conservation maps were generated using the ConSurf server (http://consurf.tau.ac.il/), and then visualized using Chimera, http://www.cgl.ucsf.edu/chimera/ [103]. Panel A: the DNAJ domain conservation map was generated using DNAJ domain sequences from 34 polyomavirus LTs in the Uniref90 collection. The black oval indicates the highly conserved HPDKGG motif. Panel B: conservation map of LT Zn-ATPase domains. The map was generated with 69 LT sequences from the Uniref90 collection. The black oval indicates the Walker motifs required for binding and hydrolysis of ATP. Fewer DNAJ domains were included in this analysis due to a stringent default E-value (0.0001) setting. This indicates a greater level of variation among the DNAJ domains in contrast to the Zn-ATPase domains of LTs.
(TIF)

S3 Fig. Analysis of LT +1 frame ORFs. The genome map depicts BKV-I as a representative example of the small clade of primate polyomaviruses encompassing SV40.
(TIF)

S4 Fig. Phylogenetic illustration of select pairwise divergences. Phylogeny.fr "one click" settings were used to draw a phylogenetic tree for the complete genomes (nucleotide) of selected members of the Almi-LT and Ortho-LT clades. The tree is arbitrarily rooted on human polyomavirus 9. The selected Almi species have only one minor capsid protein and thus belong to a "Monominor" sub-clade within clade Almi. Numbers within the nodes indicate the estimated time

(in millions of years ago) of the last common ancestor of host animals contained within the node. Branches are color-coded based on host animal families. Percentages indicate the pairwise nucleotide divergence of the complete genomes of the indicated polyomavirus species pair. Nodes that encompass possible intra-host polyomavirus divergence events are marked with asterisks.
(TIF)

S5 Fig. LT and VP1 co-divergence. SDT was used to calculate the percent divergence of LT and VP1 proteins for individual pairs of polyomaviruses. The linear relationship between LT and VP1 divergences in Ortho, Almi, and fish clades suggests that the two proteins independently diverge at a roughly similar rate. The disconnection of the Avi and Wuki clades can most easily be explained by ancient recombination events (see Fig 5).
(TIF)

S1 File. Naming key.
(XLSX)

S2 File. LT phylogenetic tree (FigTree format http://tree.bio.ed.ac.uk/software/figtree/).
(TRE)

S3 File. VP1 phylogenetic tree (FigTree format).
(TRE)

S4 File. VP2 phylogenetic tree (FigTree format).
(TRE)

S5 File. Time-resolved phylogenetic tree of the Monominor polyomavirus clade. Tabular data refers to the numbered nodes in the phylogenetic tree. The table indicates the posterior probability, node age (including 95% HPD), average (95% HPD) rate for each partition, and presence of constraints for individual nodes. The phylogenetic tree displays the evolutionary relationship between members of the Monominor clade. The tree and geological column were generated using the (APE) package within R. The scale bar indicates millions of years before the present. The inset shows the median evolutionary rate (with 95% HPD) of the 1^{st}-2^{nd} and 3^{rd} codon positions.
(XLSX)

Acknowledgments

The authors are grateful to Efrem Lim, Matt Daugherty, and other members of the ICTV Polyomavirus Study Group for an engaging and useful series of discussions about key issues in polyomavirus taxonomy. This study utilized the high-performance computational capabilities of the Helix Systems at the National Institutes of Health, Bethesda, MD (http://helix.nih.gov).

Author Contributions

Conceived and designed the experiments: CBB KVD AP PA JPK JMP AAM ACC AJM JAD ED TFFN KF CA SK WD DVP AV. Performed the experiments: CBB KVD AP EMG MJT PA JPK JAD TFFN KF CA SK DVP AV. Analyzed the data: CBB KVD AP PA JPK JMP AAM ACC AJM JAD ED TFFN KF CA SK WD DVP AV. Wrote the paper: CBB KVD AP PA JPK JMP AAM ACC AJM JAD ED TFFN KF CA SK WD DVP AV.

References

1. Schowalter RM, Buck CB. The Merkel cell polyomavirus minor capsid protein. PLoS pathogens. 2013; 9(8):e1003558. Epub 2013/08/31. doi: 10.1371/journal.ppat.1003558 PMID: 23990782; PubMed Central PMCID: PMC3749969.

2. Gross L. A filterable agent, recovered from Ak leukemic extracts, causing salivary gland carcinomas in C3H mice. Proceedings of the Society for Experimental Biology and Medicine Society for Experimental Biology and Medicine (New York, NY. 1953; 83(2):414–21. PMID: 13064287.
3. Stewart SE, Eddy BE, Gochenour AM, Borgese NG, Grubbs GE. The induction of neoplasms with a substance released from mouse tumors by tissue culture. Virology. 1957; 3(2):380–400. PMID: 13434017.
4. Dang-Tan T, Mahmud SM, Puntoni R, Franco EL. Polio vaccines, Simian Virus 40, and human cancer: the epidemiologic evidence for a causal association. Oncogene. 2004; 23(38):6535–40. Epub 2004/08/24. doi: 10.1038/sj.onc.1207877 PMID: 15322523.
5. DeCaprio JA, Garcea RL. A cornucopia of human polyomaviruses. Nature reviews Microbiology. 2013; 11(4):264–76. doi: 10.1038/nrmicro2992 PMID: 23474680; PubMed Central PMCID: PMC3928796.
6. Bouvard V, Baan RA, Grosse Y, Lauby-Secretan B, El Ghissassi F, Benbrahim-Tallaa L, et al. Carcinogenicity of malaria and of some polyomaviruses. Lancet Oncol. 2012; 13(4):339–40. Epub 2012/05/12. PMID: 22577663.
7. Padgett BL, Walker DL, ZuRhein GM, Eckroade RJ, Dessel BH. Cultivation of papova-like virus from human brain with progressive multifocal leucoencephalopathy. Lancet. 1971; 1(7712):1257–60. Epub 1971/06/19. PMID: 4104715.
8. Gardner SD, Field AM, Coleman DV, Hulme B. New human papovavirus (B.K.) isolated from urine after renal transplantation. Lancet. 1971; 1(7712):1253–7. Epub 1971/06/19. PMID: 4104714.
9. Dalianis T, Hirsch HH. Human polyomaviruses in disease and cancer. Virology. 2013; 437(2):63–72. doi: 10.1016/j.virol.2012.12.015 PMID: 23357733.
10. Feng H, Shuda M, Chang Y, Moore PS. Clonal integration of a polyomavirus in human Merkel cell carcinoma. Science. 2008; 319(5866):1096–100. PMID: 18202256. doi: 10.1126/science.1152586
11. Pantulu ND, Pallasch CP, Kurz AK, Kassem A, Frenzel L, Sodenkamp S, et al. Detection of a novel truncating Merkel cell polyomavirus large T antigen deletion in chronic lymphocytic leukemia cells. Blood. 2010; 116(24):5280–4. doi: 10.1182/blood-2010-02-269829 PMID: 20817850.
12. Rennspiess D, Pujari S, Keijzers M, Abdul-Hamid MA, Hochstenbag M, Dingemans AM, et al. Detection of human polyomavirus 7 in human thymic epithelial tumors. Journal of thoracic oncology: official publication of the International Association for the Study of Lung Cancer. 2015; 10(2):360–6. doi: 10.1097/JTO.0000000000000390 PMID: 25526237; PubMed Central PMCID: PMC4304941.
13. Ho J, Jedrych JJ, Feng H, Natalie AA, Grandinetti L, Mirvish E, et al. Human Polyomavirus 7-Associated Pruritic Rash and Viremia in Transplant Recipients. The Journal of infectious diseases. 2014. doi: 10.1093/infdis/jiu524 PMID: 25231015.
14. Mishra N, Pereira M, Rhodes RH, An P, Pipas JM, Jain K, et al. Identification of a novel polyomavirus in a pancreatic transplant recipient with retinal blindness and vasculitic myopathy. The Journal of infectious diseases. 2014; 210(10):1595–9. doi: 10.1093/infdis/jiu250 PMID: 24795478.
15. Peretti A, FitzGerald PC, Bliskovsky V, Pastrana DV, Buck CB. Genome Sequence of a Fish-Associated Polyomavirus, Black Sea Bass (Centropristis striata) Polyomavirus 1. Genome Announc. 2015; 3(1):e01476–13. doi: 10.1128/genomeA.01476-14 PMID: 25635011; PubMed Central PMCID: PMC4319505.
16. Blair JE, Hedges SB. Molecular phylogeny and divergence times of deuterostome animals. Molecular biology and evolution. 2005; 22(11):2275–84. doi: 10.1093/molbev/msi225 PMID: 16049193.
17. zur Hausen H. Red meat consumption and cancer: reasons to suspect involvement of bovine infectious factors in colorectal cancer. International journal of cancer. 2012; 130(11):2475–83. doi: 10.1002/ijc.27413 PMID: 22212999.
18. Peretti A, FitzGerald PC, Bliskovsky V, Buck CB, Pastrana DV. Hamburger polyomaviruses. J Gen Virol. 2015; 96(Pt 4):833–9. doi: 10.1099/vir.0.000033 PMID: 25568187.
19. Editorial. Horsemeat in 'beef' products: European Commission summarises progress. Vet Rec. 2014; 174(11):264. doi: 10.1136/vr.g2080 PMID: 24627501.
20. An P, Saenz Robles MT, Pipas JM. Large T antigens of polyomaviruses: amazing molecular machines. Annual review of microbiology. 2012; 66:213–36. doi: 10.1146/annurev-micro-092611-150154 PMID: 22994493.
21. Chang YP, Xu M, Machado AC, Yu XJ, Rohs R, Chen XS. Mechanism of origin DNA recognition and assembly of an initiator-helicase complex by SV40 large tumor antigen. Cell reports. 2013; 3(4):1117–27. doi: 10.1016/j.celrep.2013.03.002 PMID: 23545501; PubMed Central PMCID: PMC3748285.
22. Kim HY, Ahn BY, Cho Y. Structural basis for the inactivation of retinoblastoma tumor suppressor by SV40 large T antigen. EMBO J. 2001; 20(1–2):295–304. doi: 10.1093/emboj/20.1.295 PMID: 11226179; PubMed Central PMCID: PMC140208.

23. Woolford L, Rector A, Van Ranst M, Ducki A, Bennett MD, Nicholls PK, et al. A novel virus detected in papillomas and carcinomas of the endangered western barred bandicoot (Perameles bougainville) exhibits genomic features of both the Papillomaviridae and Polyomaviridae. J Virol. 2007; 81 (24):13280–90. PMID: 17898069.

24. Bennett MD, Reiss A, Stevens H, Heylen E, Van Ranst M, Wayne A, et al. The first complete papillomavirus genome characterized from a marsupial host: a novel isolate from Bettongia penicillata. J Virol. 2010; 84(10):5448–53. doi: 10.1128/JVI.02635-09 PMID: 20200246; PubMed Central PMCID: PMC2863809.

25. Mizutani T, Sayama Y, Nakanishi A, Ochiai H, Sakai K, Wakabayashi K, et al. Novel DNA virus isolated from samples showing endothelial cell necrosis in the Japanese eel, Anguilla japonica. Virology. 2011; 412(1):179–87. doi: 10.1016/j.virol.2010.12.057 PMID: 21277610.

26. Wen CM, Chen MM, Wang CS, Liu PC, Nan FH. Isolation of a novel polyomavirus, related to Japanese eel endothelial cell-infecting virus, from marbled eels, Anguilla marmorata (Quoy & Gaimard). J Fish Dis. 2015. doi: 10.1111/jfd.12423 PMID: 26566584.

27. Carter JJ, Daugherty MD, Qi X, Bheda-Malge A, Wipf GC, Robinson K, et al. Identification of an overprinting gene in Merkel cell polyomavirus provides evolutionary insight into the birth of viral genes. Proceedings of the National Academy of Sciences of the United States of America. 2013; 110 (31):12744–9. doi: 10.1073/pnas.1303526110 PMID: 23847207; PubMed Central PMCID: PMC3732942.

28. Feltkamp MC, Kazem S, van der Meijden E, Lauber C, Gorbalenya AE. From Stockholm to Malawi: recent developments in studying human polyomaviruses. J Gen Virol. 2013; 94(Pt 3):482–96. doi: 10.1099/vir.0.048462–0 PMID: 23255626.

29. Johne R, Buck CB, Allander T, Atwood WJ, Garcea RL, Imperiale MJ, et al. Taxonomical developments in the family Polyomaviridae. Arch Virol. 2011; 156(9):1627–34. Epub 2011/05/13. doi: 10.1007/s00705-011-1008-x PMID: 21562881.

30. Tsang SH, Wang R, Nakamaru-Ogiso E, Knight SAB, Buck CB, You J. The Oncogenic Small Tumor Antigen of Merkel Cell Polyomavirus is an Iron- Sulfur Cluster Protein that Enhances Viral DNA Replication J Virol. 2015;in press.

31. Lau L, Gray EE, Brunette RL, Stetson DB. DNA tumor virus oncogenes antagonize the cGAS-STING DNA sensing pathway. Science. 2015. doi: 10.1126/science.aab3291 PMID: 26405230.

32. Cho US, Morrone S, Sablina AA, Arroyo JD, Hahn WC, Xu W. Structural basis of PP2A inhibition by small t antigen. PLoS biology. 2007; 5(8):e202. doi: 10.1371/journal.pbio.0050202 PMID: 17608567; PubMed Central PMCID: PMC1945078.

33. Lauber C, Kazem S, Kravchenko AA, Feltkamp MC, Gorbalenya AE. Interspecific adaptation by binary choice at de novo polyomavirus T antigen site through accelerated codon-constrained Val-Ala toggling within an intrinsically disordered region. Nucleic Acids Res. 2015; 43(10):4800–13. doi: 10.1093/nar/gkv378 PMID: 25904630; PubMed Central PMCID: PMC4446436.

34. van der Meijden E, Kazem S, Dargel CA, van Vuren N, Hensbergen PJ, Feltkamp MC. Characterization of T Antigens, Including Middle T and Alternative T, Expressed by the Human Polyomavirus Associated with Trichodysplasia Spinulosa. J Virol. 2015; 89(18):9427–39. doi: 10.1128/JVI.00911-15 PMID: 26136575.

35. Fluck MM, Schaffhausen BS. Lessons in signaling and tumorigenesis from polyomavirus middle T antigen. Microbiology and molecular biology reviews: MMBR. 2009; 73(3):542–63, Table of Contents. doi: 10.1128/MMBR.00009-09 PMID: 19721090; PubMed Central PMCID: PMC2738132.

36. Jackson JA. Analysis of parasite host-switching: limitations on the use of phylogenies. Parasitology. 1999; 119 Suppl:S111–23. PMID: 11254144.

37. Sharp PM, Simmonds P. Evaluating the evidence for virus/host co-evolution. Current opinion in virology. 2011; 1(5):436–41. doi: 10.1016/j.coviro.2011.10.018 PMID: 22440848.

38. Malik HS, Burke WD, Eickbush TH. The age and evolution of non-LTR retrotransposable elements. Molecular biology and evolution. 1999; 16(6):793–805. PMID: 10368957.

39. Rethwilm A, Bodem J. Evolution of foamy viruses: the most ancient of all retroviruses. Viruses. 2013; 5(10):2349–74. doi: 10.3390/v5102349 PMID: 24072062; PubMed Central PMCID: PMC3814592.

40. Perez-Losada M, Christensen RG, McClellan DA, Adams BJ, Viscidi RP, Demma JC, et al. Comparing phylogenetic codivergence between polyomaviruses and their hosts. J Virol. 2006; 80(12):5663–9. doi: 10.1128/JVI.00056-06 PMID: 16731904; PubMed Central PMCID: PMC1472594.

41. Tao Y, Shi M, Conrardy C, Kuzmin IV, Recuenco S, Agwanda B, et al. Discovery of diverse polyomaviruses in bats and the evolutionary history of the Polyomaviridae. J Gen Virol. 2013; 94(Pt 4):738–48. doi: 10.1099/vir.0.047928–0 PMID: 23239573.

42. Krumbholz A, Bininda-Emonds OR, Wutzler P, Zell R. Phylogenetics, evolution, and medical importance of polyomaviruses. Infection, genetics and evolution: journal of molecular epidemiology and evolutionary genetics in infectious diseases. 2009; 9(5):784–99. doi: 10.1016/j.meegid.2009.04.008 PMID: 19379840.

43. Johne R, Muller H. Polyomaviruses of birds: etiologic agents of inflammatory diseases in a tumor virus family. J Virol. 2007; 81(21):11554–9. doi: 10.1128/JVI.01178-07 PMID: 17715213; PubMed Central PMCID: PMC2168798.

44. Johne R, Muller H. Avian polymavirus in wild birds: genome analysis of isolates from Falconiformes and Psittaciformes. Arch Virol. 1998; 143(8):1501–12. PMID: 9739329.

45. Katoh H, Ohya K, Une Y, Yamaguchi T, Fukushi H. Molecular characterization of avian polyomavirus isolated from psittacine birds based on the whole genome sequence analysis. Vet Microbiol. 2009; 138(1–2):69–77. doi: 10.1016/j.vetmic.2009.03.007 PMID: 19345024.

46. Lafferty SL, Fudge AM, Schmidt RE, Wilson VG, Phalen DN. Avian polyomavirus infection and disease in a green aracaris (Pteroglossus viridis). Avian Dis. 1999; 43(3):577–85. PMID: 10494430.

47. Zhuang Q, Chen J, Mushtaq MH, Chen J, Liu S, Hou G, et al. Prevalence and genetic characterization of avian polyomavirus and psittacine beak and feather disease virus isolated from budgerigars in Mainland China. Arch Virol. 2012; 157(1):53–61. doi: 10.1007/s00705-011-1138-1 PMID: 22002652.

48. Johne R, Muller H. The genome of goose hemorrhagic polyomavirus, a new member of the proposed subgenus Avipolyomavirus. Virology. 2003; 308(2):291–302. PMID: 12706079.

49. Corrand L, Gelfi J, Albaric O, Etievant M, Pingret JL, Guerin JL. Pathological and epidemiological significance of goose haemorrhagic polyomavirus infection in ducks. Avian Pathol. 2011; 40(4):355–60. doi: 10.1080/03079457.2011.582481 PMID: 21812713.

50. McGeoch DJ, Rixon FJ, Davison AJ. Topics in herpesvirus genomics and evolution. Virus Res. 2006; 117(1):90–104. doi: 10.1016/j.virusres.2006.01.002 PMID: 16490275.

51. Niewiadomska AM, Gifford RJ. The extraordinary evolutionary history of the reticuloendotheliosis viruses. PLoS biology. 2013; 11(8):e1001642. doi: 10.1371/journal.pbio.1001642 PMID: 24013706; PubMed Central PMCID: PMC3754887.

52. Nainys J, Timinskas A, Schneider J, Ulrich RG, Gedvilaite A. Identification of Two Novel Members of the Tentative Genus Wukipolyomavirus in Wild Rodents. PloS one. 2015; 10(10):e0140916. doi: 10.1371/journal.pone.0140916 PMID: 26474048; PubMed Central PMCID: PMCPMC4608572.

53. Carroll J, Dey D, Kreisman L, Velupillai P, Dahl J, Telford S, et al. Receptor-binding and oncogenic properties of polyoma viruses isolated from feral mice. PLoS pathogens. 2007; 3(12):e179. doi: 10.1371/journal.ppat.0030179 PMID: 18085820; PubMed Central PMCID: PMCPMC2134959.

54. Dayaram A, Piasecki T, Chrzastek K, White R, Julian L, van Bysterveldt K, et al. Avian Polyomavirus Genome Sequences Recovered from Parrots in Captive Breeding Facilities in Poland. Genome Announc. 2015; 3(5). doi: 10.1128/genomeA.00986-15 PMID: 26404592; PubMed Central PMCID: PMCPMC4582568.

55. Schowalter RM, Pastrana DV, Pumphrey KA, Moyer AL, Buck CB. Merkel Cell Polyomavirus and Two Previously Unknown Polyomaviruses Are Chronically Shed From Human Skin. Cell host & microbe. 2010; 7(6):509–15.

56. Kazem S, Lauber C, van der Meijden E, Kooijman S, Kravchenko AA, TrichSpin N, et al. Limited variation during circulation of a polyomavirus in the human population involves the COCO-VA toggling site of Middle and Alternative T-antigen(s). Virology. 2015; 487:129–40. doi: 10.1016/j.virol.2015.09.013 PMID: 26519899.

57. Krumbholz A, Bininda-Emonds OR, Wutzler P, Zell R. Evolution of four BK virus subtypes. Infection, genetics and evolution: journal of molecular epidemiology and evolutionary genetics in infectious diseases. 2008; 8(5):632–43. doi: 10.1016/j.meegid.2008.05.006 PMID: 18582602.

58. Gottschling M, Goker M, Stamatakis A, Bininda-Emonds OR, Nindl I, Bravo IG. Quantifying the phylodynamic forces driving papillomavirus evolution. Molecular biology and evolution. 2011; 28(7):2101–13. doi: 10.1093/molbev/msr030 PMID: 21285031.

59. Van Doorslaer K. Evolution of the papillomaviridae. Virology. 2013; 445(1–2):11–20. doi: 10.1016/j.virol.2013.05.012 PMID: 23769415.

60. Rector A, Lemey P, Tachezy R, Mostmans S, Ghim SJ, Van Doorslaer K, et al. Ancient papillomavirus-host co-speciation in Felidae. Genome biology. 2007; 8(4):R57. doi: 10.1186/gb-2007-8-4-r57 PMID: 17430578; PubMed Central PMCID: PMC1896010.

61. Siebrasse EA, Pastrana DV, Nguyen NL, Wang A, Roth MJ, Holland SM, et al. WU polyomavirus in respiratory epithelial cells from lung transplant patient with Job syndrome. Emerging infectious diseases. 2015; 21(1):103–6. doi: 10.3201/eid2101.140855 PubMed Central PMCID: PMC25531075. PMID: 25531075

62. Gaynor AM, Nissen MD, Whiley DM, Mackay IM, Lambert SB, Wu G, et al. Identification of a novel polyomavirus from patients with acute respiratory tract infections. PLoS pathogens. 2007; 3(5):e64. PMID: 17480120.

63. Allander T, Andreasson K, Gupta S, Bjerkner A, Bogdanovic G, Persson MA, et al. Identification of a third human polyomavirus. J Virol. 2007; 81(8):4130–6. PMID: 17287263.

64. Polyomaviridae Study Group of the International Committee on Taxonomy of V, Calvignac-Spencer S, Feltkamp MC, Daugherty MD, Moens U, Ramqvist T, et al. A taxonomy update for the family Polyomaviridae. Arch Virol. 2016. doi: 10.1007/s00705-016-2794-y PMID: 26923930.

65. Hedges SB, Dudley J, Kumar S. TimeTree: a public knowledge-base of divergence times among organisms. Bioinformatics. 2006; 22(23):2971–2. doi: 10.1093/bioinformatics/btl505 PMID: 17021158.

66. Hedges SB, Marin J, Suleski M, Paymer M, Kumar S. Tree of life reveals clock-like speciation and diversification. Molecular biology and evolution. 2015; 32(4):835–45. doi: 10.1093/molbev/msv037 PMID: 25739733; PubMed Central PMCID: PMCPMC4379413.

67. DeWitt HH, Heemstra PC, Gon O. Nototheniidae. Fishes of the Southern Ocean. Grahamstown: JLB Smith Institute of Ichthyology; 1990. p. 279–331.

68. Varsani A, Porzig EL, Jennings S, Kraberger S, Farkas K, Julian L, et al. Identification of an avian polyomavirus associated with Adelie penguins (Pygoscelis adeliae). J Gen Virol. 2015; 96(Pt 4):851–7. doi: 10.1099/vir.0.000038 PMID: 25537375.

69. Varsani A, Kraberger S, Jennings S, Porzig EL, Julian L, Massaro M, et al. A novel papillomavirus in Adelie penguin (Pygoscelis adeliae) faeces sampled at the Cape Crozier colony, Antarctica. J Gen Virol. 2014; 95(Pt 6):1352–65. doi: 10.1099/vir.0.064436–0 PMID: 24686913.

70. Simpson JT, Wong K, Jackman SD, Schein JE, Jones SJ, Birol I. ABySS: a parallel assembler for short read sequence data. Genome Res. 2009; 19(6):1117–23. doi: 10.1101/gr.089532.108 PMID: 19251739; PubMed Central PMCID: PMC2694472.

71. Altschul SF, Gish W, Miller W, Myers EW, Lipman DJ. Basic local alignment search tool. Journal of molecular biology. 1990; 215(3):403–10. doi: 10.1016/S0022-2836(05)80360-2 PMID: 2231712.

72. Zhang W, Li L, Deng X, Kapusinszky B, Delwart E. What is for dinner? Viral metagenomics of US store bought beef, pork, and chicken. Virology. 2014; 468-470C:303–10. doi: 10.1016/j.virol.2014.08.025 PMID: 25217712.

73. Buchan DW, Minneci F, Nugent TC, Bryson K, Jones DT. Scalable web services for the PSIPRED Protein Analysis Workbench. Nucleic Acids Res. 2013; 41(Web Server issue):W349–57. doi: 10.1093/nar/gkt381 PMID: 23748958; PubMed Central PMCID: PMC3692098.

74. Lobley A, Sadowski MI, Jones DT. pGenTHREADER and pDomTHREADER: new methods for improved protein fold recognition and superfamily discrimination. Bioinformatics. 2009; 25(14):1761–7. doi: 10.1093/bioinformatics/btp302 PMID: 19429599.

75. Bailey TL, Elkan C. Fitting a mixture model by expectation maximization to discover motifs in biopolymers. Proceedings / International Conference on Intelligent Systems for Molecular Biology; ISMB International Conference on Intelligent Systems for Molecular Biology. 1994; 2:28–36. PMID: 7584402.

76. Dereeper A, Guignon V, Blanc G, Audic S, Buffet S, Chevenet F, et al. Phylogeny.fr: robust phylogenetic analysis for the non-specialist. Nucleic Acids Res. 2008; 36(Web Server issue):W465–9. doi: 10.1093/nar/gkn180 PMID: 18424797; PubMed Central PMCID: PMC2447785.

77. Edgar RC. MUSCLE: multiple sequence alignment with high accuracy and high throughput. Nucleic Acids Res. 2004; 32(5):1792–7. doi: 10.1093/nar/gkh340 PMID: 15034147; PubMed Central PMCID: PMC390337.

78. Guindon S, Dufayard JF, Lefort V, Anisimova M, Hordijk W, Gascuel O. New algorithms and methods to estimate maximum-likelihood phylogenies: assessing the performance of PhyML 3.0. Syst Biol. 2010; 59(3):307–21. doi: 10.1093/sysbio/syq010 PMID: 20525638.

79. Darriba D, Taboada GL, Doallo R, Posada D. jModelTest 2: more models, new heuristics and parallel computing. Nature methods. 2012; 9(8):772. doi: 10.1038/nmeth.2109 PMID: 22847109.

80. Muhire B, Martin DP, Brown JK, Navas-Castillo J, Moriones E, Zerbini FM, et al. A genome-wide pairwise-identity-based proposal for the classification of viruses in the genus Mastrevirus (family Geminiviridae). Arch Virol. 2013; 158(6):1411–24. doi: 10.1007/s00705-012-1601-7 PMID: 23340592.

81. Muhire BM, Varsani A, Martin DP. SDT: a virus classification tool based on pairwise sequence alignment and identity calculation. PloS one. 2014; 9(9):e108277. doi: 10.1371/journal.pone.0108277 PMID: 25259891; PubMed Central PMCID: PMC4178126.

82. Agnarsson I, Zambrana-Torrelio CM, Flores-Saldana NP, May-Collado LJ. A time-calibrated species-level phylogeny of bats (Chiroptera, Mammalia). PLoS currents. 2011; 3:RRN1212. doi: 10.1371/currents.RRN1212 PMID: 21327164; PubMed Central PMCID: PMC3038382.

83. Cornelis G, Heidmann O, Bernard-Stoecklin S, Reynaud K, Veron G, Mulot B, et al. Ancestral capture of syncytin-Car1, a fusogenic endogenous retroviral envelope gene involved in placentation and conserved in Carnivora. Proceedings of the National Academy of Sciences of the United States of America. 2012; 109(7):E432–41. doi: 10.1073/pnas.1115346109 PMID: 22308384; PubMed Central PMCID: PMC3289388.

84. Perelman P, Johnson WE, Roos C, Seuanez HN, Horvath JE, Moreira MA, et al. A molecular phylogeny of living primates. PLoS Genet. 2011; 7(3):e1001342. doi: 10.1371/journal.pgen.1001342 PMID: 21436896; PubMed Central PMCID: PMC3060065.

85. Near TJ, Eytan RI, Dornburg A, Kuhn KL, Moore JA, Davis MP, et al. Resolution of ray-finned fish phylogeny and timing of diversification. Proceedings of the National Academy of Sciences of the United States of America. 2012; 109(34):13698–703. doi: 10.1073/pnas.1206625109 PMID: 22869754; PubMed Central PMCID: PMC3427055.

86. Lavergne A, Ruiz-Garcia M, Catzeflis F, Lacote S, Contamin H, Mercereau-Puijalon O, et al. Phylogeny and phylogeography of squirrel monkeys (genus Saimiri) based on cytochrome b genetic analysis. Am J Primatol. 2010; 72(3):242–53. doi: 10.1002/ajp.20773 PMID: 19937739.

87. Nyakatura K, Bininda-Emonds OR. Updating the evolutionary history of Carnivora (Mammalia): a new species-level supertree complete with divergence time estimates. BMC biology. 2012; 10:12. doi: 10.1186/1741-7007-10-12 PMID: 22369503; PubMed Central PMCID: PMC3307490.

88. Katoh K, Standley DM. MAFFT multiple sequence alignment software version 7: improvements in performance and usability. Molecular biology and evolution. 2013; 30(4):772–80. doi: 10.1093/molbev/mst010 PMID: 23329690; PubMed Central PMCID: PMCPMC3603318.

89. Xia X, Xie Z, Salemi M, Chen L, Wang Y. An index of substitution saturation and its application. Molecular phylogenetics and evolution. 2003; 26(1):1–7. PMID: 12470932.

90. Xia X, Lemey P. Assessing substitution saturation with DAMBE. In: Lemey P, Salemi M, Vandamme AM, editors. The Phylogenetic Handbook: A Practical Approach to DNA and Protein Phylogeny. Cambridge: Cambridge University Press; 2009. p. 615–30.

91. Xia X. DAMBE5: a comprehensive software package for data analysis in molecular biology and evolution. Molecular biology and evolution. 2013; 30(7):1720–8. doi: 10.1093/molbev/mst064 PMID: 23564938; PubMed Central PMCID: PMCPMC3684854.

92. Lanfear R, Calcott B, Ho SY, Guindon S. Partitionfinder: combined selection of partitioning schemes and substitution models for phylogenetic analyses. Molecular biology and evolution. 2012; 29(6):1695–701. doi: 10.1093/molbev/mss020 PMID: 22319168.

93. Hafner MS, Sudman PD, Villablanca FX, Spradling TA, Demastes JW, Nadler SA. Disparate rates of molecular evolution in cospeciating hosts and parasites. Science. 1994; 265(5175):1087–90. PMID: 8066445.

94. Legendre P, Desdevises Y, Bazin E. A statistical test for host-parasite coevolution. Syst Biol. 2002; 51(2):217–34. doi: 10.1080/10635150252899734 PMID: 12028729.

95. Paradis E, Claude J, Strimmer K. APE: Analyses of Phylogenetics and Evolution in R language. Bioinformatics. 2004; 20(2):289–90. PMID: 14734327.

96. Ronquist F, Huelsenbeck JP. MrBayes 3: Bayesian phylogenetic inference under mixed models. Bioinformatics. 2003; 19(12):1572–4. PMID: 12912839.

97. Huelsenbeck JP, Ronquist F. MRBAYES: Bayesian inference of phylogenetic trees. Bioinformatics. 2001; 17(8):754–5. PMID: 11524383.

98. Miller MA, Schwartz T, Pickett BE, He S, Klem EB, Scheuermann RH, et al. A RESTful API for Access to Phylogenetic Tools via the CIPRES Science Gateway. Evol Bioinform Online. 2015; 11:43–8. doi: 10.4137/EBO.S21501 PMID: 25861210; PubMed Central PMCID: PMCPMC4362911.

99. Drummond AJ, Suchard MA, Xie D, Rambaut A. Bayesian phylogenetics with BEAUti and the BEAST 1.7. Molecular biology and evolution. 2012; 29(8):1969–73. doi: 10.1093/molbev/mss075 PMID: 22367748; PubMed Central PMCID: PMCPMC3408070.

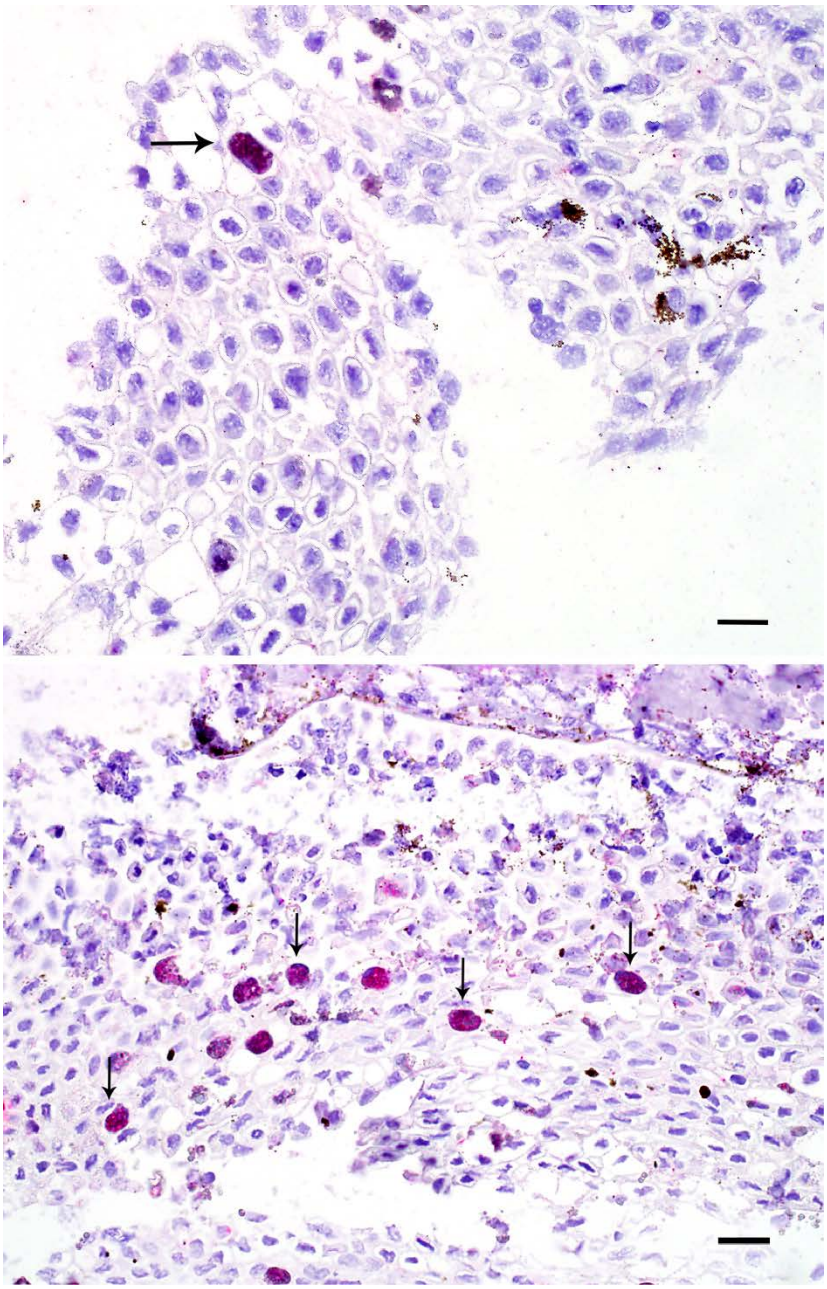
100. Teeling EC, Springer MS, Madsen O, Bates P, O'Brien S J, Murphy WJ. A molecular phylogeny for bats illuminates biogeography and the fossil record. Science. 2005; 307(5709):580–4. doi: 10.1126/science.1105113 PMID: 15681385.

101. Tate CM, Howerth EW, Mead DG, Dugan VG, Luttrell MP, Sahora AI, et al. Anaplasma odocoilei sp. nov. (family Anaplasmataceae) from white-tailed deer (Odocoileus virginianus). Ticks Tick Borne Dis. 2013; 4(1–2):110–9. doi: 10.1016/j.ttbdis.2012.09.005 PMID: 23276749; PubMed Central PMCID: PMCPMC4003554.

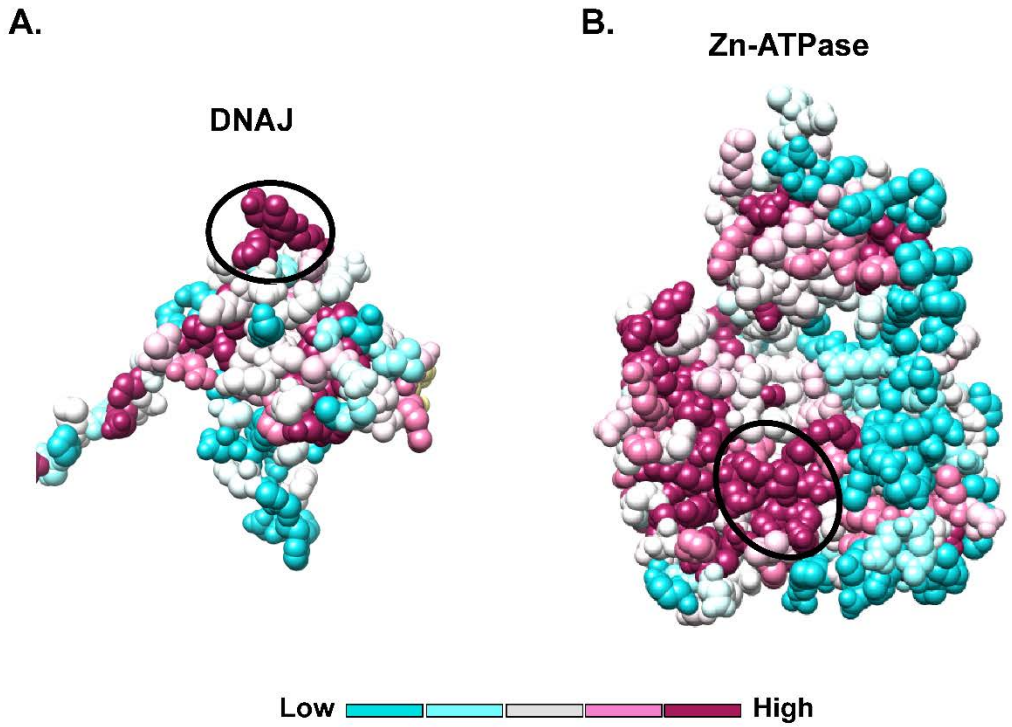
102. Dawson JE, Paddock CD, Warner CK, Greer PW, Bartlett JH, Ewing SA, et al. Tissue diagnosis of Ehrlichia chaffeensis in patients with fatal ehrlichiosis by use of immunohistochemistry, in situ hybridization, and polymerase chain reaction. The American journal of tropical medicine and hygiene. 2001; 65(5):603–9. PMID: 11716122.

103. Pettersen EF, Goddard TD, Huang CC, Couch GS, Greenblatt DM, Meng EC, et al. UCSF Chimera—a visualization system for exploratory research and analysis. Journal of computational chemistry. 2004; 25(13):1605–12. doi: 10.1002/jcc.20084 PMID: 15264254.

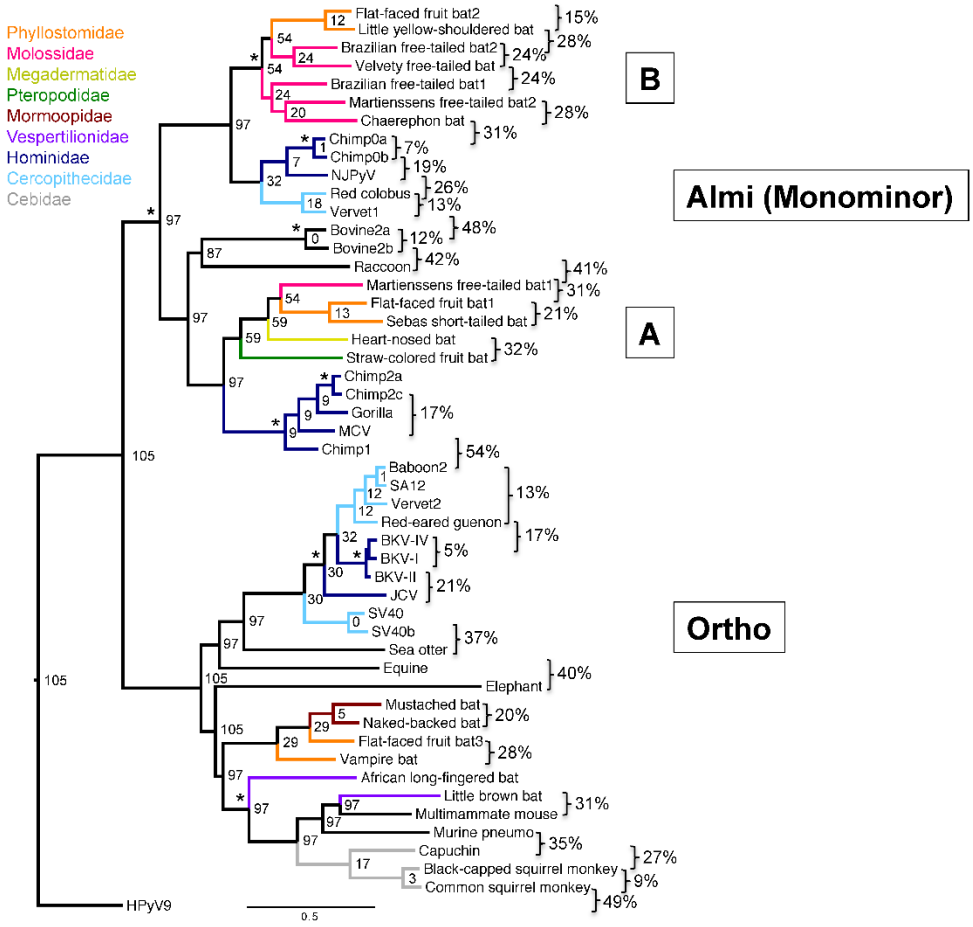
Supplemental Information



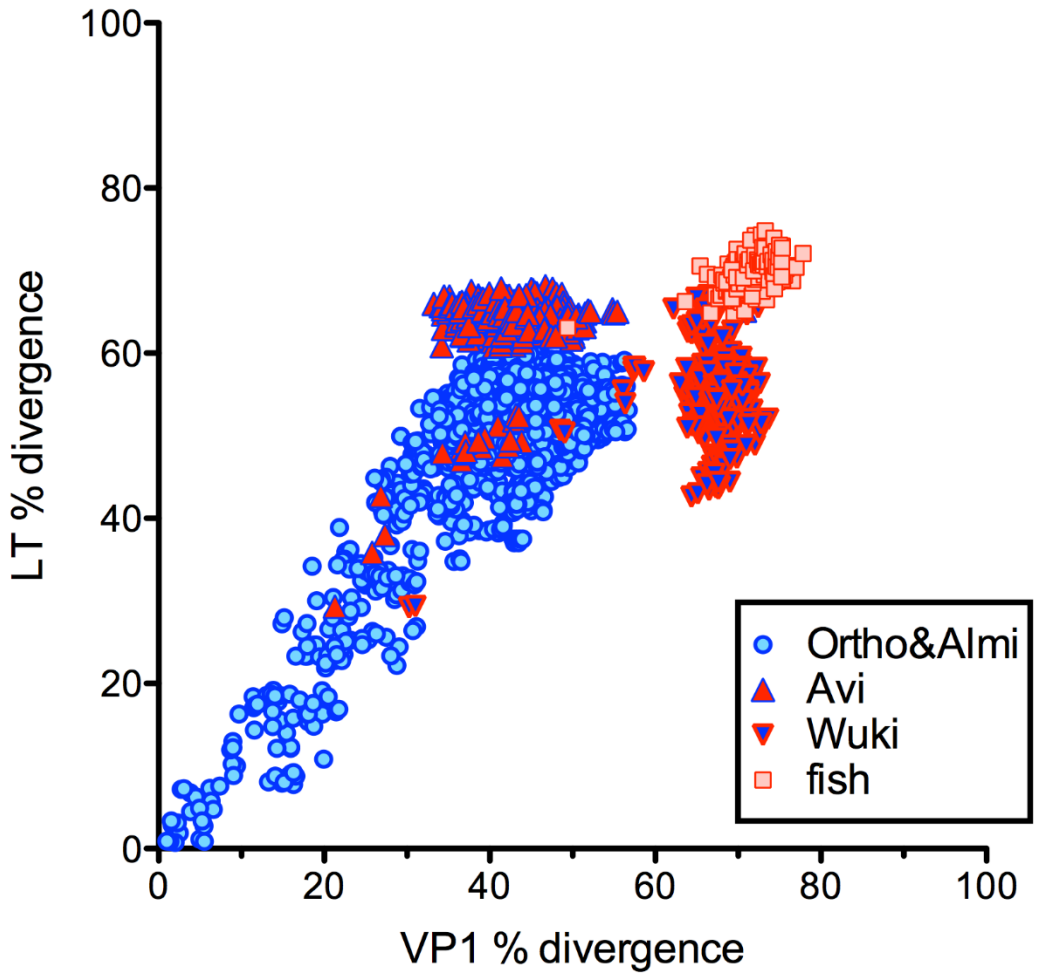
S1 Fig. In situ hybridization analysis of guitarfish polyomavirus in resolving skin lesions. A hybridization assay adapted from previously reported methods [101, 102] was used to stain sections of guitarfish skin lesions biopsied during the resolution of symptoms. Guitarfish polyomavirus VP1 probe hybridization signal (red) was observed in unidentified round cells. Arrows indicate selected positively-stained cells. The cells appear to have histiocytic or macrophage-like morphology. Free speckled brown/black patterns are attributable to melanin. Scale bar represents 20 μm.
doi:10.1371/journal.ppat.1005574.s001



S2 Fig. Conservation maps for LT DNAJ and Zn-ATPase domains.
The conservation maps were generated using the ConSurf server (http://consurf.tau.ac.il/), and then visualized using Chimera, http://www.cgl.ucsf.edu/chimera/ [103]. Panel A: the DNAJ domain conservation map was generated using DNAJ domain sequences from 34 polyomavirus LTs in the Uniref90 collection. The black oval indicates the highly conserved HPDKGG motif. Panel B: conservation map of LT Zn-ATPase domains. The map was generated with 69 LT sequences from the Uniref90 collection. The black oval indicates the Walker motifs required for binding and hydrolysis of ATP. Fewer DNAJ domains were included in this analysis due to a stringent default E-value (0.0001) setting. This indicates a greater level of variation among the DNAJ domains in contrast to the Zn-ATPase domains of LTs.
doi:10.1371/journal.ppat.1005574.s002



S4 Fig. Phylogenetic illustration of select pairwise divergences.
Phylogeny.fr "one click" settings were used to draw a phylogenetic tree for the complete genomes (nucleotide) of selected members of the Almi-LT and Ortho-LT clades. The tree is arbitrarily rooted on human polyomavirus 9. The selected Almi species have only one minor capsid protein and thus belong to a "Monominor" sub-clade within clade Almi. Numbers within the nodes indicate the estimated time (in millions of years ago) of the last common ancestor of host animals contained within the node. Branches are color-coded based on host animal families. Percentages indicate the pairwise nucleotide divergence of the complete genomes of the indicated polyomavirus species pair. Nodes that encompass possible intra-host polyomavirus divergence events are marked with asterisks.
doi:10.1371/journal.ppat.1005574.s004



S5 Fig. LT and VP1 co-divergence.
SDT was used to calculate the percent divergence of LT and VP1 proteins for individual pairs of polyomaviruses. The linear relationship between LT and VP1 divergences in Ortho, Almi, and fish clades suggests that the two proteins independently diverge at a roughly similar rate. The disconnection of the Avi and Wuki clades can most easily be explained by ancient recombination events (see Fig 5). doi:10.1371/journal.ppat.1005574.s005