INTERPRETING HIGHLY DIVERSE DSRNA VIRAL SEQUENCES FROM *PICOBIRNAVIRIDAE* WITHIN AND AMONG SPECIES FOR INTERSPECIES TRANSMISSION INFERENCE

A dissertation presented in partial fulfilment of the requirements for the degree of

Doctor of Philosophy

in

Veterinary Sciences

at the School of Veterinary Science,

Massey University, Manawatu,

New Zealand

Janelle Ruth Wierenga

2021

ABSTRACT

Cross-species transmission of infectious diseases has been at the forefront of our current reality considering the COVID-19 global pandemic and its repercussions. Now more than ever, we realise that the factors involved in these transmissions between species is both simple, such as the impact of increased contact between wildlife and humans, and complicated, such as the impact of biodiversity loss. Our understanding of these influences, and the complexity, has progressed considerably in the last decade though we are still far from predicting epidemics and pandemics. We do, though, understand that many of these infectious diseases cross species barriers frequently, with some types of pathogens more likely to cross between species.

Picobirnaviruses (genus *Picobirnavirus,* family *Picobirnaviridae*), a genus of double-stranded RNA (dsRNA) viruses, have been identified on almost every continent, though the pathogenic potential for these viruses is still debated. Picobirnaviruses are known to have massive genetic diversity. Picobirnaviruses were first identified in rabbits and then humans and since then have been reported in many mammalian species, birds, reptiles, along with wastewater, and potentially in protozoa; prokaryotes have been implicated as potential hosts, based on conserved ribosomal-binding motifs from prokaryotic messenger RNA. Although they are highly diverse, they have often been found to have very high similarity in different host species present at the same geographic site, raising the hypothesis that cross-species transmission has occurred.

The study system from Uganda described in this thesis provides a setting for the possible transmission of infectious diseases between wildlife, humans and domestic livestock. I identified multiple potential pathogens to study cross-species transmission within this system and focused on *Picobirnavirus.* I identified multiple picobirnaviruses in humans, wildlife and domestic animals from Uganda and New Zealand, for comparison, by using metagenomic and targeted PCR amplicon sequencing. Initial phylogenetic analyses revealed some host clustering among the picobirnaviruses studied, but host and geographical clustering was not upheld when known picobirnaviruses from global databases were included. Furthermore, the use of near-complete and two genogroup-specific viral sequences did not reveal host or geographic clustering on phylogenetic analyses. I identified multiple picobirnaviruses within the same host, detecting up to eight distinct viral sequence types in the samples. Despite the wide-ranging within-host diversity of the viruses, nearly identical virus sequences were also identified in different hosts from the study.

Given the high diversity, unclear host and geographic associations, and that picobirnaviruses may infect hosts from across the natural kingdoms, including bacteria and protozoa, I sought to identify picobirnaviruses from bacterial and protozoal samples, to try to determine the host range of the

viruses. I discovered picobirnavirus RNA in purified *Cryptosporidium* oocysts from New Zealand, suggesting a possible alternative Protist host. However, using an alternative genetic code from invertebrate mitochondria removed within putative open reading frame stop codons from some picobirnavirus sequences, supporting other studies. Further analyses of the picobirnavirus sequences revealed alternative genetic code usage or motifs in the untranslated regions that could indicate the virus has a prokaryotic rather than a eukaryotic host—which have been commonly assumed to be the definitive hosts. The discovery that prokaryotes may also be considered hosts of picobirnaviruses, as well as the presence of homologous sequence between species of this highly diverse RNA virus genera, will influence our understanding of these vexatious and under-studied viruses.

ACKNOWLEDGEMENTS

I am sincerely grateful for my supervisory team (Professor David Hayman, Dr. Richard Hall, Associate Professor Patrick Biggs, Dr. Kristene Gedye, Dr. Matthew Knox) for their continued support and guidance throughout the process of the PhD. Specifically, I thank: Dr. Matthew Knox for proof-reading multiple drafts of the thesis, for technical and laboratory assistance, and for your consistent guidance and advice; Dr. Kristene Gedye for extensive technical and laboratory support including, but not limited to, extraction optimisation and trouble-shooting technical complications; Associate Professor Patrick Biggs for creating the code and figures over Lockdown 1.0 (such as for Figure 16 and Figure 17), for detailed proof-reading and support on bioinformatics matters; Dr. Richard Hall for the virological guidance (especially with this dsRNA virus), proof-reading with lessons on grammar and a consistent positive energy; and, Professor David Hayman for your consistent and unwavering guidance and support through all matters PhD and career-wise through this very long journey.

I also extend special thanks to additional experts and groups that helped to make this PhD possible: Willy A. Valdivia-Granda, CEO of Orion Integrated Biosciences Inc., for the next-generation sequencing and the initial bioinformatics on the metagenomic sequence data; Conservation through Public Health including Dr. Gladys Kalema-Zikusoka, Stephen Rubanga and all of the CTPH staff that collected and sent the samples, in addition to showing us around your beautiful country of Uganda.

I would also like to thank all of the staff and students in the mEpiLab that helped me with the laboratory and statistical-related queries with special thanks to: Lynn Rogers for helping me out in the lab whenever I asked a question; Anne Midwinter for helping with all microbiological-related techniques; Simon Verschaffelt for helping with any computer issues, especially during the lockdowns; Renata de Lara Muylaert for teaching me how to use R, even when I would get frustrated.

I want to acknowledge the funds that made this PhD project possible including the Massey University School of Veterinary Science Postgraduate research fund, the Massey University McGeorge Research Fund, the Marsden Fund Fast-Start Grant and the Rutherford Discovery Fellowship.

Finally, I want to thank my friends and family for supporting me during the PhD. Specifically, thank you to my family that support me despite working and living overseas; to my dear friends in Colyton, New Zealand, I thank you for your consistent grounding and support for the creatures we have loved and lost.

ABBREVIATIONS

AGC	alternative genetic code
BLAST	Basic Local Alignment Search Tool
BLOSUM	Blocks Substitution Matrix
Bp or bp	base pairs
CTPH	Conservation Through Public Health
dH_2O	distilled water
DNA	deoxyribonucleic acid
dsRNA	Double-stranded ribonucleic acid
DRC	Democratic Republic of Congo
EID	Emerging infectious disease
GI	Genogroup I
GII	Genogroup II
MAFFT	Multiple sequence Alignment using Fast Fourier Transform
mV	millivolts
NCBI	National Center for Biotechnology Information
NGS	Next-generation sequencing
ORF	open reading frame
PBS	Phosphate-buffered saline
PC2	Physical Containment Level 2 Laboratory
PCR	Polymerase chain reaction
RBS	ribosomal-binding sequences
RdRp	RNA-dependent RNA polymerase
RNA	ribonucleic acid
SD	Shine Dalgarno
SGC	standard genetic code
TAE	Tris-acetic acid-EDTA
TBE	Tris-borate-EDTA
UTR	untranslated region

TABLE OF CONTENTS

LIST OF FIGURES

LIST OF TABLES

CHAPTER 1: BACKGROUND

1.1 INFECTIOUS DISEASE TRANSMISSION

Infectious diseases cause a substantial burden on populations, and currently account for three of the top ten causes of mortality in humans worldwide. Infectious diseases have led to some of the most significant disease events in history, including the current SARS-CoV-2 pandemic [1-6]. Infectious agents that cause clinical illness are called pathogens and can be transmitted within, and among, populations of the hosts. Pathogens include microparasites such as viruses, bacteria, and protozoa, and larger macroparasites, such as helminth worms. The mode of infection transmission among host individuals varies among pathogens and may be aerosolized, vector, water-borne, faecal-oral, direct penetration or can include a combination of modes, not necessarily pathogen-specific. Factors that can influence spread include the proportion of a population with the disease (prevalence of the pathogen), dose of pathogen, duration of exposure to pathogen and immunity of the host (Figure 1).

Some pathogens can be transmitted from one host species to another species (cross-species transmission); a process known as 'spillover'. Zoonoses are diseases that transmit from animals, the hosts, to humans, while zooanthroponosis describes transmission from humans to animals [7, 8]. The host species considered a pathogen reservoir may or may not have symptoms of clinical disease; many definitions exist for reservoir species, though usually the reservoir species is defined in relation to the "target species of interest" [9]. Reservoir species may be challenging to identify due to lack of indication of infection, rare or uncommon transmission, or unknown pathogen effects until the pathogen crosses over into another species of interest [9]. Cross-species transmission may be normal for multi-host pathogens, such as vector-borne diseases like dengue fever, or those with complex life cycles like toxoplasmosis [10-12]. However, cross-species transmission events are often called 'spillover' events, especially when they are not thought to be part of the normal infection life cycle [2, 11, 13]. The different parasite types and modes of transmission interact with host factors to determine their spillover potential (Figure 1).

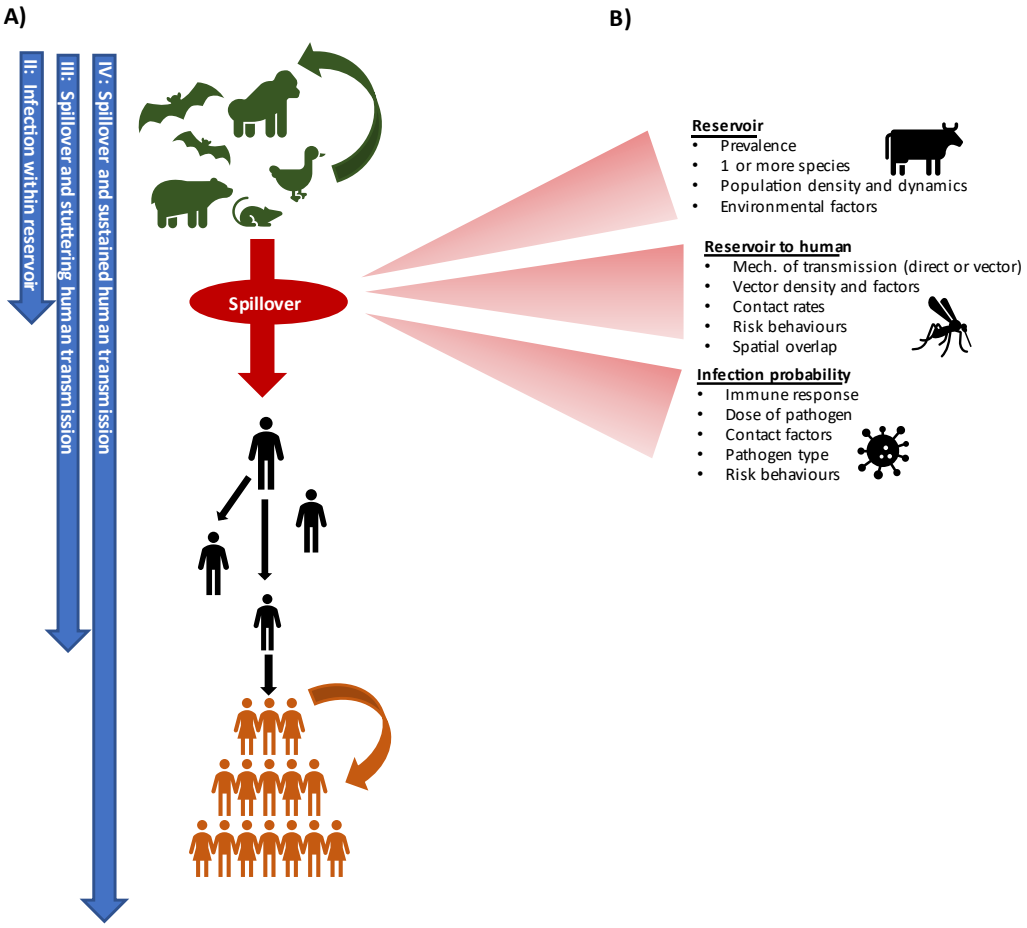


FIGURE 1 SCHEMATIC DEPICTING THE STAGES OF INFECTIOUS DISEASE TRANSMISSION WITHIN AND BETWEEN SPECIES AND THE FACTORS INVOLVED IN CROSS-SPECIES TRANSMISSION OF PATHOGENS

A) Stages of infectious disease transmission can include transmission within the species (stage 1, not shown), transmission to another species (II, III, IV) which can stutter out or persist with the potential to become a pandemic. B) The potential for spillover to occur is multifactorial, involving contributing factors within the species affected (reservoir), between species (reservoir-human contact) and the probability of infection. Contact rates also include environmental factors, in addition to those listed, such as habitat fragmentation, biodiversity alterations in the environment and encroachment of species leading to the increased potential for contact to occur between species. Probability of infection factors also include, in addition to those listed, type of pathogen (virus, bacterial, protozoal, etc), mechanism of replication (cytoplasmic, nuclear) and genetic mutation potential. Figure depicted from Lloyd-Smith, et al., Science, 2009 [2].

1.2 Cross-species transmission

The current SARS-CoV-2 pandemic, the recent (and recurring) Ebola virus epidemics, severe acute respiratory syndrome (SARS)-Coronavirus (CoV) and influenza virus pandemics and Middle East respiratory syndrome (MERS)-CoV outbreaks have revealed the need to find out more about when

and how spillover transmission events occur, rather than dealing with the transmission and consequences after they occur [5, 14, 15]. With the ability to identify reservoirs of pathogens and the creation of models to predict the spatial and temporal dynamics and subsequent impact of cross-species transmission events, we may be able to progress toward a better prediction and prevention approach, rather than a reaction approach [16-18]. We still have much to understand about the spillover of microparasites, such as viruses, bacteria and protozoa, and why it occurs when it does. Understanding the stages and factors involved in spillover events, including when and how opportunity occurs (e.g., species density, contact between species, species relatedness), the ecological and evolutionary factors associated with transmission events, and how the pathogens can continue transmission within the new host, and factors associated with sustained transmission are imperative (Figure 1) [2, 12, 19].

Stages of infectious disease transmission between species have been proposed to illustrate the different patterns and outcomes that can occur when an infectious organism infects a new species [2, 12]. These stages have been developed to understand the transmission of zoonoses, though they also apply to spillover with pathogens that result in any new cross-species transmission. The stages also help to understand emergence of the infectious disease and the subsequent expansion or spread of the pathogen within a species. Many pathogens have adapted to, and therefore stay within, the same species without obvious evidence of transmission between other species, such as measles or smallpox in humans [2, 12]. These pathogens, though species-specific, may have their ancestral origins in other species or may have the potential, in the future, to spillover to new species [20, 21]. The stages of transmission describe pathogen dynamics after the spillover events that: 1) do not continue to transmit within that new species, 2) transmit but “stutter out”, or 3) those infections that continue to spread within the new species with the potential to lead to epidemics or pandemics [2, 12]. After spillover occurs, epidemics can occur from either unsustained infection transmission with repeated transmission, or sustained transmission within the new host which can quickly result in epidemics in the newly susceptible population [2, 11]. We usually identify these spillover events after they occur, sometimes due to devastating consequences, such as the Ebola virus epidemic in West Africa or the SARS-CoV-2 pandemic resulting in millions of infections and deaths [22, 23].

Understanding the contributing factors around the actual spillover event may allow us to limit these spillover events and/or identify and intervene on risk factors more rapidly to prevent the emergent spread of disease. Lloyd-Smith et al. (2009) proposed a conceptual model to describe the factors involved in spillover and others have used similar frameworks to understand these complex and complicated systems [2, 24, 25]. These models include ecological factors such as habitat

fragmentation or de/reforestation, host factors such as immune function or genetic relatedness, pathogen traits such as type of pathogen or location of pathogen replication within the cells, and behavioural factors such as hygiene and sanitation (Figure 1) [1, 2]. Some of these factors such as habitat fragmentation, pathogen type (RNA viruses), species taxonomic relatedness and multispecies pathogens, to name just a few, have already been implicated in the increased risk to cross-species transmission [13, 19, 26-30]. Other factors that may result in cross-species transmission of pathogens include increasing density of a species resulting in more contact rates and/or different type of contact between species; habitat changes including either habitat fragmentation or re-establishment of habitat; climate changes; or inter- and intra-species genetic relatedness [11, 31]. Other factors such as mode of infection transmission (readily transmissible such as aerosolised or fomite transmission), asymptomatic as compared to symptomatic infections, increased transmissibility potential or increased basic reproductive number (R_0), increased contact rates, super-spreading potential, and lack of interventions to decrease R_o, determine whether the infectious disease will cause an epidemic or pandemic following spillover [2, 5, 32, 33].

An important risk for spillover is the increased contact with wildlife due to increasing human encroachment on previously wild areas [31, 34]. Wildlife are known reservoirs of pathogens for many emerging infectious diseases (EIDs) that have the potential to result in cross-species transmission [31, 34, 35]; on the other hand, humans and domestic animals also have the potential to transmit pathogens to wildlife, sometimes with devastating consequences, especially in endangered species or species of conservation concern [8]. Habituation of wildlife for ecotourism has resulted in further opportunity for spillover events and for known zoonoses and anthroponoses [36].

1.3 AFRICA AND EMERGING ZOONOTIC INFECTIOUS DISEASES

Equatorial regions in Africa are predicted to have a higher risk of zoonotic EIDs specifically from wildlife; while at the same time have wildlife of conservation concern [34, 37]. Vector-borne and viral diseases are predicted to emerge in central Africa [18, 34]. Viral richness within central Africa in particular, raises concerns that there are likely many “missing” zoonotic viral diseases, resulting in important EIDs [18].

Uganda, located within the central African region, has had significant EID events such as outbreaks of Ebola virus disease and the emergence and discovery of Zika virus from the Zika forest [38-40]. Uganda is situated near the Equator, sharing borders with Rwanda, Democratic Republic of Congo (DRC), South Sudan, Kenya and Tanzania (Figure 2). Uganda is known for its high biodiversity but also for its high population density, which has led to threats to ecological health and biodiversity in the country [37]. Seven national parks protect areas of ecological importance with rare and threatened species and the

biodiversity within the country that impacts the ecological health and health of the human population [37]. Uniquely biodiverse and containing species of conservation concern, both Bwindi Impenetrable National Park and Queen Elizabeth National Park lie on the border with DRC. Bwindi Impenetrable Forest is home to the endangered mountain gorilla [37, 41, 42].

1.4 Bwindi Impenetrable Forest, Uganda

Bwindi Impenetrable Forest is located in southwestern Uganda as part of the Albertine Rift montane forest and contains numerous bird, plant and mammalian species, some not found elsewhere in the world (Figure 2) [41, 43].

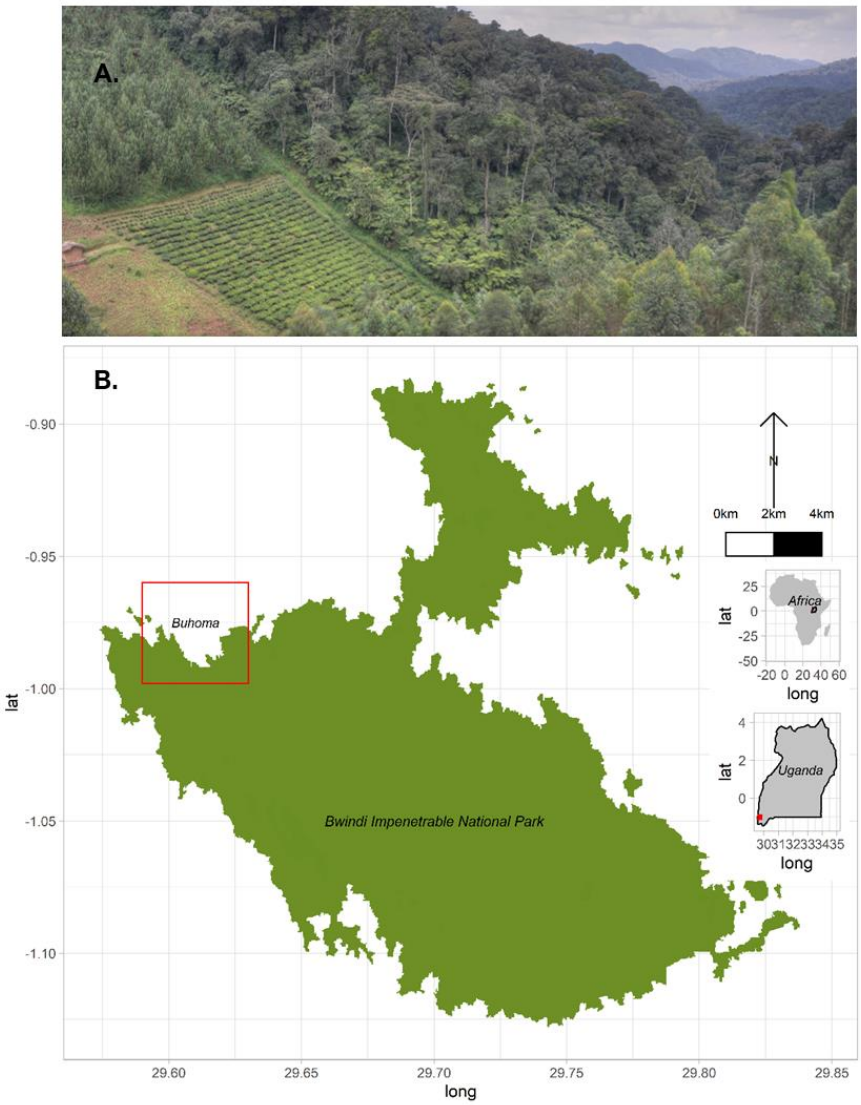


FIGURE 2 BWINDI IMPENETRABLE PARK FOREST

A) Photo of the Bwindi Impenetrable Park (background) boundary with agricultural land (foreground). B) Schematic map of Bwindi Impenetrable National Park with latitude and longitude and Buhoma on the northwestern edge of the park, with a red square around the village. The location of the park in Uganda is shown

by the red dot on the inlayed picture of Uganda on the right middle along with the location of Uganda shown in Africa by the black dot on the inlayed picture above Uganda. Photo: D. Hayman. Map: R.L. Muylaert. [44]

Bwindi Impenetrable Forest including the Bwindi Impenetrable National Park contains many wildlife species, some rare species of significant conservation concern due to near extinction of these species [41, 43]. In Bwindi Impenetrable National Park, like many national parks, wildlife face becoming endangered or extinct due to habitat loss, poaching and unnecessary culling, injury caused by human impact or by infectious disease [45]. Conservation efforts have worked to improve the situation by protecting habitat and by protecting these species, such as mountain gorillas, from the effects of human-induced injury or poaching through ecotourism, which can also benefit the local economy [36, 37, 46, 47].

1.4.1 Gorillas of Bwindi Impenetrable Forest

Mountain gorillas are critically endangered with the total population within Uganda, Rwanda and DRC of approximately 880 individuals [42]. Most of the population are within the Virunga Volcanic region located in DRC and bordering Rwanda and the remainder in Bwindi Impenetrable Forest in Uganda [42]. Mountain gorillas live within groups up to 15, usually consisting of one (silverback) or more males (silverback +/- non-dominant silverbacks or blackbacks), multiple adult females, juveniles and infants [41, 48, 49]. Habituation of the mountain gorillas in Bwindi commenced in the early 1990s, usually taking around 2-3 years to habituate a group for ecotourism [49]. The habituation process involved daily visits from researchers and trackers with monitoring of behavioural responses from the gorillas followed by exercises simulating tourists to confirm no aggression responses; in the early stages, this could result in charging, attacks and sometimes direct contact of the gorillas with the humans [50]. Within Bwindi, 3-4 groups of gorillas are habituated for tourism (Figure 3) and 1-2 for research, while there still remain a few groups of gorillas that are non-habituated [41, 49].

Besides poaching, injury from snares and habitat loss, the gorillas face injury from villagers during crop raiding of plantations, as these regions are found right on the border of the park and are easy sites to find food for the gorillas (Figure 2, Figure 4) [36, 46, 51]. Along with ecotourism, this closeness to human populations has led to increased contact between gorillas and humans and reported morbidity and mortality from infectious diseases transmitted from humans or domesticated animals such as livestock or dogs [49]. Gorillas also face morbidity and mortality from infectious diseases that are transmitted from humans, one of the more significant concerns with the introduction of ecotourism [8]. Many conservation groups have been working to assist the mountain gorillas including the Mountain Gorilla Veterinary Group and Conservation Through Public Health (CTPH), led by Dr. Gladys Kalema-Zikusoka and Stephen Rubanga. CTPH has worked to establish a database of gastrointestinal infections, particularly macroparasites, that are found in routinely collected gorilla faecal samples, and

assist in individual diagnostic and treatment of sick gorillas when intervention is necessary. They also work with the community to protect the gorillas when they go outside the park boundary, protect them from exposure to infectious diseases from humans (staff, tourists, villagers) and domestic animals, and work within the community to promote public health and wildlife conservation together [51].

FIGURE 3 MOUNTAIN GORILLAS FROM BWINDI FOREST

Adult female and infant gorilla of the habituated Rushegura gorilla family in Bwindi Impenetrable National Park. Photo: J. Wierenga.

1.4.2 People surrounding Bwindi

Populations of people live on the borders of Bwindi Impenetrable Forest and the national park and it is one of the most densely populated rural areas within Uganda, with approximately 350 people per square kilometre [52]. Based on socioeconomic indicators, the life expectancy from birth is 63 years of age, with most households having five births per woman with an under 5 years-of-age mortality rate of 109 per 1000 live births [53, 54]. Most of the people live on the northern aspect of the border of the park near Buhoma (Figure 2) or on the southern aspect in Nkuringo and Rushaga. Batwa people reside near the park borders in their own communities and villages and were historically hunter-gatherers until the establishment of the park, while Bantu people reside in their villages and rely mainly on subsistence farming [49]. The Batwa population resided within the park until it was sanctioned as Bwindi Impenetrable National Park in 1991, with eviction of their villages from the park

and restriction on the use of any resources from the park without compensation, leading to marginalisation and, at times, conflict [55, 56]. Many of the people rely on tourism in and around the park for income, but they also rely on their own agricultural sustenance. Agriculture around the park includes larger fields for tea and bananas, though people within the villages also plant additional crops and vegetables (Figure 4). Many villages have become involved in programs utilising livestock (see below). The main village of Bwindi contains the Bwindi community hospital which is a larger centre within that region with services for paediatric, adult and maternity wards, along with limited programs for mental health and family planning. Medical district offices are located further away from the park, for access to healthcare for people in outlying villages though many of these do not have inpatient facilities. Diseases of significance for the human population include acquired immunodeficiency syndrome (AIDS) from human immunodeficiency virus (HIV) infection, malarial disease, skin conditions such as scabies, and diarrheal diseases [57-59]. The human population surrounding Bwindi Impenetrable Forest utilise the park for transit to markets, harvesting or work within the park [59, 60]. Programs were set up when it became a national park to allow populations to still utilise resources within the park, while at the same time preventing over-utilisation and protection of the plants and wildlife, including the endangered mountain gorilla [60].

1.4.3 Livestock surrounding Bwindi Impenetrable Forest

Livestock are common around Bwindi forest and national park. Community organizations such as CTPH have started programs within villages to improve access to animal and animal-derived products and decrease hunting within the park. Livestock include goats, cattle and sheep, providing milk and meat. The livestock are grazed around the park boundary and grazing within the park is prohibited, though still likely occurs (Figure 4).

A) **B)**

FIGURE 4 BWINDI FOREST AGRICULTURE AND GRAZING

A) Bwindi Impenetrable Forest boundary shown on the left of the photo with clearing for agricultural planting on the right (typically banana or tea). B) Domestic cattle grazing in the village near the Bwindi Impenetrable National Park which can be seen in the background of the photo. Photos: J. Wierenga.

1.5 Cross-species transmission studies in Uganda and the surrounding regions

Considerable research has identified cross-species transmission in Uganda due to the spatial overlap and close contact of wildlife, domestic animals and humans in this region [61-64]. The national parks within Uganda are ideal systems for the study of spillover events due to humans and their domestic animals residing nearby the park with exposure to wildlife in and around the park. Close contacts can occur from the ecotourism, where humans can track habituated gorilla families or other wildlife, from domestic animals illegally roaming within the park boundaries or from wildlife raiding crops on the land surrounding the park boundaries, which is commonly converted to agricultural land [65, 66].

Research on cross-species pathogen transmission in multiple national parks (Bwindi, Mgahinga and Queen Elizabeth) in southwestern Uganda evaluated the prevalence of pathogens within the domestic dog population that may come into contact with wildlife and with frequent contact with humans and livestock in the region [61]. Serological studies in Uganda show exposure to many pathogens, some with high seroprevalence including canine distemper virus (95.9-100%), canine parvovirus (54.9-74.5%), *Leptospira interrogans* (19.0-36.1%), *Leishmania* sp. (12.3-29.2%), *Toxoplasma gondii* (83.6-95.1%) and *Neospora caninum* (19.6-36.6%). Some of these pathogens are of significant public health concern for zoonotic spread (e.g., *Leptospira* sp.) and many are of concern for potential spread to wildlife in the region with high morbidity and mortality potential (canine distemper virus, canine parvovirus, *T. gondii*, *Neospora* sp. and *Leishmania* sp.) [61].

1.5.1 Protozoal pathogens

In Kibale National Park in western Uganda, a study measured the prevalence of *Cryptosporidium* sp., a protozoal pathogen typically resulting in gastrointestinal symptoms, in humans, livestock (cattle, goats and sheep) and monkeys within the same region and possible risk factors associated with this parasite [62]. The prevalence of *Cryptosporidium* spp. in humans was high at 32%, low in livestock at 2% and was 11% in monkeys, with no significant difference noted between those monkeys in protected versus unprotected forest regions. There was no association between clinical signs (diarrhoea/soft stools or unwell appearance) and the presence of *Cryptosporidium* spp. or co-infection with *Giardia duodenalis*. Molecular analysis showed that *Cryptosporidium* spp. were similar between the primates and humans with some of the *Cryptosporidium* sequences identical between species, supporting the hypothesis that cross-species transmission of *Cryptosporidium* occurred in the region. In addition, the *Cryptosporidium* spp. found in the protected or undisturbed regions of the park were distinct from those found in the disturbed regions or outside of the park, which could indicate

contamination of the environment and/or water sources, increasing the potential for cross-species transmission [62]. In another study within the same region of Kibale, monkeys, humans and livestock were evaluated for *Giardia* [63]. The prevalence of *Giardia* was highest in humans (32-50%) and moderate in monkeys (6-20%) and livestock (7-21%). Analogous to other studies with gastrointestinal *Cryptosporidium* or *Giardia*, the presence of the organism was not associated with clinical signs (soft stool/diarrhoea), but was more likely in younger humans and cattle and when there was another infected *Giardia* human or livestock in the household. Molecular analysis showed that there was more likely human-to-monkey and livestock-to-monkey transmission of *Giardia* based on genotypic similarities in the study samples [63].

1.5.1.1 Protozoal and helminth pathogens in Gorillas

Wildlife in Africa have been extensively studied due to the significant conservation concern for many species including gorillas, especially in regard to cross-species transmission of infectious diseases [37]. Many studies have evaluated both inter- and intra-species infectious diseases involving gorillas and found evidence of cross-species transmission [36, 46]. Protozoa can be passed between mammalian species easily, especially in regions where diarrhoeal disease is prominent. In Gabon, a study evaluated the effects of human contact on the prevalence of *Cryptosporidium* and *Giardia* in western lowland gorillas comparing those with more and less contact with humans [67]. The study found that those gorillas with more human contact had a higher prevalence of *Cryptosporidium* and *Giardia* (19% and 22%, respectively) than those with little human contact (0% and 9%, respectively) [67]. A study within Rwanda looking at these same pathogens in endangered mountain gorillas, domestic cattle and forest buffalo showed a low prevalence of *Giardia* of 2-9% in all species, but a significantly increasing prevalence from 3% to 9% over a 13-year period in gorillas [68]. Molecular analysis of the *Giardia* in this study showed that the genotypes from gorillas have been historically found in humans and that there were similarities between the pathogen strains based on geographical location or spatial clustering [68].

In the Central African Republic, analyses of faecal samples during various stages of habituation of western lowland gorillas found a prevalence of 2% for *Giardia*, 0.5% for *Cryptosporidium* and 7.5% for *Encephalitozoan* spp. with no difference between the varying stages of habituation [69]. Molecular analysis of *Encephalitozoan* spp. and *Giardia* in the gorillas showed genotypes similar to human genotypes, and in particular, the genotype found in gorillas was the most common genotype of *Giardia* found in humans [69]. In another study, mountain gorillas within one of the national parks of Rwanda were found to be infected with the liver roundworm (*Capillaria hepatica*) [70]. This was of public health significance as a potential zoonotic infection risk to children within the vicinity, but thought to be likely from an alternative reservoir host in the region [70]. All of these published research studies

support increased cross-species transmission based on either increased contact between the wildlife and humans and/or contamination of the environment and/or water sources, resulting in the potential of transmission of the pathogens to wildlife and/or humans with significant impacts on public health.

1.5.2 Bacterial pathogens

In Rwanda, a high percentage (60 –85%) of faecal samples from mountain gorillas were also positive for *Campylobacter* spp. [71]. Further analysis found that 3% of the samples were consistent with *C. jejuni* and also associated with soft stools [71]. Another study from the Kibale and Bwindi National Parks, found that there were similarities between *Escherichia coli* found in livestock and humans in the region [72]. Molecular analysis showed that "bacteria from humans and livestock in the same community were virtually indistinguishable genetically" [72]. Humans who did not practice hand hygiene prior to eating had twice the odds of having *E. coli* that was genetically similar to their livestock as compared to those that practiced hand hygiene [72].

1.5.3 Viral pathogens

There is considerable concern that virus infections could result in important anthroponoses or zoonoses from wildlife or humans, specifically mountain gorillas and humans, due to their close genetic relatedness. The risk of SARS-CoV-2 infection in gorillas is known to be high. Anthroponotic SARS-CoV-2 infections were identified in gorillas in San Diego, California, USA along with high susceptibility potential based on evaluation of similar receptors for binding on the surface of the virus [73-75]. Additionally, outbreaks of respiratory diseases involving large numbers of mountain gorillas were observed in Rwanda, which resulted in most of a group showing clinical illness [76]. During this outbreak, an adult female and infant died and were found to be infected with human metapneumovirus with concurrent secondary bronchopneumonia [76]. A study in the greater Congo basin of Africa in central and east Africa, tested archived blood samples from multiple wildlife species for antibodies to alphaviruses and flaviviruses [77]. Alphaviruses were found in many wildlife species, including O'nyong-nyong and Chikungunya viruses, and mountain gorillas were found to have antibodies to some alphaviruses (non-specific) and flaviviruses (West Nile virus and non-specific) [77]. Though many of these studies demonstrate the opportunity for viral cross-species transmission, no clear patterns of cross-species transmission have been observed, indicating the complexity of the system, and requiring a more comprehensive approach to the analysis of the pathogens likely to exhibit cross-species transmission.

1.6 Cross-species transmission studies in Bwindi Impenetrable Forest

1.6.1 Ectoparasites

In Bwindi Impenetrable National Park, an outbreak of scabies resulted in the death of an infant gorilla within a human-habituated group [58]. Scabies was thought to be passed on from humans in the surrounding village as it is a common dermatological condition in the human population [58]. An outbreak of scabies was reported in another human-habituated gorilla group, necessitating invasive measures for diagnosis and treatment [57]. This outbreak was also thought to be from the human population due to the advanced stages of habituation that the gorilla members were exposed to at the time of the outbreak [57].

1.6.2 Protozoal and helminth pathogens

In a study comparing habituated to non-habituated gorilla groups, most (73%) of those positive for *Cryptosporidium* were from the habituated group or those living in closest proximity to humans [78]. The oocysts of *Cryptosporidium* were found to be genotypically suggestive of a bovine genotype of *C. parvum* [79]. Human faecal samples were also found to have the same genotype of *C. parvum*, highest in park staff (21%) as compared to the community (3%) [80]. It was suspected that *Cryptosporidium* originated from cattle grazing on park land, but other risk factors may include drinking from streams or not properly burying faecal material within the park. Cattle, gorillas and people also shared the same *Giardia* genotype, the most common genotype found in humans rather than cattle or gorillas [81]. Additionally, habituated gorilla groups also have been found to have a higher burden of gastrointestinal nematode parasites (*Strongyloides*, *Ascaris* and *Oesophagostomum* spp.) and helminths [82-84].

1.6.3 Bacterial pathogens

Nizeyi and colleagues identified that, in addition to a higher burden of parasites in habituated gorillas, gorillas within Bwindi had a high prevalence of enteric bacteria such as *Campylobacter, Salmonella* and *Shigella* spp. Additionally, in another study *Escherichia coli* detected in humans and cattle were found to be genetically similar as were the isolates in gorillas and humans, especially those gorillas with more human contact [64]. Antibiotic susceptibility testing of the *E. coli* isolates identified 35% with resistance to one antibiotic in humans, 27% in cattle and 17% in gorilla isolates. Cattle would not routinely be administered antibiotics and even less in gorillas, whereas humans can commonly obtain over-the-counter antibiotics, suggestive of “sharing” of the gastrointestinal bacteria or genes and passage of antibiotic susceptibility [64].

1.6.4 Bwindi Impenetrable Forest as a study system

Bwindi Impenetrable National Park provides a model site for identification and evaluation of cross-species transmission events due to the interactions between the wildlife, domesticated animals and humans, the perturbations to the environment and ecological impacts, along with the ecotourism that

increase the contact between the various species residing within the region [37, 45, 46]. These unique factors, in addition to the predicted increased risk for EIDs in equatorial Africa, create an ideal study site for the opportunity for cross-species transmission [18, 34]. The use of traditional techniques to identify and characterize cross-species transmission is limited in sensitivity and specificity in the complex landscape of transmission events; therefore, the use of more advanced molecular techniques such as metagenomics and polymerase chain reaction (PCR) can facilitate a better understanding of cross-species transmission events.

1.7 Use of metagenomics and application in Africa

1.7.1 Metagenomics

Metagenomics is the assessment of all of the organisms within a system, usually from environmental samples through the evaluation of genetic material. The use of metagenomics has expanded and has been used to identify the diversity of microorganisms within any given system, from evaluation of the human microbiome (microorganisms within a species) to the virome (viruses within a species) of wildlife or contamination of food products, through the use of next generation sequencing (NGS, also known as high throughput sequencing) [85, 86]. NGS has been shown to be beneficial in situations where multiple organisms are present and difficult to identify through more conventional methods (e.g., culture, viral isolation) through the sequencing of "millions of DNA fragments in parallel" [87]. NGS is the tool used in metagenomics to analyse nucleic acids, from potentially every organism within the system of interest, with limited or no *a priori* knowledge of what specific organisms are in the sample. Metagenomics can be applied with great effect when examining environmental samples such as soil, water or faeces where multiple organisms and multiple types of organisms can be present [87].

Metagenomics identifies the genomic material present in these samples by sequencing segments of DNA in an 'unbiased' way— it being unbiased in that organism or gene-specific primers are not specifically used for amplification. Bioinformatic algorithms compare these sequence fragments to databases to identify the organisms within a sample [87]. Primary databases such as GenBank® synchronise the data with quality control and assurance which includes computer-based "syntactic rule" checks on the proposed sequences from researchers inputting the data/sequences and "trained curators and annotators" for review [88]. Protozoal samples can be challenging as they may need vigorous physical disruption during DNA extraction (e.g., of oocyst walls) to expose the genetic material. Viral genomic evaluation can be challenging due to the different genomic formats (RNA, DNA, single versus double-stranded), the potential degradation of viral genetic material prior to evaluation, the confounding effect of abundant host genome, and the lack of redundant elements among viral genetic sequences [85, 86, 89-94] .

Methods employed for metagenomics involve nucleic acid extraction from the sample, conversion into complementary DNA (cDNA) for RNA samples, and then amplification. The amplified DNA is then sequenced without *a priori* knowledge of the organisms present; sequencing by synthesis metagenomics (MiSeq or HiSeq next-generation sequencing through Illumina®) is the term used to sequence segments ('reads') of DNA fragments in the sample with the prospect of identifying genetic material of all organisms [87]. The smaller DNA sequences obtained are then compared to known sequences from organisms in a variety of databases using search algorithms [95]. In general, the longer the read, the more specific it will be for an organism. A range of procedures are performed on the raw sequence reads for quality control, including trimming and longer contig assembly. Computational procedures use different types of software because multiple steps are required to analyse data, and because of the large amount of data obtained [95]. Assembly based methods are preferred in low diversity and/or high sequence depth samples. Genome fragments or near-complete genomes are assembled together into bins of overlapping reads, called contigs, and then assigned to the highest taxonomic level. Alternatively, read-based identification is used in high diversity and/or low sequence depth samples or when the taxa of interest are rare within the samples. In this method, each individual read is assigned to a taxonomic classification where possible and grouped accordingly for later analyses [96]. The sequences can then be compared on a global database of genomic sequences for known organisms, such as U.S. National Center for Biotechnology Information [97].

1.7.2 Application of NGS in Africa

To date, identification of pathogens with the potential to cross species has primarily been pursued by methods such as microscopy or PCR and few studies have been published that apply metagenomics to African apes, or within Uganda [84, 91, 98]. A study from Ghana used metagenomics to identify possible zoonotic pathogens from the African straw-coloured fruit bat (*Eidolon helvum*), identifying poxviruses, adenoviruses and polyomaviruses that were closely related to both human and other primate viruses [99]. Similarly, a study in Cameroon evaluated the faecal virome from two different fruit bats to identify the potential for cross-species transmission [93]. Whole genome sequencing was utilized to identify a known human rotavirus strain in pigs in Africa, not previously identified within the continent [100].

In Uganda, NGS was utilized in an outbreak of suspected viral haemorrhagic fever, but conventional diagnostic methods did not reveal any typical viral haemorrhagic fever disease agents [101]. Instead, NGS was able to ascertain the aetiological agent in the outbreak and found the patients were suffering from yellow fever, showing the capability of using NGS in hard-to-diagnose outbreak situations. NGS was also used in a case study within Uganda and Sudan to identify a novel paramyxovirus after all prior tests failed to identify the disease [102]. NGS has also been used in Uganda for identification of

possible reservoir species. With the use of NGS, Ndumu virus, an alphavirus, previously only identified in the vector of the mosquito, was found in pigs in Uganda, indicating the potential for a reservoir species and cross-species transmission [103]. Another study from Uganda found both a parvovirus and torque teno virus for the first time in the African continent in bushpigs (*Potamochoerus larvatus*) by NGS [104]. Additionally, Hepatitis A virus was identified by NGS in wild baboons (*Papio anubis*) in Kibale National Park National Park in Uganda, highlighting the concern for cross-species transmission [105]. Recently, viral metagenomics was used to evaluate faecal samples from gorillas in the DRC, finding an abundance of novel RNA viruses [98].

1.8 Characteristics or factors for spillover organisms

Characteristics that are known drivers for cross-species transmission have included pathogen factors, anthropogenic factors, and ecological factors [11, 13, 35, 106-108]. Multiple theories and models have been proposed in an attempt to explain which factors relating to the organisms, determine cross-species transmission events—though no one model appears to work in all cases of pathogen transmission. A more holistic approach to investigation of cross-species transmission has identified that specific, or isolated factors (e.g., host factors, reservoir factors, pathogen factors or immune system factors), all contribute to the potential for cross-species transmission, though these factors are dynamic, variable and, sometimes, not predictable [18, 25, 108, 109].

1.8.1 Cross-species transmission 'model' organisms

Pathogen factors may include the type of organism such as viruses, bacteria, protozoa, fungi, helminth, or ectoparasites. Additionally, multiple host pathogens are more likely to spillover to new hosts compared to single host pathogens.

The mutation rates of certain organisms, including deletions or insertions, can be a factor as some pathogens mutate slowly or more quickly due to proof-reading in the genetic code and replication processes. Also, the mode of transmission can result in increased likelihood of cross-species transmission: for example, pathogens that spread via aerosolized or fomite transmission would be more likely to be transmitted across different species compared to pathogens that spread through bodily fluids or direct penetration (e.g., via bites). In addition, high stability of the pathogen in the environment or within the host would increase transmission potential. Anthropogenic and ecological factors may include the potential for host-jumping events which is based on contact rates, including particular hosts that routinely do not have contact with certain species. For example, humans may not come into contact with wildlife unless certain environmental alterations occur or other resource considerations (such as planting nutritional resources or seeking medicinal resources). In addition, multiple behavioural and host factors including host abundance can play a role in increasing contact

rates or transmission potential. Moreover, symptomatic as compared to asymptomatic spread of the pathogen can contribute significantly to transmission potential as well as duration of the infection [11, 13, 35, 106-108].

Looking at pathogen considerations alone, the potential for different classifications of pathogens such as viruses, bacteria, protozoa, fungi and helminths, has been considered with the likelihood of cross-species transmission and, though all have the potential for cross-species transmission, viruses appear most suited for cross-species transmission [13, 27, 109]. In particular, RNA viruses have been proposed to have an increased likelihood of cross-species transmission [13, 29]. Proposed mechanisms for the higher likelihood of RNA viruses to cross-species barriers include high mutability of the virus, as RNA viruses have less stringent proof-reading during replication than DNA viruses (higher mutation rates and evolve more quickly); frequent cytoplasmic replication as compared to replication within the nucleus compared to DNA viruses; potential to infect multiple hosts or species; and multiple as compared to single genomic segments along with the opportunity for reassortment or recombination of genetic material [13, 18, 27-29, 110, 111]. Evaluation of host and viral evolution through co-divergence has also supported the probability that RNA viruses actually perform host-switching (or host-jumping) much more frequently than initially thought with the likelihood that these events could but do not commonly result in sustained transmission [112]. Also, RNA viruses may be less host-specific and therefore more likely to cross species barriers to cause infection and segmented viruses may also be more likely to cross species barriers [108, 112]. In a recent paper, thirty-one factors, including 9 host, 16 virus and 6 environmental factors, were assessed for known viral pathogens to produce a risk analysis framework and quantify the potential for spillover to occur [108]. This framework could be beneficial to predict what the next viral pathogen to spillover will be like, and theoretically lead to sustained transmission with epidemic potential. In contrast, though, these are known viral pathogens, and unknown or novel viral pathogens would be less likely to be predicted prior to spillover [108].

Due to these considerations, I chose to study a ubiquitous virus that is a suspected pathogen, known to infect multiple host species and also of considerable risk for cross-species transmission—an RNA virus genus *Picobirnavirus* (family *Picobirnaviridae*).

1.9 Overall aim and structure of the thesis

Due to the potential for pathogenic viruses to transmit between species in and around the Bwindi Impenetrable Forest system, I sought to further evaluate the RNA virus, picobirnavirus, as a 'model organism' of spillover within this ecosystem. RNA viruses such as picobirnaviruses are challenging to identify with culture or virus isolation methods, but with next-generation sequencing (NGS) and PCR

techniques I aimed to identify and confirm the presence of picobirnavirus in multiple species. In order to positively identify and confirm the presence of picobirnavirus, I specifically aimed to test and develop molecular tools. Further taxonomic evaluation of picobirnaviruses is important to clarify and understand the genetic diversity of this dsRNA virus. Once molecular tools were refined for picobirnavirus, including the further understanding of the taxonomy for *Picobirnavirus* [113], my original aims were to:

- examine the suitability of picobirnavirus as a model organism for understanding cross-species transmission among humans, their livestock and gorillas in south-west Uganda
- develop and test tools to validate NGS data on the identification of picobirnavirus
- determine the genetic and evolutionary relationships among these picobirnaviruses.

These objectives are addressed in the thesis.

Synopsis of the thesis is as follows:

The next chapter, Chapter 2, is a literature review of *Picobirnavirus*, comprising information on the structure of the virus and genetic composition, the host spectrum and systems the virus has been identified within to date, the classifications of the virus and the diversity with a focus on the components for evaluation of the potential for cross-species transmission. I also discuss some of the debate surrounding *Picobirnavirus* as a classified vertebrate virus and introduce additional theories of the virus as a protozoal and/or bacterial virus. The hypotheses will be stated at the end of this chapter for the dsRNA virus within the overall research aim and specified objectives.

Chapter 3 is the description of the Materials and Methods. The relevant ethical components, sample collection and storage of the samples from Uganda are described. Further samples were collected from New Zealand for additional geographical comparisons and the possible identification of picobirnavirus in bacterial and/or protozoal samples. Both DNA and RNA extractions were performed on the original samples, and these are described along with next-generation sequencing preparation and then validation with polymerase chain reaction screening and experimental controls. The molecular detection of picobirnavirus and sequencing of the virus from samples, metagenomic sequencing results from the samples and finally cloning of a select set of samples that may contain multiple viruses is described. Finally, the bioinformatic methods for the sequence results, comparisons and analyses are described in detail.

Chapter 4 contains the results and discussion of the picobirnavirus identification from the various samples from the study. It also describes the further characterization of the picobirnaviruses with the evaluation of nucleotide versus amino acid sequences and trimmed versus untrimmed sequences.

Phylogenetic analyses of the picobirnaviruses are evaluated by genogroup and by aligning and comparing conserved regions of the RNA-dependent RNA polymerase gene or near-complete picobirnavirus sequences.

Chapter 5 contains the results and discussion of the phylogenetic evaluation of the picobirnaviruses by looking at whether they cluster by host and/or by geography. This was done by genogroup with the conserved region of the RNA-dependent RNA polymerase gene and between genogroups with the near-complete picobirnavirus sequences.

Chapter 6 contains the results and discussion of within host diversity of picobirnaviruses, evaluating the samples where multiple picobirnaviruses were identified and compared. The diversity of picobirnaviruses is discussed and compared, especially within the same host when multiple viruses were identified.

Chapter 7 contains the results and discussion of the debate of whether picobirnaviruses are truly vertebrate viruses as previously thought, as compared to potential viruses of protozoa or bacteria. This chapter describes the evaluation of various bacterial and protozoal samples for the identification of picobirnavirus. The chapter also describes the evaluation of the picobirnaviruses by different methods including the use of alternative genetic codes, for the presence of open-reading frames (ORFs) and for the presence of bacterial motifs in the sequences.

Chapter 8 is the general discussion highlighting some of the unexpected results, the potential conclusions, limitations and further work to develop the findings of this project.

CHAPTER 2: PICOBIRNAVIRUS LITERATURE REVIEW

2.1 CHARACTERISTICS AND CLASSIFICATIONS

Picobirnaviruses are bisegmented (segmented virus with two segments) double-stranded RNA (dsRNA) viruses in the family *Picobirnaviridae*. Picobirnaviruses were first reported in a mammalian species after identification of two bands on gel electrophoresis distinct from rotavirus during a gastroenteritis outbreak in children [114]. Bisegmented dsRNA viruses were previously only identified in fish and birds, and all were classed within family *Birnaviradae*. These novel bisegmented dsRNA viruses similar to the human gastroenteritis strain, were also discovered in rat intestines and later in rabbits [114-116]. Comparison of the most conserved region of the RNA-dependent RNA-polymerase (*RdRp*) gene confirmed *Picobirnaviridae* as a distinct family, with a single *Picobirnavirus* genus identified [117]. *Picobirnavirus* was later identified in Human-immunodeficiency virus (HIV)-infected humans, possibly associated with gastroenteritis and diarrhoea, and primers were developed for detection of an approximately 200 base pair segment of the *RdRp* gene by PCR [117]. Since those initial reports, picobirnaviruses have been detected in faecal samples of many mammalian species including wildlife and domestic animals along with wastewater samples [118-125].

2.1.1 VIRUS FAMILY CHARACTERISTICS

The characterization of *Picobirnavirus* as a virus family has been wrought with challenge and speculation. When *Picobirnavirus* was initially identified as a novel, bisegmented dsRNA virus, it was thought to be related to the *Birnaviridae* virus family as bisegmented dsRNA viruses were previously only identified in avian and piscine host species. Further evaluation of the structure of picobirnaviruses over the decades has shown that both of the segments of *Picobirnavirus* are more similar to the *Partitiviridae* family while some separate components are similar to elements of the *Cystoviridae* and *Birnaviridae* families [126, 127]. Looking at phylogenetic analyses and protein structure of picobirnaviruses in a recent study by Knox et al. (2018), the authors found that the high diversity within some of the picobirnavirus groups were similar to the diversity within other dsRNA virus families such as *Partitiviridae* but very high compared to other potentially similar virus families such as the *Birnaviridae* virus family [113] (Table 1).

TABLE 1 DSRNA VIRUS CLASSIFICATION

Name	Classification	Host	Genome structure
Riboviria	Realm		
Orthornavirae	Kingdom		
Other (incertae sedis)	Phylum		
Birnaviridae	Family	Birds, molluscs, IV, other V	Segmented
Buplornaviricota	Phylum		
Chrymotiviricetes	Class	Fungi, prot	
Chrysoviridae	Family	Fungi	Segmented

Megabirnaviridae	Family	Fungi	Segmented
Totiviridae	Family	Prot, fungi	Unsegmented
Resentoviricetes	Class		
Reoviridae	Family	V, IV, plant, prot, fungi	Segmented
Vidaverviricetes	Class		
Cystoviridae	Family	Bacteria	Segmented
Psiuviricota	Phylum		
Duplopiviricetes	Class		
Amalgaviridae	Family	Plant	Unsegmented
Hypoviridae	Family	Fungi	Unsegmented
Partitiviridae	Family	Fungi, plant, prot	Segmented
Alphapartitivirus	Genus	Fungi, plant	
Betapartitivirus	Genus	Fungi, plant	
Cryspovirus	Genus	Prot	
Deltapartitivirus	Genus	Fungi, plant	
Gammapartitivirus	Genus	Fungi, plant	
Picobirnaviridae	Family	?, IV, V, prot, bacteria	Segmented
Picobirnavirus	Genus		

IV: Invertebrates; V: Vertebrates; Prot: Protozoa

[113, 128-130]

2.1.2 GENOMIC CHARACTERISTICS

The genomes of picobirnaviruses are small. There are two genome segments that range from 2.2 to 2.5 kilobases (kbp) for segment 1, which codes for the capsid protein, and 1.6 to 1.95 kbp for segment 2, which codes for the *RdRp* gene [116, 117, 119, 131]. Segment 1 contains three open reading frames (ORFs) with ORF1 +/- 2 that code for unknown proteins and ORF3 (though some studies report it as ORF2) that codes for the capsid protein at 552 amino acids (Figure 5) [116, 127, 132-134].

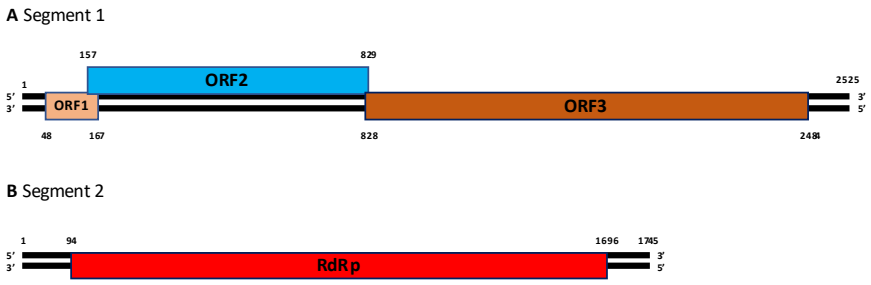


FIGURE 5 *PICOBIRNAVIRUS* SEGMENTS AND OPEN-READING FRAMES

Segment 1 and segment 2 of *Picobirnavirus* with corresponding open-reading frames (ORFs) within each segment. Figure depicted from Ghosh et al. Frontiers in Veterinary Science [134].

Conserved motifs (Table 2) have been identified and confirmed in segment 1 with the repetitive motif ExxRxNxxxE which repeats up to 8 –10 times or more [135]. Segment 2 contains one ORF that codes for the *RdRp* at 534 amino acids in length [126, 132]. Conserved *RdRp* polymerase motifs and conserved domains have been identified and confirmed in segment 2 of picobirnaviruses (Table 2);

these include seven polymerase motifs: PMG, PMF PMA, PMB, PMC, PMD, PME and three conserved domains: CD1, CD2, CD3 [126, 136].

TABLE 2 CONSERVED REGIONS ON SEGMENT 1 (CAPSID) AND SEGMENT 2 (*RDRP*) OF DSRNA VIRUS, *PICOBIRNAVIRUS*

Conserved Region	Segment location	Amino Acid (AA) Sequence	Location of Sequence by amino acid position (AB186898.1)	Location of Sequence by nucleotide position (AB186898.1)
Conserved domain 1	S2, *RdRp*	TDFSKFD	266	798
Conserved domain 2	S2, *RdRp*	SGSGGT	329	987
Conserved domain 3	S2, *RdRp*	GDD	365	1095
Polymerase motif G	S2, *RdRp*	KKSTNSGSPXF	146	438
Polymerase motif F	S2, *RdRp*	DVKQRVVWMFP	203	609
Polymerase motif A	S2, *RdRp*	ICTDFSKFDQH	262	786
Polymerase motif B	S2, *RdRp*	GXHGMGSGSGGTNXDETLXHRALQYEAA	320	960
Polymerase motif C	S2, *RdRp*	NSXCLGDDGXLXY	360	1080
Polymerase motif D	S2, *RdRp*	VXXXYXXHGXEMNXDKQXXS	380	1140
Polymerase motif E	S2, *RdRp*	CTYLRRWHH	425	1356
Conserved repeat domain	S1, Capsid	EXXRXNXXXE	ORF1	ORF1

S2=segment 2; S1=segment 1; *RdRp*=RNA-dependent RNA polymerase; ORF=open reading frame [126, 131, 137-140].

Segment 2 has been characterized further into genogroup I and genogroup II and more recently into an additional genogroup III (GIII) [117, 131, 141]. More recent studies have proposed further genogroups within the segment 2 (*RdRp*) of picobirnaviruses due to separate branches and clades dissimilar from the genogroup I and genogroup II clades [123, 141-145]. Overall, most studies have found higher quantities of genogroup I as compared to genogroup II of segment 2 [146] with a review article by Malik et al. in the mid-2010s reporting that 83% of the picobirnaviruses in the National Center for Biotechnology Information (NCBI) were genogroup I, only 2.5% were reported as genogroup II and the remainder not designated by genogroup [136].

2.2 IDENTIFICATION OF PICOBIRNAVIRUSES

Picobirnaviruses were discovered with the use of polyacrylamide gel electrophoresis (PAGE) for identification of other viruses (typically rotavirus) and the detection of additional bands confirmed the two segments of *Picobirnavirus* [114, 147, 148]. PAGE was used to identify picobirnaviruses for many

years but were found to only identify the virus when the viral load was high and were unable to detect them with lower viral loads that could only be identified with PCR-specific primers [119, 149-152].

PCR primers were developed to identify the *RdRp*, segment 2, of picobirnaviruses by genogroup with genogroup I (25F/43R) and genogroup II (23F/24R) by Rosen [117](2000), which were used in many early studies. Additional primers have been used in other studies for non-genogroup specific identification of the *RdRp*, segment 2 [131, 153, 154] and for segment 1 [125, 131]. Many other studies have developed their own primers and/or utilized other methods such as high-throughput sequencing methods [123, 155, 156]. Since the initial report of picobirnaviruses, the number of picobirnavirus sequences registered in the NCBI database has grown exponentially from 300 to over 2000 in the last 2 decades [97]. The standard nomenclature for picobirnaviruses in NCBI is *Picobirnavirus* strain name/GI/GII/non-GI/non-GII, PBV, common name of host species, three letter country code, strain name and year of isolation (e.g., Porcine picobirnavirus/GI/PBV/pig/BRA-07/2006) [157]. The initial primers by Rosen (2000) provided a useful basis for taxonomic discrimination of different picobirnavirus types and most of the picobirnavirus representatives on NCBI are based on this partial fragment of the segment 2 of the *RdRp*, ranging from 200 base pairs (bp) to 300-400 bp. Other, more recent studies [153], used different primer sets to amplify longer regions of the *RdRp*, ranging from 600-800 bp. More recent studies have identified longer genomes either based on designed primers, the single primer cloning or amplification method and/or the use of metagenomics which has uncovered higher quantities of near complete to complete picobirnavirus genomes from both the *RdRp* and capsid segments [117, 123, 131, 148, 158-161].

2.3 Pathogenicity of picobirnaviruses

The pathogenicity of picobirnaviruses remains unclear. Picobirnaviruses have been identified in immunocompromised people (e.g., those with human immunodeficiency virus (HIV), the elderly and children). The association of picobirnaviruses with clinical signs of diarrhoea have been evaluated in humans and other mammals [114, 117, 162]. The association with diarrhoea and gastroenteritis in wildlife, livestock, captive animals and horses with picobirnavirus infections has been inconsistent [163-166]. Some studies have not shown any association with clinical signs like diarrhoea or morbidity [143, 151]. A meta-analysis [167] did find increased odds of diarrhoea in people with picobirnavirus, an effect that was most pronounced in children and in immune-compromised individuals such as those with HIV [163-167]. Picobirnaviruses have also been associated with diarrhoea in broiler chickens in Brazil [168]. There is also evidence of co-infection of picobirnaviruses with other viruses such as rotavirus, bocavirus and astroviruses [161, 169-171], which may make it difficult to distinguish causality, or an opportunistic effect.

2.4 PICOBIRNAVIRUS DIVERSITY

The genetic diversity of picobirnaviruses is substantial when compared to many other virus families. The discovery of novel viruses by metagenomics has expanded the known diversity of viral genera, suggesting that the genetic diversity of many viral families may be far more extensive than previously thought [129, 172]. There is considerable genetic variability between *Picobirnavirus* sequence types or strains, even within the same *RdRp* genogroups (segment 2). Within the same genogroup and host species, picobirnavirus sequence pairwise-nucleotide identities can range from 50–100% — while between genogroups, the pairwise nucleotide identities within the *RdRp* can range between 25–60% [117, 119, 131]. Segment 1 sequences of picobirnaviruses have also been found to be highly diverse, especially the ORF coding for the capsid protein, with pairwise identities ranging from 45–50% in some studies [98, 131, 173].

The high levels of diversity and the lack of similarity between genogroups has led to some authors to propose that further genogroup classifications for *Picobirnavirus*, at least within the *RdRp* portion of segment 2 [120, 123, 141, 144, 145, 157, 174]. A recent study proposed that picobirnaviruses be classified into three species (R1-R3) with multiple genogroups within each species (R1: G1-G5, R2: G1-G8, R3: G1-G3) [145]. Some studies have even identified fused picobirnavirus segment 1 and segment 2 genomes (both segments joined into one) in both horses and marmots [144, 174] and recombination suspected of segmented genomes [174]. Reassortment has also been suspected in picobirnaviruses and may contribute to further genetic diversity [125, 155, 160, 163]. In a study evaluating picobirnaviruses and circoviruses in camels using metagenomics, picobirnaviruses were found to be highly diverse and different picobirnaviruses, sometimes of different genogroups, were present in "different animals of the same species" [123]. Picobirnaviruses in wastewater also demonstrated substantial diversity with nucleotide identities from 42–100% [142]. Within respiratory samples, equally high genetic diversity has been reported: within-genogroup diversity ranging from 58–80% percent identity for genogroup I and 69–96% for genogroup II in pigs and overall, 58–97% percent identity in human respiratory samples [175, 176] .

Virome studies have shown that we know very little of the vastness of viral diversity. In a study evaluating viral richness and diversity, it was estimated that there can be as many as five or more virus species per sample, or per host [177]. The addition of metagenomic data has increased both the number of identified viral sequences, along with understanding of the dynamics of the viral sequences within hosts [178]. The use of phylodynamics for analysing viral sequences has added to the knowledge of how factors such as "host immunity, transmission bottlenecks and epidemic dynamics" affect the genetics of the pathogen within the host [111]. Within-host diversity of the pathogen from either *in vivo* mutations in the host (such as in influenza), or within a vector (such as dengue), can be

due to many different processes: positive selection, stochastic processes such as genetic drift, and negative selection resulting in purifying selection [179]. In addition to within-host diversity, in studies of causative agents of human diarrhoea, it is not unusual to identify co-infections of multiple viral pathogens in the same sample: primary pathogens and opportunists [173, 180]. “Viral co-occurrence” was identified in a virome study of rhesus macaques in Bangladesh with detection of multiple types of viruses including picobirnaviruses in the same individual macaque [153].

Many studies have identified multiple different picobirnaviruses within the same sample or individual host [123, 152, 155, 175]. Picobirnavirus co-infections vary in genetic similarity, even within the same individual [136, 173, 175, 180]. One study of both respiratory and gastrointestinal samples for picobirnaviruses found that multiple samples and hosts were positive for more than one *Picobirnavirus* species, and these picobirnaviruses had 70% nucleotide similarity even within the same sample [160]. Another study in a diarrheic chicken found 8 different picobirnaviruses (five segment 1s and seven segment 2s) within one sample [133].

2.5 Host and geographic characteristics

Numerous research studies have shown that certain factors such as host genetic-relatedness, increased contact rates, increased transmission potential, and a similar geographical region of distribution all contribute to similarities between viral pathogens [111, 181, 182]. Host clustering of viral pathogens is common and increased genetic-relatedness in hosts has been associated with viral similarities [19, 20, 182-185]. Viruses are typically more similar among hosts in geographical regions due to either increased contact potential between different hosts sharing multihost viruses and/or contamination of the environment leading to increased similarity in the viruses [111, 181, 186-189].

2.5.1 Host of picobirnaviruses and associations

Picobirnaviruses were first reported in a mammalian species during a gastroenteritis outbreak in children [114] and later identified in rats and rabbits [115, 116]. Since those initial reports, picobirnaviruses have been detected in faecal samples of many mammalian species including wildlife and domestic animals [117, 131, 154]. Picobirnaviruses have been identified in marine mammals and dromedary camels, in humans and other primates, in environmental water samples and in birds and reptiles [118, 120, 122-124, 142, 149, 190-192]. The sample origins have been primarily faecal, but have also included intestinal contents, cloacal samples from birds, wastewater and river water samples [119, 121, 123, 142, 168, 193]. Additionally, picobirnaviruses have also been identified in respiratory specimens from cattle and monkeys [160] , along with pigs and humans [175, 176].

With the widespread distribution of picobirnaviruses and identification in various sample types, many studies have sought to understand whether picobirnaviruses are more similar within the same host

species, also termed host clustering. There is wide variation in the findings from such studies, with some showing consistent similarities between picobirnaviruses in certain species [119, 120, 150, 154, 163] though some show inconsistent clustering [194-197] or no clustering at all [120, 151, 154, 155, 160, 176, 194] dependent on host species. The identification of picobirnaviruses in terrestrial and marine mammalian species, along with avian and reptile species, has allowed further evaluation of host clustering but with no clear association being apparent [120, 124, 146, 155].

Despite the studies showing a lack of host clustering, there may be some support for the possibility of sharing of picobirnaviruses between species in close proximity [119, 142, 146, 163]. Also, some studies have shown that porcine picobirnaviruses tend to cluster with human picobirnaviruses [119, 120, 150, 154, 196]; additionally, horse or equine picobirnaviruses also show more similarity to human picobirnaviruses [163](Figure 6).

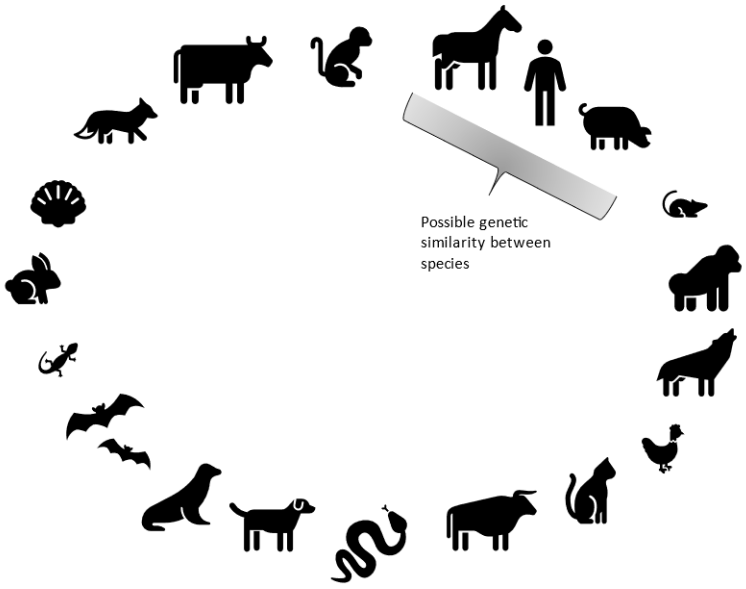


FIGURE 6 HOST AND ENVIRONMENTAL SAMPLES FOR PICOBIRNAVIRUS

Depiction of the various host and environmental samples in which picobirnaviruses have been identified. Picobirnavirus similarities between human and porcine and human and equine along with wastewater samples have been reported [119, 120, 142, 150, 154, 163, 196]. Depicted from figure from Ganesh et al. Animal Picobirnaviruses 2014. [124]

In contrast to the high diversity of picobirnaviruses, many other studies have detected picobirnavirus sequences with very high similarities or even identical sequences from different hosts species. This may either conflict with the extensive diversity noted within the virus family, or may confirm the potential for cross-species transmission of the virus between different host species, especially considering the high diversity. A study evaluating picobirnaviruses in pigs on a farm in Argentina found

that approximately one third of the picobirnaviruses identified in different aged pigs on the farm at different time points were 100% identical [198]. Additional studies have shown picobirnaviruses with, or close to, 100% pairwise identity within the same study and different host species [119, 142, 163, 196, 198] and also persistent picobirnavirus identification of the same virus in the same host for prolonged time periods [198]. All of these studies, though, identified homologous sequences in the shorter approximately 200 bp fragment of the *Picobirnavirus RdRp* gene. In addition, the lack of host clustering and the identification of high similarity picobirnaviruses in different host species from the same study, could also support cross-species transmission of picobirnaviruses [142]. In the SpillOver online tool (mentioned in Section 1.8), picobirnaviruses rank overall mid-range in spillover risk (49–61 out of a risk score of 155) with the highest risk position of all viruses in the tool at 199 out of 887 virus types [108, 199].

2.5.2 Geographical distribution and associations

The known global distribution of picobirnaviruses has been expanding because of the use of molecular techniques. Picobirnaviruses have been reported in North America, South America, Europe and Asia [136] but had not previously been reported on the African continent until 2016 [93, 98, 200]. A study from the USA tested wastewater for viruses to detect contamination from sewage and found picobirnaviruses in 100% of the raw sewage and 33% of the final effluent from various states in the USA [121]. One of the earlier picobirnavirus studies, which generated the most widely-used picobirnavirus primers, detected them from faecal samples from humans with gastroenteritis in the USA and also in Argentina, Venezuela and China [117]. Further studies have identified picobirnaviruses in the European countries of Italy [201], Spain [85], Portugal [125], Hungary [119, 202], Belgium [203], the United Kingdom [204], the Netherlands [141, 173, 176], Croatia [205] and Slovenia [194]. Picobirnaviruses have also been identified in the United Arab Emirates [123], India [163, 196, 206], Bangladesh [153], Thailand [131, 169, 207], Brazil [149], and Uruguay [146], Hong Kong [122, 160], Japan [144, 156], South Korea [208], Australia [90, 209], and New Zealand [210]. Picobirnaviruses were first identified on the African continent in the mid-2010s, initially reported in Kenya and Uganda [200] and later in the Democratic Republic of Congo, Ethiopia, Cameroon and again in Uganda [91, 93, 98].

In the early reports of picobirnaviruses, detection of atypical picobirnaviruses, which appeared to be associated with protozoal-positive faeces (see below, Section 2.6.1), did show similar genome profiles between the different locations all within the United Kingdom, though overall they were also found to be more genetically similar to one another as compared to typical picobirnaviruses [204]. In a study on viral diversity, picobirnaviruses were identified in multiple rhesus macaques in Bangladesh at multiple sites, though they were more similar to each other within sites than between sites [153]. The identification of picobirnaviruses in Thailand in pigs from 22 farms found viruses from the same farms

were more similar than to picobirnaviruses from other farms, clustering together on the same branch of the phylogenetic tree [169].

In contrast, most studies have shown only weak or no associations or clustering of picobirnaviruses based on geographical origin. In the Rosen (2000) study, with the identification of picobirnaviruses in the USA, Argentina, Venezuela and China, little to no geographical clustering was noted on the phylogenetic tree of the sequences from the segment 2 of picobirnaviruses. On phylogenetic analysis, the picobirnaviruses in wastewater samples in the USA did not cluster or group by states or even by regions based on similarity [121]. Other studies have shown little to no clustering by geography of the picobirnaviruses identified [131, 132, 163, 173, 190, 197, 211]. Even in the same host (broiler chickens from various farms in Brazil), the picobirnaviruses were more likely to have a higher pairwise-nucleotide and amino acid identity to picobirnaviruses from other sites than picobirnaviruses from the same site, even in the pooled samples [149].

2.6 What is the actual host of picobirnavirus?

Due to the extensive diversity of picobirnaviruses between genogroups and even within genogroups, the ubiquitous nature of this dsRNA virus geographically, and the identification of picobirnaviruses in a growing number of hosts and sample types, theories have been proposed that question whether picobirnaviruses are truly a "simple" vertebrate virus. Is their detection in eukaryotes just coincidental?

2.6.1 A protozoal host?

When first identified, genetic analysis of picobirnaviruses showed similarity to known picobirnavirus sequences based on PAGE characteristics, which were reported in approximately one third of human faecal samples in association with *Cryptosporidium parvum,* later termed atypical picobirnaviruses [116]. These picobirnaviruses were found to have slightly smaller genomes and less genomic variation as compared to the divergent genome of the typical picobirnavirus [116]. To explore this association, purified protozoal oocyts were evaluated but no picobirnaviruses, either typical or atypical, were identified in the purified samples. The identification of the atypical picobirnaviruses in faecal samples with *C. parvum* represented a possible association between the two pathogens, or that the virus was a potential viral pathogen of protozoa. Just a couple of years later, *Cryptosporidium* isolated from faecal material from calves, goats and mice were found to contain dsRNA viruses with evidence of polymerase activity and possible capsid protein encoding [212]. Another study by the same authors found that it was only the *C. parvum* that was found with the dsRNA and when comparing them between the calves and human samples, they were very similar (>92–93% similarity) [213]. The authors were able to isolate virus-like particles and identify proteins that were encoded for in ORFs in

dsRNA for RNA polymerase showing "direct evidence for the expression of dsRNA genes in the oocysts of *C. parvum*" [212]. A larger dsRNA genome (L-dsRNA, 1786 bp), thought to be associated with the polymerase, and a smaller dsRNA genome (S-dsRNA, 1374nt), thought to be associated with the capsid were identified. The polymerase and properties of the genes was most similar to the family *Partitiviridae*, though viruses within this family had only been associated with plants and fungi up until this time [212] while dsRNA viruses of protozoa were in the *Totiviridae* and *Reoviridae* families only. Later, the virus family *Cryspovirus* was approved within the *Partitiviridae* family as a virus of the protozoan *Cryptosporidium parvum*, but within the family of similar viruses infecting plants and fungi [214]. The bisegmented dsRNA virus was similar, in the attribute of size for the two segments, to picobirnavirus*,* with a larger segment at 1700 bp and a smaller segment at 1400 bp, though different to picobirnavirus. The larger segment contained the ORF for the *RdRp* and the smaller segment for the capsid protein [214]. It was likely that the prior atypical picobirnaviruses (associated with *Cryptosporidium*) were actually cryspoviruses. The presence of picobirnaviruses in respiratory samples provides contrasting evidence that these viruses may be associated with gastrointestinal protozoa, although *Cryptosporidium* has been reported from human and other animal respiratory tracts [160, 175, 176, 215].

In addition to *Cryptosporidium*, a study looking for rotaviruses in diarrheic and non-diarrheic canine faeces by PAGE found picotrirnavirus in combination with *Giardia* and *Ancylostoma* [216]. Picotrirnavirus was first detected in 1990 in chickens, similar to picobirnaviruses but with a trisegmented rather than bisegmented genome (2.7, 2.3 and 0.9kpb) [217] and also found in feces of children with diarrhea [218]. Five of these human samples were found to be consistent with picotrirnavirus, all non-diarrheic and young in age with variable sizes and patterns on PAGE.

2.6.2 A BACTERIAL HOST?

Due to the extensive diversity of the picobirnaviruses, the inability to draw a firm conclusion on pathogenesis, and the speculation that the closest related viruses to picobirnaviruses may be the partitiviruses which infect fungi, plants and now protozoa [212, 214, 219]—it has been hypothesized that picobirnaviruses are prokaryotic viruses rather than eukaryotic viruses. Possible prokaryotic motifs or identifiers in picobirnaviruses have been identified. From the work of Boros et al. (2018), multiple segment 1 and segment 2 picobirnaviruses were identified from a single diarrheic cloacal sample from a chicken and close inspection of these components of the viruses found conserved motifs in the untranslated regions (UTRs) preceding ORF1s of both the segment 1 and segment 2 [133]. These conserved motifs were consistent with ribosomal-binding motifs (RBS) from prokaryotic messenger RNA, termed Shine-Dalgarno sequences [220] and found in all of the picobirnaviruses from this host [133]. Similarly, Krishnamurthy in 2018 evaluated picobirnaviruses from the NCBI database

and from a prior study of macaque picobirnaviruses, and found that all of the viruses had at least four of the RBS (4-mer; AGGA/GGAG/GAGG) and 75% of them had five (5-mer; AGGAG/GGAGG) to the complete six (6-mer; AGGAGG) of the RBS in the UTRs preceding the ORFs in either segment of the picobirnaviruses [203]. The authors then evaluated various eukaryotic and prokaryotic viruses for RBS quantities and found that "81% of prokaryotic RNA viral genomes had at least 50% of their ORFs preceded by a 4-mer RBS while only 6.5% of eukaryotic RNA viruses had this property", similar for DNA viruses [203]. They concluded that the absence of the RBS motifs does not indicate whether a virus is likely a prokaryotic or eukaryotic virus though the presence of at least 30% of the virus family having RBS sequences is more consistent with a prokaryotic virus [203]. A subsequent study that evaluated picobirnaviruses in a mongoose species found that all of the segment 2 genogroup I picobirnaviruses that were identified did possess the entire RBS in the UTRs upstream of the start codon [195].

The genetic analyses of picobirnaviruses, especially in respect to the genetic diversity and genogroup classification, has led to further theories on the characterization of picobirnaviruses. A virome study from Cameroonian fruit bats found picobirnaviruses that did not cluster with the known genogroup I or genogroup II segment 2 picobirnaviruses, nor were they able to identify the ORF in the segment 2 of the *RdRp* [93]. Translation of nucleotides into amino acids utilises various translation tables with the standard genetic code (translation table_1) the most common codon translation table (Table 15, Appendix A, Table 4) [221]. Only when they used an alternative genetic code (AGC) from invertebrate mitochondria (using translation table_5), were they able to identify the ORF in the picobirnaviruses. Further evaluation of the AGC picobirnaviruses found that they were more similar to the mitoviruses that also use the same AGC [93]. In addition, the translation of the ORF of a novel near-complete to complete segment 2 *Picobirnavirus* found in the mongoose on the island of Saint Kitts was identified only with the use of an AGC [195]. In this study, the authors inadvertently identified the ORF with the use of both the mould (transl_table 4) and invertebrate (transl_table 5) mitochondrial genetic codes, still with the known conserved motifs of the *RdRp* and clustering with other genogroup I picobirnaviruses [195].

2.6.3 Other host?

The phyla that is the "true host" of picobirnaviruses has not been resolved, and a recent review paper summarises the possibilities [134](Ghosh et al. 2021). As noted in Table 1, only one other dsRNA virus is known to have a bacterial host, the *Cystoviridae* in a different phylum to *Picobirnaviridae*, and bacterial hosts for other RNA viruses are uncommon (*Leviviridae*, ssRNA+ virus) [130]. In contrast, bacteriophages are the most-commonly known viruses of bacteria, either of DNA or RNA genome, and have been studied for use as antimicrobials, for fermentation or for diagnostic purposes [222-224].

2.7 Specific Aims

The first aim, addressed in Chapter 4, was to identify picobirnaviruses in the various samples from Uganda and New Zealand, to evaluate different methods of sequence generation, and to evaluate different analytical tools for the characterization of the virus. In a system where the possibility for within- or between-host transmission is probable, the objectives were to evaluate picobirnaviruses for evidence of clustering-based genogroup structure, phylogeny and/or the potential for cross-species transmission. My initial hypotheses were that I would find evidence of clustering based on genogroup but find little to no evidence of cross-species transmission.

The second aim, addressed in Chapter 5, was to identify if picobirnaviruses from the study cohort samples from Uganda and New Zealand (along with previously identified picobirnaviruses) would reveal any evidence of clustering within host species or based on geographical location. The research objectives were to evaluate by phylogenetic analyses, the picobirnaviruses from the study samples as well as from a database for evidence of clustering. The null hypothesis was that there would be no host, or geographic structure, in the various picobirnavirus groups analysed.

The third aim, addressed in Chapter 6, was to identify if multiple picobirnaviruses existed within the same sample, and to analyse the diversity of these viruses within- and between- those hosts. The research objectives were to evaluate for multiple picobirnaviruses from select study samples and the range of diversity by phylogenetic analysis and heatmaps. My hypotheses were that I would find multiple picobirnaviruses within the same host; and that further analysis of these viruses would identify within-host diversity.

Finally, due to the expected high genetic-diversity and lack of host- or geographic-clustering of the picobirnaviruses, the forth aim from the study was to assess picobirnaviruses further in my study, addressed in Chapter 7, by the specific objectives of: 1) identification of picobirnaviruses in samples from various bacterial colonies; 2) identification of picobirnaviruses in protozoal samples from previously-associated protozoal species; 3) evaluation of an alternative genetic code for the translation of the picobirnaviruses; and 4) the detection of RBS in near-complete picobirnaviruses from the samples. These objectives sought to further clarify the possible hosts of picobirnaviruses. The null hypotheses were that we would not find picobirnaviruses in these samples, that there would be no difference between the different genetic code usage, and we would not detect RBS motifs in the study sequences.

CHAPTER 3: MATERIALS AND METHODS

3.1 ETHICS AND IMPORTATION APPROVAL

Ethical approval was obtained through the Massey University Human Ethics Committee, application number Southern A Application-14-44. Further consultation with the New Zealand Ministry of Health and Disability Committee and the Massey University Animal Ethics Committee was sought and official approval not deemed to be needed due to anonymous samples and absence of manipulations of animals performed. The Ministry for Primary Industries Animal Welfare Senior Policy Advisor was consulted because of ape involvement, but no further consultation was necessary because of the above reasons. Kaupapa Kura Taiao (KKT) was consulted via the Environmental Protection Agency as part of the discussions relating to the importation of potentially infectious materials into New Zealand. Permission to import samples was given and Uganda National Council for Science and Technology committee approval was given locally.

After consultation with the Massey University Animal and Human Ethics Committees regarding the convenience sampling from New Zealand, ethical approval was not deemed necessary for the collection of anonymized and convenience faecal samples for analysis of viral components in mammalian, bacterial and surveillance protozoal samples.

3.2 SAMPLES

3.2.1 UGANDAN SAMPLES

Faecal samples were collected from mountain gorillas, cattle and humans from the region of Bwindi Impenetrable Forest in south-western Uganda (Figure 2, Figure 3 from Chapter 1, Section 1.4). Sample collection was undertaken by my collaborators in Uganda, Conservation through Public Health (CTPH) between 2014 to 2015. CTPH routinely collects mountain gorilla faecal samples from the night nests with gorilla names and age when known for surveillance purposes. Habituated gorilla groups are tracked daily. Samples were categorized as infant, juvenile, sub-adult, adult or silverback based on the size of the sample and presence of hair samples; faeces were then placed into containers for their analysis (Figure 7) [225].

A) B)

FIGURE 7 GORILLA FAECAL SAMPLE COLLECTION

A) Gorilla faecal sample collection from night nests of the Rushegura gorilla family in Bwindi Impenetrable Forest; B) measuring diameter of the faecal sample for distinguishing age of gorilla [225] and also included evaluation for hair particles for further classification of gorilla sample origin. Photos: J. Wierenga.

Cattle faeces were collected using convenience sampling from community cattle in the region surrounding Bwindi Impenetrable Forest. Both cattle and gorilla faeces were collected and placed into 5 mL cryovials containing RNAlater® (Sigma, MO, USA). Samples from clinically-unwell humans were collected from the local health centres in the region, Kanungu District Medical Office and Kasese District, for evaluation of macroparasites. Samples were anonymised. Samples were placed in RNAlater® within 24 hours of collection and stored in a -20°C freezer until shipment; samples were shipped on ice. For the study presented here, three faecal samples from clinically-unwell humans, 16 cattle faecal samples and 44 mountain gorilla faecal samples were imported into New Zealand and were stored until analysis at -80°C freezer in a PC2 containment facility at the Hopkirk Research Institute, Massey University, New Zealand. Frozen faecal samples in RNAlater® were thawed and separated as duplicate samples (reference sample and working sample) once they were in the PC2 facility in New Zealand.

3.2.2 New Zealand Samples

3.2.2.1 Additional vertebrate host samples

In addition to the faecal samples from Uganda, faecal samples were also collected in New Zealand for identification of picobirnaviruses, and for phylogenetic analyses by host and geography. New Zealand

is a useful comparative system. New Zealand is an isolated island system in the Pacific, approximately 2000 kilometres from the next nearest continent, Australia. This geographic isolation means its fauna and flora are often unique, with a high level of endemism [226, 227]. New Zealand was one of the last places on earth that people colonised, including with domesticated animals [228, 229]. Moreover, while people move freely (pre-COVID-19) globally, most live livestock is now domestically bred with very little importation. For example, most cattle were imported into New Zealand during the 1860s and 1950s–1990s, but since 1991 fewer than 100 live cattle have been imported each year [228, 229]. These unique characteristics make it a useful natural experiment [226, 230].

Power calculations determined that a sample size of just 10 samples per host species was required to detect one positive sample with 95% confidence, given the estimated prevalence of picobirnavirus identification in prior studies (on average approximately 30% for PCR-based studies from animal hosts) [136]. Faecal samples were collected from clinically-well domestic cattle, sheep and horses from local farms in the Whanganui-Manawatu region of New Zealand. Faecal samples were also collected from both clinically-well and clinically-unwell domestic cats and dogs from convenience sampling; all samples were collected after defecation. Samples were placed in RNAlater® within 24 hours of collection and placed in a -80°C freezer until extraction. For the phase of the New Zealand study, 10 faecal samples were collected from domestic cattle (all clinically well), sheep (all clinically well), horses (all clinically well), cats (2 clinically unwell, 8 well) and 12 samples from dogs (4 clinically unwell, 8 well).

3.2.2.2 Additional samples for investigation of the actual host of Picobirnavirus

With the aim to further elucidate the actual host of *Picobirnavirus*, colonies of various enteric bacteria and purified oocysts of *Cryptosporidium* and *Giardia* were obtained to investigate if picobirnaviruses were present. Bacteria were selected through convenience sampling of recently grown colonies in the Molecular Epidemiology and Public Health Laboratory at Massey University, and protozoa were purified using established techniques. Specifically, oocysts from *Cryptosporidium* were purified from faeces from a modified method described by Meloni & Thompson (1996) with an additional 2% Ficoll layer added to the flotation step and PBS to replace dH_20/0.02% w/v Tween-20 for all of the washes [231, 232]. *Giardia* cysts were purified from faecal samples using the protocol recently published by Obguigwe et al. (2020). Thirty colonies from enteric bacteria were collected and 10 purified oocyst samples from *Cryptosporidium* (*C. hominis, C. parvum, C. cuniculus*) and 10 cyst samples from *Giardia intestinalis* were collected and stored until analysis at -80°C in a freezer in a PC2 containment facility at the Hopkirk Research Institute, Massey University, New Zealand.

3.3 Nucleic acid extraction and amplification

The extraction of nucleic acids from the collected faeces was performed following the protocol as described by Hall et al. (2014) [233]. Approximately 200 mg of faeces was placed into a 2 mL safelock tube (Eppendorf, Hamburg, Germany) with 1 mL of 0.01 M PBS pH 7.3 and incubated for one hour at 4˚C. The suspended faeces were centrifuged and the supernatant passed through a sterile 0.45 µm filter (Macherey-Nagel GmbH & Co. KG, Düren, Germany) and the filtrate was used as the source material for nucleic acid extraction with the Roche High Pure Viral Nucleic Acid kit (Roche, Basel, Switzerland) or the Macherey-Nagel NucleoMag Vet Viral RNA/DNA Isolation Kit (Macherey-Nagel GmbH & Co. KG, Düren, Germany) with the KingFisher Flex Purification system (ThermoFisher Scientific, MA, USA). Nucleic acids were then treated by either one of two methods depending upon if RNA or DNA was to be extracted (Figure 8).

3.3.1 For RNA extraction

For the extraction of RNA (Figure 8), the DNA was removed with Ambion DNA-free™ (ThermoFisher, Massachusetts, USA) following manufacturer's instructions. First-strand cDNA synthesis was performed using Invitrogen SuperScript™ III (ThermoFisher, Massachusetts, USA). The resulting single-stranded DNA was used as the template for whole transcriptome amplification using the QuantiTect Whole Transcriptome kit with the substitution of Superscript III for the reverse transcriptase (RT) enzyme (Qiagen, Hilden, Germany; ThermoFisher, Massachusetts, USA). Quantification of cDNA was performed using the DNA Qubit kit (ThermoFisher, Massachusetts, USA).

3.3.2 For DNA extraction

For the extraction of the DNA (Figure 8), whole genome amplification was undertaken to bring the DNA yield up to the amount required for sequencing. Whole genome amplification was performed on the extracted nucleic acids using the Illustra GenomiPhi V2 DNA Amplification Kit (GE Healthcare Life Sciences, Illinois, USA) as per the manufacturer's instructions. Quantification of DNA was performed using the dsDNA HS Qubit kit (ThermoFisher, Massachusetts, USA).

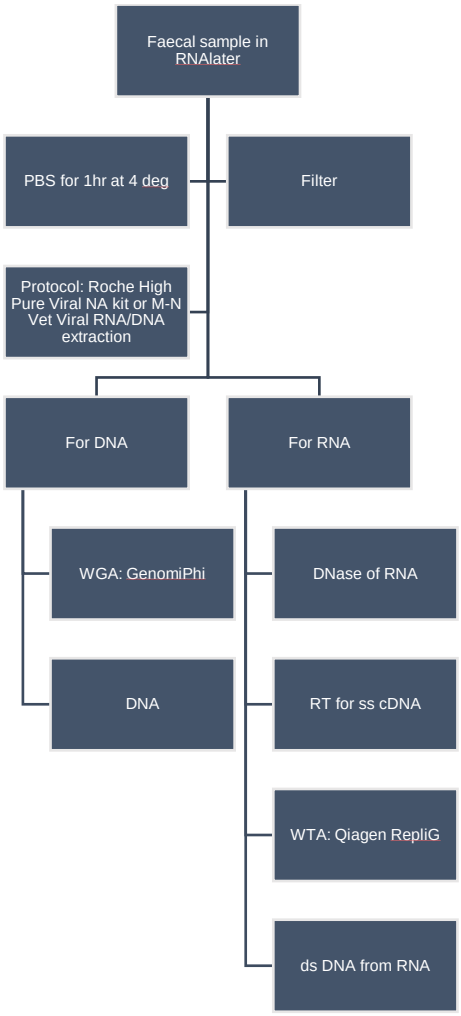


FIGURE 8 NUCLEIC ACID PROCESSES FOR SAMPLES SELECTING FOR DNA OR RNA PATHOGENS

PBS=phosphate buffered saline; NA=nucleic acid; RNA=ribonucleic acid; DNA=deoxyribonucleic acid; M-N=Macherey-Nagel; WGA=whole genome amplification; RT=reverse transcriptase; ss cDNA=single stranded copy DNA; WTA=whole transcriptome amplification; ds DNA=double-stranded DNA.

3.3.3 NEXT-GENERATION SEQUENCING PREPARATION

Aliquots of 5µg of whole transcriptome and whole genome amplified nucleic acids were dried down in a hooded cabinet under negative pressure at room temperature overnight in a DNAstable® plate (Biomatrica, California, USA) and sent overnight in room temperature conditions via courier to Orion Integrated Biosciences Inc. (Manhattan, Kansas, USA) for next generation sequencing. Next-generation sequencing (NGS) was performed on an Illumina® MiSeq with Nextera XT DNA Library Prep (Illumina®, California, USA) and preliminary bioinformatics were performed using a novel k-mer based system developed at Orion that uses protein motif fingerprint discovery methods [234]. For NGS, a procedural negative control of RNAlater® alone was included in the nucleic acids extraction, for both RNA and DNA amplification methods; these negative-template control results were then subtracted from the sample results.

3.4 CONFIRMATION OF NEXT-GENERATION SEQUENCING

3.4.1 PATHOGEN POLYMERASE CHAIN REACTION (PCR) SCREENING

Analysis of the NGS results from both extracted RNA and DNA with the Orion k-mer approach identified target pathogenic organisms which required further confirmation. The presence of these target organisms, organisms considered as potential zoonotic pathogens, (Table 3) was confirmed by pathogen-specific PCR methods. Pathogen-specific PCRs were performed with forward and reverse primers (Table 3). Primers were sourced for the specific organisms based on the NGS results (Appendix A, Table 11) (Integrated DNA Technologies, Iowa, USA). Organisms that required confirmatory PCR, based on NGS data (Appendix A, Table 11) included: *Coxiella burnetii*, Lassa virus, *Haemophilus parainfluenzae*, Hepatitis C virus, *Salmonella* species, *Escherichia coli*, *Plasmodium* species and *Picobirnavirus*. Both DNA, and cDNA from RNA, were screened with pathogen-specific PCR assays. DNA samples were tested for *C. burnetti*, *H. parainfluenzae, Salmonella* species, *E. coli*, and *Plasmodium* species by using PCR; RNA (as cDNA) was tested for all the pathogens by using PCR assays. PCR testing was performed using 1X HOT FIREPol Blend Master Mix (10 mM $MgCl_2$, Solis BioDyne, Estonia), 300 nM of each primer and 1 µL of template made to a total volume of 20 µL with nuclease free water. The thermocycler protocol for each pathogen is specified in Table 3.

TABLE 3 PCR PRIMERS AND THERMOCYCLER PROTOCOLS

Organism (genus, species, virus)	Type of pathogen	Forward primer	Reverse primer	Alternate primers	Thermocycler protocol	Reference(s)
Coxiella burnetii	Obligate intracellular bacteria	IS1trg-f	IS1trg-r		15 min 95°C; 35 cycles of 95°C 30s, **60°C** 30s, 72°C 60s; 72°C 10 min	[235] (de Bruin 2011)
Haemophilus parainfluenzae	Gram-negative, pleomorphic, coccobacilli bacteria	H-para-F	H-para-R-1	H-para-R-2	15 min 95°C; 45 cycles of 94°C 30s, **56°C** 30s, 72°C 30s; 72°C 10 min	[236] (Tian Guo 2012)
Hepatitis C	Positive-sense single-stranded RNA virus	NS5B-F	NS5B-R	NS5B-Nest F; NS5B-Nest R	15 min 95°C; 50 cycles of 95°C 30s, **63°C** 30s, 72°C 30s; 72°C 10 min	[237, 238] (Sandres-Saune 2003, Margall 2015)
Lassa virus	Ambisense RNA virus	36E2-F	LVS526-revR	LVS-339-revR	15 min 95°C; 45 cycles of 95°C 30s, **52°C** 30s, 72°C 30s; 72°C 10 min	[239] (Asogun 2012)
Picobirnavirus	Double-stranded RNA virus	PicoB25	PicoB43	PicoB23 / PicoB24	15min 95°C; 45 cycles of 94°C 30s, **52°C** 30s, 72°C 30s; 72°C 10min	[117] (Rosen 2000)
Picobirnavirus	Double-stranded RNA virus	327F/ 662F/ 732F/ 758F	751R/ 777R/ 1243R/ 1358R	1224F/1339F 1923F 1940R/1944R 2125R/2343R	15min 95°C; 7 cycles of 94°C 30s, 62**°C-1°C each cycle for** 30s, 72°C 1min;	NA; designed primers performed on Geneious®

					38 cycles of 94°C 30s, 56**°C** 30s, 72°C 1min; 72°C 10min	software program
Picobirnavirus	Double-stranded RNA virus	PicoF3/ PicoF5	PicoR5/ PicoR8		15min 95°C; 45 cycles of 96°C 30s, **48°C** 30s, 72°C 60s; 72°C 4min	[153] (Anthony 2015)
Plasmodium spp.	Parasitic protozoa	rPLU 3	rPLU 4		15min 94°C; 35 cycles of 94°C 30s, **62°C** 60s, 72°C 60s; 72°C 10min	[240] (Singh 1999)

PCR products were separated by gel electrophoresis with fluorophore (RedSafe, iNtRON Biotechnology, South Korea) in a 1–2% agarose gel (Meridian Bioscience, London, UK) on 70–90 mV for 30–45 minutes, depending on expected size and indicated quality of the PCR products, and included negative controls, positive controls (Section 3.5) and 1 kb ladder (ThermoFisher Scientific, MA, USA) on 0.5X TBE running buffer. PCR bands were visualised by UV light using a Gel Doc XR and Quantity One 4.6.2 software program (Bio-Rad Laboratories, California, USA). After pathogen-specific PCR screening, bands of the expected size were excised, and DNA was eluted with 40μL 10 mM Tris, pH 8.0, for 12–24 hours at 4°C and then sent for bi-directional Sanger sequencing to the Massey Genome Service (Massey University, Palmerston North, New Zealand). See Appendix A: supplementary material for Chapter 3.

3.5 Experimental controls

Positive and negative controls were always included in the laboratory methods and analysis. Negative controls included both RNAlater® (procedural control) with the above protocol for RNA and DNA extraction (Figure 8), as well as PCR-grade water (non-template control) for PCR. For PCR methods, positive controls (synthetic DNA; gBlocks®) for the target organisms were run for all PCR assays to confirm that primers, 1xHOT FIREPol Blend Master Mix and thermocycler protocols were all functional; in addition, positive control amplicon size was determined as further confirmation prior to sequencing. gBlocks® are double-stranded DNA fragments synthetically produced allowing positive controls to be easily acquired, and standardised. Positive controls (gBlocks Gene Fragments®, IDT, Iowa, USA) for the organisms were designed to contain mutations outside the primer-binding sites, which ensured that there were stop codons in all six reading frames, to allay concerns of pathogenicity of the clone and to provide a means to identify laboratory contamination. Figures below (Figure 10, Figure 11, Figure 12) show the picobirnavirus gBlock® sequence with the six translation frames with the tested primers. Each PCR for a specific organism was trialled with a corresponding positive control (gBlock®) and negative control (non-template). The gBlocks® were cloned into an *E. coli* host using the Invitrogen TOPO® TA Cloning Kit for Sequencing (ThermoFisher, Massachusetts, USA), and

transformed plasmids were extracted using an alkaline lysis extraction protocol and were resuspended in elution buffer (10 mM TrisHCl, pH7.5). Extracted plasmids were tested by PCR with the corresponding primer sets for that organism (Table 3) and sequenced. Optimisation for the adequate quantity of positive control was performed with serial dilutions of the gBlock® positive controls with PCR to identify a clear, well-defined band without excess template.

3.6 *Picobirnavirus* Detection

3.6.1 Molecular detection of *Picobirnavirus* by PCR

3.6.1.1 Published Picobirnavirus primers

For the specific evaluation of picobirnavirus, multiple detection methods were utilized for detecting the organism. Based on published protocols for picobirnaviruses, I used the following primers: PBV forward primer 23 (5'-CGGTATGGATGTTTC-3') & reverse primer 24 (5'-AAGCGAGCCCATGTA-3') and forward primer 25 (5'-TGGTGTGGATGTTTC-3') & reverse primer 43 (5'-ARTGYARTGYTGGTCGAACTT-3') [117]. Both sets of primers amplify picobirnavirus genomic segment 2 which codes for the RNA-dependent RNA polymerase (*RdRp*) gene, with primers 23/24 detecting a 368 bp region in the *RdRp* open reading frame (ORF) from nucleotides 685–1053 of genogroup II (GII) viruses and primers 25/43 detecting a 200 bp region in the ORF from nucleotides 665–865 in genogroup I (GI) [117] (Figure 9, Figure 10, Figure 11). The thermocycler protocol for picobirnavirus genogroup I and genogroup II primers was as per Rosen et al. (2000) and optimised with minor modifications to increase the denaturation time from 5 min to 15 min for the 1xHOT FIREPol Blend Master Mix (Table 3).

Further PCR assays for picobirnavirus became available during this study and were used to obtain longer amplicons: PicoF3 (5'-GTDRTDTGGATGTTYCC-3'), PicoF5 (5'-GTHTGGATGTWYCCIATG-3'), PicoR5 (5'-GGRTGRTAYTTVCARTTYTC-3'), PicoF8 (5'-GGRTBRTCHACACARTTYTC-3') [153] (Figure 9, Figure 12). All four primers, two forward and two reverse primers, amplify picobirnavirus segment 2, coding for the *RdRp* gene for both genogroup I and genogroup II, resulting in amplicon sizes of approximately 800 bp. PCR amplification was performed with all four primers as above with the thermocycler protocol: 15 min at 95°C; 45 cycles of 96°C for 30 seconds, 48°C for 30 seconds and 72°C for 60 seconds; and 72°C for 4 minutes (Table 3).

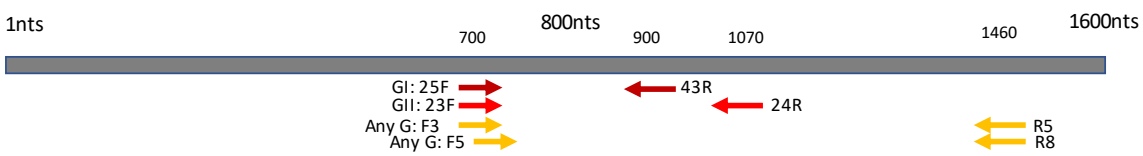


Figure 9 Binding sites for all published picobirnavirus primers used in this study

Primers are designated by the forward and reverse arrows as forward and reverse primers. Dark red forward (25F) primer and reverse (43R) primer for genogroup I (GI) picobirnaviruses; red forward (23F) primer and reverse (24R) primer for genogroup II (GII) picobirnaviruses; orange forward (F3/F5) primers and reverse (R5/R8) primers for non-genogroup (Any G for genogroup) picobirnavirus detection [117, 153]

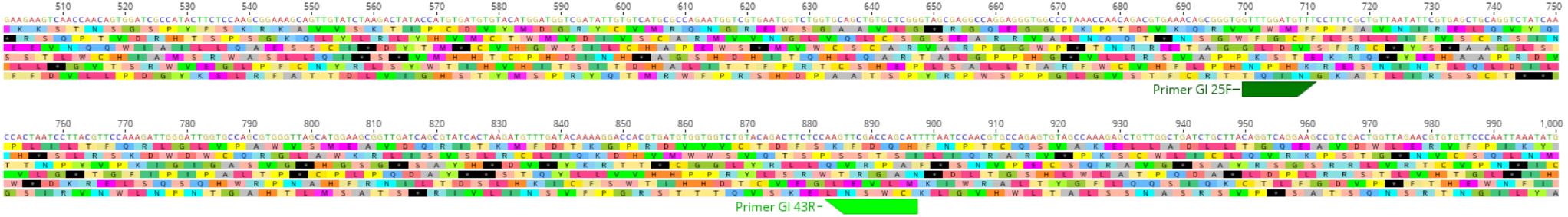


FIGURE 10 GENOGROUP I PRIMERS FOR PICOBIRNAVIRUSES

Genogroup I picobirnavirus primers on the positive control, the picobirnavirus gBlock® sequence, showing multiple reading frames cover all 6 possible combinations of translation from mRNA to protein for the *RdRp* gene fragment. Stop codons are in black and present in all 6 frames. Annotations below the sequence include the forward primer (first one below the sequence and amino acids-pico25F) and reverse primer (second one below the sequence and amino acids-pico43R).

FIGURE 11 GENOGROUP II PRIMERS FOR PICOBIRNAVIRUSES

Genogroup II picobirnavirus primers on the positive control, picobirnavirus gBlock® sequence, showing multiple reading frames cover all 6 possible combinations of translation from mRNA to protein for the *RdRp* gene fragment. Stop codons are in black and included in all 6 frames. Fragments below the sequence include the forward primer (first one below the sequence and amino acids-pico23F) and reverse primer (second one below the sequence and amino acids-pico24R).

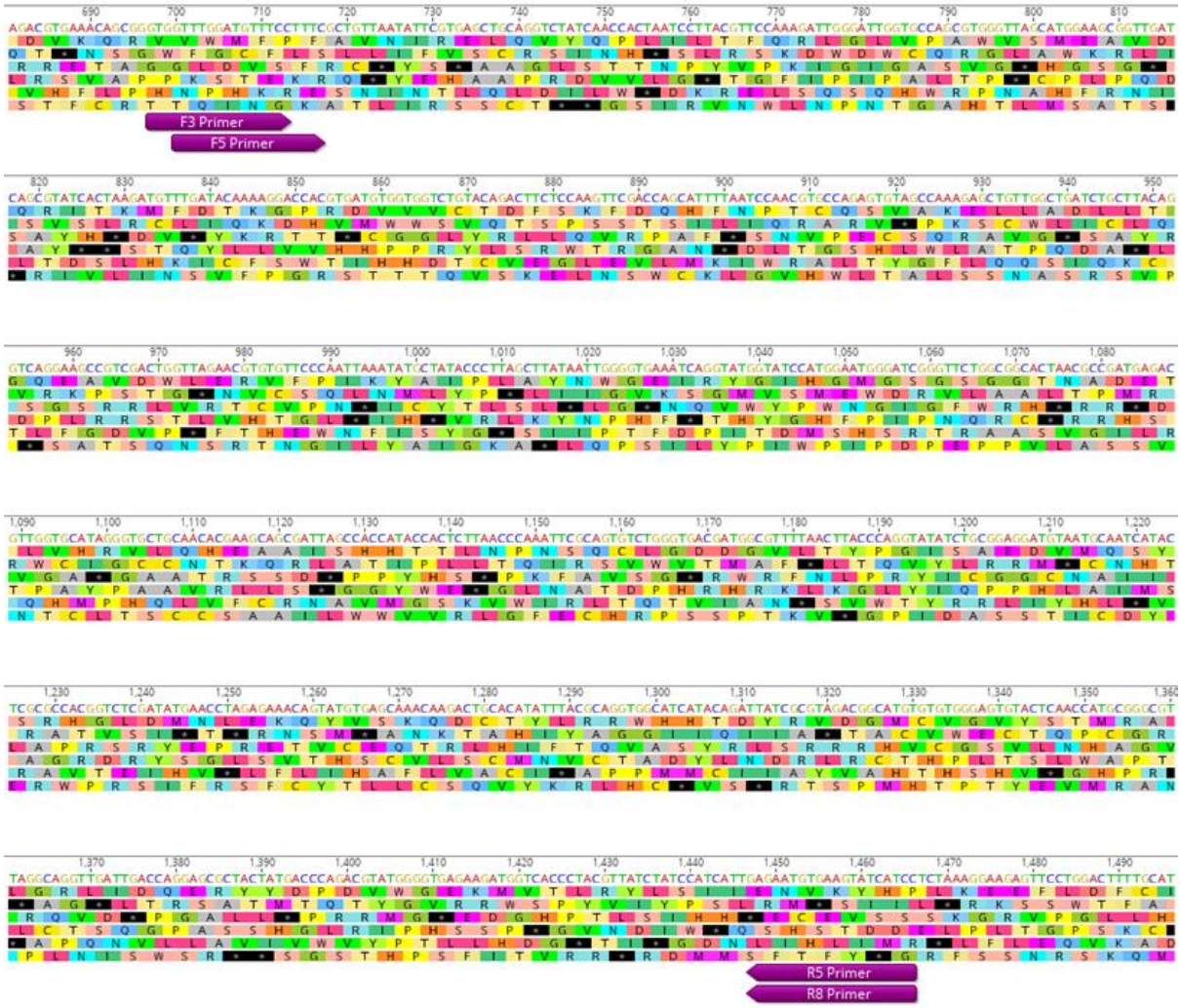


FIGURE 12 ADDITIONAL PRIMERS FOR PICOBIRNAVIRUSES

Four primer set for picobirnavirus shown on the positive control, picobirnavirus gBlock® sequence, showing multiple reading frames cover all 6 possible combinations of translation from mRNA to protein for the *RdRp* gene fragment. Fragments below the longer sequence include the forward primers (first and second one below the sequence and amino acids-F3 and F5) and reverse primers (third and fourth one below the sequence and amino acids-R5 and R8). The four primers can sequence both genogroup I and genogroup II of *Picobirnavirus* to sequence an 600-800 bp amplicon.

3.6.1.2 Designed Picobirnavirus primers

Additional primers were designed to detect and amplify longer amplicons of *Picobirnavirus* using conventional PCR. Primers were designed using Geneious® (Biomatters Ltd.) from the multiple alignment of complete/near-complete picobirnavirus sequences for the segment 2 *RdRp* gene, on the NCBI database [241]. Complete/near-complete picobirnavirus sequences longer than 1000 bp of both genogroup I and genogroup II of the *RdRp*, segment 2, were selected from NCBI. Multiple alignment using the default settings for Geneious® multiple alignment was performed on the 33 downloaded

picobirnavirus sequences (download date: 05 Dec 2017) and a consensus sequence of conserved nucleotides was created (Figure 13). In the Geneious® Primer Design function [242, 243], seven forward primers and eight reverse primers were designed with each primer being 20 bp in length, with a GC content of 45-57.9% and a T_m of 49.7–53.8˚C. Primers were designed with the criteria of having the fewest ambiguities based off the consensus sequence, and selected based on identifying a product size of 200–800 bp with overlap and no repetition, i.e. if the designed primers overlap and both were forward primers, only one was selected. The forward and reverse primers were named based on the nucleotide location and designed to identify and amplify sequences of approximately 200–800 bp by staggering primers with overlapping regions in an attempt to identify and amplify longer *Picobirnavirus* PCR products from the samples (Figure 14). PCR amplification was performed as above (Table 3) with the thermocycler touchdown protocol: 15 minutes at 95°C; 7 cycles of 94°C for 30 seconds, 62°C minus 1°C for each subsequent cycle for 30 seconds and 72°C for 1 minute; 38 cycles of 94°C for 30 seconds, 56°C for 30 seconds and 72°C for 1 minute; 72°C for 10 minutes. Optimisation included utilising the touchdown PCR by altering the annealing temperature by 1–2˚C in the PCR as noted above to obtain adequate and appropriate-sized bands of the product [244, 245].

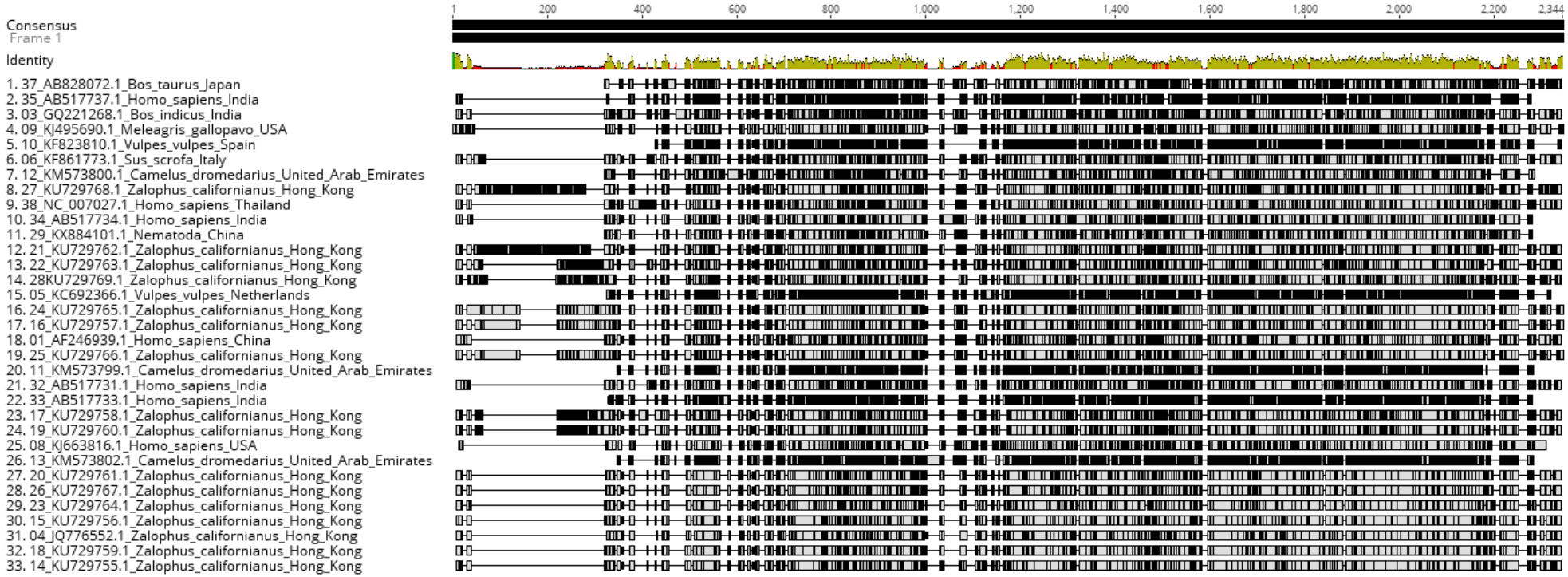


FIGURE 13 PICOBIRNAVIRUS ACCESSIONS AND ALIGNMENT FROM THE NCBI DATABASE OF THE 33 COMPLETE TO NEAR-COMPLETE PICOBIRNAVIRUSES FOR PRIMER DESIGN

Accession number is followed by host species and geographical origin of picobirnavirus. Aligned using the Geneious® alignment tool with default settings. Boxes denote nucleotides that are aligned with the darker-coloured boxes for less consensus between sequence alignments and lighter-coloured boxes with consensus between sequence alignments. Lines between boxes are gaps in the multiple alignment.

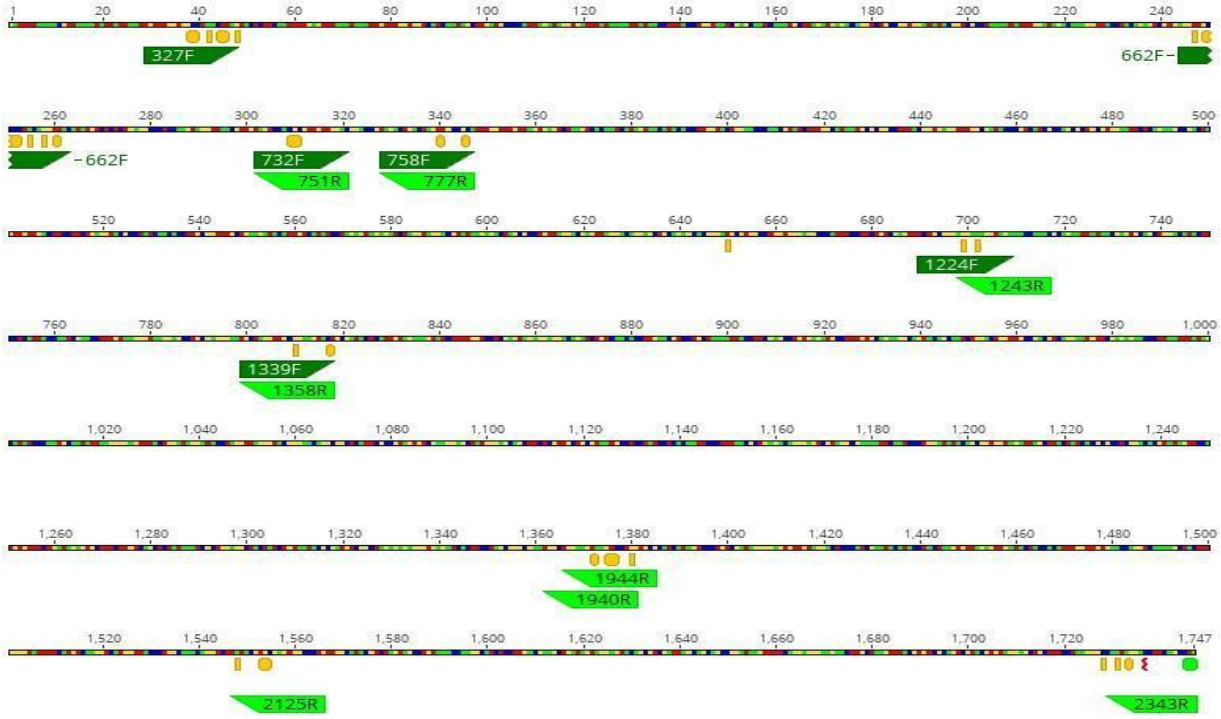


FIGURE 14 DESIGNED PRIMERS FOR PICOBIRNAVIRUSES

Designed primers on Geneious® software (Biomatters Ltd.) from the multiple alignment of complete/near complete picobirnavirus sequences for the segment 2, *RdRp* gene, on the NCBI database [241]. Seven forward primers and eight reverse primers were designed and utilized through conventional and nested PCR.

3.6.2 CLONING FOR PICOBIRNAVIRUS

3.6.2.1 CLONING SELECTION

Cloning of picobirnavirus PCR products was used to investigate the presence of multiple viruses present in individual samples. Samples were selected for cloning based on: 1) those that were positive for picobirnavirus with both 'ambiguous' (e.g., multiple peaks, suggesting polymorphisms) and 'unambiguous' sequences on Sanger sequencing; 2) those that were positive for picobirnavirus with multiple primer sets; and 3) samples that were identified to have more than two different picobirnavirus sequence types in the NGS data.

Cloning was performed on amplicons from three gorilla samples and four cattle samples from Uganda along with a single cattle sample from New Zealand. Primers for the initial PCR were selected based on prior positive picobirnavirus results in those samples. The amplicons from the gorilla samples and primers used are as follows: 1) Ug15: both F3/F5/R5/R8 and 25F/43R primers; 2) Ug22: both F3/F5/R5/R8 and 25F/43R primers; and 3) Ug43: 25F/43R primers. The amplicons from the cattle samples and primers used are as follows: 1) Ug49: both F3/F5/R5/R8 and 25F/43R primers; 2) Ug55:

both F3/F5/R5/R8 and 23F/24R primers; 3) Ug59: 23F/24R primers; 4) Ug60: F3/F5/R5/R8 primers; 5) NZC01: 25F/43R primers.

3.6.2.2 Cloning protocols

PCR was performed on the cDNA from the samples with the associated primers from the previously positive picobirnavirus identifications; PCR products were separated on gel electrophoresis with the fluorophore RedSafe (iNtRON Biotechnology, South Korea) and 1 kb ladder (ThermoFisher Scientific, MA, USA) on a 1.5% w/v agarose gel (Meridian Bioscience, London, UK) in 0.5X TBE running buffer at 70 mV for 70 min. PCR bands were visualised by UV light using a Gel Doc XR and Quantity One 4.6.2 software program (Bio-Rad Laboratories, California, USA). Distinct bands were cut out and placed in elution buffer to clean the PCR product and to be used for cloning.

The PCR products were cloned into an *E. coli* host (One Shot®TOP10F' Chemically Competent *E. coli*) using the Invitrogen TOPO® TA Cloning Kit for Sequencing as per the protocol for cloning and rapid chemical transformation (ThermoFisher, Massachusetts, USA) with growth of colonies on Luria-Bertani (LB) agar, Isopropyl beta-D-1-thiogalactopyranoside (IPTG) 0.16mM, 5-bromo-4-chloro-3-indolyl-ß-D-galactopyranoside (X-gal) 0.064mg/mL and ampicillin 0.04mg/mL on plates with 40µL and 80µL of samples at 37°C for overnight growth (ThermoFisher Scientific, Massachusetts, USA). Seven to ten colonies were selected the following day (day two) based on white to light-blue growth on plates and grown in 1.5 mL LB broth and 0.075 mg ampicillin at 37°C with shaking at 100 rpm for overnight growth. On the third day, 10 µL of separated colonies grown in LB broth and ampicillin were placed into long-term storage solutions of LB broth, 25% glycerol and ampicillin; the remainder of the separated colonies in broth were used for extracting plasmids using the NucleoSpin® Plasmid EasyPure kit (Macherey-Nagel GmbH & Co. KG, Duren, Germany). Extracted plasmids were tested by PCR and sequenced using the M13 primers (M13 Forward: GTAAAACGACGGCCAG; M13 Reverse: CAGGAAACAGCTATGAC) from the TOPO® TA Cloning Kit (ThermoFisher, Massachusetts, USA). PCR amplification was performed using 1xHOT FIREPol® Blend Master Mix (10mM $MgCl_2$, Solis BioDyne, Estonia), 300 nM of each primer and 1 µL of template made to a total volume of 20 µL with nuclease free water. The thermocycler protocol for the cloned PCR products for picobirnaviruses was: 15 min at 95°C; 40 cycles of 95°C for 40 seconds, 50°C for 40 seconds and 72°C for 90 seconds; and 72°C for 7 minutes. Gel electrophoresis was performed as described previously, positive amplicons were excised from the gel matrix, placed in elution buffer overnight and sent for sequencing to the Massey Genome Service for Sanger sequencing.

3.7 Bioinformatics

3.7.1 Sequence identification and nomenclature

3.7.1.1 Sequence identification

Sanger sequences from PCR amplicons were trimmed (poor sequence quality and vector removal), assembled and aligned, as described below, using Geneious® software (version 10.2.6) [241]. Forward and reverse sequences were pair-end assembled and assembled contigs or unique singletons (forward or reverse reads that could not be assembled either due to low quality and/or being too short in length) were matched to known sequences on the NCBI database using the Basic Local Alignment Search Tool (BLAST) [97, 241]. BLAST was either performed in Geneious® (Biomatters Ltd.) or directly on the NCBI website using the nt/nr database using Basic Local Alignment Search Tool for nucleotides (BLASTn) [97, 241]. The database for all BLASTn searches from NCBI included the default databases (GenBank®, RefSeq, EMBL, DDBJ, PDB) [97, 137, 246, 247] and termed NCBI database for the remainder of the thesis. Details of the contigs or singletons to the highest NCBI database BLASTn nt/nr hit including accession number, length of accession, e-value, maximum bit score, query cover, percent identity and query and subject alignment were documented. The highest hit with an e-value 1e-10 was reported along with the description of the hit, the length of the hit, the maximum bit score, the e-value, the query coverage and the alignment of the query with the subject.

Assembled consensus sequences, including *de novo*, and high-quality reads (singletons) from picobirnavirus-specific primers and cloning were aligned (where applicable) to further assess the diversity among samples. Sequences with >97% similarity were classified as belonging to the same *Picobirnavirus* taxon (considered a duplicate, if from the same sample). Multiple sequence Alignment using Fast Fourier Transform (MAFFT) with default parameters in Geneious® was used for pairwise and multiple alignments (Mafft Multiple Alignment, Version 1.4.0., Biomatters Ltd). MAFFT was selected for alignment of picobirnaviruses based on a recent comparison of alignment methods that found that MAFFT (with the default E-INS-I option, used in this study) was the most accurate for alignment of picobirnavirus sequences [145].

3.7.1.2 Sequence nomenclature

Primary designation of the contig nomenclature was based on the country of origin, Ug for Uganda and NZ for New Zealand, followed by the number designation for the sample for Ug samples, Ug#, and then the host species (H=Human, G=Gorilla, C=Cattle), Ug#_G, or followed by the host species and number designation for NZ samples (C#=Cattle, S#=Sheep, H#=Horse, D#=Dog, CT#=Cat, B#=Bacteria, CR#=*Cryptosporidium*, GD#=*Giardia*). Secondary designation for contigs were based upon whether they were obtained from conventional PCR and Sanger sequencing (_S#), cloning (_C#) or metagenomics (_M#). For the cloned picobirnavirus contigs that were mapped to the metagenomic reads for creation of longer contigs, the secondary designation number was followed by a decimal point and number (_C#.#).

Additional labels utilised for identification of gorilla family group, gender and age in the following chapters (Chapter 4) were included (Gorilla family: R=Rushegura, M=Mubare; gender and age: INF=infant, SA=subadult, BB=blackback male, SB=silverback male, ADF=adult female) [225].

3.7.2 METAGENOMIC SEQUENCE ANALYSES

Double-stranded RNA (dsRNA) virus reads from the NGS were identified using the Orion k-mer approach [234]. dsRNA virus families were separated by sample source and reads were 150 bp or less and designated to a taxon (Table 17). Metagenomic reads were *de novo* assembled separately into multiple, variable-length contigs and the consensus sequence was compared to BLASTn for confirmation of picobirnaviruses. Additionally, some of the conventional PCR generated-sequences from the cloning were used as a map to reference for the metagenomic (Figure 15). The Geneious® assembler was used within Geneious® for *de novo* assemblies and various sensitivity methods were compared for the Sanger sequences, cloned sequences and metagenomic sequences. All sequences were trimmed for low quality (error probability limit of 0.05) and vectors prior to assemblies with annotation and visual inspection of the trimmed regions prior to removal.

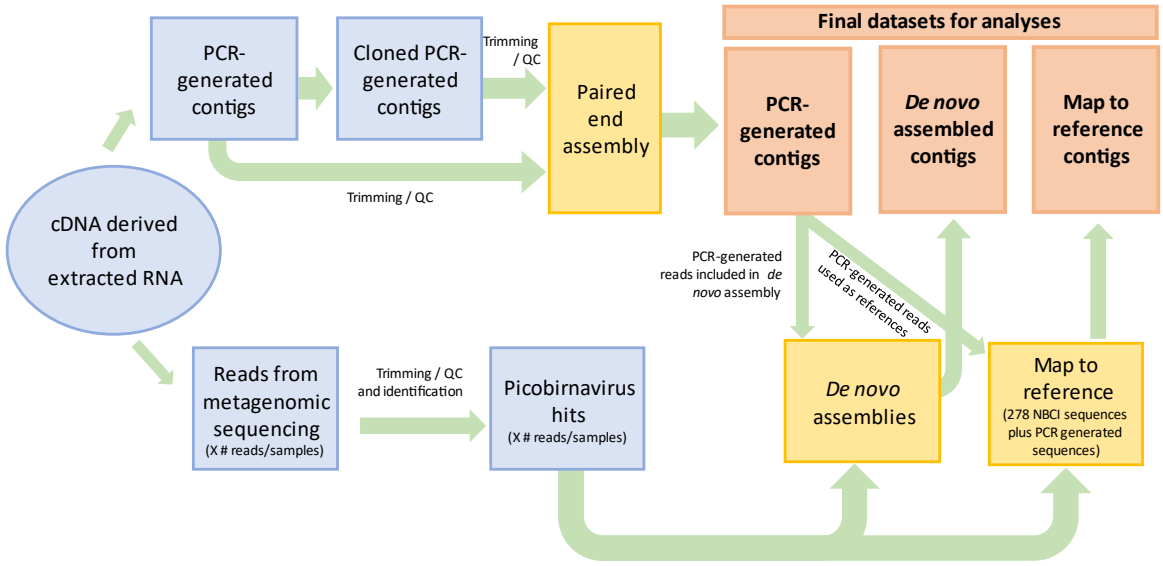


FIGURE 15 WORKFLOW FOR BIOINFORMATICS ON PICOBIRNAVIRUS SEQUENCES

The workflow shows the various bioinformatic methods performed on the picobirnavirus sequences. The blue-coloured boxes indicate the initial sequences from PCR, cloning and NGS; the yellow-coloured boxes indicate the bioinformatic methods performed on Geneious® including alignments and assemblies; the red-coloured boxes indicate the final sequence datasets. The final datasets were used for further phylogenetic analyses. Arrows indicate the flow of the work; some sequences were analysed by multiple methods and various assembly

methods which were then used to create longer contigs. X = the number (#) of contigs per study sample, which was variable for each sample. The first PCR-generated contigs blue box indicates those contigs generate by initial PCR as per Section 3.7.1.1. The final dataset PCR-generated contigs red box indicates those contigs that were not extended by assembly or map to reference and were used for further analyses as is.

3.7.2.1 Contig generation using de novo assembly

Sensitivity methods for assemblies included the default parameters for the lowest-sensitivity setting for the strictest criteria:

- exclusion of circularizing contigs with matching ends;
- allowance for the maximum number of gaps that can be inserted into each read if less than 10% of the size of the overlap between reads;
- twenty-four consecutive bases at minimum needed for a match between two sequences;
- exclusion of words (sub-sequences equal in length to the index word length) repeated more than 100 times;
- allowance for less than or equal to 10% of the size of the overlap between two reads to be mismatched;
- allowance for the maximum gap size of one that can be inserted into the reads;
- allowance for 14 consecutive bases in each sequence to be used as an index for finding all other sequences containing the same consecutive bases;
- a re-analysis threshold of 16; and a maximum ambiguity allowed in word matches of 4 [241].

3.7.2.2 Contig generation using map to reference

Metagenomic reads were mapped to picobirnavirus reference sequences. Picobirnavirus reference sequences were selected from the NCBI nucleotide default database as segment 2 for the *RNA-dependent RNA polymerase* gene if they were >1000 bp in length. Reference sequences were aligned with MAFFT and distances between sequences were evaluated with those sequences with >90% similarities to others deleted to create a database of 278 reference sequences for segment 2 (date of download: 05 May 2020) (See Table 5 at end of this chapter), which accounts for the current known diversity of near-complete *RNA-dependent RNA polymerase* gene picobirnaviruses on NCBI. Any highest-hit BLASTn picobirnavirus accession was included (not deleted), if it met the above criteria. Metagenomic reads were then mapped to the n = 278 segment 2 reference sequences. The NCBI picobirnavirus accessions are listed in Appendix A (Table 14).

Metagenomic reads were also mapped to the cloned picobirnavirus contigs by sample to create longer contigs and confirm identification of both diverse, and similar, picobirnaviruses through high-throughput sequencing and conventional PCR. The Geneious® mapper was used to 'map to reference' and various sensitivity methods were compared on the sequences. Sensitivity methods for map to reference assemblies included the default parameters for the lowest sensitivity setting for the strictest criteria as listed above except with the following different criteria:

- map multiple best matches randomly;

- exclusion of words (sub-sequences equal in length to the index word length) repeated more than eight times;
- an allowance for the maximum gap size of three that can be inserted into the reads [241].

3.7.3 SEQUENCE ANALYSIS

Translation of nucleotide to amino acid sequences of the protein were performed manually for each individual sample with the standard genetic code (translation table_1) in the correct translational based on conserved regions (polymerase motifs and conserved domains-see below and Table 4) or identification of the primer nucleotides in the correct translational frame if no conserved regions were identified and annotated (25F: TGGTGTGGATGTTTC translated into VWMF; 43R: AAGTTCGACCARCAYT translated into KFCQH; 23F: CGGTATGGATGTTTC translated into VMMF; 24R: TACATGGGCTCGCTT translated into HGM/LG/A) [117, 153] (Table 2, Table 4). Additional translation was performed for Chapter 7: Picobirnavirus genogroup I sequences were translated to amino acid sequences by the standard genetic code (SGC) (translation table_1) as noted above, and the alternative genetic code (AGC) (translation table_4), identified as the genetic code for mould protozoan mitochondrial (Table 15, Appendix A) [93, 195]. Stop codons were edited to 'X' as an unknown amino acid in protein sequences in the standard genetic code (see Chapter 7, Section 7.1.2) though this was not indicated when translating in the AGC.

All sequences were also annotated with the *RdRp* polymerase motifs and conserved domains if present (Table 2). Conserved regions included seven conserved motifs and three conserved domains of the *RdRp* per Collier and Da Costa [98, 126, 135]. Segment 2 *RdRp* conserved motifs included polymerase motif G (amino acids (AA): KKSTNSGSPXF at nucleotide (NT) 438, AA 146), motif F (AA: DVKQRVVWMFPX at NT 609, AA 146), motif A (AA: XXICTDFSKFDQH at NT 786 and AA 262), motif B (AA: XXXGXHGMGSGSGGTNXDETLXHRALQYEAAX at NT 960 and AA 320), motif C (AA: NSXCLGDDGXLXY at NT 1080 and AA 360), motif D (AA: XXVXXXYXXHGXEMNXDKQXXS at NT 1140 and AA 380), and motif E (AA: XCTYLRRWHHXX at NT 1356 and AA 452); and also include *RdRp* conserved domain 1 (AA: TDFSKFD at NT 798 and AA 266, domain 2 (AA: SGSGGT at NT 987 and AA 329) and domain 3 (AA: GDD at NT 1095 and AA 365) [135, 195] (Table 2, Table 4). The repetitive (4–10x) capsid domain EXXRXNXXXE was identified and annotated in open reading frame (ORF) 1 on segment 1 [135].

Picobirnavirus genogroup I sequences were initially identified from all sequencing methods. The genogroup I designation was based on: 1) the use of the genogroup I primers (25F/43R) in the PCR of the sample for conventional PCR and cloning; and, 2) the presence of genogroup I primers (25F/43R) within the sequences from conventional PCR and cloning with either the F3/F5/R5/R8 primers or the metagenomic sequences identified as picobirnavirus; or, 3) a top hit match with sequences identified as genogroup I picobirnavirus on the NCBI database through BLASTn.

Picobirnavirus genogroup II sequences were initially identified from all sequencing methods. The genogroup II designation was based on: 1) the use of the genogroup II primers (23F/24R) in the PCR of the sample for conventional PCR and cloning; and, 2) the presence and annotation of genogroup II primers (23F/24R) within the sequences from conventional PCR and cloning with either the F3/F5/R5/R8 primers or the metagenomic sequences identified as picobirnavirus; or, 3) a top-hit match with sequences identified as genogroup II picobirnavirus on the NCBI database through BLASTn.

TABLE 4 NUCLEOTIDE ABBREVIATIONS AND STANDARD AMINO ACID ABBREVIATIONS

Abbreviations	Nucleotide
A	*Adenine*
G	*Guanine*
C	*Cystosine*
T (DNA)	*Thymine*
U (RNA)	*Uracil*
R	*Guanine* or *Adenine (purine)*
Y	*Cytosine* or *Thymine/Uracil (pyrimidine)*
K	*Guanine* or *Thymine/Uracil*
M	*Adenine* or *Cytosine*
W	*Adenine* or *Thymine/Uracil*
S	*Cystosine* or *Guanine*
D	*Adenine* or *Guanine* or *Thymine/Uracil*
B	*Cytosine* or *Guanine* or *Thymine/Uracil*
V	*Adenine* or *Guanine* or *Cytosine*
H	*Adenine* or *Cytosine* or *Thymine/Uracil*
N	*Any base*

Abbreviations	Amino Acid
A	Alanine
C	Cysteine
D	Aspartic acid
E	Glutamic acid
F	Phenylalanine
G	Glycine
H	Histidine
I	Isoleucine
K	Lysine
L	Leucine
M	Methionine
N	Asparagine
P	Proline
Q	Glutamine
R	Arginine
S	Serine
T	Threonine
V	Valine
W	Tryptophan
Y	Tyrosine
U	Selenocysteine
O	Pyrrolysine
B	D (Aspartic acid) or N (Asparagine)
J	I (Isoleucine) or L (Leucine)
Z	E (Glutamic acid) or Q (Glutamine)
X	Unknown/Any amino acid

Left table includes nucleotide abbreviations and right table includes standard amino acid abbreviations. Some "ambiguous" abbreviations are not standardized for amino acids so may utilize multiple abbreviations as listed in the table also [137-139]

All of the picobirnavirus reference sequences (see below and Table 5) as well as the Sanger, clones and metagenomic sequences matching to picobirnaviruses were annotated for seven conserved motifs and three conserved domains on the *RdRp* per Collier et al. 2016, primer binding sites per Rosen et al. (2000) and Anthony et al (2015) and the conserved repeated motif on the capsid segment per Da Costa et al. (2011) [98] as noted above [117, 126, 135, 153].

3.7.4 NEAR-COMPLETE GENE SEQUENCES

De novo assembly of metagenomic reads and map to reference of the metagenomic reads to the clones assembled longer sequences of near-complete picobirnaviruses. Near-complete picobirnavirus

sequences from the study samples were selected based on the following criteria: 1) sequence greater than 900 bp; 2) primers, if present, are in correct order (forward primer before reverse primer) and distance apart; 3) conserved domains in correct configuration (PMG, PMF, PMA/D1, PMB/D2, PMC/D3, PMD, PME) and distances apart; though do not all have to be present; 4) +/- methionine as the first amino acid, which is the start codon.

NCBI *Picobirnavirus* accessions to add to the near-complete picobirnaviruses from this study for phylogenetic analyses were selected based on the following criteria: 1) 30 of 38 complete to near-complete *Picobirnavirus RdRp* sequences from the Knox et al. 2018 paper that contain methionine as the first amino acid; along with 2) 81 of 149 complete picobirnavirus *RdRp* sequences from the NCBI database selecting at least three of every host species or geographical region in the list (date of download: 27 Oct 2020); 3) primers, if present, are in correct order (forward primer before reverse primer) and distance apart; 4) conserved domains in correct configuration (PMG, PMF, PMA/D1, PMB/D2, PMC/D3, PMD, PME) and distances apart though do not all have to be present.

3.7.5 Sequence alignments and phylogenetic analyses

3.7.5.1 Separate reference sequence collection for specific analyses

Most nucleotide and amino acid sequences were trimmed to the approximately 200 bp region of the genogroup I amplicon (approximately 66 amino acids) or to the approximately 370 bp region of the genogroup II amplicon (approximately 123 amino acids). All picobirnavirus accessions were collected from NCBI and included the default databases (GenBank®, RefSeq, EMBL, DDBJ, PDB) [97, 137, 246, 247] and termed NCBI database for the remainder of the thesis.

For Chapter 4, separate multiple alignments (see below) of trimmed picobirnavirus *RdRp* genogroup I and genogroup II sequences from the study cohort were performed. Alignments were based on amino acid sequences of approximately 65 (GI; or 195 bp) or 120 (GII; or 360 bp) amino acids aligned between the forward and reverse primers detailed in Chapter 3, Sections 3.6.1.1 and 3.7.3. Additional trimming was performed with the trimAl® program using the Gappyout and no gaps selections and sequences were then imported back into Geneious® for multiple alignment and phylogenetic tree creation as described below. Top-hit BLAST accessions of the sample genogroup I or genogroup II picobirnaviruses were included in the trimmed versus untrimmed analyses.

For Chapter 5, in order to compare the host and geographic structure in the picobirnaviruses found in my samples with a broader dataset, I downloaded picobirnavirus sequences from NCBI using the search term “picobirnavirus AND RNA-dependent RNA polymerase” on the NCBI database filtered for a custom range of sequence length 1000 to 3000 bp; the exclusion of similar picobirnavirus sequences (described in Section 3.7.2.2) resulted in a final dataset of n = 278 picobirnavirus *RdRp* accessions that

was then representative of the total known *Picobirnavirus RdRp* diversity to date including any highest hit picobirnavirus BLASTn if indicated (date of download: 05 May 2020).

3.7.5.2 Alignments

Sequence alignment and phylogenetic tree building were performed for phylogenetic analyses. MAFFT was selected for alignment of picobirnaviruses based on a recent comparison of alignment methods (CLUSTAL, MUSCLE, MAFFT) that found that MAFFT (with the default E-INS-I option) was the most accurate for alignment of picobirnavirus sequences [145]. Alignments were done with MAFFT (Mafft Multiple Alignment, Version 1.4.0., Biomatters Ltd) on Geneious® [241] and visually inspected for gaps. Gaps in the alignment greater than one amino acid were inspected to identify if they occurred due to multiple sequence alterations or from a single sequence. Alterations causing gaps from a single accession resulted in exclusion of that sequence; if gaps were due to multiple sequences with repeated and consistent alterations, the sequences were kept in the alignment. If stop codons were identified in the amino acid translation in the correct translation frame (occurring only in 2 samples), stop codons were edited into 'X' for 'unknown amino acid'. Percent identity of the alignment was reported from Geneious® and stated in the results for either pairwise alignments for two sequences aligned or average pairwise sequence similarity for multiple alignments as indicated.

3.7.5.3 Phylogenetic tree construction

Phylogenetic trees were constructed with PhyML, version 3.0 (ATGC, Montpellier Bioinformatics Platform) [248] which builds trees using maximum-likelihood principles, comprising a distance-based method with Neighbour-Joining algorithm and optimised with Nearest Neighbor Interchanges [248, 249]. Automatic model selection by Smart Model Selection (SMS) was used within PhyML with the Akaike Information Criterion (AIC) selected [250]. Tree searching in PhyML selected a starting tree (BIONJ), the type of tree improvement was Nearest Neighbor Interchange (NNI) and "no" random starting tree. Maximum likelihood is the most appropriate tree-building method as evidenced by Perez et al. (2021), which found that maximum likelihood resulted in the optimal topology for phylogenetic analyses with picobirnaviruses as compared to other commonly used tree-building methods [145]. Phylogenetic trees were rooted if an outgroup from a different genogroup was included (Accession for genogroup II outgroup for genogroup I tree: AF246940_Human PBV_USA; accession for genogroup I outgroup for genogroup II tree: AB186898_Human PBV_Thailand). Branch support was performed by the Fast-likelihood-based method of aLRT SH-like [251].

3.7.5.4 Phylogenetic tree annotations and further analyses

Phylogenetic trees were annotated with colour-coding in Evolview online editor [252-254] and/or manually in Inkscape (https://inkscape.org/).

Co-phylogenetic analyses, typically utilised to compare host-parasite relationships [255], were utilised instead to compare congruence based on nucleotide and amino acid phylogeny, trimmed and untrimmed phylogeny of picobirnaviruses or picobirnaviruses translated with different genetic codes. Co-phylogenetic trees were created in R [256] for comparison. Phylogenetic trees were exported as newick files and co-phylogenetic comparisons were done in RStudio® [256-267]. The use of the package, PACo or Procrustes Approach to Cophylogeny, was used to test "the congruence between two given phylogenies by using a permutation approach to test for significance" between phylogenetic trees [257, 261, 268].

3.7.5.5 Additional plots and ORF and RBS identification

Heatmaps of the pairwise identity between sequences and phylogenetic trees were annotated with colour-coding in Evolview online editor [252-254]. Further column plots were also created in Evolview for identification of sample and host species. Additional heatmaps were created in RStudio [256].

ORFs were identified in Geneious® for the genogroup I picobirnavirus sequences from the study samples using both the SGC and AGC as indicated.

The near-complete picobirnaviruses were evaluated for evidence of ribosomal binding sites/motifs (RBS) on Geneious® -4 to -18 nucleotides upstream of the start codon, methionine, in 4-mer (AGGA/GGAG/GAGG), 5-mer (AGGAG/GGAGG) or 6-mer (AGGAGG) windows [133, 203].

TABLE 5 SAMPLE AND NCBI PICOBIRNAVIRUS SEQUENCES ALONG WITH ANALYSES USED IN THE SUBSEQUENT CHAPTERS

Chapter/Section	Sample PBVs Used	Primers Used	Accessions PBVs Used	Analyses
4: Structure and characteristics	NT v AA Cophylo: GI: 44, GII: 15	GI: 25F/43R[1] GII: 23F/24R[1] Any genogroup: F3/F5/R5/R8[2] Designed primer: 1358R	None	Multiple alignment (Geneious)[3]; Phylogenetic tree (PhyML)[4]; Co-phylogenetic trees (R)[5]; trimAl[6]; Additional annotations (Evolview[7] and Inkscape)
	trimAl Cophylo: GI: 47, GII: 16		trimAl Cophylo: GI: 81, GII: 40 Top PBV accession hit for each study PBV sample (date of download: 08 Sept 2020)	
	Sample trees: GI: 51 + GII root GII: 18 + GI root NC/C: 12		None	

5: Host and geography	GI: 44 GII: 16 NC/C: 17 All trimmed except labelled as NC/C PBVs	GI: 25F/43R[1] GII: 23F/24R[1] Any genogroup: F3/F5/R5/R8[2] Designed primer: 1358R	278 *RdRp* PBV sequences based on criteria in Chp 3, Section 3.7.5.1 (date of download: 05 May 2020) GI: 82 GII: 43 NC/C: 81	Multiple alignment (Geneious)[3]; Phylogenetic tree (PhyML)[4]; Additional annotations (Evolview[7] and Inkscape)
6: Diversity and WHD	Near-complete PBVs untrimmed: Cloned PBVs: Total 28: Ug15, Ug22, Ug43, Ug49, Ug55, Ug59, Ug60 Additional PBVs through initial PCR and metagenomics also with above samples	GI: 25F/43R[1] GII: 23F/24R[1] Any genogroup: F3/F5/R5/R8[2] Designed primer: 1358R	None	Multiple alignment (Geneious)[3]; Phylogenetic tree (PhyML)[4]; Heatmaps (R and Evolview)[7]; Additional annotations (Evolview)[7]
7: Actual host?	SGC v AGC: 44 Near-complete PBVs untrimmed for ORF ID and RBS: 18	GI: 25F/43R[1] GII: 23F/24R[1] Any genogroup: F3/F5/R5/R8[2]	None	Multiple alignment (Geneious)[3]; Phylogenetic tree (PhyML)[4]; Co-phylogenetic tree (R)[5]; ORF identification (Geneious)[3]

[1] [117]; [2] [153]; [3] [241]; [4] [248], [250], [251], [269]; [5] [256], [263], [260], [264], [261], [265, 266], [267], [258, 259, 268]; [6] [270], [271]; [7] [252], [253], [254]. PBV: Picobirnavirus; GI: Genogroup I; GII Genogroup II; Ug Uganda; WHD: Within host diversity; ORF: open reading frame; ID: identification; RBS: ribosomal binding motif; PCR: polymerase chain reaction; *RdRp*: RNA-dependent RNA polymerase; NC: near-complete sequences; R: RStudio

CHAPTER 4: PICOBIRNAVIRUS CHARACTERISTICS, SEQUENCE PROFILE AND PHYLOGENY

4.1 RESULTS

Due to the high diversity of picobirnaviruses expected in the samples, multiple laboratory approaches were used to identify and confirm the detection of picobirnaviruses within the samples. Conventional PCR and Sanger sequencing were used with different primer sets to identify the presence of picobirnaviruses as noted in Section 3.6.1 along with metagenomic methods as described in Section 3.7.2. Also, due to the initial Sanger sequences showing multiple peaks on the chromatograms, potentially indicating multiple picobirnaviruses in the same sample, cloning was performed (see Chapter 6) to identify the presence of multiple picobirnaviruses.

4.1.1 *PICOBIRNAVIRUS* SEQUENCE IDENTIFICATION AND CHARACTERISTICS

A total of 102 picobirnaviruses were identified in the Ugandan and New Zealand samples using sequences from segment 2, *RdRp*. Seven picobirnaviruses were from the three clinically-unwell Ugandan humans, 40 picobirnaviruses were identified from the 44 habituated Ugandan gorilla samples, 53 picobirnaviruses were identified from the 16 domestic Ugandan cattle samples and one picobirnavirus each was identified in the New Zealand cattle and purified *Cryptosporidium* oocyst samples (to be discussed in Chapter 7 in more detail) (Table 6).

Of the 102 picobirnaviruses identified, 59 (58%) were from the initial conventional PCR Sanger sequencing, 21 (20.5%) were from cloning and 33 (32%) were from the metagenomic sequences with some duplicated picobirnaviruses found in multiple methods (Table 6; Appendix B, Table 16). The median sequence length from all types was 347 bp (1st quartile 196 bp; 3rd quartile 742 bp) with the minimum sequence length of 126 bp and the maximum sequence length of 1887 bp. The median/mean guanine-cytosine (GC%) content of the picobirnavirus sequences was 44% (min 36.8%, max 51.2%). Seventy percent of the picobirnaviruses contained conserved regions in the sequences, many sequences with multiple polymerase motifs and conserved domains. Polymerase motifs were identified in the picobirnaviruses to varying extents with the most common polymerase motifs (PM) identified as PMF (25 times) and PMA (22 times); PMB (10 times), PMC (10 times) and PMG (8 times) were identified uncommonly and PMD (4 times) and PME (2 times) identified rarely. All conserved domains (CD) were identified commonly in the picobirnaviruses; CD1 was identified most commonly at 51 times, CD3 the second most common at 30 times and CD2 the least common at 19 times.

TABLE 6 SUMMARY OF PICOBIRNAVIRUS *RDRP* SEQUENCES IDENTIFIED FROM RT-PCR AND METAGENOMICS FROM UGANDA AND NEW ZEALAND

Sample	Species	Methods employed	Genogroups (if known)	Sequence lengths	Total PBVs per sample	Highest PBV BLAST match (description)
Ug01	Human	S	GI	181–183	2	Human, raw sewage
Ug02	Human	S	GI	182–207	2	Bovine, Himalayan Goral
Ug03	Human	S	GI, GII	211–380	3	Porcine, Mongoose
Ug04	Gorilla	S	GI	182	1	Wastewater
Ug06	Gorilla	S, M	GI	378–798	2	Human
Ug07	Gorilla	M	GI	1009	1	Marmot
Ug08	Gorilla	S	GI	157–158	2	Porcine, Bovine
Ug10	Gorilla	S	GI	196	1	Human
Ug14	Gorilla	S	GI	188	1	Macaque
Ug15	Gorilla	S, M, C	GI	206–1887	5	Chicken, Wastewater, Equ3 PBV
Ug16	Gorilla	S, M	GI	447–473	2	Human, Microtus
Ug17	Gorilla	S, M	GI	212–654	4	Human, Wastewater, Microtus, Marmot
Ug22	Gorilla	M, C	GI	204–1427	4	Human, Wastewater, Fox
Ug23	Gorilla	S	GI	188	1	Bovine
Ug25	Gorilla	M	GI	742	1	Wolf
Ug27	Gorilla	S	GI	179	1	Bovine
Ug28	Gorilla	M	GII	1035	1	Human
Ug39	Gorilla	S	GI	182	1	Ovine
Ug41	Gorilla	S	GI	185	1	Wastewater
Ug42	Gorilla	S	GI	153	1	Avian
Ug43	Gorilla	S, M, C	GI	185–732	6	Porcine, Macaque, Wastewater, Bovine, Gorilla, Human
Ug44	Gorilla	S	GI	187	1	Wastewater
Ug45	Gorilla	M	GI	315–1131	2	Chicken
Ug47	Gorilla	S	GI	1671	1	Bovine
Ug48	Cattle	M	GI	1027	1	Porcine
Ug49	Cattle	S, C, M	GI	206–1525	7	Ovine, Wastewater, Bovine, Human, Dromedary, Chicken
Ug50	Cattle	M	GI	392–1063	3	Bovine, Simian, Monkey
Ug51	Cattle	S	GI	126	1	Horse
Ug52	Cattle	S	GI	197	1	Bovine
Ug53	Cattle	S, M	GI	187–1223	2	Horse, Mongoose
Ug54	Cattle	S, M	GI, GII	173–557	6	Porcine, Marmot, Chicken
Ug55	Cattle	S, M, C	GI, GII	158–1363	6	Ovine, Bovine, Porcine, Human
Ug56	Cattle	S	GI, GII	180–760	2	Porcine, Bovine
Ug57	Cattle	S, M	GI	188–558	1	Human, Chicken
Ug58	Cattle	S, M	GI	148–679	3	Bovine, Porcine
Ug59	Cattle	S, M, C	GI, GII	176–704	3	Wastewater, Bovine
Ug60	Cattle	S, M, C	GI	171–942	9	Horse, Tas devil, Genet, Bovine, Dromedary, Human, Marmot, Chicken
Ug61	Cattle	S	GI, GII	196–344	2	Goat, Deer
Ug62	Cattle	S, M	GI, GII	347–1688	4	Human, Porcine, Simian
Ug63	Cattle	S, M	GI	165–1472	2	Dromedary
NZC01	Cattle	S, C	GI	203	1	Goat
NZCR03	Cryptosporidium	S	GII	148	1	Human

Ug: Uganda; NZ: New Zealand. Method: S=Sanger sequencing from rt-PCR; C=Cloned and then Sanger sequencing from rt-PCR; M=Metagenomics. Genogroups: GI for genogroup I; GII for genogroup II. Sequence length is the range of base pairs of all of the picobirnavirus sequences identified.

4.1.2 Evaluation of methodologies

All primer sets, both previously developed [117, 153], and the primers designed in this study, were successful in identifying picobirnaviruses from the samples evaluated. The use of the genogroup I primers (25F/43R) resulted in the identification of a higher number of picobirnaviruses; the four-primer set (F3/F5/R5/R8) also identified picobirnaviruses, though less than the genogroup I primers [117, 153]. The genogroup II primers (23F/24R) also resulted in the identification of picobirnaviruses from the samples though less often than with the genogroup I primers [117]. Seventy-two picobirnaviruses (68 not including duplicates) were identified with the genogroup I primer set and 10 picobirnaviruses (10 not including duplicates) were identified with the genogroup II primer set from the samples. Only one picobirnavirus sequence was detected using the primers designed in this study. Most of the separate forward or reverse sequences were pair-end assembled (77%) while less than a quarter (23%) were only from the forward or reverse primer. Seventy-three (71.5%) of the picobirnaviruses were characterized or designated as genogroup I and 11 (10.7%) as genogroup II based on the criteria in Section 3.7.3; the remainder of the picobirnaviruses were not designated within genogroup I or genogroup II.

The dataset of the number of reads for the metagenomic sequencing is included in Table 17. The table shows the number of dsRNA virus reads per sample along with the number of designated picobirnavirus reads in each sample.

4.1.3 Sequence profiling with BLAST

All picobirnavirus sequences from the study samples were compared using BLASTn on the nr/nt database collection at NCBI (accessed September 22, 2020) for the closest match to known picobirnavirus sequences. Most query picobirnaviruses had BLAST matches against multiple picobirnaviruses in the database with most having a greater e-value than 1×10^{-10}, with the highest match for each sample documented in Table 7. The most common top BLAST hit was the accession number KY120170, a bovine picobirnavirus that was the most similar hit to nine different picobirnaviruses collected from three gorilla and six cattle Ugandan samples. Table 7 contains only the most common top BLAST accession hit but I also looked at the top 20 hits for the samples which then included approximately 450 sample contig hits of different lengths to KY120170 (Figure 16 and Figure 17). Figure 16 and Figure 17 shows the sample contig matches, the overlap with the accession and the bit score and percent identity, respectively, associated with that match. The figure also shows that contigs were sometimes split and matched, overlapping various lengths of the accession. Multiple additional bovine picobirnavirus accessions were also the highest match for study samples, some with multiple study samples matched to the accession (Table 7). Many different porcine picobirnavirus accessions were most similar to the picobirnaviruses from the samples (Table 7); additionally, various

porcine accessions were similar to cattle Ugandan samples not listed in the table (MK378865, MK378854, KP868555, KP984805, GU230540, GU230525) (Appendix B, Table 18). Many human picobirnavirus accessions were similar to sample picobirnaviruses (Table 7) with some additional human picobirnavirus accessions (MH933822, MH933831, MH933824) similar to gorilla picobirnavirus samples and human picobirnavirus accessions (MH933806, MH933821, KR827415, KR827416) similar to Ugandan cattle samples (Appendix B, Table 18). In addition, a human picobirnavirus accession (AB214978) was most similar to a *Cryptosporidium* picobirnavirus New Zealand sample (Appendix B, Table 18). All three species/hosts sampled from Uganda matched picobirnavirus accessions from raw sewage or wastewater picobirnaviruses including a human sample (to EU938815), gorilla samples (KJ135823) and to a cattle sample (KJ135919) (Table 7 and Table 18). Many other different host picobirnavirus accessions were most similarly matched to the Ugandan and New Zealand picobirnavirus samples including, but not limited to chicken picobirnaviruses, marmot picobirnaviruses, horse picobirnaviruses, vole/microtus picobirnaviruses and sheep picobirnaviruses (see Table 7, Table 16 and Table 18 for details). Median maximum bit scores for all hits was 239 (range 45.5 – 1468; quartiles 159 – 420); median query coverage was 92% (range 17 – 100%; quartiles 75 – 98%); median and mean percent identity to matches was 81% (range 67–98%).

TABLE 7 TABLE OF THE MOST COMMON/HIGHEST MATCHED PICOBIRNAVIRUSES ON BLAST

Accession #	Accession description	# top hits from study samples	Study sequences
KY120170	Bovine (Cattle) PBV	9	Ug23_G_S1, Ug27_G_S1, Ug47_G_S1, Ug49_C_C3.1, Ug49_C_M1, Ug52_C_S1, Ug55_C_S2, Ug55_C_M1, Ug58_C_M1
KJ135902	Wastewater PBV	4	Ug17_G_S1, Ug22_G_C2, Ug43_G_C2, Ug49_C_C2
JF755420	Microtus (Vole) PBV	4	Ug16_G_M1, Ug17_G_M1, Ug22_G_C2.1, Ug22_G_M1
KY120178	Bovine (Cattle) PBV	3	Ug56_C_S4 Ug59_C_S2 Ug59_C_C1.1
KY120176	Bovine (Cattle) PBV	3	Ug08_G_S3, Ug50_C_M1, Ug60_C_M2
KJ135922	Wastewater PBV	3	Ug41_G_S1, Ug43_G_C1, Ug59_C_C2
MK378865	Porcine (Pig) PBV	3	Ug48_C_M1, Ug51_C_S1, Ug54_C_S3
KC846784	Porcine (Pig) PBV	2	Ug03_H_S2, Ug08_G_S1
MK378853	Porcine (Pig) PBV	2	Ug43_G_S1, Ug54_C_M2
KT720491	Human PBV	2	Ug01_H_S2, Ug10_G_S1
KJ135796	Wastewater PBV	2	Ug04_G_S1, Ug15_G_S1

MH933835	Human PBV	2	Ug06_G_S1, Ug06_G_M1
MH425585	Chicken PBV	2	Ug15_G_S2, Ug15_G_C1.1
MH933803	Human PBV	2	Ug16_G_S1, Ug17_G_S2
KX964659	Bovine (Cattle) PBV	1	Ug02_H_S1
KY120175	Bovine (Cattle) PBV	1	Ug43_G_M1
KY120177	Bovine (Cattle) PBV	1	Ug58_C_M2
MN196313	Bovine (Cattle) PBV	1	Ug49_C_C5
MN196315	Bovine (Cattle) PBV	1	Ug55_C_C3
KC841460	Porcine (Pig) PBV	1	Ug03_H_S3

Abbreviations: Ug=Uganda, S#=Sanger Sequence, C=Cloned Sequence, M#=Metagenomic Sequence, PBV=Picobirnavirus.

FIGURE 16 PLOT OF CONTIG HIT FROM UGANDAN SAMPLES WITH THE MOST COMMON ACCESSION, KY120170, BASED ON BIT SCORE

A selection of Ugandan samples that matched with KY120170, a bovine picobirnavirus, in the top 20 hits within each sample. Contig samples are on the y-axis and the match overlap with KY120170 by base pairs is on the x-axis. The legend and colour-coding shows the bit score of the contig with the accession with the darker colour, the higher the bit score.

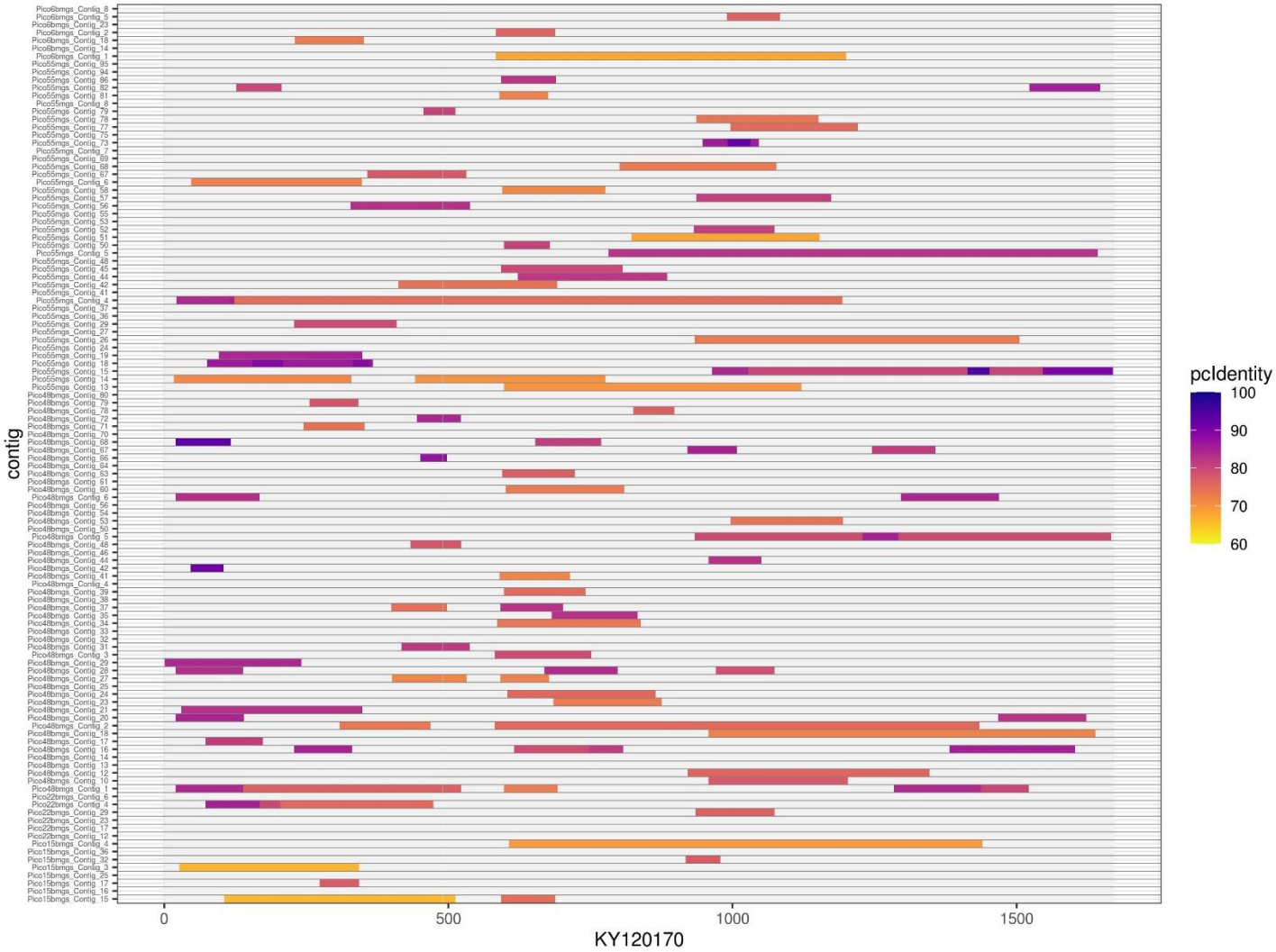


FIGURE 17 PLOT OF CONTIG HIT FROM UGANDAN SAMPLES WITH THE MOST COMMON ACCESSION, KY120170, BASED ON PERCENT IDENTITY

A selection of Ugandan samples that matched with KY120170, a bovine picobirnavirus, in the top 20 hits within each sample. Contig samples are on the y-axis and the match overlap with KY120170 by base pairs is on the x-axis. The legend and colour-coding shows the percent identity of the contig with the accession with the darker colour, the higher the percent identity or similarity.

4.1.4 Sequence profiling with BLAST and multiple alignments

Of the total picobirnaviruses identified, 71 paired-end and *de novo* assembled picobirnaviruses were submitted to GenBank® (GenBank® accession numbers MZ560630–MZ560700). Appropriate translational frames were selected along with the translation table based on Chapter 3, Section 3.7.3. Ten of the 76 sequences were translated with the alternative genetic code (AGC, translation table_4) with the remainder (66 sequences) translated with the standard genetic code (SGC, translation table_1). A small subset (7) of the sequences were identified as having insertions and/or deletions (indels) resulting in frameshifts of the protein sequence. Further analysis of these sequences with the CDS alignment tool on BLASTn from NCBI [272] identified sites with apparent indels resulting in frame-shifting, such as with Ug10_G_S1 though this was indeterminate in others, such as Ug49_C_M2, presented in the following paragraphs.

Figure 18 shows the position of an apparent insertion at position 70 and 77 in the nucleotide sequence of Ug10_G_S1 (query sequence). Examination of the nucleotide alignment (made up of all genogroup I study sequences from Figure 19) shows a similar result, with insertions of two bases at position 142 (Figure 19) and a potential reading frame shift in the amino acid alignment resulting in a stop codon using the SGC (Figure 18, Figure 20). However, examination of the Sanger sequence chromatograms at these sites did not show any clear evidence for insertions or deletions (Figure 21). Nonetheless, due to the discrepancy between the nucleotide alignment and translated amino acid sequences (Figure 19 and Figure 20), I elected to remove sequence Ug10_G_S1 from subsequent analyses.

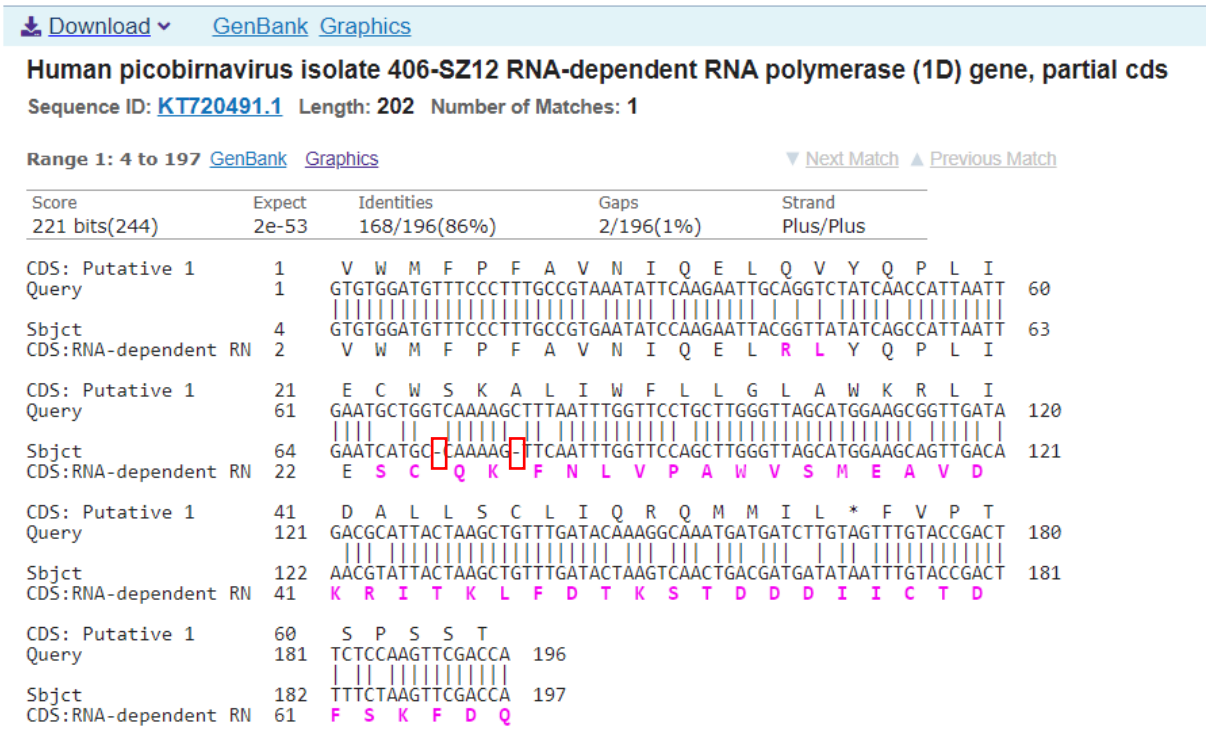


FIGURE 18 TOP HIT PICOBIRNAVIRUS ACCESSIONS FROM BLASTN OF UG10_G_S1 WITH THE CDS ALIGNMENT TOOL

The query is Ug10_G_S1 while the subject is the accession. The pink/purple amino acids below the nucleotides for the subject shows disagreements in the amino acid sequence potentially due to frameshifting from an indel in the query.

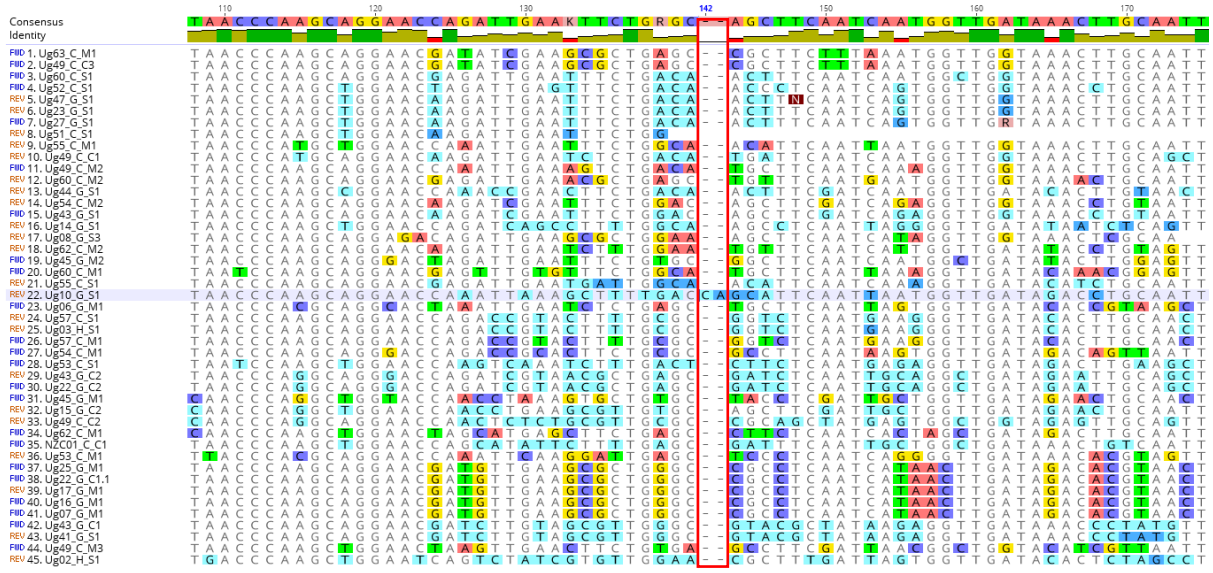


FIGURE 19 MULTIPLE NUCLEOTIDE ALIGNMENT OF THE GENOGROUP I PICOBIRNAVIRUSES FROM THE STUDY

The multiple alignment shows a gap at nucleotide position 142 outlined in red due to the possible insertion of two nucleotides from ug10_G_S1.

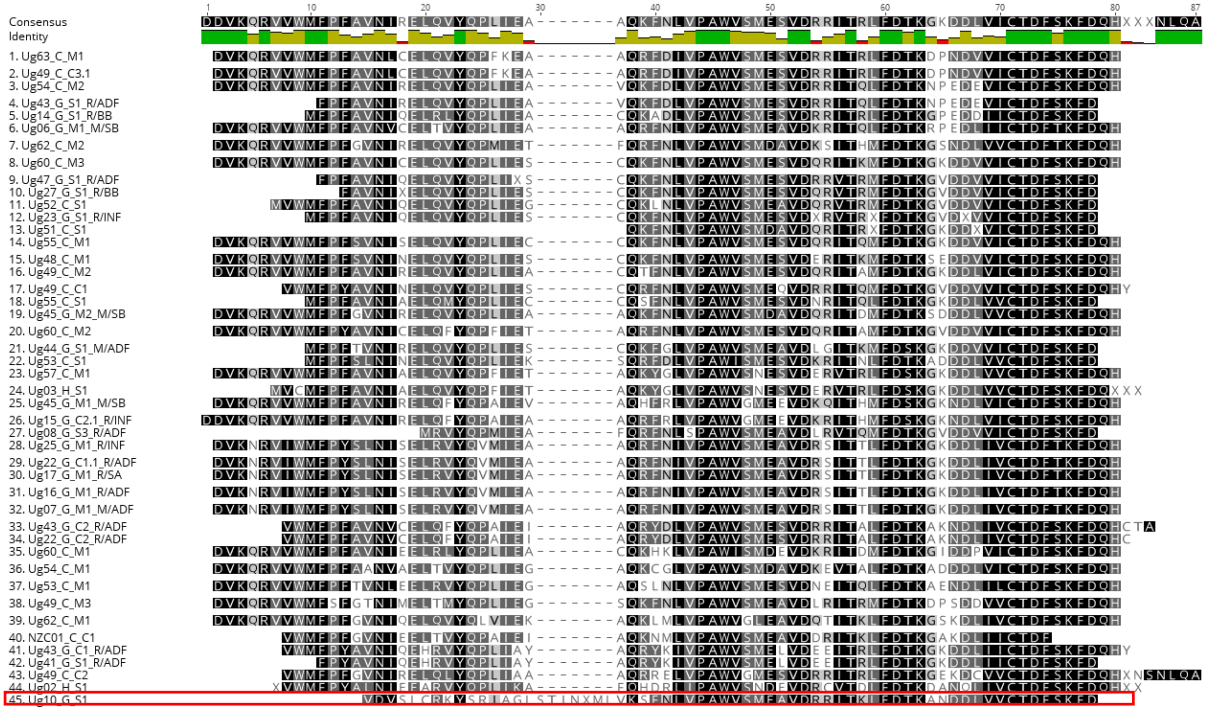


FIGURE 20 MULTIPLE ALIGNMENT OF 45 GENOGROUP I PICOBIRNAVIRUS SEQUENCES WITH UG10

The multiple alignment shows the gap created by the attempted alignment with Ug10_G_S1 in the other aligned sequences. Ug10 at the bottom of the alignment shows the least similarity with the other genogroup I picobirnaviruses in the alignment.

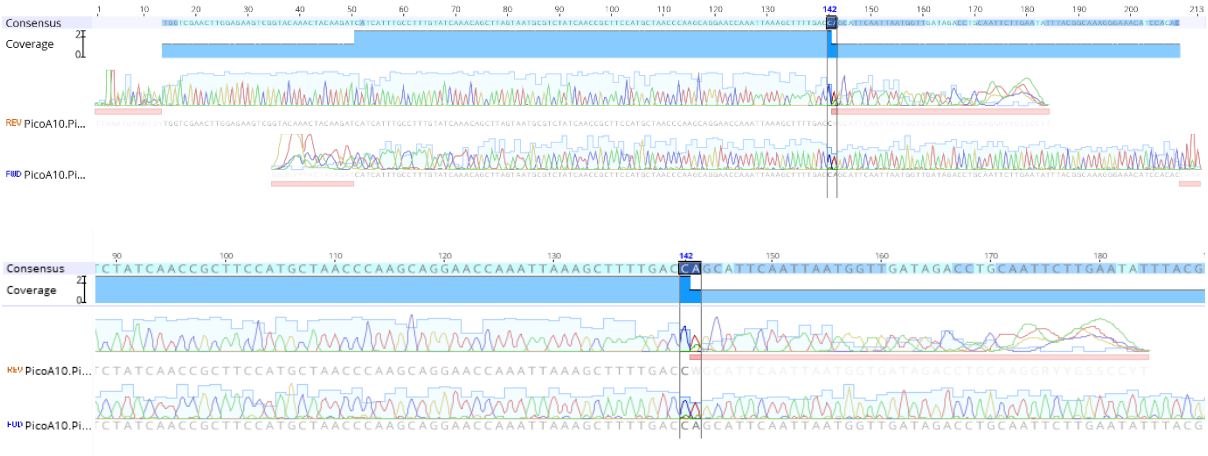


FIGURE 21 UG10_G_S1 PAIRED-END ASSEMBLED CHROMATOGRAM

The top picture contains the complete paired-end sequence with the potential indel region in the rectangle at nucleotide 142. The bottom picture is the region zoomed in for evaluation of the peak quality and agreement between the paired-end sequences at the potential indel site.

Sequence Ug49_C_M2 was similarly identified as having potential indels. Figure 22 shows the position of apparent deletions at position 362 and 369 in the nucleotide sequence of Ug49_C_M2 (query sequence in Figure 22). Examination of the nucleotide alignment and amino acid alignments show high similarity with other picobirnaviruses and no gaps in the alignments (Figure 23). The nucleotide and amino acid alignments did not include the region of interest as the sequences were trimmed to the approximately 200 bp region of the *RdRp* gene. The potential region was approximately 30 nucleotides or 10 amino acids from the end of the sequence (at position 280 on Figure 24). Examination of these positions in the Ug49_C_M2 assembly shows no evidence for insertion or deletions and a reasonable read depth (>20) (Figure 24). Due to the lack of evidence for indels in the raw data and because these issues do not appear to result in a sustained reading frame shift, I have elected to retain this sequence in further analyses. Additional sequences were similarly affected (Ug22_G_C2.1, Ug55_C_M1, Ug60_C_M1) but retained in the analyses but not uploaded to NCBI GenBank®.

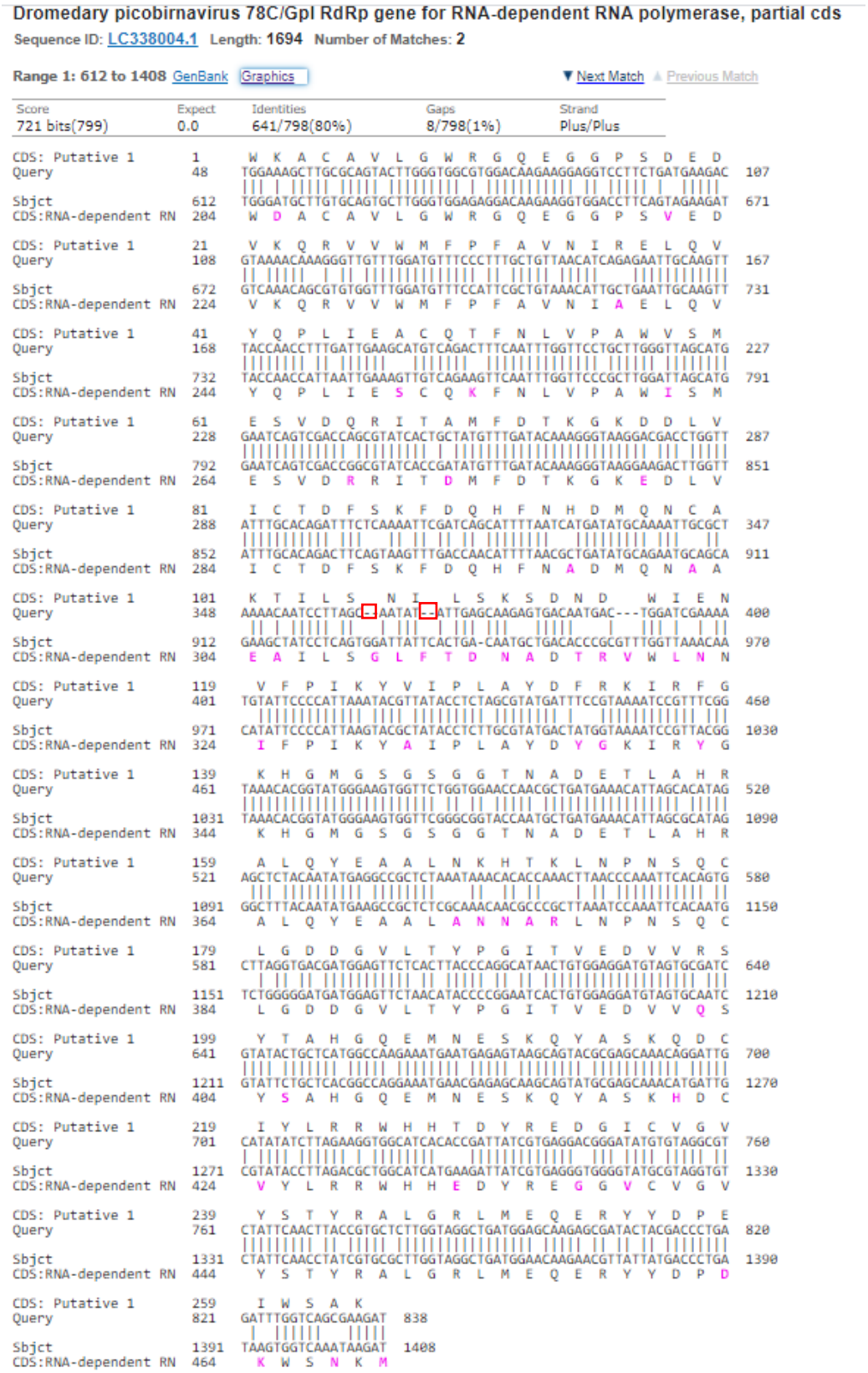


FIGURE 22 TOP HIT PICOBIRNAVIRUS ACCESSIONS FROM BLASTN OF UG49_C_M2 WITH THE CDS ALIGNMENT TOOL

The query is Ug49_C_M2 while the subject is the accession. The pink/purple amino acids below the nucleotides for the subject shows alterations in the amino acid sequence.

FIGURE 23 MULTIPLE ALIGNMENT OF 44 GENOGROUP I PICOBIRNAVIRUS SEQUENCES WITH UG49

The upper picture shows the nucleotide alignment with Ug49 outlined in red while the lower picture shows the protein alignment with Ug49 outlined in red. The multiple alignment shows high similarity of Ug49_C_M2 with other picobirnaviruses from the study and no gaps in the alignments. Ug10_G_S1 was excluded from these alignments.

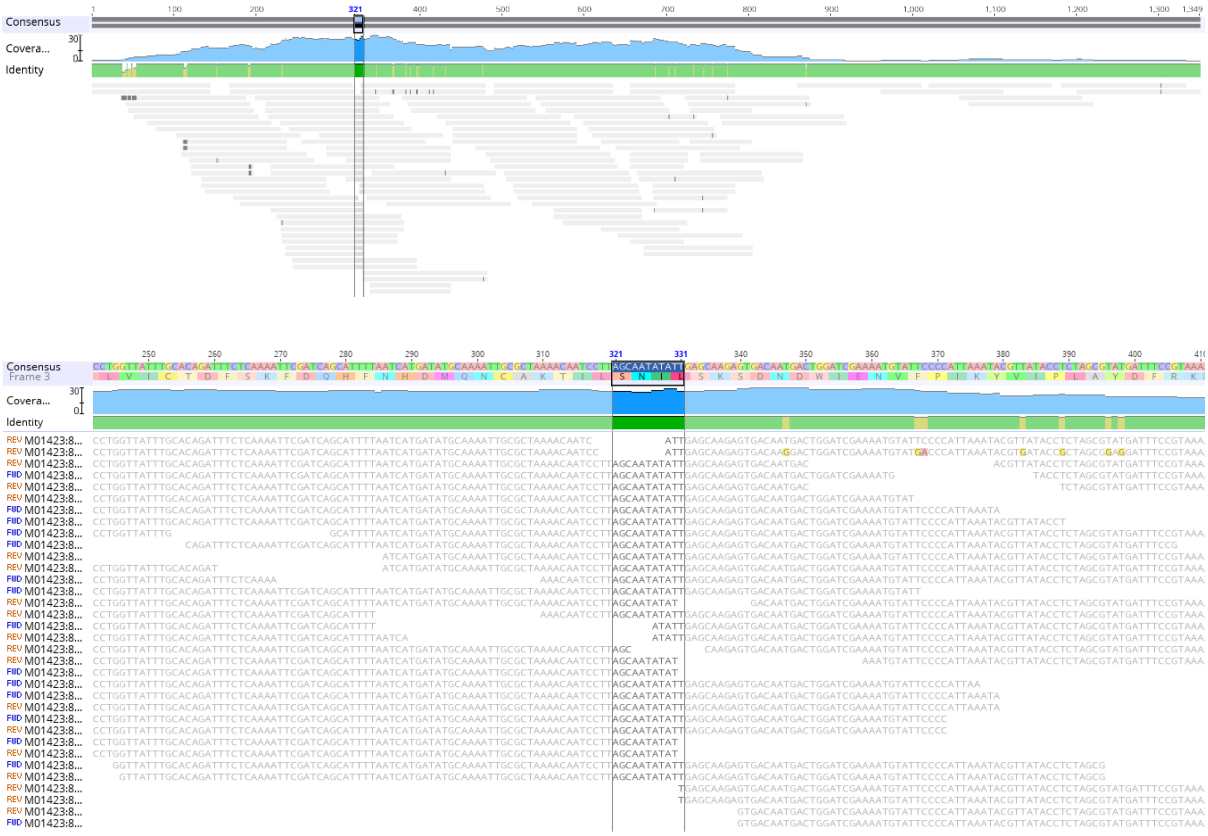


FIGURE 24 UG49_C_M2 *DE NOVO* ASSEMBLED METAGENOMIC SEQUENCE

The top picture contains the complete *de novo* sequence for the metagenomic data for Ug49_C_M2 sequence with the potential indel region in the rectangle. The bottom picture is the region zoomed in for evaluation of the read depth and nucleotide agreement in the reads peak quality and agreement between the reads at the potential indel site.

4.1.5 SEQUENCE PROFILING BY PHYLOGENETIC RELATIONSHIPS

4.1.5.1 NUCLEOTIDE VERSUS AMINO ACID SEQUENCES

Phylogenetic tree comparisons (co-phylogenies) of the picobirnavirus *RdRp* sequences from this study were performed to compare the use of nucleotide and/or amino acid sequences to assess picobirnavirus diversity patterns. Co-phylogenetic comparisons for genogroup I include: n = 44 picobirnavirus sequences from Uganda and New Zealand with n = 2 human samples, n = 20 gorilla samples and n = 22 cattle samples (Figure 25). Co-phylogenetic comparisons for genogroup II include n = 15 picobirnavirus sequences from Uganda with: n = 1 human sample, n = 5 gorilla samples and n = 9 cattle samples (Figure 25).

The co-phylogenetic comparison of genogroup I picobirnaviruses demonstrates that the use of nucleotide as compared to amino acid sequences does alter the overall tree topology. Beyond single sequence shifts of samples between nucleotide and amino acid tree topologies, there were shifts of groups within a clade, suggesting changes in tree topology though with consistency in some of the

groups remaining together (see Figure 25, designation grey brackets). The overall nucleotide percent identity was at 64.2% in the multiple alignment of the genogroup I nucleotide sequences. The amino acid genogroup I tree displays shorter branch lengths than the nucleotide sequences, indicative of less variation or diversity of the translated picobirnaviruses. The amino acid percent identity was higher at 73.4% in the multiple alignment of the genogroup I amino acid sequences. The amino acid sequences suggest more clustering by hosts than the nucleotide-based phylogenetic tree though the clustering is not strong. Analysis of the congruence of the co-phylogenies, despite the above changes in tree topologies between nucleotide and amino acid sequences did show that there was a statistically significant (p-value = <0.01) association between the two trees that suggests they are congruent [257, 268].

The co-phylogenetic comparison of genogroup II picobirnaviruses demonstrates that the use of nucleotide, as compared to amino acid sequences, does not alter the tree topology or clustering of the sequences. Longer branch lengths are noted on the nucleotide sequences of genogroup II picobirnaviruses as compared to the amino acid sequences, as noted also in the genogroup I picobirnaviruses, indicative of less variation or diversity of the translated picobirnaviruses. The nucleotide and amino acid percent identities were similar with the nucleotide percent identity of 55.6% and the amino acid percent identity was 54.5% in the multiple alignment of the genogroup II amino acid sequences. The genogroup II phylogenetic trees display clustering by host though fewer sample numbers are included in the genogroup II picobirnavirus tree.

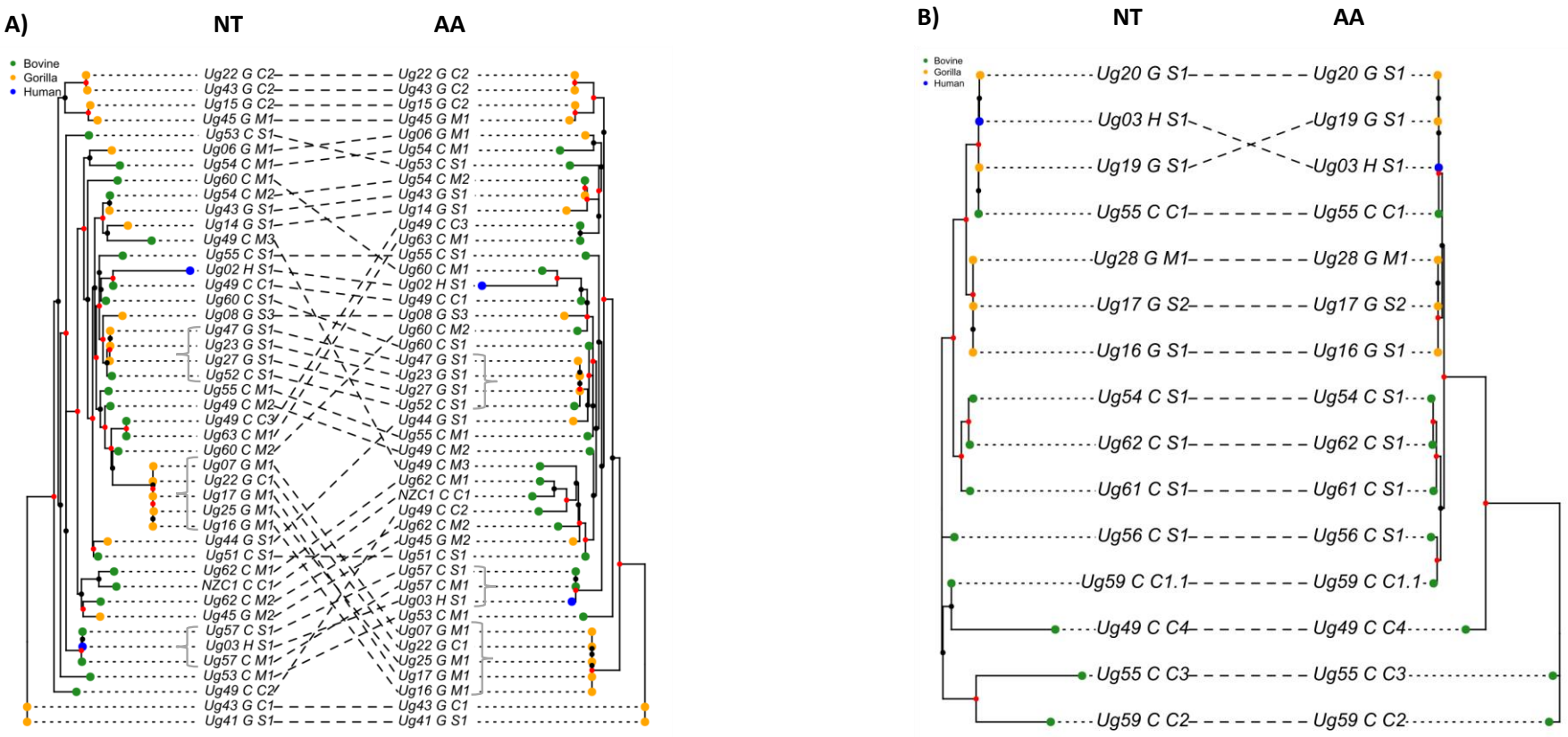


FIGURE 25 CO-PHYLOGENY OF GENOGROUP I AND GENOGROUP II NUCLEOTIDE VERSUS AMINO ACID PICOBIRNAVIRUS SEQUENCES

Co-phylogeny of the picobirnavirus *RdRp* sequences from this study only comparing nucleotide to amino acid sequences. A) The left co-phylogeny is the reported sequences of genogroup I *RdRp* of picobirnavirus. Gray brackets show shifts of groups in the tree topology though still remaining within the same clade as discussed in the text. Best tree model was selected from PhyML (Section 3.7.5.3) as HKY85 + G + I for nucleotide and LG + G for amino acid**.** Nucleotide sequences were trimmed to between 195-200 bps; amino acid sequences were trimmed to between 65-66 amino acids (Section3.7.5.1). **B) The right co-phylogeny phylogeny is the reported sequences of genogroup II *RdRp* of picobirnavirus.** Nucleotide sequences (NT) are on the left-hand side of the co-phylogeny trees and amino acid sequences (AA) are on the right-hand side of the co-phylogeny trees. Best tree model was selected from PhyML (Section 3.7.5.3) as GTR + G for nucleotide and WAG + G for amino acid. Nucleotide sequences were trimmed to between 360-370 bps; amino acid sequences were trimmed to between 120-123 amino acids (Section3.7.5.1). Fast-

likelihood-based methods of aLRT SH-like were used for branch support values (Section 3.7.5.3). Black nodes denote branch support values >0.8 and red nodes <0.8. Tip colours denote host species with green tips for cattle, orange for gorilla and blue for human. Ug# refers to the sample number from Uganda and NZ# from the sample number from New Zealand. G after the Ug# or NZ# refers to gorilla samples and C refers to cattle samples. S#, C# or M# at the third site on the labels designates S#=Sanger sequence, C#=cloned sequence, M#=metagenomic sequence.

4.1.5.2 TRIMMED VERSUS UNTRIMMED SEQUENCES

Phylogenetic tree comparisons (co-phylogenies) of picobirnavirus *RdRp* amino acid sequences from this study along with the corresponding highest BLASTn picobirnavirus accessions (Chapter 3, Section 3.7.5, Table 3.4) were performed to compare the use of trimmed versus untrimmed sequences. Co-phylogenetic comparisons for genogroup I include n = 47 study sequences from Uganda and New Zealand and n = 81 NCBI database picobirnavirus accessions (Figure 26, Figure 48). Co-phylogenetic comparisons for genogroup II include 16 study sequences from Uganda and n = 40 NCBI database picobirnavirus accessions (Figure 27, Figure 49).

The co-phylogenetic comparisons of untrimmed versus trimmed sequences does little to alter the grouping or clustering of the genogroup I picobirnavirus sequences and accessions with very few sequences that shift major clades (Figure 26). Overall, there are two sequences and one group of two sequences out of 130 that are significantly altered. Most of the genogroup I picobirnavirus sequences and accessions remain within the major clades, although they still shift tree topology. The percent identity of the untrimmed genogroup I sequences was lower at 65.2% than the trimmed genogroup I at 73.8% in the multiple alignments. Analysis of the congruence of the co-phylogenies, despite the above changes in tree topologies between untrimmed and trimmed sequences did show that there was a statistically significant (p-value = <0.01) association between the two trees and suggests they are congruent [257, 268].

The co-phylogenetic comparisons of untrimmed versus trimmed amino acid sequences for the genogroup II sequences (Figure 27) and accessions is not altered by trimming with only a minor rotation of sequences within the same clades. The percent identity of the untrimmed and trimmed genogroup I sequences was very similar at 61.7% for the untrimmed and 60.1% for the trimmed amino acid sequences on the multiple alignments. Analysis of the congruence of the co-phylogenies, despite the above changes in tree topologies between untrimmed and trimmed sequences did show that there was a statistically significant (p-value = <0.01) association between the two trees and suggests they are congruent [257, 268].

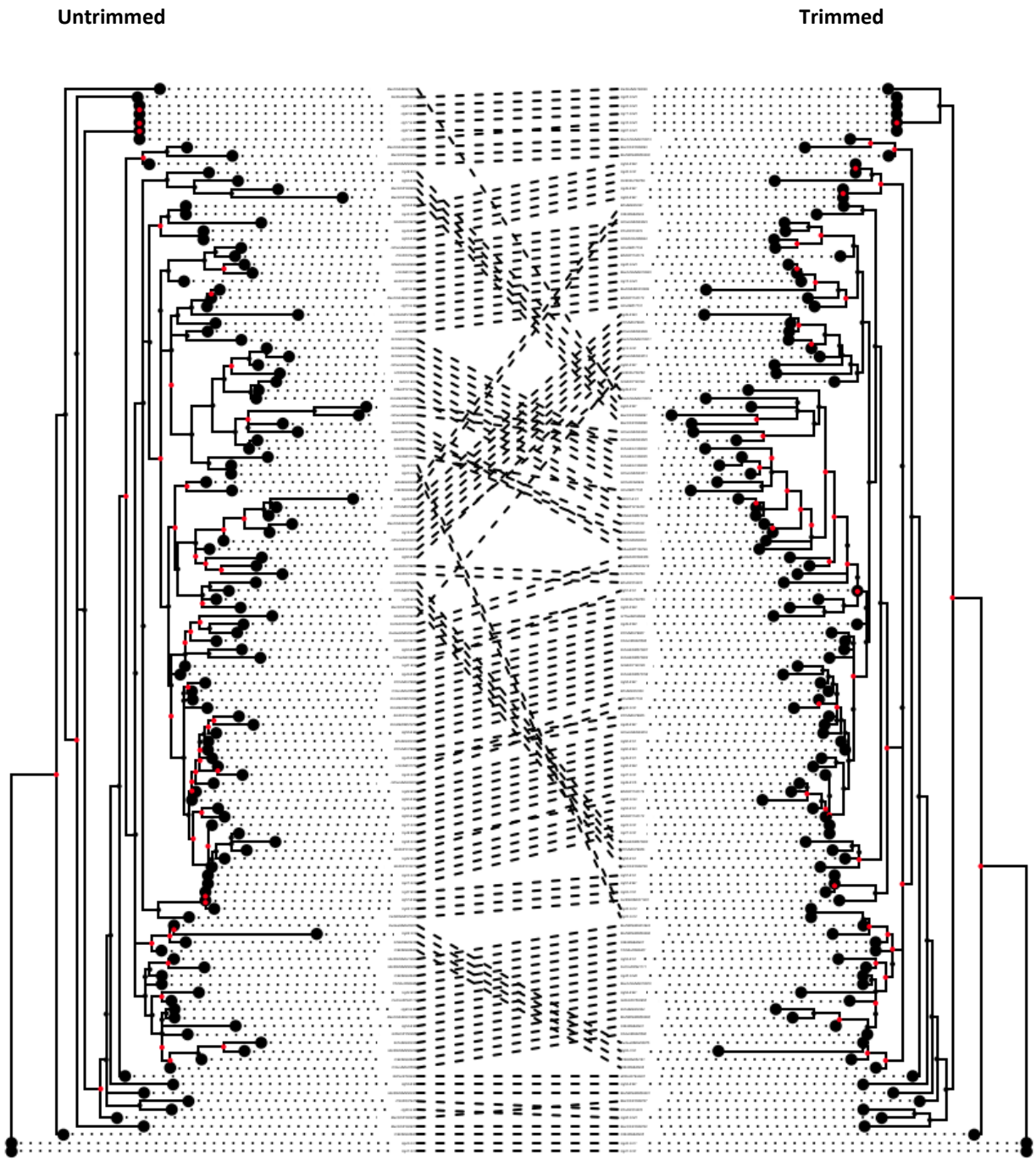


FIGURE 26 CO-PHYLOGENY OF UNTRIMMED TO TRIMMED GENOGROUP I PICOBIRNAVIRUS SEQUENCES

Co-phylogeny of the picobirnavirus *RdRp* genogroup I amino acid sequences from this study and highest BLAST match comparing untrimmed to trimmed sequences. Untrimmed sequences are on the left-hand side of the co-phylogeny trees and trimmed sequences are on the right-hand side of the co-phylogeny trees. Total of 130 sequences with the untrimmed sequences of 80–88aa in length and the trimmed sequences of 60–72aa in length. Best tree model was selected from PhyML (Section 3.7.5.3) as LG + G + I for untrimmed and trimmed trees. Fast-likelihood-based methods of aLRT SH-like were used for branch support values (Section 3.7.5.3). Black nodes denote branch support values >0.8 and red nodes <0.8. Labels from the NCBI picobirnaviruses are host species as noted above/country/accession number. Labels for the samples from the study are: Ug# refers to the

sample number from Uganda and NZ# from the sample number from New Zealand. G after the Ug# or NZ# refers to gorilla samples, H refers to human samples and C refers to cattle samples. S#, C# or M# at the third site on the labels designates S#=Sanger sequence, C#=cloned sequence, M#=metagenomic sequence.

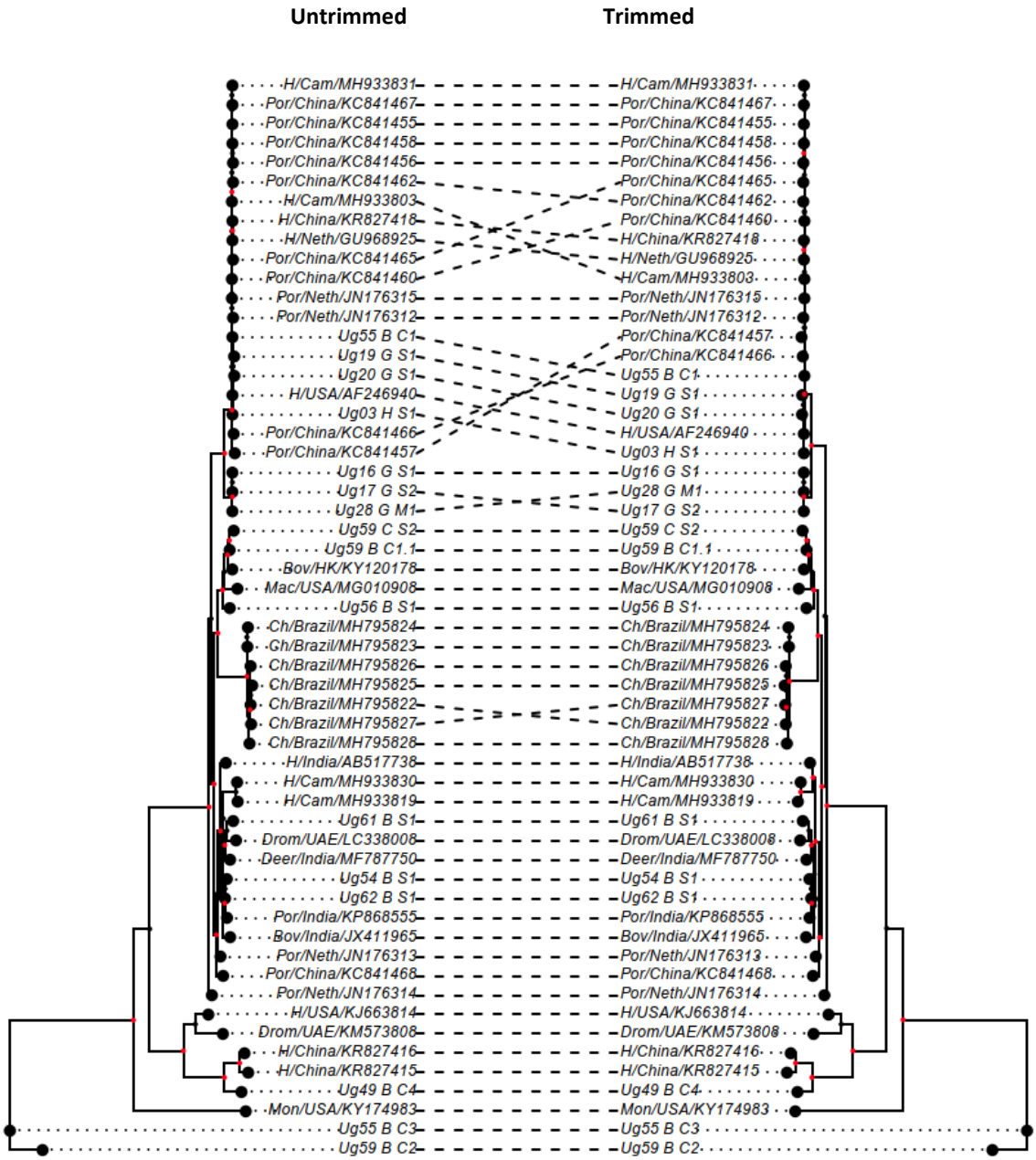


FIGURE 27 CO-PHYLOGENY OF UNTRIMMED TO TRIMMED GENOGROUP II PICOBIRNAVIRUS SEQUENCES

Co-phylogeny of the picobirnavirus *RdRp* genogroup II amino acid sequences from this study and highest BLAST match comparing untrimmed to trimmed sequences. Untrimmed sequences are on the left-hand side of the co-phylogeny trees and trimmed sequences are on the right-hand side of the co-phylogeny trees. Total of 56 sequences with the untrimmed of 137aa in length and the trimmed of 67–121aa in length. Best tree model was selected from PhyML (Section 3.7.5.3) as LG + G for untrimmed and trimmed trees. Fast-likelihood-based methods of aLRT SH-like were used for branch support values (Section 3.7.5.3). Black nodes denote branch support values >0.8 and red nodes <0.8. Labels from the NCBI picobirnaviruses are host

species/country/accession number. Labels for the samples from the study are: Ug# refers to the sample number from Uganda. G after the Ug# or NZ# refers to gorilla samples, H refers to human samples and C refers to cattle samples. S#, C# or M# at the third site on the labels designates S#=Sanger sequence, C#=cloned sequence, M#=metagenomic sequence. Labels for the NCBI picobirnaviruses are: Species: H=human, Drom=dromedary, Ch=chicken, Por=porcine, Mon=monkey, Bov=bovine, Mac=macaque; Country of origin: USA=United States of America, HK=Hong Kong, UAE=United Arab Emirates, Cam=Cameroon, Neth=Netherlands/NCBI Accession number.

4.1.5.3 GENOGROUP I PHYLOGENY

Phylogenetic analysis of all of the *RdRp* genogroup I picobirnavirus sequences from Uganda and New Zealand study was performed, including duplicate sequences derived from the same sample independently using a different methodology (Sanger, cloning, metagenomics) (Figure 29). Forty-four sequences were included in the *RdRp* genogroup I picobirnavirus amino acid sequences with the alignment, not including duplicates (Figure 28). The phylogenetic analysis with the duplicates included shows that the multiple methods of obtaining picobirnavirus sequences were effective (Figure 29). Two sequences, a cloned sample C1.1 and the metagenomic sample M1, from gorilla sample Ug22 were 100% identical. Three sequences from gorilla sample Ug15 were 100% identical; the two cloned samples C2 and C2.1, which included metagenomic reads mapped to the C2 to extend the contig, and metagenomic sample, M1.

Within the genogroup I picobirnaviruses, samples in the same host species within the same gorilla family shared >99% percent identity (Ug43_G_C1_R/ADF and Ug41_G_S1_R/ADF; Ug47_G_S1_R/ADF and Ug23_G_S1_R/INF; Ug43_G_C2_R/ADF and Ug22_G_C2_R/ADF; Ug17_G_M1_R/SA and Ug22_C1.1_R/ADF and Ug25_G_M1_R/INF and Ug16_G_M1_R/ADF and others) (clade a, Figure 29). Some picobirnavirus sequences shared >99% percent identity from different samples from different host species (see Ug03_H_S1 and Ug57_C_S1; Ug43_G_S1_R/ADF and Ug54_C_M2) (clade b, Figure 29).

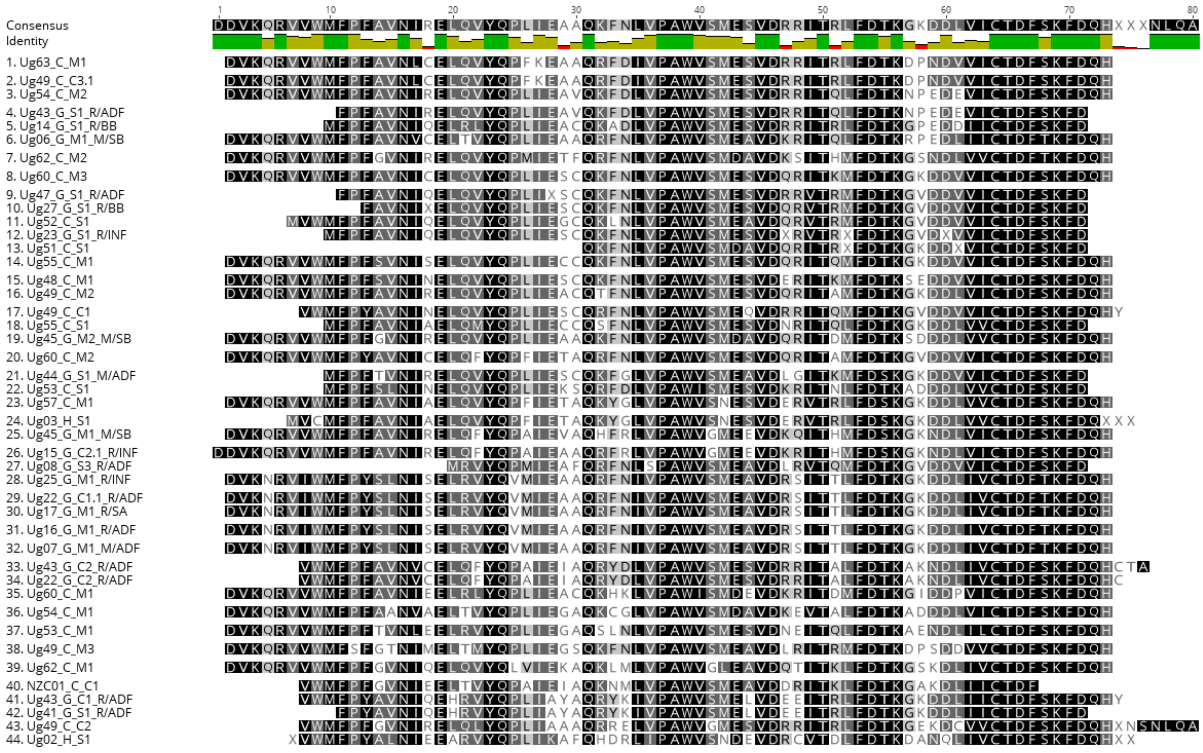


FIGURE 28 MULTIPLE ALIGNMENT OF THE GENOGROUP I PICOBIRNAVIRUSES FROM THIS STUDY

Multiple alignment on Geneious® of the picobirnavirus *RdRp* genogroup I amino acid sequences from Uganda and New Zealand samples with the exclusion of Ug08 and Ug10. Sequence lengths ranged from 126 to 1887 bp and were manually trimmed to around 200 bp for the multiple alignment. Percent identity for the multiple alignment was 86.4% with 37.3% of the sites identical. Labels for the samples from the study are: Ug# refers to the sample number from Uganda and NZ# from the sample number from New Zealand. H after the Ug# or NZ# refers to human samples, G refers to gorilla samples and C refers to cattle samples. S#, C# or M# at the third site on the labels designates S#=Sanger sequence, C#=cloned sequence, M#=metagenomic sequence. The fourth site for the Gorilla samples designates the gorilla family and gender and age if applicable: R=Rushegura gorilla family, M=Mubare gorilla family/INF=infant, SA=sub-adult, BB=blackback male, SB=silverback male, ADF=adult female [225].

Fifty-two sequences were included in the *RdRp* genogroup I picobirnavirus amino acid sequence phylogenetic analysis. The genogroup II outgroup was also included. The percent identity in the multiple alignment was 82.7% for all sequences, including the genogroup II root (without the root, the percent identity was 85.2%; Figure 50). Some of the same picobirnavirus sequences from above shared >99% percent identity from different samples in the same host species, especially one group of 5 different gorillas (seeUg16_G_M1 and Ug17_G_M1 and Ug22_G_M1 and Ug07_G_M1 and Ug25_G_M1) (a clades, Figure 30). Clustering based on host was noted, although the human samples did not cluster together but within other cattle picobirnavirus samples. In some instances, despite the clustering by host in this tree, there remains almost identical picobirnaviruses shared by hosts (Ug57_C_M1 and Ug03_H_S1; Ug43_G_S1 and Ug54_C_M2) (b clades, Figure 30).

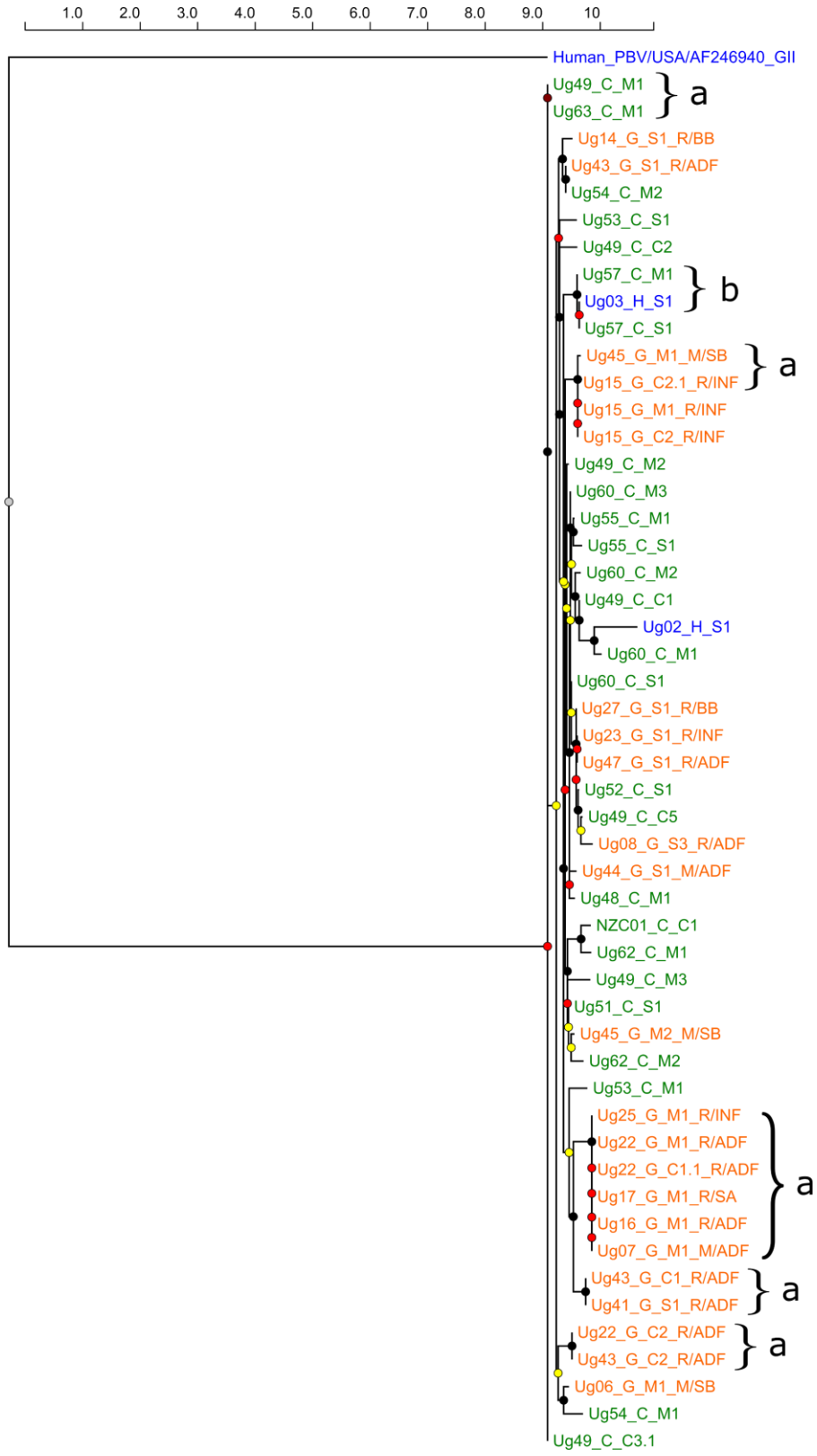


FIGURE 29 PHYLOGENTIC TREE OF THE GENOGROUP I PICOBIRNAVIRUSES FROM THIS STUDY

Phylogenetic tree of the *RdRp* genogroup I picobirnavirus amino acid sequences from Uganda and New Zealand with colour-coding of hosts. Fifty-one genogroup I sequences and one genogroup II sequence included. Amino acid sequences trimmed to between 65-66 amino acids (Section 3.7.5.1). The tree does not include duplicates from the same sample and is rooted to a picobirnavirus genogroup II sequence from the NCBI

database (AF246940). Similar viruses in the same host groups within clades are designated by the brackets labelled 'a' and between different hosts by the bracket labelled 'b' is discussed in the text. Best tree model was selected from PhyML (Section 3.7.5.3) as LG + G. Fast-likelihood-based methods of aLRT SH-like were used for branch support values (Section 3.7.5.3). Black nodes denote branch support values >0.8, yellow nodes between 0.5 and 0.8 and red nodes <0.5. Labels for the samples from the study are: Ug# refers to the sample number from Uganda and NZ# from the sample number from New Zealand. H after the Ug# or NZ# refers to human samples, G refers to gorilla samples and C refers to cattle samples. S#, C# or M# at the third site on the labels designates S#=Sanger sequence, C#=cloned sequence, M#=metagenomic sequence. The fourth site for the Gorilla samples designates the gorilla family and gender and age if applicable: R=Rushegura gorilla family, M=Mubare gorilla family/INF=infant, SA=sub-adult, BB=blackback male, SB=silverback male, ADF=adult female [225]. Orange: gorilla, green: cattle, blue: human.

4.1.5.4 GENOGROUP II PHYLOGENY

Phylogenetic analysis of all of the *RdRp* genogroup II picobirnavirus sequences from Uganda and New Zealand study was performed, including duplicate sequences derived from the same sample independently using different methodology (Sanger, cloning, metagenomics) (Figure 30). Eighteen genogroup II sequences, including two duplicates (i.e., >97% percent identity between picobirnaviruses from the same sample) (clade a, Figure 30), were included in the *RdRp* genogroup II picobirnavirus amino acid sequence analysis along with a genogroup I sample as a root for the tree. The percent identity in the multiple alignment was 63.2% across all 18 sequences. Two groups or clades of similar picobirnaviruses were identified in different samples (Ug16_G_S1, Ug17_G_S2 and Ug28_G_M1; Ug19_G_S1, Ug55_C_C1, Ug03_H_S1 and Ug20_G_S1) (clade b, Figure 30). The gorilla only group had a 98% percent identity between the three amino acid sequences; they were all from the same family group (the Rushegura family) with a blackback (Ug28), adult female (Ug16) and sub-adult (Ug17) (clade b with asterisk, Figure 30). The second similar clade had an overall percent identity of 91.2% with highly similar genogroup II picobirnaviruses between different individuals of different host species including a human, gorilla and cattle (clade c, Figure 30). In the genogroup II analysis (Figure 30), Ug59_C_C1 was 100% identical to Ug59_C_S1. Additionally, Ug59_C_C1 was near identical to Ug59_C_C1.1 as C1.1 was the cloned sequence (C1) with the metagenomic reads from sample Ug59 used to create a longer contig (clade a, Figure 30).

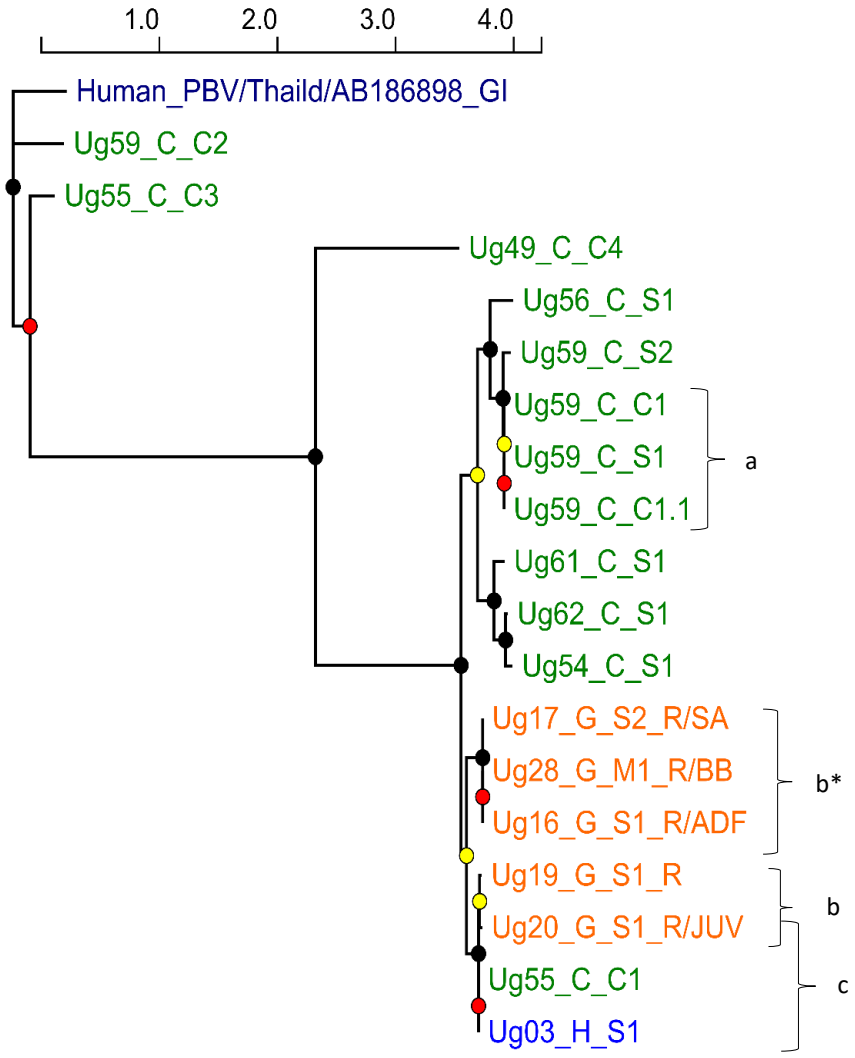


FIGURE 30 PHYLOGENETIC TREE OF THE GENOGROUP II PICOBIRNAVIRUSES FROM THIS STUDY

Phylogenetic tree of the 18 *RdRp* genogroup II picobirnavirus amino acid sequences from Uganda with root to genogroup I (AB186898). Amino acid sequences trimmed to between 120-130 amino acids (Section 3.7.5.1). Designation of similar viruses in the same host groups within clades by the brackets labelled b and between different hosts by the brackets labelled c is discussed in the text; duplicates in the same sample is designated by brackets labelled a and discussed in the text. Best tree model was selected from PhyML (Section 3.7.5.3) as LG + G. Fast-likelihood-based methods of aLRT SH-like were used for branch support values (Section 3.7.5.3). Black nodes denote branch support values >0.8, yellow nodes between 0.5 and 0.8 and red nodes <0.5. Labels for the samples from the study are: Ug# refers to the sample number from Uganda and NZ# from the sample number from New Zealand. H after the Ug# or NZ# refers to human samples, G refers to gorilla samples and C refers to cattle samples. S#, C# or M# at the third site on the labels designates S#=Sanger sequence, C#=cloned sequence, M#=metagenomic sequence. The fourth site for the Gorilla samples designates the gorilla family and gender and age if applicable: R=Rushegura gorilla family, M=Mubare gorilla family/INF=infant, SA=sub-adult, BB=blackback male, SB=silverback male, ADF=adult female [225]. Colour-coding of hosts; orange: gorilla, green: cattle, blue: human.

4.1.5.5 Near-complete Picobirnavirus phylogeny

Near-complete (NC) picobirnaviruses from this study, together with complete picobirnaviruses from NCBI, were analysed to assess the effect of longer contigs on phylogenetic analyses. Twelve NC

picobirnaviruses from the study were selected based on the criteria in Chapter 3, Section 3.7.4, which consisted of 6 gorilla samples and 6 cattle samples, all from Uganda. The NC picobirnaviruses were either *de novo* assembled metagenomic sequences (8/12 or 67%) or cloned sequences that mapped to a reference/clone of metagenomic reads to lengthen the cloned contig (4/12 or 33%). The median number of reads assembled in the contigs was 291 reads (range: 69–2540 reads; quartiles: 142–526 reads). The median nucleotide length of the NC picobirnaviruses was 1361 bp (range 987–1887 bp; quartiles: 1105–1418 bp) and the median amino acid length was 453 aa (range 329–628 aa; quartiles: 368–472aa).

Four cattle samples and three gorilla samples in the same branch all contained both genogroup I primers that were annotated (Ug53_C_M1, Ug22_G_C1.1, Ug07_G_M1, Ug45_G_M1, Ug55_C_M1, Ug48_C_M1, Ug49_C_C3.1) (labelled GI, dark blue bracket, Figure 31); a separate branch within the larger clade of the suspected genogroup I picobirnaviruses contained just the reverse primer for the genogroup I and conserved domain 2 and 3 (Ug58_C_M1) (labelled GI?, light blue bracket, Figure 31). A gorilla sample with the genogroup II primers was on a separate branch (Ug28_G_M1) (labelled GII, purple bracket, Figure 31); two further gorilla samples and a cattle sample all were on their own separate branches (Ug15_G_C1.1, Ug39_G_M1, Ug49_C_C4.1) (labelled other genogroup, bracket, Figure 31).

The only potentially complete picobirnavirus sequence from the samples was a gorilla sample, Ug15 obtained from the cloned contigs from the four-primer set (F3/F5/R5/R8) and assembled into a 795 bp sequence. The sequence was augmented with a map to a reference of 1925 reads from the metagenomic data to make a contig of 1887 bp and 628 amino acids. The complete contig contains only the conserved domain 3 and does not have methionine as the first amino acid in the translation based on the standard genetic code (translation_table 1). Percent identity of the multiple alignment for the NC samples was 30%, though they contained both genogroup I and genogroup II (and possibly other genogroups) in the alignment. Percent identity for the NC samples and selected complete picobirnaviruses from the NCBI database increased to 45% (Chapter 5, Section 5.1.1.3, Figure 34).

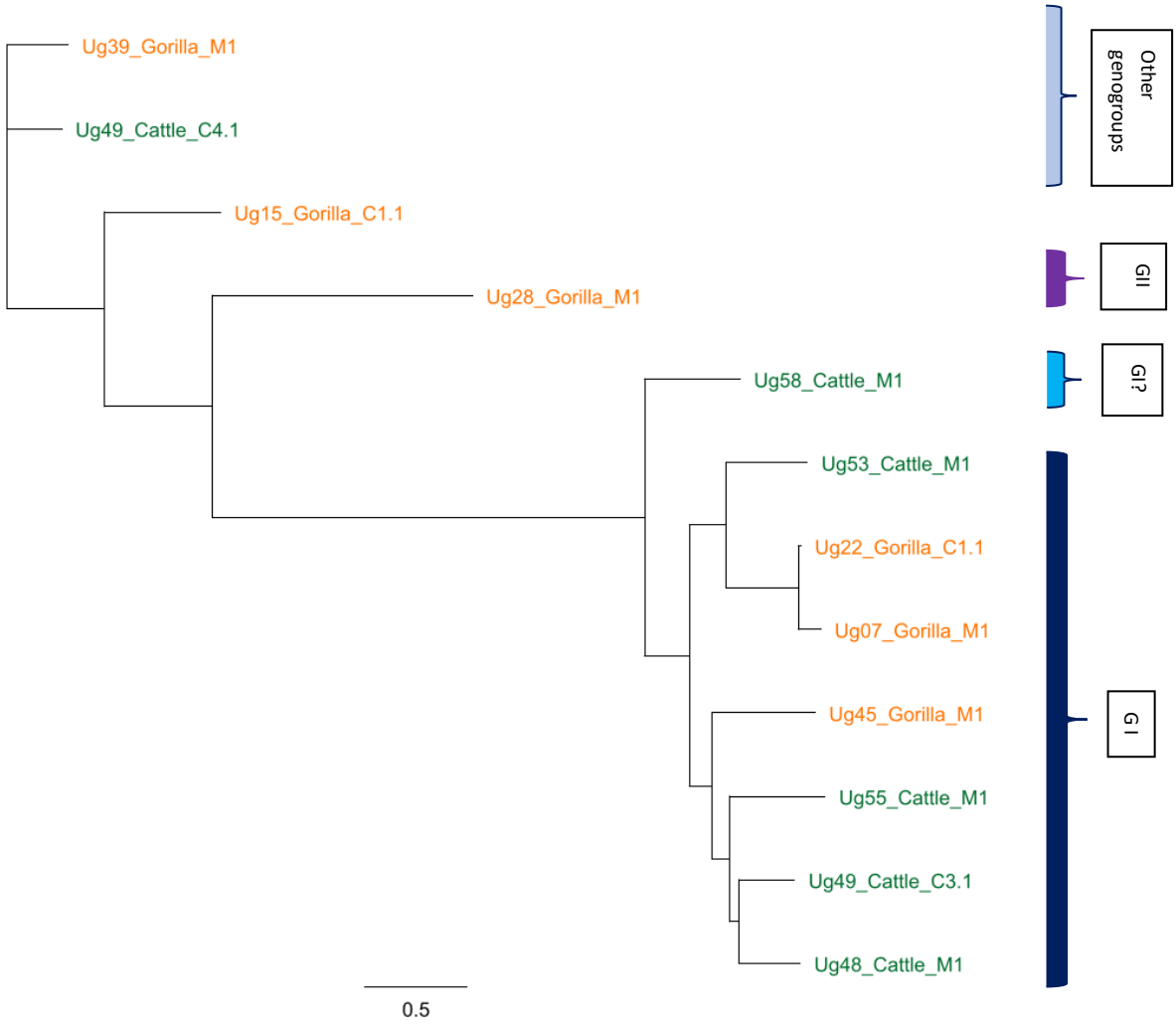


FIGURE 31 PHYLOGENETIC TREE OF THE NEAR COMPLETE PICOBIRNAVIRUS SEQUENCES FROM THIS STUDY

Phylogenetic tree of near-complete picobirnavirus *RdRp* amino acid sequences from 6 gorilla and 6 cattle samples from Uganda. Amino acid sequences were between 300-600 amino acids (Section 3.7.4). Selected tree model was GTR**.** Labels for the samples from the study are: Ug# refers to the sample number from Uganda. G refers to gorilla samples and C refers to cattle samples. C# or M# at the third site on the labels designates C#=cloned sequence, M#=metagenomic sequence; #.1 designates a clone sample with a map to reference/clone of metagenomic reads for longer contigs. Orange=gorilla samples; dark green=cattle samples. GI=*RdRp* genogroup I PBVs; Non-GI=Non-genogroup I *RdRp* PBV; GII=*RdRp* genogroup II PBV; Other genogroups=possibly other separate genogroups not in GI, GII or non-GI.

4.2 Discussion

4.2.1 Picobirnavirus identification and confirmation

Picobirnaviruses were readily identifiable in the samples from Uganda with a small number of additions from the New Zealand samples. Three different methodologies were utilized to identify the picobirnaviruses including both conventional PCR and metagenomic methods which allowed for both confirmation of the methods in obtaining duplicates and also the addition of more picobirnaviruses and longer contigs from the various methods. In some instances for Sanger sequences, forward and reverse reads could not be assembled, which could have been due to inadequate primer binding or

from the high diversity common amongst picobirnaviruses. Typically, if the pair-end reads could not be assembled, they would be excluded but I decided to include the high-quality singletons determined by chromatogram evaluation in order to analyse the diversity, phylogeny and potential for cross-species transmission of the dsRNA viruses. Other methods such as deeper sequencing, more cloning or additional extractions to attempt to obtain better quality RNA could have reduced the occurrence of these non-assembled reads. The published primer sets were more successful in the identification of picobirnaviruses from my samples than the primers designed in Geneious®. This could be due to many factors such as non-specific primer binding, the primer looping back on itself due to complementary bases within the primer (these are usually screened for in the primer design software), primer dimers (also usually noted in the primer design program), primers that are too long or short, GC content that is too high for PCR, or the consensus sequence used for primer design was inappropriate. The primers that were designed with the criteria from Section 3.6.1.2 were within the appropriate length (recommended 18–22 bp) and GC content (recommended 40–60%) though they may have been slightly low in the melting temperature at 50-53 ˚C, which reduces the annealing efficiency of the primers (recommended 52–58˚C); a touchdown protocol was used to overcome this lower melt temperature, and five of the 15 primers, did produce amplicons that could be detected by gel electrophoresis with either the positive controls or samples. Only one amplicon was of high-quality for Sanger sequencing and produced a match on BLASTn with the criteria listed in Section 3.7.1.1. Overall, both genogroup I and genogroup II picobirnaviruses were identified and many picobirnaviruses that did not group within the typical genogroups were also identified, potentially suggesting more genogroups in these samples.

4.2.2 Nucleotide and amino acid co-phylogenetic relationships

The results show that the use of nucleotide versus amino acid sequences for analysing picobirnaviruses can produce differing patterns of phylogenetic structure. Many picobirnavirus phylogenetic studies use both nucleotide and amino acid sequences [119, 145, 154, 195, 201], though less-common just one type of sequence [113, 131, 155, 174]. One study did evaluate picobirnaviruses comparing the use of nucleotide and amino acid sequences but only within the known genogroups and these were found to be similar whether comparing nucleotide or amino acid sequences [201].

Co-phylogenetic relationships have routinely been used to analyse host and parasite (or pathogen) relationships to evaluate co-evolution, either through co-speciation or co-divergence [255]. I sought to use this technique to understand the effect that different approaches to picobirnavirus comparison have on phylogenetic interpretation, such as: nucleotide versus amino acid sequences, or untrimmed versus trimmed amino acid sequences. This method was previously utilised for analysis of picobirnavirus in Knox et al. 2018. My results show that there is a difference in co-phylogeny in the

genogroup I but less so in the genogroup II picobirnaviruses in this study. The difference in the genogroup I picobirnaviruses was estimated to be less than 25% of the total genogroup I picobirnavirus sequences based on congruence of nucleotide as compared to amino acid sequences with the co-phylogeny. In addition, there was a difference in the diversity of the multiple alignments of the nucleotide (percent identity 70.9% with 36.7% identical sites) compared to the amino acid (percent identity 73.4% with 37.3% identical sites) sequences with the amino acid alignment indicating much less diversity. In contrast, there was little to no difference in the co-phylogeny tree for the genogroup II picobirnavirus samples and the diversity of the nucleotide (percent identity 55.6%) as compared to amino acid (54.5% identity) sequences was similar.

The higher mutation rates of RNA viruses, due to lack of proof-reading, can result in high genetic diversity in these viruses and potentially higher evolutionary rates [130, 179]. Due to the faster timescale that mutations occur in these viruses, information on viral evolution may be evaluated more easily than for bacterial or eukaryotic organisms, including the effects of positive and/or negative selection on the organism [179]. RNA viruses also have a smaller genome size on average, thought to be related to the lack of proof-reading, therefore, high mutation rates and the 'error threshold' for deleterious mutations that would result in extinction of that virus [111]. These high mutation rates may also lead to high diversity within the viral family if the mutations are not deleterious, possibly something that is occurring in the picobirnaviruses. Though I did not analyse synonymous to non-synonymous ratios in the picobirnaviruses from my samples, one hypothesis for the incongruence in the nucleotide and amino acid sequences in the genogroup I picobirnaviruses may be from synonymous substitutions.

Furthermore, due to the known high diversity within the picobirnaviruses, the reduced diversity of the amino acid as compared to the nucleotide sequences would allow for a better evaluation of clustering, by host or geography, if identified, in the analyses (Chapter 4 and 5). I would then be more likely to evaluate clustering by factors such as host species or geography with amino acid sequences due to less diversity. In a recent study to discover novel or highly diverse viruses, methods in the alignment algorithms used conserved protein segments rather than nucleotide sequences, which gave increased sensitivity to identify these challenging viruses [89]. In the situation of the highly diverse picobirnavirus, synonymous mutations would alter the nucleotide phylogenetic analyses but would not alter the amino acid phylogenetic analyses. Due to the points above, amino acid sequences were used for further analyses for both genogroup I and genogroup II picobirnaviruses.

Limitations in the use of amino acid, as compared to nucleotide sequences, for the analyses were considered. It was possible that an incorrect translation could be used when translating nucleotide

sequence into protein, which could result in incorrect alignments and phylogenetic analyses. Known conserved gene regions were checked for each individual translation, to ensure no stop codons were present. If incorrect reading frames were chosen, this could result in an artefactual higher non-synonymous to synonymous mutation ratio. An additional method to investigate this possibility would be, as mentioned above, to estimate the synonymous to non-synonymous ratios in my viruses, which I did not perform. This could only be performed on the near-complete picobirnaviruses (of which I had limited numbers) and only then on the ones with identified open-reading frames (ORFs) of which there were only 12 sequences identified (see Chapter 7, Section 7.2.1.2). The use of amino acid sequences also excludes the possibility of evaluating the nucleotides specifically for insertions or deletions, though divergent sequences on protein alignments could still be assessed on the raw reads as described above. Another limitation is that the standard genetic code (SGC) may not necessarily be correct for these viruses [93, 195]. This is further addressed in Chapter 7 (The Actual Host of Picobirnavirus).

4.2.3 Untrimmed and trimmed co-phylogenetic relationships

The effect of additional modes of analysing and assessing picobirnaviruses such as the use of trimAl® for spurious sequence trimming was also evaluated [271]. The removal of sequences from poorly aligned regions in a multiple alignment can help to improve the phylogenetic analyses [271]. The use of trimming with trimAl® in a recent study on picobirnaviruses did show altered tree topology ("rearrangement of the phylogeny") of the picobirnaviruses, more so in the incomplete or partial *RdRp* picobirnavirus sequences as compared to the complete sequences that were trimmed [113]. In my analysis, I analysed the effects of trimming on partial *RdRp* sequences on the phylogenetic tree topology. In contrast to the prior study and using similar *RdRp* picobirnavirus sequences from the NCBI database, the use of trimmed or untrimmed sequences did not significantly alter the phylogenies. The most likely reason for this difference is that I used an ~200–300 bp region of the *RdRp* ORF on segment 2 as opposed to the Knox et al. 2018 paper that used complete genomes. The sequences from the study cohort may not have been trimmed to the same extent, thus not affecting the co-phylogenetic analyses between the untrimmed and trimmed sequences. As mentioned in 4.1.3.2, only 3% (4 out of 130 sequences) of the genogroup I picobirnaviruses shifted within the topology of the tree, outside of their group or clade. Most of the picobirnaviruses remained within the major branches. For the genogroup II phylogenies, there are only minor shifts of branches. Both co-phylogenies showed similar patterns of association between the untrimmed and trimmed trees. Overall, even with analysing partial *RdRp* sequences to evaluate the effects on the phylogeny with- and without-trimming, trimming did not substantially alter or rearrange the phylogeny. There is still the possibility of an effect

on the *RdRp* genogroup I picobirnavirus sequences when not trimming, but it was concluded that the impact was low and less likely to affect the remainder of the analyses.

4.2.4 HOMOLOGOUS SEQUENCES WITHIN SIMILAR HOSTS

In addition to similar sequences (100% identical amino acid residues) within the same sample using different sequencing techniques, similar sequences were also found in different samples derived from the same host species, specifically within the gorillas. There are multiple clades of similar to near-similar gorilla genogroup I picobirnaviruses from different gorillas (Figure 29 and Figure 30). Samples from the mountain gorillas for this study were taken from any of the three habituated mountain gorilla families in the Bwindi forest, Uganda. The identification of the gorilla family is based on the park rangers judgement, they track the gorilla families and can identify the night nest from those families where the samples are collected from; the identification of the age and gender is based on the specimen size, the knowledge of the mountain gorilla family demographics along with the identification of hair which can identify a silverback versus a blackback [78, 225]. As the phylogenetic trees show (Figure 29 and Figure 30), many of the similar gorilla picobirnaviruses from different gorillas are from the same family group, the Rushegura family, which could indicate sequence homology. The picobirnaviruses found in the specimens from the Mubare family did not usually cluster, with the exception of one virus. Given how highly divergent this virus can be, the finding of the homologous picobirnaviruses in different gorillas could indicate sharing of the virus between the gorillas, more commonly within the same family group. This could provide further support for picobirnaviruses spreading through those individuals with increased or higher contact rates. Limitations of these conclusions include that I do not have quantitative contact rate information to support this presumption. Additionally, most of the samples where I found picobirnaviruses were from one family group, the Rushegura family (15 picobirnaviruses out of 30 total samples) as compared to the other family group from whom specimens were collected, the Mubare family (5 picobirnaviruses out of 15 total samples); overall though, the percentage of picobirnaviruses from the available specimens is comparable (45% from the Rushegura family, 33% from the Mubare family). There was no significant difference between these two family groups (χ^2= 0.55, p value = 0.46).

4.2.5 HOMOLOGOUS SEQUENCES BETWEEN HOSTS

Multiple occurrences of similar or near-similar picobirnavirus sequences between different host species were also detected in the phylogenetic analyses. The recognition of picobirnaviruses with 100% percent identity in different host species within the same environment is unexpected, especially given these are highly divergent dsRNA viruses. This observation could indicate sequence homology and cross-species transmission events of the picobirnaviruses in this ecosystem. Multiple different risk factors such as increased contact rates, genetic relatedness between species, host distribution,

pathogen prevalence and immunological response, to name a few, may all contribute to cross-species transmission events and ongoing transmission [2, 18, 25, 28]. As mentioned above, the homologous picobirnaviruses in the clades of the habituated gorillas from the same family group could have been seen as evidence for increased contact rates as a factor for cross-species transmission of picobirnaviruses.

The homologous picobirnaviruses between different species in the system may also support the possibility of cross-species transmission with factors such as increased contact between species as a potential driver. Homologous genogroup I picobirnaviruses were identified in human and cattle and in gorilla and cattle with the smaller *RdRp* fragment trimmed to ~200 bp. Homologous genogroup II picobirnaviruses were identified in human, cattle and gorilla with the *RdRp* trimmed to ~400 bp. The sequences were manually trimmed to improve the multiple alignment but also could have trimmed out regions that were not homologous. It is, also, possible that there are high rates of contact between humans and cattle as there are domesticated cattle in the villages that come into contact with humans. A paper recently published from authors in my group [44] investigated contact potential between the various host species from the same study system, though the investigation did not occur during the same time period as my sample collection. The authors found using "contact and health survey data" that the most common contact was between similar species (human to human) or domestic livestock and humans; less commonly, contact was noted between gorillas and humans or gorillas and livestock though it was still reported in the short timeframe of a week of self-reporting [44]. Even though the time periods are different for the two studies, I can presume that the contact potential and behaviour may be similar and, therefore, the findings of this study could indicate that the increased contact between cattle and humans or contact between cattle and gorillas in the park (cattle roaming into the park) or when gorillas come out of the park to forage, could result in the potential for pathogen sharing. The less common, though still reported, contact of the wildlife with humans also has the potential to result in cross-species transmission. Close evolutionary relationships among host species such as with humans and gorillas might translate into similar immunological responses and life-history traits [273-275], and hence increase the likelihood of successful cross-species infection. The majority of the evidence does not support more similar viruses within similarly-related hosts in my samples with one or two exceptions (Figure 30).

4.2.6 Clustering and challenges of picobirnaviruses from my samples

My results show that designated genogroup I and genogroup II picobirnaviruses cluster together within the study samples (Figure 31), though many picobirnaviruses could not be designated to a genogroup and did not cluster, potentially showing there are more genogroups in these samples. Rooting the phylogenetic trees with other genogroup picobirnaviruses helped to show that divergent

possible-genogroup I picobirnaviruses, or picobirnaviruses without a known genogroup, belonged in separate clades. It is also possible that these sequences, along with Ug62_M1, a cattle sample, may be different genogroups altogether. Additionally, in the NC picobirnaviruses (which included both known genogroup I and genogroup II picobirnaviruses), the genogroup I picobirnaviruses all clustered in the same clade (Figure 31). A slightly divergent branch, though not in the same clade as the other genogroup I picobirnaviruses, was the cattle sample Ug58, which did not cluster with any other picobirnavirus. The one known genogroup II picobirnavirus was on a separate branch and three other picobirnaviruses were also on their own separate branches which may indicate additional genogroups (Figure 31) [123, 141-145]. Additionally, the phylogenetic analyses of the rooted genogroup I tree (Figure 29) and the genogroup II tree (Figure 30) do show some clustering by host though this is examined in greater detail in the next chapter, Chapter 5.

The extensive diversity of the picobirnaviruses generates challenges in the interpretation of the picobirnavirus sequences. Picobirnaviruses within the RNA viruses have high substitution rates with estimated substitution rates of 0.004 to 0.014 substitutions per site per year [142, 276]; that study also showed that some of the picobirnaviruses in different hosts may evolve at different rates with the substitution rates of human, avian and porcine strains similar, and monkey strains lower [142]. Mutations in RNA viruses are common due to lack of proof-reading with the *RdRp* gene, resulting in point mutations, insertions or deletions, some which can result in the shifting of the reading frame for translation. These indels can result from normal biological processes in viruses but can also result from sequencing error.

The high degree of diversity among known picobirnaviruses, as well as the potential under-representation in NCBI for this diverse group, presents challenges in the identification of mutations in newly-generated sequences. Consequently, multiple methods of assessment of mutations within the sequences and in the analyses was utilised. These included manual assessment of chromatograms of Sanger sequences and read depth and nucleotide agreement in metagenomic sequences after trimming and assembly. In addition, gaps in the multiple alignments or aberrant sequences with long branches on phylogenetic analyses were also reassessed. The comparison of the sequences on BLASTn and the use of the CDS tool [272] also helped to identify indels of concern. Agreement and good quality peaks on the chromatogram for Ug10, for example, could indicate biological indels, though indels from processing cannot be excluded. Regardless, frameshifting of the amino acid sequence resulted in persistent gaps on the multiple alignments with the inclusion of Ug10. I, therefore, elected to exclude Ug10 from further analyses as inappropriate indels or mutations could not be dismissed. In contrast, further evaluation with the other sequences such as Ug49, as shown in Figure 22, Figure 23, and Figure

24, could not conclusively dismiss inappropriate indels or mutations as they did not result in problems based on the above criteria. Further analyses with programs to identify indels could be utilised on the sequences though were not pursued in this project. In addition, due to the inability to culture these viruses, evaluation of the functionality of viral proteins could not be evaluated.

4.3 SUMMARY

I did identify many picobirnaviruses in the samples from Uganda and New Zealand from various hosts. I confirmed the consistency of these findings across the different methods of sequence generation by conventional PCR and initial Sanger sequencing, cloning and metagenomic evaluation. Once identified, the picobirnaviruses were further analysed to determine what appropriate analytical approach should be taken in further analyses. Due to the known high diversity of the picobirnaviruses and the co-phylogenetic results, the use of amino acid sequences without the need for trimming was deemed appropriate for further analyses. The phylogenetic analyses of the picobirnaviruses from the samples showed genogroup clustering, homologous sequences in similar host species and, unexpectedly, homologous sequences in different hosts. Further investigation on clustering by host and geography, as well as the potential for host-sharing of this dsRNA virus is covered in the next chapter (Chapter 5).

CHAPTER 5: PICOBIRNAVIRUS HOST AND GEOGRAPHIC STRUCTURE

5.1 RESULTS

After identification of multiple picobirnaviruses from my study samples in Uganda and New Zealand (Chapter 4: Picobirnavirus characteristics, sequence profile and phylogeny), I sought to analyse their phylogenetic structure and test for associations with host and geography.

5.1.1 HOST ASSOCIATIONS

Clustering of the *RdRp* picobirnaviruses by host species was assessed more extensively with both the genogroup I and genogroup II picobirnaviruses from the study samples along with picobirnaviruses from the NCBI database (selection criteria from Chp 3, Section 3.7.5.1 and Table 5).

The addition of the picobirnaviruses from the database to the study samples was done to increase the number and diversity of the host species. In Chapter 4, there was some clustering by host noted in the genogroup I and genogroup II phylogenetic analyses of the picobirnaviruses from my study (See Figure 29 and Figure 30).

5.1.1.1 GENOGROUP I PICOBIRNAVIRUSES

The phylogenetic analyses for the genogroup I picobirnaviruses for host clustering, showed little to no clustering by host (Appendix C, Figure 51), as well as in the cladogram (Figure 32) presented here for clarity. The overall percent identity for the multiple alignment of the genogroup I sequences for both the study samples and accessions was 71.2%.

Overall, even with the addition of picobirnaviruses from the NCBI database, there appears to be little to no clustering by host species in picobirnaviruses (Figure 32 for genogroup I, Figure 33 for genogroup II). The study samples (noted with the black and brown circles) showed little to no clustering as expected. Several of the gorilla samples from the study clustered (clade a, Figure 32 and Figure 51). Similarly, several cattle samples from my study clustered together with another cattle/bovine sample from NCBI (clade b, Figure 32 and Figure 51). Only one small group of NCBI human and porcine picobirnaviruses clustered together (clade c, Figure 32 and Figure 51), but otherwise no other consistent clustering by host was seen with human and porcine or equine picobirnaviruses. There was also no suggestion of clustering by similarly related host species (seen in Figure 32 and Figure 51 by similarly coloured host species). There is also a grouping of study sequences (two cattle and one human) between the b and c clusters which could indicate potential spillover or at least geographical association, though this was not supported by the evidence described in Section 5.2.1.1 (Figure 32). A small quantity (8/124; 6%) of the nodes had branch support values that were greater than 0.8 in the genogroup I picobirnaviruses.

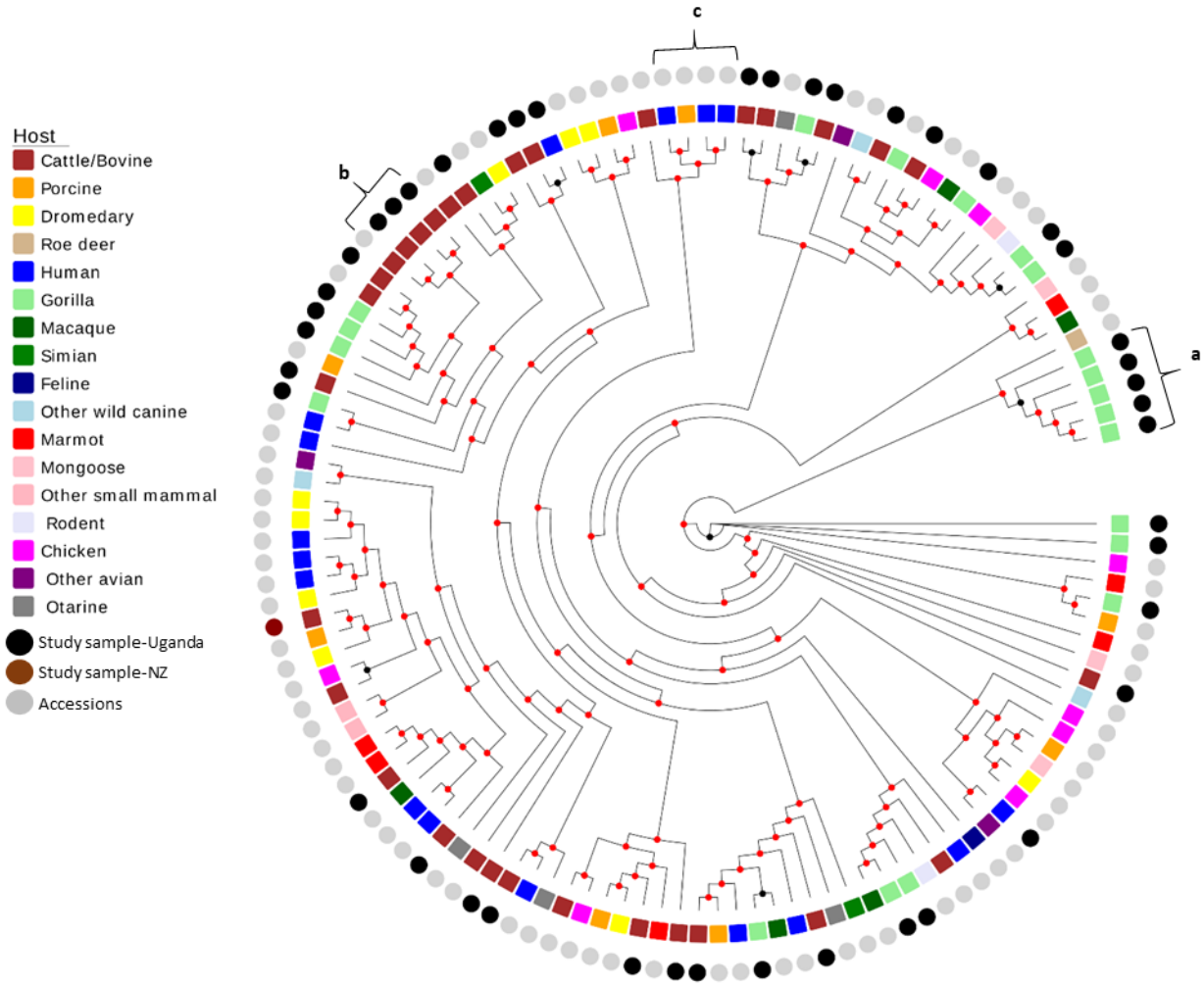


FIGURE 32 CLADOGRAM OF THE GENOGROUP I PICOBIRNAVIRUSES BY HOST

Circular cladogram from the multiple alignment of 126 amino acid sequences of *RdRp* genogroup I picobirnaviruses from this study and accessions from the NCBI database. Amino acid sequences trimmed to between 65-66 amino acids (Section 3.7.5.1). Branch lengths are uniform in the circular tree. Duplicates from the study within the same sample were excluded but possible duplicates from NCBI accessions top hits of picobirnaviruses were not excluded. Bracket designated with a is a cluster of gorilla samples from this study; bracket b is a cluster of cattle samples from this study and NCBI picobirnaviruses; bracket c is a cluster of NCBI human and porcine picobirnaviruses. Color-coding in first set of squares for the host species as noted in the legend; other wild canine: wolf, fox; other small mammal: rabbit; rodent: rat, murine or vole; other avian: shelduck, duck, turkey. Color-coding for the second set of circles: black: study samples from Uganda; brown: study sample from New Zealand; light grey: NCBI accession picobirnaviruses. Best tree model was selected from PhyML (Section 3.7.5.3) as LG + G + I. Fast-likelihood-based methods of aLRT SH-like were used for branch support values (Section 3.7.5.3). Black nodes denote branch support values >0.8 and red nodes <0.8.

5.1.1.2 Genogroup II picobirnaviruses

The evaluation of host clustering of the *RdRp* picobirnaviruses by host species was repeated with the genogroup II picobirnaviruses along with a selection of known genogroup II picobirnaviruses from the

NCBI database (Figure 33 and Figure 52). The genogroup II picobirnaviruses show moderate clustering based on host species though it is not consistent. The overall percent identity of the multiple alignment of the genogroup II sequences for both the study samples and accessions was 72.4%.

The phylogenetic tree and cladogram show four host clusters of two or more picobirnaviruses in similar host species (Figure 33 and Figure 52). One cluster of porcine and human picobirnaviruses can be seen (clade c, Figure 33 and Figure 52). There is evidence of minimal host clustering by the same host species with one cluster of 3 primate genogroup II picobirnavirus sequences (3 gorilla samples from my study, clade b, Figure 33 and Figure 52) and two human picobirnaviruses (1 human from my study and 1 human accession, clade a, Figure 33 and Figure 52). Minimal clustering by similarly-related host species is seen in the genogroup II picobirnaviruses with two clusters of livestock/ruminant genogroup II picobirnavirus sequences (clades d, Figure 33 and Figure 52). Approximately one third (21/56; 38%) of the nodes had branch support values that were greater than 0.8 in the genogroup II picobirnaviruses.

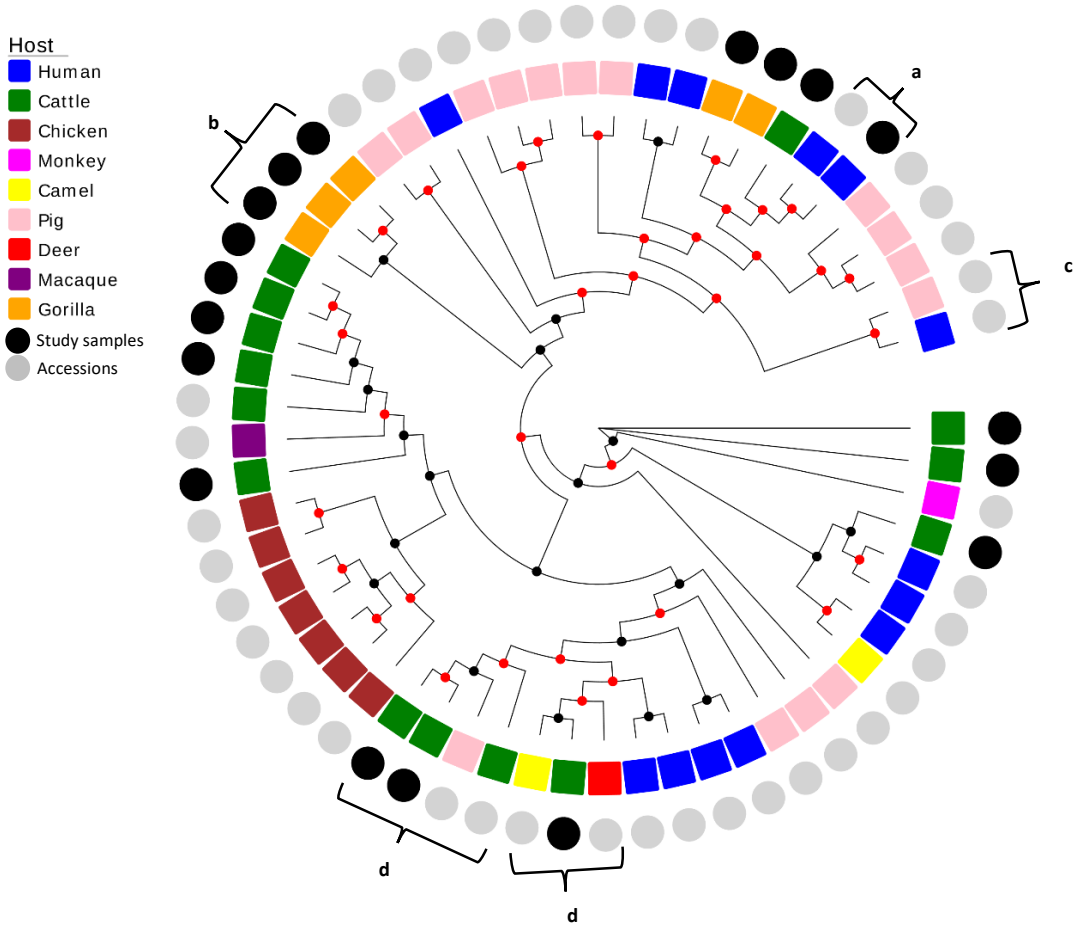


FIGURE 33 CLADOGRAM OF THE GENOGROUP II PICOBIRNAVIRUSES BY HOST

Circular cladogram from multiple alignment of 59 amino acid sequences of *RdRp* genogroup II picobirnaviruses from this study and a selection of genogroup II picobirnaviruses from the NCBI database. Amino acid sequences trimmed to between 120-130 amino acids (Section 3.7.5.1). Bracket designated with a include a human sample from this study with a human NCBI picobirnavirus sequence; bracket b is a cluster of gorilla samples from this study; bracket c is a cluster of NCBI human and porcine picobirnaviruses. Color-coding in first set of squares for the host species are designated in the legend. Color-coding for the second set of circles: black: study samples from Uganda; light grey: NCBI accession picobirnaviruses. Labels from the NCBI picobirnaviruses are host species/country/year collected/accession number. Labels for the samples from the study are: Ug# refers to the sample number from Uganda. G after the Ug# or NZ# refers to gorilla samples, H refers to human samples and C refers to cattle samples. S#, C# or M# at the third site on the labels designates S#=Sanger sequence, C#=cloned sequence, M#=metagenomic sequence. Labels for the NCBI PBVs are: Country of origin: USA=United States of America, UAE=United Arab Emirates/NCBI Accession number. Best tree model was selected from PhyML (Section 3.7.5.3) as WAG + G. Fast-likelihood-based methods of aLRT SH-like were used for branch support values (Section 3.7.5.3). Black nodes denote branch support values >0.8 and red nodes <0.8.

5.1.1.3 NEAR-COMPLETE PICOBIRNAVIRUSES

To explore if the host clustering analyses above were biased because of the short sequence lengths used, the near-complete (NC) picobirnavirus sequences generated in my study (both genogroup I, genogroup II, other genogroups), together with a selection of complete picobirnaviruses from the NCBI database (known genogroup I, genogroup II, other genogroups, unknown genogroups, Section 3.7.5.1 and Table 5, Chapter 3) were analysed. Phylogenetic analysis was performed to assess whether longer contig lengths may delineate grouping or clustering differently to those seen with shorter length contigs. Figure 34 (below) shows only minimal clustering by host species in a few small clades. Minimal clustering is noted of the same host species (clades a, Figure 34) and similarly-related host species (clades b, Figure 34). The percent identity of the multiple alignment of the NC picobirnaviruses from the study and the selection of complete picobirnavirus accessions was 45%. Most (69/96; 72%) of the nodes had branch support values that were greater than 0.8 in the NC picobirnaviruses.

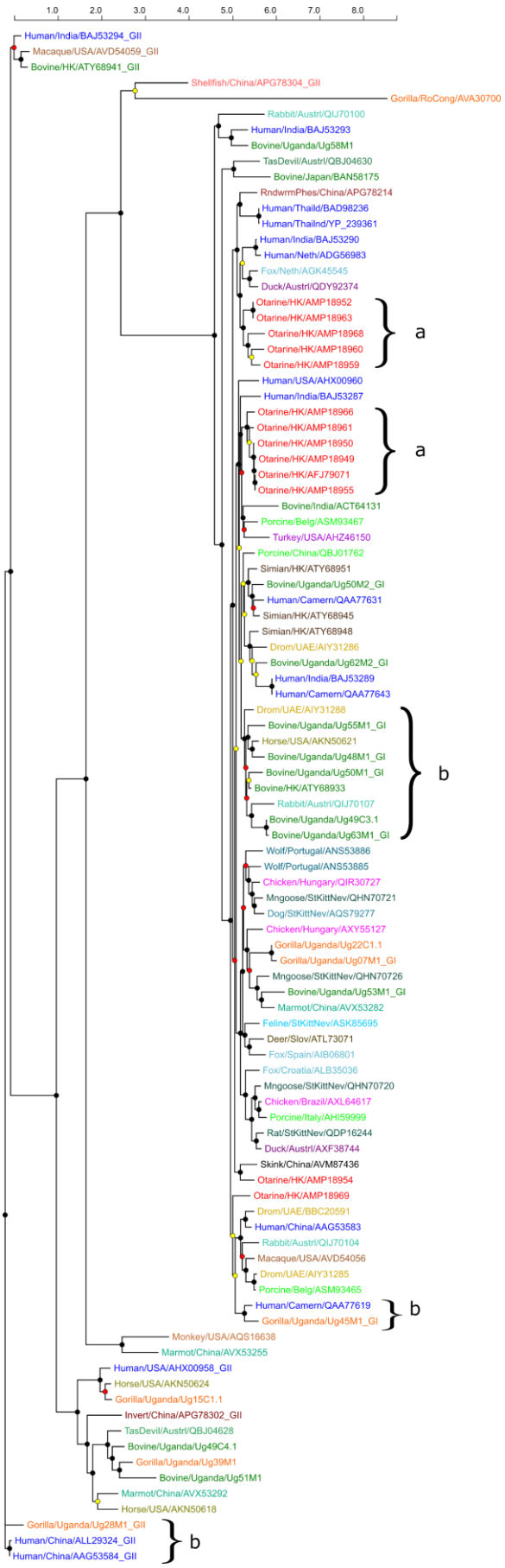

1.0
2.0
3.0
4.0
5.0
6.0
7.0
8.0
Human/India/BAJ53294_GII
Macaque/USA/AVD54059_GII
Bovine/HK/ATY68941_GII
Shellfish/China/APG78304_GII
Gorilla/RoCong/AVA30700
Rabbit/Austrl/QIJ70100
Human/India/BAJ53293
Bovine/Uganda/Ug58M1
TasDevil/Austrl/QBJ04630
Bovine/Japan/BAN58175
RndwrmPhes/China/APG78214
Human/Thaild/BAD98236
Human/Thailnd/YP_239361
Human/India/BAJ53290
Human/Neth/ADG56983
Fox/Neth/AGK45545
Duck/Austrl/QDY92374
Otarine/HK/AMP18952
Otarine/HK/AMP18963
Otarine/HK/AMP18968
Otarine/HK/AMP18960
Otarine/HK/AMP18959
a
Human/USA/AHX00960
Human/India/BAJ53287
Otarine/HK/AMP18966
Otarine/HK/AMP18961
Otarine/HK/AMP18950
Otarine/HK/AMP18949
Otarine/HK/AFJ79071
Otarine/HK/AMP18955
a
Bovine/India/ACT64131
Porcine/Belg/ASM93467
Turkey/USA/AHZ46150
Porcine/China/QBJ01762
Simian/HK/ATY68951
Bovine/Uganda/Ug50M2_GI
Human/Camern/QAA77631
Simian/HK/ATY68945
Simian/HK/ATY68948
Drom/UAE/AIY31286
Bovine/Uganda/Ug62M2_GI
Human/India/BAJ53289
Human/Camern/QAA77643
Drom/UAE/AIY31288
Bovine/Uganda/Ug55M1_GI
Horse/USA/AKN50621
Bovine/Uganda/Ug48M1_GI
Bovine/Uganda/Ug50M1_GI
Bovine/HK/ATY68933
Rabbit/Austrl/QIJ70107
Bovine/Uganda/Ug49C3.1
Bovine/Uganda/Ug63M1_GI
b
Wolf/Portugal/ANS53886
Wolf/Portugal/ANS53885
Chicken/Hungary/QIR30727
Mngoose/StKittNev/QHN70721
Dog/StKittNev/AQS79277
Chicken/Hungary/AXY55127
Gorilla/Uganda/Ug22C1.1
Gorilla/Uganda/Ug07M1_GI
Mngoose/StKittNev/QHN70726
Bovine/Uganda/Ug53M1_GI
Marmot/China/AVX53282
Feline/StKittNev/ASK85695
Deer/Slov/ATL73071
Fox/Spain/AIB06801
Fox/Croatia/ALB35036
Mngoose/StKittNev/QHN70720
Chicken/Brazil/AXL64617
Porcine/Italy/AHI59999
Rat/StKittNev/QDP16244
Duck/Austrl/AXF38744
Skink/China/AVM87436
Otarine/HK/AMP18954
Otarine/HK/AMP18969
Drom/UAE/BBC20591
Human/China/AAG53583
Rabbit/Austrl/QIJ70104
Macaque/USA/AVD54056
Drom/UAE/AIY31285
Porcine/Belg/ASM93465
Human/Camern/QAA77619
Gorilla/Uganda/Ug45M1_GI
b
Monkey/USA/AQS16638
Marmot/China/AVX53255
Human/USA/AHX00958_GII
Horse/USA/AKN50624
Gorilla/Uganda/Ug15C1.1
Invert/China/APG78302_GII
TasDevil/Austrl/QBJ04628
Bovine/Uganda/Ug49C4.1
Gorilla/Uganda/Ug39M1
Bovine/Uganda/Ug51M1
Marmot/China/AVX53292
Horse/USA/AKN50618
Gorilla/Uganda/Ug28M1_GII
Human/China/ALL29324_GII
Human/China/AAG53584_GII
b

FIGURE 34 PHYLOGENETIC TREE OF THE NEAR-COMPLETE PICOBIRNAVIRUSES BY HOST

Phylogenetic tree of 98 C/NC *RdRp* picobirnavirus amino acid sequences from the study and from a selection of complete to near complete *RdRp* picobirnavirus genomes from the NCBI database. Amino acid sequences were between 300-600 amino acids (Section 3.7.4). Best tree model was selected from PhyML (Section 3.7.5.3) as LG + G + I + F. Fast-likelihood-based methods of aLRT SH-like were used for branch support values (Section 3.7.5.3). Black nodes denote branch support values >0.8, yellow nodes 0.8 to 0.5 and red nodes <0.5. Brackets designated as a are clusters of the same host species; b brackets are clusters of similarly-related host species. Picobirnaviruses are color-coded by host: orange=gorilla, bright blue=human, magenta=porcine, yellow=Drom=dromedary, brick red=macaque, bright green=rabbit, grey=otarine, red=chicken, hot pink=marmot, purple/pink=Mngoose=mongoose, dark red=duck, purple=porcine, dark green=bovine, light red=turkey, brown=simian, dark blue=TasDevil=Tasmanian devil, light orange=monkey, aqua=horse, black=other (fox, roundworm from a pheasant, feline, deer, fox, dog, wolf, skink, shellfish, invertebrate)/Country of origin: Camern=Cameroon, Belg=Belgium, UAE=United Arab Emirate, USA=United States of America, Austrl=Australia, HK=Hong Kong, Thailnd=Thailand, Neth=Netherlands, StKittNev=Saint Kitts and Nevis island, Slov=Slovenia, RoCong=Democratic Republic of Congo/Accession number if from NCBI database or Ug# if from study sample with M#=metagenomic or C#.1=cloned sample with map-to-reference or clone of metagenomic reads. Lastly, samples will have GI or GII if known genogroup I or genogroup II picobirnaviruses, others are either unknown or another genogroup.

5.1.2 GEOGRAPHIC ASSOCIATIONS

As noted in Chapter 2, Section 2.5.2, the growing number of countries where picobirnaviruses have been identified has been increasing, especially in the last decade. Due to my findings, New Zealand has been added to the list and to the map of countries with identified picobirnaviruses (Figure 35). As mentioned previously, seventy-one picobirnavirus sequences from both Uganda and New Zealand have been submitted to the NCBI database, GenBank® accession numbers: MZ560630–MZ560700.

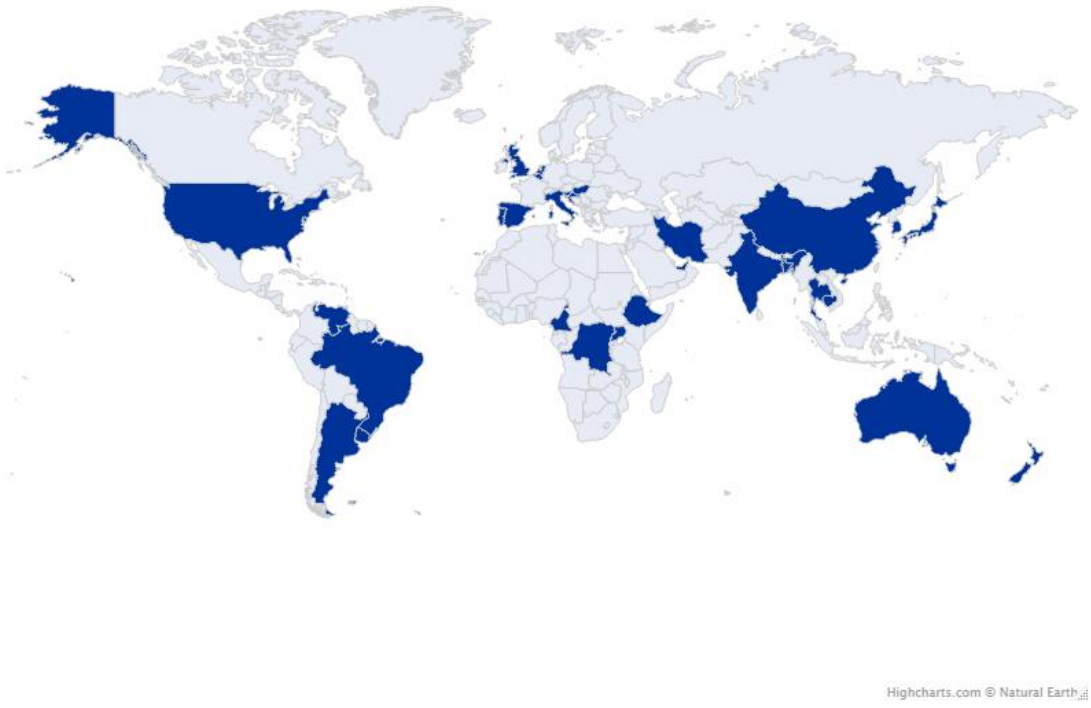


FIGURE 35 CURRENT GLOBAL PICOBIRNAVIRUS DETECTION

Updated world map of countries with identified or reported picobirnaviruses found in samples, typically faecal or gastrointestinal samples but also includes respiratory samples, wastewater and river samples. Picobirnaviruses were from countries identified from NCBI database of picobirnavirus nucleotide sequences from either segment 1 and/or segment 2 and partial or complete genomes.

5.1.2.1 GENOGROUP I PICOBIRNAVIRUSES

The evaluation of geographic clustering of the *RdRp* picobirnavirus sequences by geographic location was with the use of the *RdRp* genogroup I picobirnaviruses (and then genogroup II picobirnaviruses separately, see below), representing the samples and picobirnavirus accessions from NCBI (Section 3.7.5.1 and Table 5, Chapter 3). The picobirnaviruses were designated based on country or geographic region and also on whether they were a study sample or picobirnavirus from NCBI (Figure 36 and Figure 53). Even with the addition of picobirnavirus sequences from the NCBI database to increase the number and diversity of the picobirnaviruses in the various countries or regions, there appears to be little to no clustering by geography (Figure 36 and Figure 53). The percent identity of the multiple alignment of the genogroup I sequences for both the study samples and accessions was 71.2%. The samples from Uganda and New Zealand were dispersed throughout the phylogenetic tree with limited (clade a; clades b, Figure 36) to no geographic clustering (Figure 36 and Figure 53).

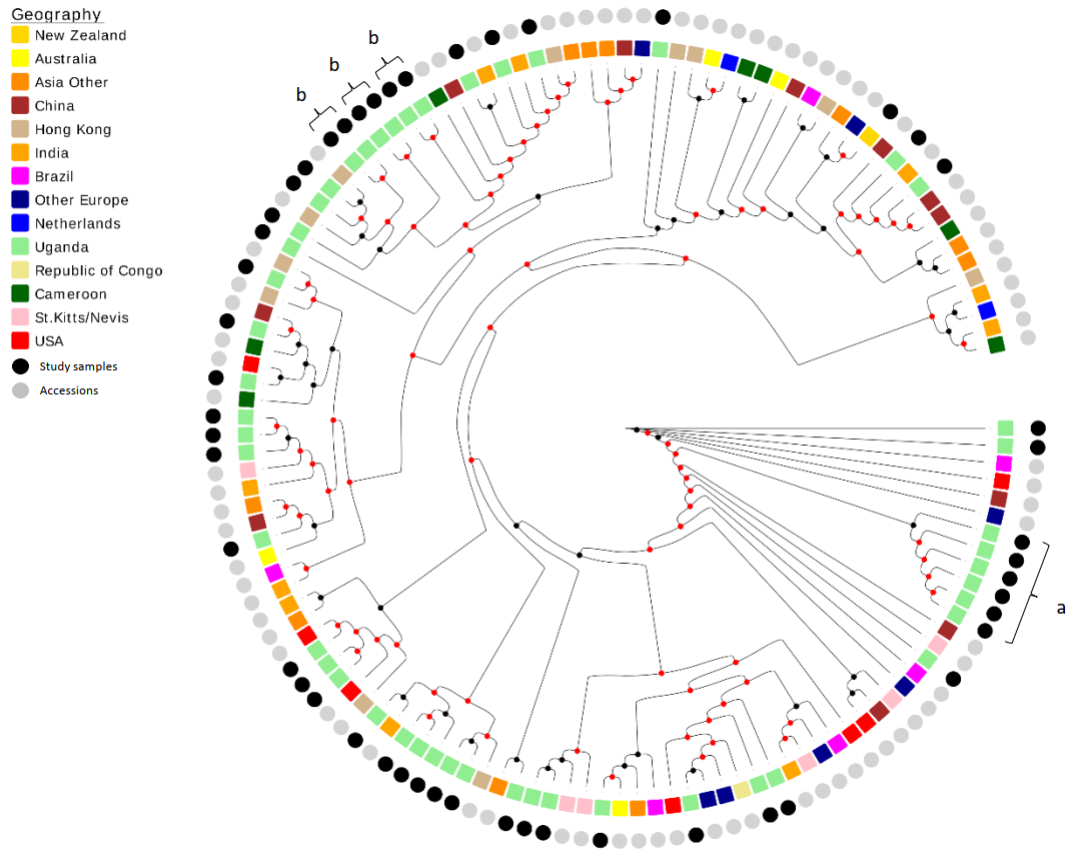


FIGURE 36 CLADOGRAM OF THE GENOGROUP I PICOBIRNAVIRUSES BY GEOGRAPHY

Amino acid *RdRp* genogroup I picobirnaviruses from study and accessions. Amino acid sequences trimmed to between 65-66 amino acids (Section 3.7.5.1). Branch lengths are uniform in the circular tree. Best tree model was selected from PhyML (Section 3.7.5.3) as LG + G + I. Fast-likelihood-based methods of aLRT SH-like were used for branch support values (Section 3.7.5.3). Black nodes denote branch support values >0.8 and red nodes <0.8. Bracket designated with a is a larger cluster of Ugandan samples from this study; brackets b are smaller clusters of Ugandan samples from this study. Colour-coding in first set of squares for the geographical regions are indicated in the legend with additional information: yellow/brown for Asia/Oceania Pacific region; blue for Europe; green for Africa; pink/red for Americas. Color-coding for the second set of circles: black: study samples from Uganda and New Zealand; light grey: NCBI accession picobirnaviruses.

5.1.2.2 GENOGROUP II PICOBIRNAVIRUSES

The evaluation of geographic clustering of the genogroup II picobirnaviruses was repeated with the genogroup II *RdRp* gene sequences using 18 study sequences and 41 known genogroup II viruses from the NCBI database (Section 3.7.5.1 and Table 5, Chapter 3). The percent identity of the multiple alignment of the genogroup II sequences was 72.4%. Overall, minimal to no evidence was noted of clustering based on geography. Two small geographic clusters were noted from Uganda (this study, clades a) a larger one from Brazil (clade b) and one small clade from China (clade c), all from the same study (Figure 37 and Figure 54).

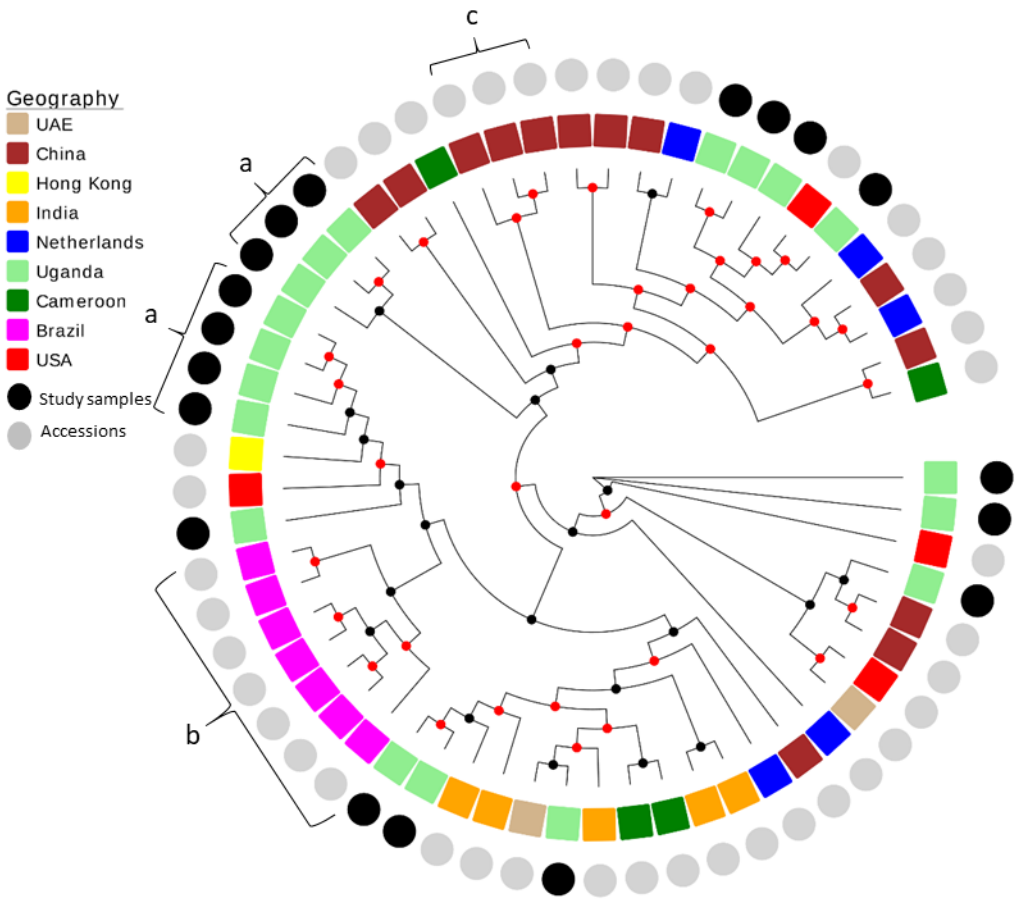


FIGURE 37 CLADOGRAM OF THE GENOGROUP II PICOBIRNAVIRUSES BY GEOGRAPHY

Cladogram of *RdRp* genogroup II picobirnaviruses from study and a selection of genogroup II picobirnaviruses from the NCBI database from various geographical regions. Amino acid sequences trimmed to between 120-130 amino acids (Section 3.7.5.1). Best tree model was selected from PhyML (Section 3.7.5.3) as WAG + G. Fast-likelihood-based methods of aLRT SH-like were used for branch support values (Section 3.7.5.3). Black nodes denote branch support values >0.8 and red nodes <0.8. Brackets designated with a include two small clusters from Uganda; bracket b is a larger cluster from Brazil; bracket c is a smaller cluster from China. Color-coding in first set of squares for the geographical regions are indicated in the legend with additional information: yellow/brown for Asia/Oceania Pacific region; blue for Europe; green for Africa; pink/red for Americas. Color-coding for the second set of circles: black: study samples from Uganda and New Zealand; light grey: NCBI accession picobirnaviruses. Labels from the NCBI picobirnaviruses are host species/country/year collected/accession number. Labels for the samples from the study are: Ug# refers to the sample number from Uganda. G after the Ug# or NZ# refers to gorilla samples, H refers to human samples and C refers to cattle samples. S#, C# or M# at the third site on the labels designates S#=Sanger sequence, C#=cloned sequence, M#=metagenomic sequence. Labels for the NCBI PBVs are: Country of origin: USA=United States of America, UAE=United Arab Emirates/NCBI Accession number.

The NC picobirnavirus sequences (genogroup I, genogroup II, other genogroups) with a selection of complete picobirnaviruses from the NCBI database (genogroup I, genogroup II, other genogroups,

unknown genogroups) were also evaluated for geographic clustering to assess if longer contig lengths may delineate grouping not seen with shorter length contig analyses. Figure 38 (below) shows only minimal to no clustering by geography or geographical region. In a couple of clades, there was minimal clustering with the largest group from Hong Kong though these were all from the same study (clade a, Figure 38) with a small clade from east Asia (clade b). The percent identity of the multiple alignment of the C/NC picobirnaviruses from the study and the selection of complete picobirnavirus accessions was 57.3%. Overall, there does not appear to be much if any geographical clustering with the use of longer contigs.

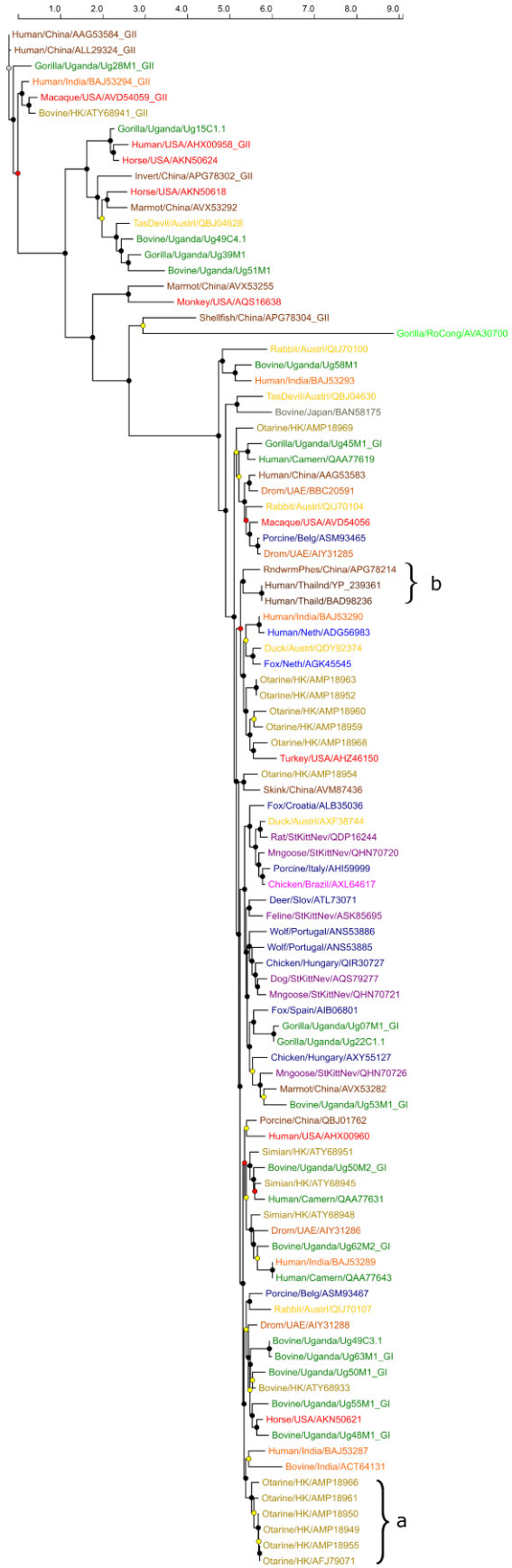


FIGURE 38 PHYLOGENETIC TREE OF THE NEAR-COMPLETE PICOBIRNAVIRUS GENOMES BY GEOGRAPHY

Phylogenetic tree (ladderised) for the geographic distribution of near-complete picobirnavirus genome sequences from the study and accessions from NCBI. Amino acid sequences were between 300-600 amino acids (Section 3.7.4). Best tree model was selected from PhyML (Section 3.7.5.3) as LG + G + I + F. Fast-likelihood-based methods of aLRT SH-like were used for branch support values (Section 3.7.5.3). Black nodes denote branch support values >0.8, yellow nodes 0.8 to 0.5 and red nodes <0.5. Bracket designated as a is a cluster from Hong Kong from the same study; b bracket is a small cluster from east Asia. Color-coding by geography and geographical region: green=Africa (dark green: Cameroon or Camern, light green: Uganda, olive green: Democratic Republic of Congo or RoCong); brown/orange=East Asia (dark brown: China, light brown: Hong Kong or HK, orange: India, dark orange: United Arab Emirates or UAE, burnt orange: Thailand or Thailnd); red/purple=Americas (magenta: Brazil, purple: St. Kitts and Nevis or StKittNev, red: United States of America or USA); blue for Europe (dark blue: other European, blue: Netherlands or Neth); yellow=Australia or Austrl; dark grey=Japan. Accessions are labelled as: host species/country/accession +/- genogroup I or II or not known. Samples are labelled similarly but will be in red as: host species/Country/Sample # as Ug# with sequence type and # as C for clone, M for metagenomic and S for Sanger sequence, if decimal in number, designates map to reference with clone sequence used as reference and metagenomic reads mapped to extend the contig or sequence length. All sequences >900 base pairs in length.

5.2 Discussion

The effect of various factors such as host species, geographical location and genetic relatedness of host species on the clustering of picobirnaviruses was evaluated. Prior studies have shown mixed results for picobirnaviruses clustering based on either host species or geographical location [119, 120, 150, 151, 154, 155, 160, 163, 176, 194, 195]. Overall, the picobirnaviruses clustered minimally to moderately by host, though it was likely related to clustering of picobirnaviruses from a particular study more so than true host clustering. The host clustering was not altered or even improved with the use of longer sequences, suggesting these were robust results. The picobirnaviruses clustered even less by geographical region with no noticeable difference noted between the different genogroups or the different-length sequences.

5.2.1 Host associations

There was minimal host clustering or grouping in my samples with the included picobirnavirus genomes from NCBI, which supports most other studies in the observation of inconsistent, weak or no evidence for host clustering in picobirnaviruses [120, 132, 151, 154, 155, 160, 176, 194, 195, 197]. I did find that there were multiple porcine accessions that were the highest hit in the BLASTn for some of the picobirnaviruses found in my human and gorilla samples. In the genogroup I host analysis (Figure 32 and Figure 51), only one small group of human and porcine picobirnaviruses clustered together in the overall host phylogenetic tree (Figure 32 and Figure 51, clade c), though this cluster did not include my study samples. The reason that the phylogenetic tree and cladogram did not include some top hit porcine NCBI database accession sequences by BLASTn were due to the accession sequences not meeting the criteria in Sections 3.7.2.2 and 3.7.5.1. In the genogroup II host analysis, 3 clusters of porcine and human picobirnaviruses were noted (Figure 33 and Figure 52, clade c), which did include my study samples; otherwise no other consistent clustering was seen with human and

porcine picobirnaviruses. The limited clustering of porcine and human picobirnaviruses did not strongly support the prior studies showing similarities between porcine and human picobirnaviruses [119, 120, 150, 154, 196]. The identification of RNA viruses that are similar between porcine and human hosts is not unusual as similar influenza A viruses have been identified between the hosts, considered a single gene pool for the virus [178, 277]. When simply visualising associations, using similar colour schemes to evaluate host associations based on taxonomic-relatedness, clustering based on similar host species did not improve the strength of the clustering by host species. Additionally, evaluation of host clustering with longer contigs did not improve or reveal stronger host clustering of the *RdRp* picobirnavirus sequences.

Weak clustering based on host was noted in my samples (Chapter 4, Figure 29 and Figure 30), though the human samples did not cluster together but within other cattle picobirnavirus samples. This may indicate that similar picobirnaviruses may be more likely to occur through cross-species transmission and potential increased contact between the humans and cattle though this cannot be supported based on the currently available evidence for this study. In some instances, despite the moderate clustering by host in this tree, there remains almost identical picobirnaviruses between hosts (Ug57_C_M1 and Ug03_H_S1; Ug43_G_S1 and Ug54_C_M2). This is an important finding which provides further support for possible spillover of this RNA virus. In a group as diverse and under-sampled as picobirnaviruses are, it is less likely that close matches would be identified between viruses. The identification of the same sequence in different hosts within my system of study suggests spillover. Other studies have also detected picobirnavirus sequences with very high similarities or even identical sequences from different host species [119, 142, 163, 198]. Four studies [119, 197, 198, 278] that identified picobirnaviruses in pigs and one study in monkeys [118] found homologous picobirnavirus sequences among different pigs and monkeys, respectively, in the same environment, suggesting animal-to-animal transmission. Another study that detected a *Picobirnavirus* in a horse in India, identified a homologous picobirnavirus sequence from a human also in the same region, though no further information was given on the potential for close host contact or environmental sharing [163]. Additionally, studies detecting picobirnaviruses in wastewater samples have also identified homologous picobirnaviruses sequences from the same study system though true host identification from wastewater samples cannot be confirmed nor can cross-species transmission [121, 142]. Similarly to these other studies of same species transmission, the identification of the homologous picobirnavirus sequences I found among the gorillas may be from gorilla-to-gorilla transmission either through direct contact or through environmental contamination.

Limited studies have shown homologous picobirnaviruses among host species in the same study

system. A virome study from Cameroon found human picobirnaviruses from their pooled human faecal samples that were 99% identical to a Cameroonian bat strain picobirna-like virus from the pooled bat samples from the same region [279]. Giordano et al. (2011) similarly studied human and porcine picobirnaviruses and detected a nucleotide identity between 95–100% between porcine and human picobirnaviruses from the same region [150]. In contrast to my study though, the human samples were from immunocompromised (kidney-transplant) patients from the hospital in Cordoba, Argentina, while the pigs were from the outskirts of the city in breeding farms with no known contact between the hosts. Though other studies have shown homologous picobirnavirus sequences between different host species in the same region with unknown or variable contact rates, my work appears to be the first report of homologous picobirnavirus sequences between hosts with high contact potential in the same study system.

Host clustering was stronger in the genogroup II picobirnaviruses as compared to the genogroup I picobirnaviruses (Chapter 4, Figure 29 and Figure 30; Figure 51 and Figure 52). This could be due to the lessened diversity of genogroup II picobirnaviruses as compared to genogroup I picobirnaviruses overall [117, 119, 131, 175, 176], which may help to show more host clustering due to decreased diversity of these sequences. This is not supported, though, in the similarity of the picobirnaviruses (percent identity within the multiple alignments) when comparing genogroup I to genogroup II picobirnaviruses. For example, the overall percent identity of the shorter fragments for the Uganda and New Zealand genogroup I samples was 85.2% as compared to the genogroup II at 67.5%; the Uganda and New Zealand samples and NCBI database viruses had an overall percent identity for the genogroup I of 71.2% and for the genogroup II of 72.4% over the smaller *RdRp* fragment. The presence of more host clustering may be due to fewer genogroup II picobirnaviruses in the database which could result in an under-sampling of genogroup II picobirnaviruses in an already hyper-diverse group of viruses. Also, the host clustering, as noted above, was more evident when looking at my samples alone as compared to my samples and the database picobirnaviruses, perhaps suggesting sampling bias and highlighting the caution needed when interpreting sequence results from single studies. I may expect the overall similarity of the genogroup II sequences to be higher as compared to the genogroup I sequences based on the finding of less diversity with the genogroup II in some studies [117, 119, 131, 175, 176]. Not all studies, including my study, detect less diverse genogroup II picobirnaviruses as compared to genogroup I picobirnaviruses. I also evaluated the genogroup I picobirnaviruses for host clustering by amino acid sequences rather than nucleotide sequences to remove the effect of silent mutations and lessen the diversity. Based on Chapter 4, there were some alterations noted on evaluation of nucleotide versus amino acid genogroup I picobirnavirus sequences so this may have adjusted the clustering or grouping of the picobirnaviruses.

5.2.2 GEOGRAPHIC ASSOCIATIONS

I did, for the first time, find picobirnaviruses in New Zealand domestic animals, which had not been previously reported [210]; I also did find many picobirnaviruses in Uganda, in which the sample collections date to 2014, with only three reports of picobirnaviruses in Africa in the last five years [93, 98, 200]. I also evaluated the *RdRp* sequences for clustering or grouping based on geography. The genogroup I phylogenetic analysis showed minimal to no clustering by geography when compared with a global dataset with the study samples distributed throughout the phylogenetic tree (Figure 53). Similarly, the genogroup II picobirnaviruses did not show any real evidence of geographical associations, though the study samples did cluster together a bit more in the genogroup II phylogenetic analysis (Figure 54). Additionally, the use of longer contigs also did not show any geographical clustering in the phylogenetic analysis of the NC picobirnaviruses.

As noted previously (Chapter 3: Materials and Methods), I did use the segment 2, *RdRp* segment of the picobirnaviruses for evaluation of clustering by host species or geography. The segment 2, *RdRp* segment has less diversity than segment 1 (the capsid protein), making it a potentially better choice for the evaluation of clustering due to the extremely high sequence diversity seen in picobirnaviruses [98, 131, 173]. I did not evaluate capsid sequences of the picobirnaviruses due to the higher diversity, along with my identification of only short (<150 bp) capsid fragments/sequences that mapped to multiple reference capsid sequences with no consistency in positioning on the ORF. Beyond this initial identification, no further work was done with the capsid sequences. I am then making the assumption that the use of the segment 1 of the picobirnavirus or the whole genome from my samples would not show clustering based on host species and/or geography. I did not perform these analyses, though, so I cannot confirm this statement. I did, though, see minimal clustering by geography in the genogroup II picobirnaviruses, though the clustering that was seen were mainly sequences from the same study. The clustering may not actually be from geographic associations, but more a matter of clustering based on similar study environments.

5.2.3 LIMITATIONS

It is possible that by excluding the NCBI database sequences with the 10% similarity threshold (Chapter 3, Section 3.7.5.1), I removed potential host or geographic clustering groups. This pre-analysis step was performed to decrease the overall number of sequences in the dataset for the multiple alignments and provide representative sequences for clustering analyses. In many of these exclusions, it was more common to exclude highly similar sequences (>98–99% sequence similarities) that were submitted from the same study (same host and same geographical region, typically) and may or may not have been excluded as duplicate sequences prior to submission. Additionally, the exclusion criteria resulted in 278 sequences from the initial 375 sequences and presumed to still provide a sufficient

range of host and geographical regions for clustering analyses. As I noted in the phylogenetic trees, potentially inaccurate host or geographical clustering was noted from within study sequence similarities which was expected to be lessened with the exclusion criteria. Furthermore, sampling bias due to limited or heterogenous collection of samples or viruses could result in inaccurate conclusions on host or geographical associations [111, 178]. The support for many nodes in the trees was also lower than 0.8, especially within the phylogenetic analyses of the genogroup I picobirnaviruses (Figure 32). However, even when near complete and complete genomes were used (Figure 34) and node support was >0.8 for 72% of the nodes, there was still approximately one third of the nodes that had branch support values below the cut-off. The low branch support values, then, especially within the genogroup I picobirnaviruses, could question the results of the phylogenetic analyses. In actuality, I can only report on how total picobirnavirus diversity related to my samples and further work would be required to evaluate for the current total picobirnavirus diversity and the impact on associations, which is beyond the scope of this work.

Additional considerations for the lack of association with geography or host could be due to mixed infections due to high within-host variability. Viruses or pathogens with high mutation rates, such as RNA viruses, evolve within the host due to *in vivo* bottlenecks and purifying selection [179]. The mutation rate of picobirnaviruses has been reported on the higher end of the range for RNA viruses (substitution rates of 4×10^{-3} to 1.4×10^{-2} substitutions per site per year) [142, 276]. Additionally, recombination or more likely reassortment for segmented viruses may result in further viral variability and diversity. Evidence of reassortment of segmented RNA viruses such as picobirnaviruses has been reported, which could contribute even further to the genetic variability of the virus and challenge in identification of factor associations [125, 155, 160, 163]. I did not evaluate for recombination in the picobirnaviruses identified in this study due to the high diversity of the viruses and the challenges associated with identification of similar genetic regions. I also did not evaluate for reassortment in the picobirnaviruses due to evaluating only the segment 2 of the virus (see Section 6.2.1 for further discussion).

In epidemiological and ecological contexts, RNA virus diversity in regard to geographic clustering can be described by phylogeographic patterns [111]. These patterns may help to explain the presence, or lack, of clustering of RNA viruses by geographical regions due to host movements and movement constraints. As Holmes states, “RNA viruses exhibit rapid evolutionary dynamics and shallow genetic diversity which means that their phylogeography is shaped by the movement and growth of human populations” [111]. Based on these assumptions, picobirnaviruses would appear to have no clear spatial structure, rather than a pattern of geographic association that could be linked to a source or

higher population densities. Furthermore, some viruses cluster based on host species, or geographical region, in phylogenetic analyses that may indicate evolutionary processes or transmission dynamics which include the possibility of cross-species transmission [111].

5.3 SUMMARY

In conclusion, I did not find strong host or geographical associations of picobirnaviruses with my samples, nor with the addition of picobirnaviruses from the NCBI database. My findings support the null hypothesis of no host or geographic structure among the various picobirnavirus sequences analysed. More importantly though, I did find homologous sequences (Chapter 4, Section 4.2.5) in different host species, which may be stronger support for cross-species transmission considering the lack of host or geographic associations.

CHAPTER 6: PICOBIRNAVIRUS WITHIN-HOST DIVERSITY

6.1 RESULTS

6.1.1 MULTIPLE PICOBIRNAVIRUSES WITHIN SAMPLES

Cloning of PCR products was performed to further investigate initial Sanger sequences that showed multiple peaks on the chromatograms, those being suggestive of multiple picobirnaviruses within the same sample (Chapter 3, Section 3.6.2). The aim was to identify the presence of multiple, genetically-distinct picobirnaviruses and to use some of these as reference sequences to assemble longer contigs from metagenomic reads (Chapter 3, Section 3.7.4).

Specifically, three gorilla samples (Ug15, Ug22, Ug43) and four cattle samples (Ug49, Ug55, Ug59 and Ug60) from Uganda were cloned (Chapter 3, Sections 3.6.1, 3.6.2, 3.7.2). In addition to the picobirnaviruses identified through the cloning, I also identified and included the initial conventional PCR and Sanger sequenced picobirnavirus sequences from these three gorilla and four cattle samples, along with NZC01 from any primer sets (25F/43R, 23F/24R, F3/F5/R5/R8, and the designed primers). I also identified and included metagenomic picobirnavirus sequences from these samples. Metagenomic reads were mapped to reference/clone (Section 3.6.2 and Section 3.7.2.2) to generate longer contigs; additionally, metagenomic reads not utilised in the map to reference were *de novo* assembled into contigs as distinct picobirnaviruses.

Collectively, the gorilla samples, Ug15, Ug22 and Ug43 possessed two, two and four different picobirnaviruses, respectively; the cattle samples, Ug49, Ug55, Ug59 and Ug60 possessed seven, six, two and eight different picobirnaviruses, respectively. The total number of different picobirnaviruses for all samples was 31 (see Table 8 below). On average, the gorilla samples had 2.7 different picobirnaviruses per sample and the cattle samples had 5.8 different picobirnaviruses per sample. The clones were the most successful in identifying picobirnaviruses with 21 identified and the initial PCR with Sanger sequencing and metagenomics each identified 6 and 5 picobirnaviruses, respectively. One picobirnavirus in cattle sample Ug59 was identified through the initial Sanger sequencing and also through the cloning; each method extended the contig in different directions so the contig was combined with the addition of metagenomic reads that mapped to the clone and extended the contig further (Ug59_C_CS1.1). The primers used to identify the 31 different picobirnaviruses from these select samples included F3/F5/R5/R8 (13 samples), 25F/43R (11 samples), 23F/24R (4 samples) and a designed primer, DP1358R (1 sample). Most (26/31, 84%) of the picobirnaviruses had conserved regions identified in the *RdRp* sequences, with the conserved domains found twice as commonly (24 times) as the conserved polymerase motifs (13 times). The most common conserved domain was CD1 (16 times), and the most common polymerase motif was PMA (12 times). Sequence lengths ranged

from 158 bp to 1525 bp with a median length of 578 bp. Based on the primers used, the picobirnaviruses from the samples would be a range of genogroup I, genogroup II and other genogroups, but only 6 of the 31 picobirnaviruses had identified and annotated primer sets on the sequences (all genogroup I).

BLASTn search was performed on the picobirnavirus sequences for the top hit among the NCBI database accessions. Thirteen different types of picobirnaviruses were found to be the highest match to the samples with the most common match of bovine, human and wastewater picobirnaviruses (see Table 8 below). The gorilla samples matched most closely to chicken, human, marmot, microtus, wastewater, porcine and macaque picobirnaviruses in the database; the cattle samples matched to bovine, ovine, wastewater, human, dromedary, chicken, porcine, Tasmanian devil, macaque, genet and horse picobirnaviruses from the database. The range of maximum bit scores for the matches ranged from 72 to 778 with a median of 253; the range of query cover ranged from 28 to 100% with a median of 94%; and the range of pairwise identity to the matches ranged from 67 to 98% with a mean of 79%.

TABLE 8 SUMMARY OF THE MULTIPLE PICOBIRNAVIRUSES PER SAMPLE

Sample	Species	Method	Sequence name	Primers used	Sequence length	Conserved regions	Accession highest match	Accession description	Bit score	Query cover	Pairwise ID
Ug15	Gorilla	C1	Ug15_C1/G1.1	F3/F5/R5/R8	795	D3	MH425585	Chicken PBV	778	99	82
Ug15	Gorilla	C2.1	Ug15_C2.1/G1.2	25F/43R	1086	PMG,PMF,PMA,PMB,D1,D2	MH933801	Human PBV	561	96	72.11
Ug22	Gorilla	C1.1	Ug22_C1.1/G2.1	F3/F5/R5/R8	1427	D2,D3	KY928700	Marmot PBV	390	75	69.1
Ug22	Gorilla	C2.1	Ug22_C2.1/G2.2	25F/43R	589		JF755420	Microtus PBV	190	49	74.5
Ug43	Gorilla	C1	Ug43_C1/G3.1	25F/43R	204	PMA,D1	KJ135922	Wastewater PBV	253	98	88
Ug43	Gorilla	C2	Ug43_C2/G3.2	25F/43R	208	D1	KJ135902	Wastewater PBV	190	96	81
Ug43	Gorilla	S1	Ug43_S1/G3.3	25F/43R	195	D1	MK378853	Porcine PBV	203	94	85
Ug43	Gorilla	S2	Ug43_S2/G3.4	F3/F5/R5/R8	675	D3	MG010912	Macaque PBV	656	94	82.9
Ug49	Cattle	C1	Ug49_C1/C1.1	25F/43R	204	PMA,D1	MF693847	Ovine PBV	259	97	88.9
Ug49	Cattle	C2	Ug49_C2/C1.2	25F/43R	221	D1	KJ135902	Wastewater PBV	153	90	77.5
Ug49	Cattle	C3.1	Ug49_C3.1/C1.3	F3/F5/R5/R8	1367	PMG,PMF,PMA,PMB,PMC,D1,D2,D3	KY120170	Bovine PBV	626	93	70.9
Ug49	Cattle	C4.1	Ug49_C4.1/C1.4	F3/F5/R5/R8	1525	D3	KR827415	Human PBV	696	91	71.8
Ug49	Cattle	C5	Ug49_C5/C1.5	25F/43R	206	PMA,D1	MN196313	Bovine PBV	255	97	88.1
Ug49	Cattle	M2	Ug49_M2/C1.6	NA/GI	954	PMF,PMA,PMB,PMC,D1,D2,D3	LC338004	Dromedary PBV	721	95	80.3
Ug49	Cattle	M3	Ug49_M3/C1.7	NA/GI	967	D1,D2	MG846408	Chicken PBV	242	54	71.5
Ug55	Cattle	C1	Ug55_C1/C2.1	23F/24R	490		KC841460	Porcine PBV	627	75	97.6
Ug55	Cattle	C2	Ug55_C2/C2.2	F3/F5/R5/R8	933	D3	MH933810	Human PBV	288	87	68.5
Ug55	Cattle	C3	Ug55_C3/C2.3	23F/24R; F3/F5/R5/R8	578	PMA,D1,D2	MN196315	Bovine PBV	187	55	73.3
Ug55	Cattle	M1	Ug55_M1/C2.4	NA/GI	1363	PMF,PMA,PMC	KY120170	Bovine PBV	682	98	74.2
Ug55	Cattle	S1	Ug55_S1/C2.5	25F/43R	188	D1	MF693849	Ovine PBV	187	100	81.9
Ug55	Cattle	S2	Ugg55_S2/C2.6	DP1358R	158	PMF	KY120170	Bovine PBV	153	67	91.6
Ug59	Cattle	C2	Ug59_C2/C3.1	23F/24R	205	PMA,D1	KJ135922	Wastewater PBV	252	97	87.5
Ug59	Cattle	CS1.1	Ug59_CS1.1/C3.2	23F/24R; F3/F5/R5/R8	852		KY120178	Bovine PBV	550	99	91.8
Ug60	Cattle	C1	Ug60_C1/C4.1	F3/F5/R5/R8	236		KR827416	Human PBV	123	90	73.2
Ug60	Cattle	C2	Ug60_C2/C4.2	F3/F5/R5/R8	335	D3	MK521924	Tas Devil PBV	244	97	76.9
Ug60	Cattle	C4.1	Ug60_C4.1/C4.3	F3/F5/R5/R8	715	D3	KM573809	Dromedary PBV	149	73	66.7
Ug60	Cattle	C5.1	Ug60_C5.1/C4.4	F3/F5/R5/R8	288		MG010917	Macaque PBV	71.6	32	76.6
Ug60	Cattle	C6.1	Ug60_C6.1/C4.5	F3/F5/R5/R8	885	D3	MH425586	Chicken PBV	178	39	72.4
Ug60	Cattle	M1	Ug60_M1/C4.6	NA/GI	942	PMF,PMA,D1	KF823812	Genet PBV	204	28	77
Ug60	Cattle	M2	Ug60_M2/C4.7	NA/GI	540	PMF,PMA,PMB,PMC,D1,D2,D3	KY120176	Bovine PBV	438	99	78.4
Ug60	Cattle	S1	Ug60_S1/C4.8	25F/43R	209	PMA,D1	GU230508	Horse PBV	280	95	91
NZC01	Cattle	C1	NZC01_C1	25F/43R	203		MH835431	Goat PBV	234	100	88.3

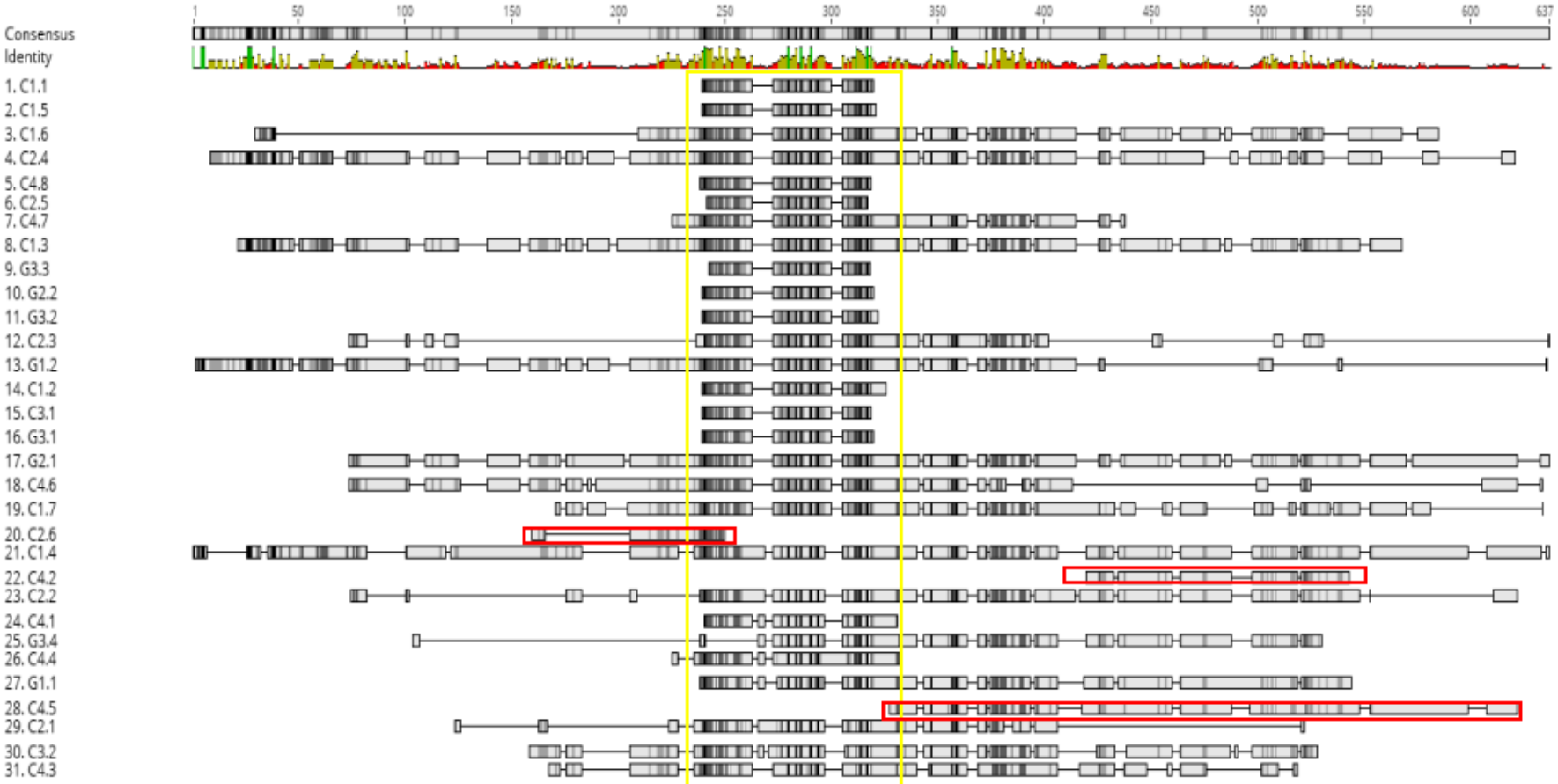
Ug: Uganda, NZ: New Zealand. Method: S=Sanger sequencing from RT-PCR; C=Cloned and then Sanger sequencing from RT-PCR; M=Metagenomics. Sequence name Ug# or NZ# followed by S#, C# or M#; After forward slash is the renamed sequence shown in the phylogenetic tree below (FIGURE 40). Primers: 25F/43R for *RdRp* GI, 23F/24R for *RdRp* GII from Rosen et al. 2000; F3/F5/R5/R8 primers uses all four primers for identification of any genogroup PBV from Anthony et al. 2015; DP is designed primer; If metagenomic sequence, primers will be not applicable (NA) though may find GI or GII primers amino acids in sequence so annotated. Seq length is length of the sequence in base pairs or bp. Conserved regions: PMA=Polymerase motif A; PMB=Polymerase motif B; PMC=Polymerase motif C; PMD=Polymerase motif D; PME=Polymerase motif E; PMF=Polymerase motif F; PMG=Polymerase motif G; D1=Conserved domain 1;

D2=Conserved domain 2; D3=Conserved domain 3. Accession numbers are reported from the NCBI database based on the highest BLASTn match. Bit score is “derived from the raw alignment score, taking the statistical properties of the scoring system into account.” E-value is representative of “the number of different alignments with scores equivalent to or better than is expected to occur in a database search by chance.” Query cover is the amount of nucleotides or amino acids that the subject shares with the input sequence, expressed as a percentage of the total number of nucleotides or amino acids in the query. Pairwise ID is pairwise identity and is the “extent to which two (nucleotide or amino acid) sequences have the same residues at the same positions in an alignment, expressed as a percentage.” [280]

6.1.2 WITHIN-HOST DIVERSITY

In order to evaluate the within-host diversity of the range of picobirnaviruses in the seven selected samples, a phylogenetic analysis was performed. Four of the seven samples were from cattle (designated C1.# –C4.#) and three from gorilla (designated G1.# –G3.#); numbers beyond the decimal designates the number of different picobirnaviruses in that sample. The multiple alignment of the seven samples (Figure 39a, red rectangles) revealed three cattle sequences (C2.6: Ug55_S2, C4.2: Ug60_C2 and C4.5: Ug60_C6.1) that did not overlap with the other sequences. These three sequences were excluded and the other sequences were trimmed (Figure 39b, yellow region) and a second multiple alignment was performed (Figure 39).

A)


1
50
100
150
200
250
300
350
400
450
500
550
600
637
Consensus
Identity
1. C1.1
2. C1.5
3. C1.6
4. C2.4
5. C4.8
6. C2.5
7. C4.7
8. C1.3
9. G3.3
10. G2.2
11. G3.2
12. C2.3
13. G1.2
14. C1.2
15. C3.1
16. G3.1
17. G2.1
18. C4.6
19. C1.7
20. C2.6
21. C1.4
22. C4.2
23. C2.2
24. C4.1
25. G3.4
26. C4.4
27. G1.1
28. C4.5
29. C2.1
30. C3.2
31. C4.3

B)

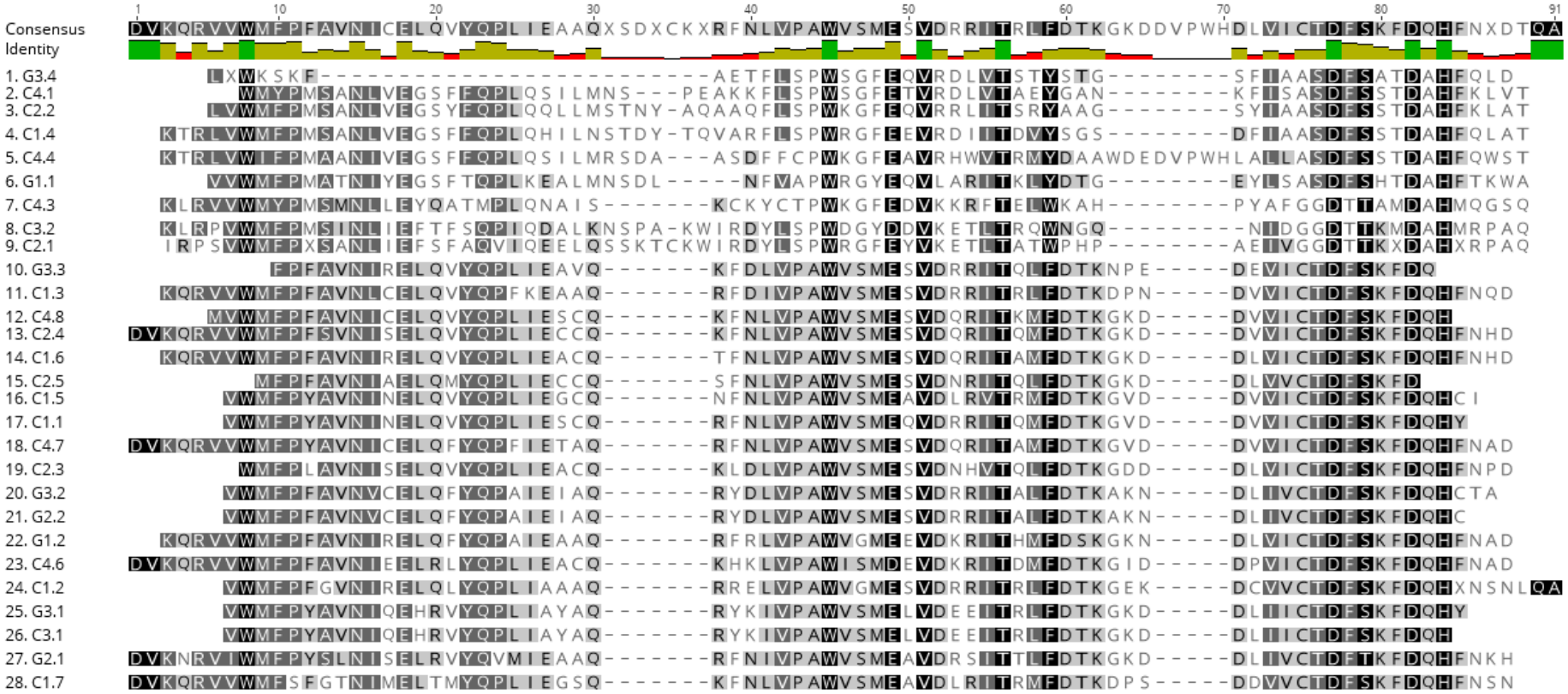


FIGURE 39 MULTIPLE ALIGNMENT OF THE PICOBIRNAVIRUS SEQUENCES ANALYSED FOR WITHIN HOST DIVERSITY

A) All 31 picobirnavirus sequences showing the diversity, length and alignments of all of the sequences. B) Multiple alignment of 28 trimmed sequences used in the phylogenetic analysis and heatmap below. Four of the seven samples were from cattle (designated C1.#-C4.#) and three from gorilla (designated G1.#-G3.#); numbers beyond the decimal designates the number of different picobirnaviruses in that sample (i.e.. G1.1 and G1.2 has two different picobirnaviruses identified from the same gorilla sample; C1.1-C1.7 has seven different picobirnaviruses identified from the same cattle sample

A phylogenetic tree and heatmap show the similarities and differences of the picobirnavirus amino acid sequences between- and within- the same hosts. The heatmap (Figure 40 on right) shows percent identity between the picobirnaviruses with the higher similarities (more similar picobirnaviruses) in red/orange, mid-range similarities in yellow and low similarities in blue. The phylogenetic tree shows two major clusters with three sequences outside of those clusters (C3.2, C2.1, and C4.3, Figure 40). It can be seen in the heatmap of the percent identity that those three sequences have a lower percent identity to all of the other sequences, with the highest similarity of 45 to 62% (not including themselves); the average percent identity of these three separate sequences was 27.6%. Interestingly, there are two sets of two sequences from different samples that are 100% identical, G3.1 (Ug43_C1) to C3.1 (Ug59_C2) and G2.2 (Ug22_C2) to G3.2 (Ug43_C2). The gorilla samples, G2.2 and G3.2, are from the same species of host but different individual hosts of gorillas. The first set, G3.1 and C3.1, are from a cattle and gorilla sample.

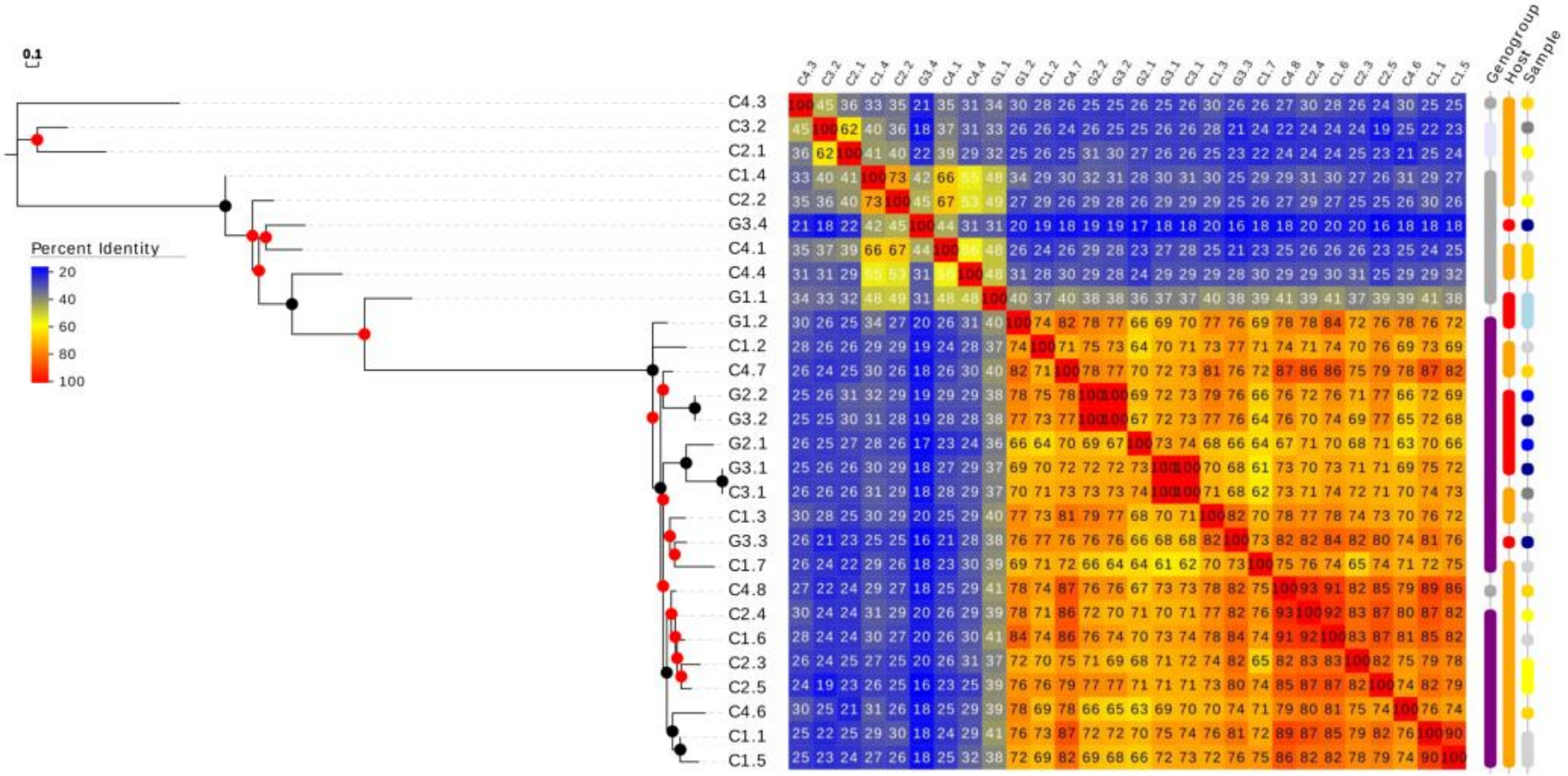


FIGURE 40 PHYLOGENETIC TREE AND HEATMAP OF THE MULTIPLE PICOBIRNAVIRUS AMINO ACID SEQUENCES IDENTIFIED IN SEVEN SAMPLES FROM UGANDA

Tree and heatmap show the similarities and differences of the picobirnaviruses between and within the same host. Amino acid sequences were trimmed to between 74-91 amino acids in length. The heatmap shows percent identity (percent similarity) between the picobirnaviruses with the higher similarities in red/orange, mid-range similarities in yellow and low similarities in blue with the unit of the numbers as percentages. Four of the seven samples were from cattle (designated C1.#-C4.#), three from gorilla (designated G1.#-G3.#); numbers beyond the decimal designates the number of different picobirnaviruses in that sample (i.e. G1.1 and G1.2 has two different picobirnaviruses identified from the same gorilla sample; C1.1-C1.7 has seven different picobirnaviruses identified from the same cattle sample). The column plots on the far right of the heatmap show genogroup, sample and host clustering. Color-coding for the genogroups: lavender=GII, purple=GI, grey=unknown. Color-coding for the hosts: orange=cattle; red=gorilla. Color-coding for samples: light grey=C1=Ug49; yellow=C2=Ug55; grey=C3=Ug59; gold=C4=Ug60; light blue=G1=Ug15; blue=G2=Ug22; dark blue=G3=Ug43. Selected tree model was LG. Branch support values: black nodes=>0.8, red nodes=<0.8

Based on this heatmap, the least similar sequences in each sample are Ug15_C1 (G1.1), Ug43_S2 (G3.4), Ug49_C4.1 (C1.4), Ug55_C1 (C2.1), Ug59_CS1.1 (C3.2), Ug60_C4.1 (C4.3), Ug55_C2 (C2.2), Ug60_C1 (C4.1), Ug60_C5.1 (C4.4). The most similar sequences in each sample are Ug15_C2.1 (G1.2), Ug22_C2.1 (G2.2), Ug43_C1 (G3.1), Ug49_C1 or C2 or C5 (C1.1, 1.2 or 1.5), Ug55_S1 (C2.5), Ug60_S1 or M1 or M2 (C4.6, 4.7 or 4.8).

The column plots on the far right of the heatmap show genogroup, sample and host clustering, or lack thereof (Figure 40). There is weak to moderate clustering based on host species with two to four gorilla sequences clustering into two groups; the other single gorilla sequences are isolated and scattered. The cattle samples show moderate clustering although they are broken up by the gorilla samples. There does not appear to be any clustering based on sample. The known genogroups did cluster into mainly genogroup I for the major bottom clade and some known genogroup II in the top sequences with the remainder unknown in regard to genogroup designations.

6.2 Discussion

6.2.1 Multiple intra-host picobirnaviruses

I identified multiple picobirnaviruses in the seven samples from Uganda. With the focus on a selection of gorilla and cattle samples that were positive for picobirnaviruses from multiple different primer sets or showed evidence of multiple peaks on the sequence chromatogram, I was able to identify between two and eight different picobirnaviruses in each sample. The cattle samples had evidence of higher quantities of picobirnaviruses as compared to the gorilla samples and this was also noted in the number of metagenomic reads identified as picobirnavirus between the gorilla and cattle samples. The metagenomic reads, though, did not contribute solely to these increased quantities in the cattle as I noted that the cloning resulted in the highest identification of picobirnaviruses in these select samples. This is not surprising as I selected samples that were likely to have multiple picobirnaviruses present and also isolated and propagated up to 10 colonies for cloning in an attempt to identify as many different picobirnaviruses in the samples as possible (Chapter 3, Section 3.6.2). The metagenomic reads did result in some distinct and separate picobirnaviruses in the samples but were of more benefit in mapping the reads to already known picobirnaviruses from the initial conventional PCR and Sanger sequencing, or the clones to create longer and more robust contigs of the picobirnaviruses.

I was not the first to identify multiple picobirnaviruses in the same sample. Many of the other studies found between two and eight picobirnaviruses per sample, commenting that this presents the possibility for recombination and/or reassortment with multiple picobirnaviruses within the same host/individual [123, 133, 152, 155, 160, 174, 175]. As noted, I also found between two and eight picobirnaviruses within the same individual, but I also found up to eight picobirnaviruses in the same individual (Ug49 and Ug55) in some cases. This has not been reported as frequently, though in a recent study Boros et al. showed 8 different picobirnaviruses in one sample, with both segment 1s and segment 2s [133]. I did not focus on segment 1 of the picobirnaviruses, despite finding evidence of their presence in the metagenomic data, because the assembled contigs were short in length and provided little phylogenetic insight (see Chapter 5, Section 5.2.2). Capsid reads, then, may not have been detected as often as segment 2 from the *RdRp* gene due to a higher number of *Picobirnavirus RdRp* genes in the NCBI database. In fact, over half of the picobirnavirus segment 1, capsid, gene sequences in the database identified from one study in marmots [174]. In addition, the reported marked diversity of segment 1 may generate more challenges in identification of these segments of the picobirnaviruses [98, 131, 173, 174]. I did, though, identify many of the conserved regions in segment 2 of the picobirnaviruses, with over half of those sequences with multiple conserved regions.

6.2.2 Within-host genogroup diversity

I attempted to find different genogroups in the same individual by using multiple different primer sets and/or using the four primer set that was not genogroup-specific for cloning [153]. Genogroup designation was challenging as the translated genogroup I and genogroup II primers were not commonly present in the amino acid sequence in all six reading frames using translation table_1 (Chapter 3, Section 3.7.3). Moreover, even though many of the sequences were identified with the genogroup I or genogroup II primers, this does not necessarily result in obtaining a genogroup I or genogroup II picobirnavirus. Instead, the identification of these primers in the sequence, the high matching of the sequence with a known genogroup I or genogroup II picobirnavirus on the NCBI database and/or the clustering of the genogroup I or genogroup II picobirnaviruses in the phylogenetic analyses would be better at confirming the genogroup designation. Though this was noted in the complete to near-complete picobirnaviruses described in Chapters 4 and 5, clustering of the picobirnaviruses by genogroup did show the distinct major clade of the genogroup I picobirnaviruses (Figure 40) and the unknown and genogroup II picobirnaviruses outside of the genogroup I clade (Figure 40). One exception is C4.8 which was originally classified as a genogroup I picobirnavirus due to the detection of the virus with the genogroup I primers (25F/43R), but it was then subsequently classified as unknown as it did not fulfil the remaining criteria, with the exception of clustering with the genogroup I picobirnaviruses. It is then questionable whether it is a genogroup I picobirnavirus or unknown genogroup though likely genogroup 1. The other unknown picobirnaviruses that did not cluster within the genogroup I clade or with the only known genogroup II picobirnaviruses likely would be other genogroups though they did not cluster together into distinct clades and are therefore, impossible to classify with the information I have at this stage. I did, though, suspect that many samples did contain multiple picobirnaviruses from different genogroups which is supported by the clades in the phylogenetic analysis. I attempted to find different genogroups in the same individual by using multiple different primer sets and/or using the four primer set that was not genogroup-specific for cloning [153]. This finding of different genogroups in the same sample is not unusual with one study finding different genogroup picobirnaviruses in the same sample of a human with diarrhoea [173].

6.2.3 Within-host diversity

The finding of multiple picobirnaviruses in the same sample highlights the variability of this dsRNA virus. The percent identity shown within the heatmap (Figure 40) illustrates an extensive range of diversity of these picobirnaviruses with minimal to moderate clustering based on host species (Chapter 5). The small number of samples assessed means it is hard to evaluate the association, though as seen in Chapter 5, even with larger numbers of sequences, separating them into distinct genogroups or evaluating longer length sequences, host clustering was not strong. Additionally, there

was no clustering based on the sample or individual despite expectations that within the same host, the picobirnaviruses may be more similar. The similarity of the picobirnaviruses within the same individual ranged from very low at 16% to moderate at 72% for the gorilla sample (G3) Ug43 to between 29% to 90% for the cattle sample Ug49 with seven picobirnavirus sequences (Appendix D, Table 19). Some sequences were from different parts of the *RdRp* gene and therefore had 0% identity with other picobirnaviruses in the same sample such as within the cattle sample Ug60; these samples were excluded in the subsequent analyses due to lack of overlap on the multiple alignment making comparisons inaccurate.

In the Anthony et al. 2015 study where the four primer set (F3/F5/R5/R8) was used for identification of picobirnaviruses, the authors attempted to predict changes in viral communities either through stochastic, as compared to deterministic, processes [153]. In this referenced study, "the maximum observed percent identity between any 2 picobirnavirus sequences found in the same individual was 85.8% for genogroup I and 88.7% for genogroup II; in contrast, the maximum percent identity for any 2 non-identical sequences found in different individuals at the same site was 99.8% for both genogroup I and genogroup II….suggesting that deterministic mechanisms do exist to limit the co-occurrence of closely related viruses in the same animal….including virus: virus interactions such as competitive exclusion or virus: host interactions like immune recognition" [153]. My study found that the maximum percent identity between picobirnaviruses in the same gorilla sample was 76% and the maximum percent identity between picobirnaviruses in the same cattle sample was 87–93% for the cattle samples with 6–8 sequences in the same genogroup. I also found that the maximum percent identity of picobirnaviruses between individuals was high (100%) as in G3.1 (gorilla sample Ug43) and C3.1 (cattle sample Ug59) along with G2.2 and G3.2 (gorilla samples Ug15 and Ug22, respectively).

6.2.4 Additional limitations

Cloning was the most successful method for identifying different picobirnaviruses with 21 identified and the initial PCR with Sanger sequencing and metagenomics each identified 6 and 5 picobirnaviruses, respectively. Due to time and financial constraints, I was not able to clone all of the samples for evaluation of the overall quantity and diversity of picobirnaviruses. I also did not perform deep sequencing of the samples in which I may have identified higher quantities of picobirnaviruses, higher quality of the assembled picobirnaviruses and possibly longer sequences or more complete to near-complete picobirnaviruses. This may have resulted in an even higher diversity within the picobirnaviruses than I did find, especially increasing the within-host diversity. Additionally, known components of sequencing that result in sequencing errors such as within PCR amplification and mapping of reads or read assembly could also lead to inaccurate conclusions of diversity within the viruses [281]. I also did not evaluate for reassortment of the bi-segmented RNA virus which has been

reported in other picobirnavirus studies [125, 174] and the potential contribution of these processes on the diversity of my picobirnaviruses.

6.3 SUMMARY

I identified multiple picobirnaviruses within the same sample as expected and also an extensive range of diversity of the picobirnaviruses, even within the same host. Despite the diversity, I still identified picobirnaviruses that were identical between hosts, raising the question about cross-species transmission of this dsRNA virus. Before I can answer this question of cross-species transmission, though, I need to delve into the questions about the actual host of picobirnavirus that have plagued our understanding of this virus from the first day of its discovery.

CHAPTER 7: WHAT IS THE HOST OF PICOBIRNAVIRUS? TESTING POSSIBLE PROTOZOA OR PROKARYOTIC HOSTS

7.1 RESULTS

Due to the extensive diversity of picobirnaviruses, the lack of any clustering by host or geography and the continual debate on the true host of this virus, including reported Shine-Dalgarno sequences [220], I attempted to identify picobirnaviruses from bacterial and protozoal hosts. In addition, I attempted to identify motifs and perform further genetic analyses to clarify the possible host of this virus.

7.1.1 BACTERIAL AND PROTOZOAL HOSTS

I identified a number of amplicons in the conventional PCR with the use of the three picobirnavirus primer sets: 11 bacterial species from 30 bacterial samples; 4 *Cryptosporidium* from 10 samples; and 3 *Giardia* from 10 total samples (Table 9 and Table 20). PCR amplicons that were similar in size to the expected picobirnaviruses were obtained from the 4 primer set (F3/F5/R5/R8) and 25F/43R [117, 153]; PCR amplicons were obtained for the *Cryptosporidium* samples were from all three of the *Picobirnavirus* primer sets including 23F/24R [117] (Figure 41, Figure 42, Figure 43); and similar PCR amplicons were obtained for the *Giardia* were only from the Rosen primers (23F/24R and 25F/43R). Sanger sequencing was performed on the amplicons with the bacterial samples resulting in sequence lengths ranging from 165–621 bp with a mean length of 376 bp; sequence lengths for the *Cryptosporidium* samples ranged from 148–378 bp with a mean length of 253 bp; sequence lengths for the *Giardia* samples ranged from 64–476 bp with a mean length of 280 bp (Table 9). Sequencing results from the amplicons from the bacterial samples were compared by BLASTn to the NCBI database and revealed different bacterial species but no evidence of picobirnavirus (Table 9). Sequencing results from the amplicons from the *Giardia* samples were also compared on the database and either resulted in no matches or bacterial species but no picobirnavirus (Table 9). Amplicons from the *Cryptosporidium*-derived sequences were also searched for through BLASTn and identified sequences from 2 *Cryptosporidium* species (*C. hominis* and *C. parvum*), a *Penicillium* bacteria and a Human picobirnavirus (Table 9). The match to the *Picobirnavirus* was high with a query cover of 100% and percent identity of 100% for the *Human picobirnavirus* AB214978, characterised as a *Human picobirnavirus* pseudogene for RNA-dependent RNA polymerase (accession length: 193 bp) from a diarrheic stool from a study of human picobirnaviruses from children in Kolkata, India [282]. Comparison of the picobirnavirus from the Cryptosporidium showed dissimilarity with <15% of the amino acids similar to other picobirnaviruses from my study on a multiple alignment.

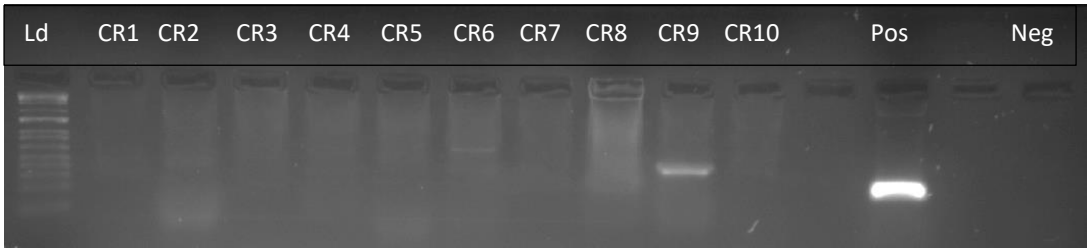


FIGURE 41 GEL ELECTROPHORESIS OF *CRYPTOSPORIDIUM* SAMPLES WITH GENOGROUP I PRIMERS

10 *Cryptosporidium* samples tested with picobirnavirus *RdRp* genogroup I primers (25F43R from Rosen et al. 2000) [117]. CR: *Cryptosporidium*, Pos: positive control, Neg: negative control, Ld: 1 kb standard ladder.

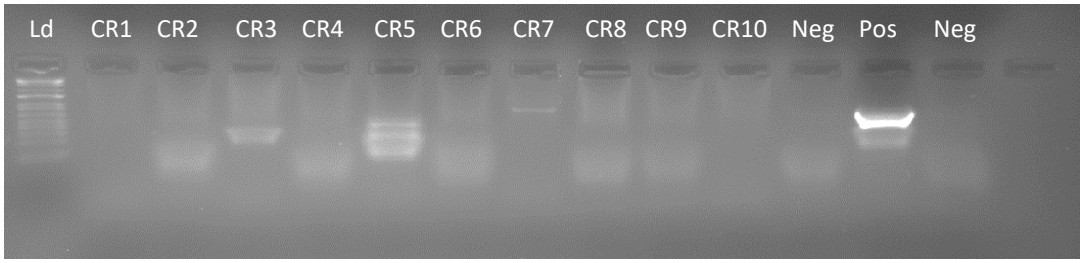


FIGURE 42 GEL ELECTROPHORESIS OF *CRYPOTOSPORIDIUM* SAMPLES WITH GENOGROUP II PRIMERS

10 *Cryptosporidium* samples tested with picobirnavirus *RdRp* genogroup II primers (23F24R from Rosen et al. 2000) [117].

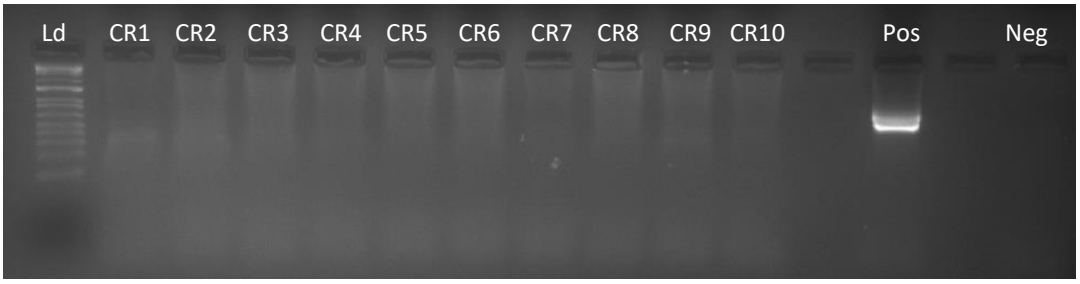


FIGURE 43 GEL ELECTROPHORESIS OF *CRYPTOSPORIDIUM* SAMPLES WITH ADDITIONAL PRIMERS

10 *Cryptosporidium* samples tested with picobirnavirus four primers (F3/F5/R5/R8 from Anthony et al. 2015) [153]. CR: *Cryptosporidium*, Pos: positive control, Neg: negative control, Ld: 1 kb standard ladder.

TABLE 9 PCR RESULTS OF THE BACTERIA AND PROTOZOA TESTED FOR PICOBIRNAVIRUS RNA

Sample	Isolate/Sample source	Primer set	Contig length	Accession number	Description of accession	Bit score	E-value
B2	Unknown/YEP 75	25F/43R	564	CP007156	*Corynebacterium falsenii*	577	1.7e-160
B2	Unknown/YEP 74	F3/F5/R5/R8	462	CP011312	*Corynebacterium kutscheri*	457	4e-124
B13	*C. petrophilum*/Unknown	25F/43R	238	LS483309	*Staphylococcus aureus*	228	2.6e-55
B15	Unknown/Unknown	F3/F5/R5/R8	165	LR213458	*Shigella sonnei*	250	1.1e-62
B17	*E. coli*/Unknown	F3/F5/R5/R8	441	LR213458	*Shigella sonnei*	324	1.9e-84
B18	Unknown/Unknown	F3/F5/R5/R8	250	CP033242	*Klebsiella pneumonia*	243	3e-60
B20	Unknown/Human	F3/F5/R5/R8	621	LR213458	*Shigella sonnei*	308	2.7e-79
B22	Unknown/Unknown	F3/F5/R5/R8	283	CP033242	*Klebsiella pneumonia*	254	1.6e-63
B23	Unknown/Unknown	F3/F5/R5/R8	317	CP031321	*Escherichia coli* Es_ST2350	93	4e-15
B30	Unknown/YEP 76	25F/43R	397	CP026501	*Corynebacterium pseudotuberculosis*	401	2.3e-107

B30	Unknown/YEP 76	F3/F5/R5/R8	397	No matches			
CR1	*C. hominis*/Human	F3/F5/R5/R8	276	KY882333	*Cryptosporidium hominis*	156	2.2e-34
CR3	*C. hominis*/Human	23F/24R	148	**AB214978**	***Human picobirnavirus***	268	5.1e-68
CR5	*C. parvum*/Human	23F/24R	211	KU530219	*Penicillium polonicum*	339	5e-89
CR9	C. unknown/Human	25F/43R	378	AF040725	*Cryptosporidium parvum*	579	6.3e-161
GD4	*G. intestinalis*/Human	23F/24R	476	AP019189	*Escherichia coli*	731	0
GD5	*G. intestinalis*/Human	25F/43R	299	CP010537	*Cupriavidus basilensis*	139	8.5e-29
GD7	*G. intestinalis*/Human	25F/.43R	64	No matches			

Bacteria were named based on the sample name B for bacteria and # of the samples 1–30; *Cryptosporidium* were named based on the sample name CR for *Cryptosporidium* and # of the samples 1–10; and *Giardia* were named based on the sample name GD for *Giardia* and # of the samples 1–10. Isolate is the genus and species of the bacteria or protozoa, if known; sample source is the host of the sample, if known (e.g., YEP = Yellow-eyed penguin, Human). All three primer sets were evaluated on all samples (see Chapter 3, 3.6.1.1) with 25F/43R for the GI picobirnaviruses, 23F/24R for the genogroup II picobirnaviruses [117] and the four primers [153] for any genogroup picobirnavirus but only the positive results and primers are listed in the table [117, 153]. Contig length is the length of the sequence in base pairs or nucleotides from the conventional PCR and Sanger sequencing. Accession number and description of the accession are based upon BLASTn on the NCBI database done through Geneious® blast search; maximum bit score and e-value are also based on the highest blast hit for each sequence.

7.1.2 Open-Reading Frame identification and genetic code translation

Near-complete picobirnavirus *RdRp* genogroup I sequences from both Uganda and New Zealand were evaluated for the presence of ORFs with the use of the standard genetic code (SGC), translation table 1, and an alternative genetic code (AGC), translation table 4, identified as the genetic code for mould protozoan mitochondrial [93, 195] (Chapter 3, Section 3.7.3). Nucleotide sequences were translated into protein sequences either using SGC and AGC and co-phylogenetic analysis was performed to compare the effect of the use of different genetic codes (Figure 44).

Co-phylogenetic evaluation revealed a few amino acid sequences that shifted clades or clusters by altering the genetic code but, overall, the use of the AGC as compared to the SGC only mildly altered the clustering of the sequences. It is also important to note that stop codons did exist in some of the translated amino acid sequences from the SGC whereas no stop codons were identified after translation into the AGC. Stop codons were edited to 'X' as an unknown amino acid in protein sequences in the standard genetic code (3 samples), and this was not required when translating in the AGC as previously noted. Sequences suspected to be translated with the AGC based on the presence of stop codons in the SGC are identified in Figure 44 with rectangles. Those with green rectangles have altered the tree topology, but not shifted clades, while those with yellow rectangles have altered the tree topology and shifted clades between the use of the SGC and AGC. Additionally, no obvious clustering by host is noted with the use of the different genetic codes; no evaluation of geographical clustering can be made with the use of these samples only.

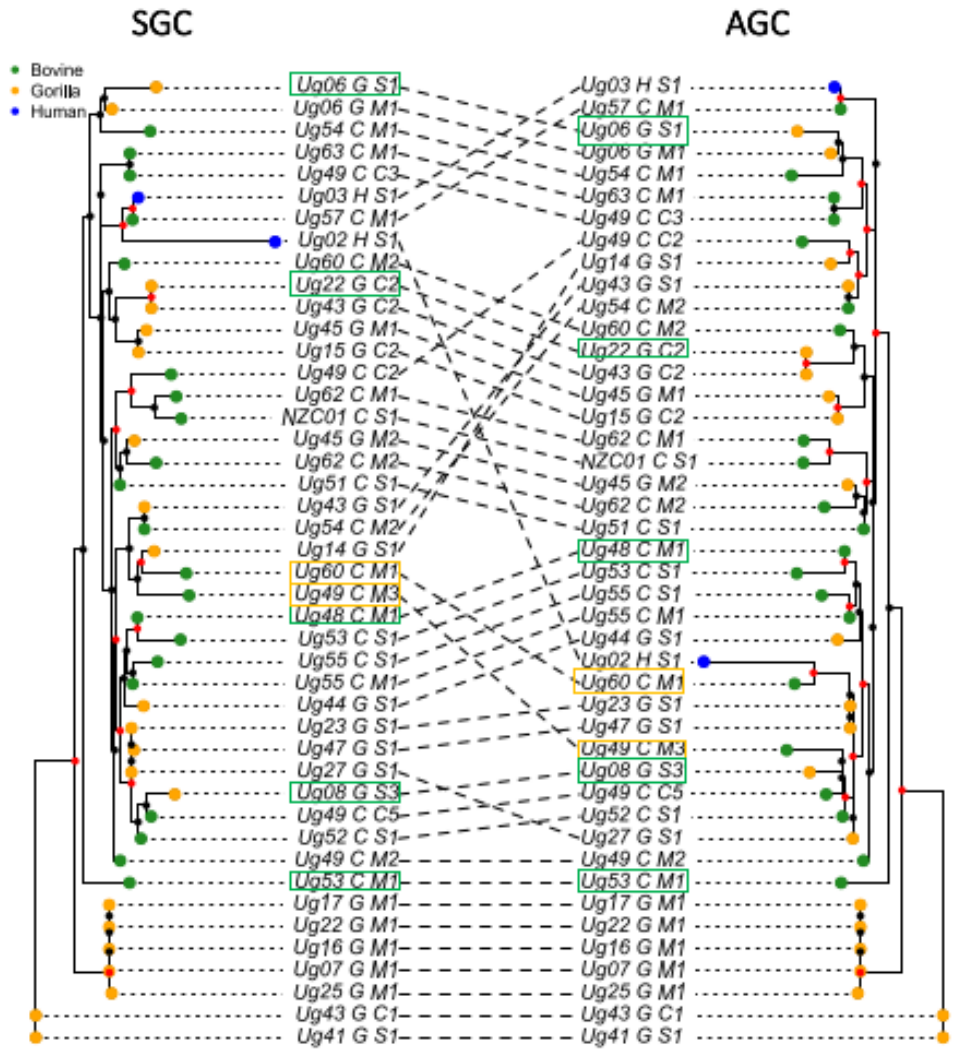


FIGURE 44 CO-PHYLOGENY OF AMINO ACID SEQUENCES OBTAINED USING TWO DIFFERENT GENETIC CODES FOR PICOBIRNAVIURSES FROM STUDY SAMPLES

Picobirnavirus *RdRp* genogroup I amino acid sequences from this study only from both Uganda and New Zealand comparing an alternative genetic code. Amino acid sequences trimmed to between 65-66 amino acids (Section 3.7.5.1). Amino acid sequences from the standard genetic code (SGC) or translation table_1 are on the left-hand side of the co-phylogeny trees and amino acid sequences from an alternative genetic code (AGC) for mould and protozoal mitochondria or translation table_4 are on the right-hand side of the co-phylogeny trees. Green rectangles show sequences that have altered the tree topology but not shifted clades; yellow rectangles show sequences that have altered the tree topology and shifted clades. Best tree model was selected from PhyML (Section 3.7.5.3) as LG + G**.** Fast-likelihood-based methods of aLRT SH-like were used for branch support values (Section 3.7.5.3). Black nodes denote branch support values >0.8 and red nodes <0.8. Tip colours denote host species with green tips for cattle, orange for gorilla and blue for human. Ug# refers to the sample number from Uganda and NZ# from the sample number from New Zealand. G after the Ug# or NZ# refers to gorilla samples and C refers to cattle samples and H refers to human samples. S#, C# or M# at the third site on the labels designates S#=Sanger sequence, C#=cloned sequence, M#=metagenomic sequence.

Out of the 18 sequences evaluated for the presence of ORFs on both of the SGC and AGC, 12 (67%) identified ORFs on either of the genetic codes (Table 10). Six gorilla samples and 12 cattle samples were evaluated with four (67%) of the gorilla samples and eight (67%) of the cattle samples with ORFs identified on sequences either with the SGC and/or AGC. Of the four picobirnavirus partial amino acid sequences from gorillas, two (50%) of them were found to have identical ORFs with the use of both genetic codes; the SGC identified an ORF on one protein sequence (Ug45_M1) that the AGC did not identify an ORF, while the AGC ORF was longer on the sequence from another gorilla sample (Ug07_M1) as compared to the SGC ORF (Table 10). On the cattle samples, none of the ORFs were identical on the sequences between the SGC and AGC translations utilised. On all eight of the samples where sequences had ORFs that were identified in the cattle samples, the AGC found longer ORFs than the SGC.

TABLE 10 TABLE OF NEAR-COMPLETE PICOBIRNAVIRUS GENOMES WITH IDENTIFICATION OF ORF1 FROM *RDRP*

Sequence	Species	Length bp (aa)	Translational frame	Conserved regions	Primers	ORF SGC Length (bp)	ORF AGC Length (bp)
Ug07_M1	Gorilla	1009 (336)	R Frame 2		25F/43R	912	960
Ug15_C1.1	Gorilla	1887 (628)	R Frame 2	D3		1445	1445
Ug22_C1.1	Gorilla	1415 (471)	Frame	D2, D3	25F/43R	1389	1389
Ug28_M1	Gorilla	987 (329)	R Frame 1	D3	23F	No ORF	No ORF
Ug39_M1	Gorilla	1425 (475)	R Frame 1	D3		No ORF	No ORF
Ug45_M1	Gorilla	1131 (377)	R Frame 1	PMG, PMF, PMA, PMB, D1, D2, D3	25F/23F/43R	1026	No ORF
Ug48_M1	Cattle	1027 (342)	R Frame 2	PMF, PMA, PMB, D1, D2, D3	25F/23F/43R	918	1017
Ug49_C3.1	Cattle	1367 (455)	R Frame 1	PMG, PMF, PMA, PMB, PMC, D1, D2, D3	25F/43R	No ORF	No ORF
Ug49_C4.1	Cattle	1525 (507)	Frame 3	D3	25F/23F	1410	1449
Ug49_M1	Cattle	1291 (430)	R Frame 1	PMG, PMF, PMA, PMB, PMC, D1, D2, D3	25F/43R	1104	1212
Ug50_M1	Cattle	1032 (343)	R Frame 3	PMG, PMF, PMA, D1	25F/23F/43R	No ORF	No ORF
Ug50_M2	Cattle	1063 (353)	R Frame 3	PMF, D1	25F/23F/43R	990	1026
Ug51_M1	Cattle	1052 (350)	R Frame 1	D3		828	882
Ug53_M1	Cattle	1223 (407)	Frame 1	PMF, D1, D2, D3	25F/23F/43R	990	1059
Ug55_M1	Cattle	1363 (454)	Frame 1	PMF, PMA, PMC	25F/23F/43R	1065	1173
Ug58_M1	Cattle	1359 (453)	R Frame 1	D2, D3	43R	No ORF	No ORF
Ug62_M2	Cattle	1688 (562)	Frame 3	PMG, PMF, PMC, PME, D2, D3	25F/23F/43R	No ORF	No ORF
Ug63_M1	Cattle	1472 (490)	R Frame 2	PMG, PMF, PMA, PMB, PMC, D1, D2, D3	25F/43R	1281	1356

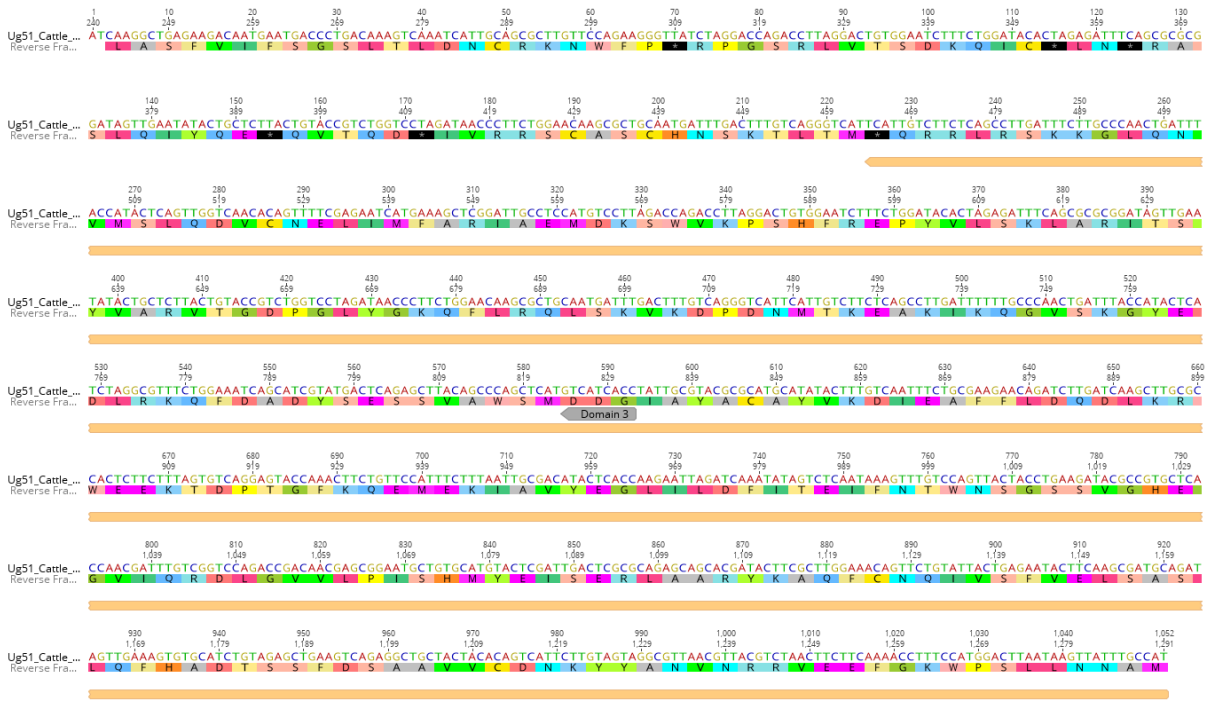
ORFs were identified based on either the standard genetic code (SGC) (translation table_1) and the alternative genetic code (AGC) (translation table_4). Sequence names are Ug for Uganda, # of the sample with the designation of C for clones and M for metagenomic sequences of PBVs. Lengths are in base pairs (bp) and amino acids (aa). Translational frames are based on

the conserved domains and/or the primers that are annotated on the sequence. Conserved regions are: polymerase motif G (PMG), polymerase motif F (PMF), polymerase motif A (PMA), polymerase motif B (PMB), polymerase motif C (PMC), polymerase motif D (PMD) and polymerase motif E (PME); and conserved domain 1 (D1), conserved domain 2 (D2) and conserved domain 3 (D3). Primer designations are 25F/43R for genogroup I (GI) or 23F/24R for genogroup II (GII).

Figure 45 and Figure 46 show the ORFs from both the SGC and AGC on the sequence from the gorilla sample, Ug07, and cattle sample, Ug51. The extension of the ORFs on the AGC are the result of the change in translation of the stop codon from the SGC to the AGC. Start codons in both genetic codes are methionine.

FIGURE 45 OPEN-READING FRAME (ORF) 1 FROM SEGMENT 2 (*RDRP*) OF PICOBIRNAVIRUS FROM GORILLA SAMPLE 07

The upper sequence is the standard genetic code (SGC) (translation table_1) and the lower sequence is from the alternative genetic code (AGC) for mould protozoan mitochondria (translation table_4). The AGC is a longer ORF1 due to a change in the stop codon from the SGC to the AGC. Neither sequence contained conserved domains or polymerase motifs though did contain the correct amino acid sequences for the forward and reverse primers, F25 and R43 for genogroup I picobirnaviruses.



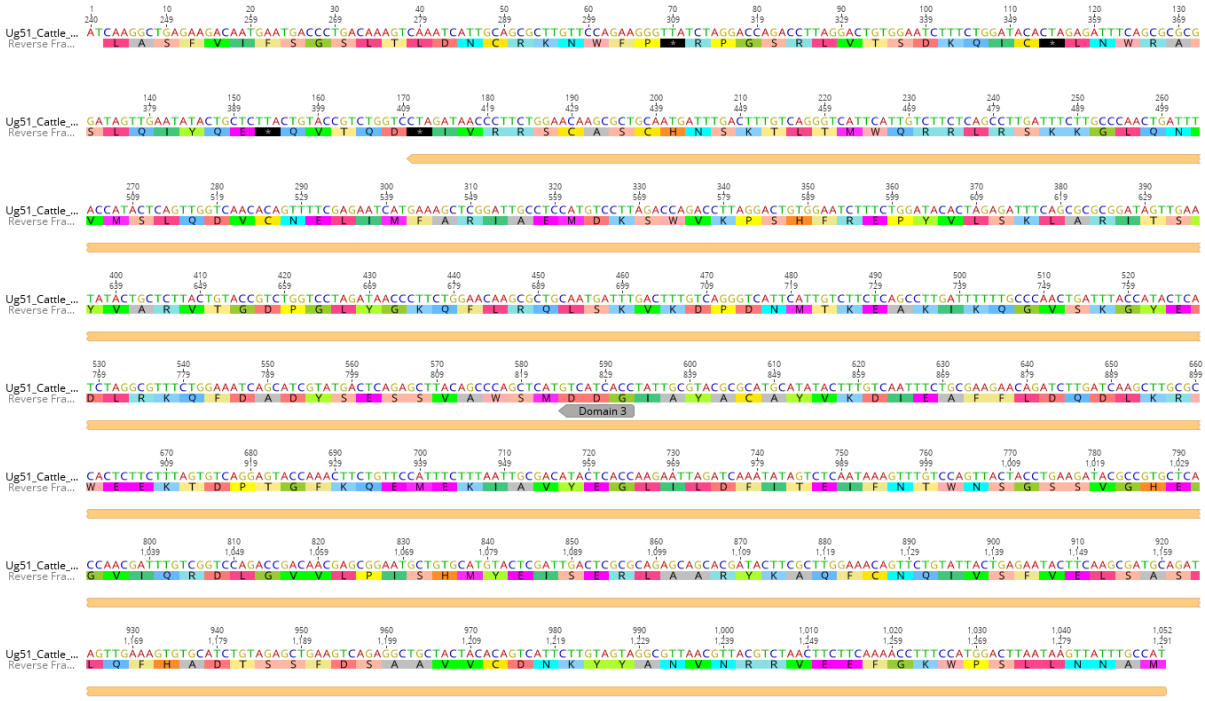


FIGURE 46 OPEN-READING FRAME (ORF) 1 FROM SEGMENT 2 (*RDRP*) OF PICOBIRNAVIRUS FROM CATTLE SAMPLE 51

This picobirnavirus sequence was found to have one conserved region, conserved domain 3, noted above. The upper sequence is the standard genetic code (SGC) (translation table_1) and the lower sequence is from the alternative genetic code (AGC) for mould protozoan mitochondria (translation table_4). The AGC is a longer ORF1 due to a change in the stop codon from the SGC to the AGC. Both sequences contained the conserved domain 3 in the amino acid sequence.

7.1.3 Ribosomal binding motifs (RBS) for prokaryotic host

Similarly to Krishnamurthy and Boros [133, 203], the near-complete picobirnavirus genomes were evaluated for evidence of RBS -4 to -18 nucleotides upstream of the start codon, methionine, in 4-mer (AGGA/GGAG/GAGG), 5-mer (AGGAG/GGAGG) or 6-mer (AGGAGG) windows. Two 6-mer RBS were identified from the near-complete picobirnavirus genomes in those that contained an identified ORF and were within the correct position (Ug63M1 at -14 to -8; Ug50M2 at -6 to 0) (Figure 47). The 5-mer RBS was also found on the sequences in the prior two samples. Four possible 4-mer RBS were identified from the near-complete picobirnavirus genomes, though all of them were greater than -18 nucleotides upstream of the start codon (Ug55M1 at -39 and -75; Ug53M1 at -22; Ug49C4.1 at -51; Ug45M1 at -42) (Figure 47).

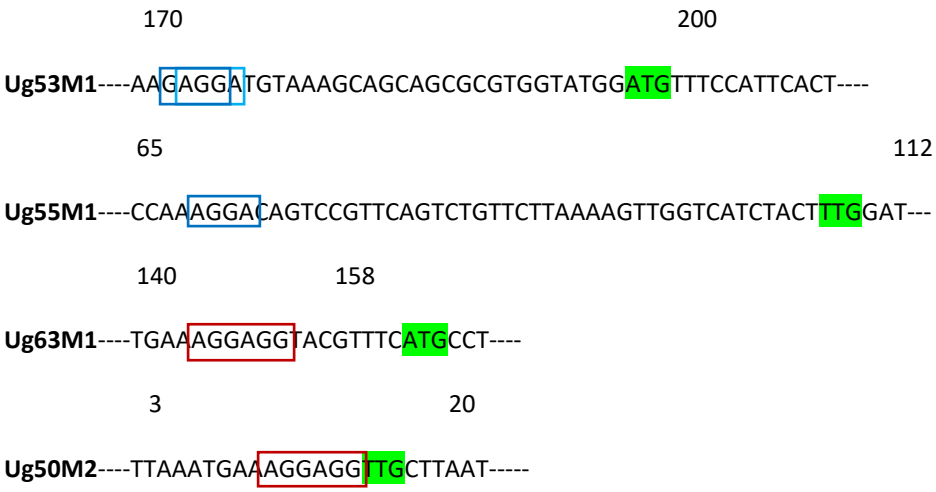


FIGURE 47 RBS ON NEAR-COMPLETE PICOBIRNAVIRUSES FROM UGANDA

All picobirnaviruses were evaluated for ORFs and the RBS upstream from the ORF, reportedly up to -18 upstream of the start codon. Ug = Uganda; # is the number of the sample; M = metagenomic contig; M# = number of the metagenomic contigs from that sample. Start codons are highlighted in green. The two lower contigs (Ug50M2 and Ug63M1) had the 6-mer RBS identified within -18 nucleotides upstream of the start codon (Ug63M1 from -14 to -8; Ug50M2 from -6 to 0). The two upper contigs had the 4-mer RBS identified upstream of the start codon and ORF though all were greater than -18 nucleotides from the start codon. Ug53M1 and Ug55M1 are two examples from the 4-mer RBS identified (Ug53M1 at -22 and Ug55M1 at -39).

7.2 Discussion

I tested various bacterial genera and protozoa for the presence of the dsRNA virus, picobirnavirus to identify a potential association of the virus with protozoa or bacteria. I identified one picobirnavirus from a protozoan sample and none from bacterial samples. I also identified ORFs from different genetic codes and the presence of some RBS from the picobirnavirus sequences from the Ugandan samples.

7.2.1 Protozoal host?

7.2.1.1 Identification from protozoal samples

I evaluated purified cysts of *Giardia* and oocysts of *Cryptosporidium* for the presence of picobirnaviruses using various picobirnavirus primers (Chapter 3, Section 3.6.1.1). None of the *Giardia* cysts evaluated were positive for picobirnavirus RNA after PCR and Sanger sequencing. The lack of picobirnavirus detection in *Giardia* does not necessarily confirm the absence of an association with this protozoan, though only limited data has supported this possible association [216]. The association of *Giardia* with similar RNA viruses has been limited. An unclassified trisegmented RNA virus called picotrirnavirus, was initially found in chicken faeces and later in the faeces of diarrheic children [217, 218]. A study investigating rotaviruses, dsRNA viruses in the family *Reoviridae*, in diarrheic canine faeces found picotrirnavirus in canine faecal samples that were also positive for *Giardia* and *Ancylostoma* sp. [216, 218]. Other than these few studies, no further association has been reported of either picotrirnavirus or picobirnavirus with *Giardia*.

Picobirnavirus was identified, though, from one of the *Cryptosporidium* samples from the purified oocysts. Remarkably, the picobirnavirus matched with 100% query cover and percent identity with BLASTn to a human *Picobirnavirus* (accession AB214978), identified after conventional PCR and Sanger sequencing. Picobirnavirus has previously been identified in the early reports in samples where *Cryptosporidium* was also identified [116, 204]. Picobirnaviruses were initially identified in *Cryptosporidium* positive human faeces but could not be identified at that time from purified oocysts of *Cryptosporidium* [204]. In contrast, not all studies that have investigated the association of picobirnaviruses with *Cryptosporidium* found evidence of co-occurrence [170]. I used a new method of purification for the *Cryptosporidium* oocysts (Chapter 3, Section 3.2.2.2) to decrease the possible contamination of the sample from faecal material that could contain enteric bacteria. Furthermore, the first report of the possible association of the virus with *Cryptosporidium* identified what was termed atypical picobirnaviruses as the size of the two segments was smaller and the diversity of the sequences less than the typical picobirnaviruses identified at that time [204]. I only found one picobirnavirus from the *Cryptosporidium* samples which was from a *C. hominis* oocyst sample which usually infects human hosts. Further study with increased quantity and variable species of *Cryptosporidium* oocysts from various host species would provide additional support for this association.

In the years after the discovery of picobirnaviruses in 1988 [114, 115], associations with protozoal organisms such as *Cryptosporidium* were hypothesized [116, 204]. Additional studies evaluating *Cryptosporidium* identified dsRNA viral material most similar to *Partitiviridae* based on the polymerase characteristics [212, 213]. Viruses in *Partitiviridae* were historically viruses of plant or fungi though a new genera was designated in 2009 within the family for viruses of protozoa called *Cryspovirus* which

are viruses of *Cryptosporidium* (*parvum, hominis, meleagridis* or *felis*) [214]. The cryspoviruses were similar to the previously identified atypical picobirnaviruses from the early reports [116, 204]. This may further explain the previous association of these atypical picobirnaviruses with a protozoal host. Based on only limited and early reports, and my finding of a typical picobirnavirus in purified *Cryptosporidium* oocysts, further investigation is needed to add to, or clarify, the actual association of picobirnaviruses with *Cryptosporidium*.

7.2.1.2 Identification of ORFs

In order to evaluate the possibility of an AGC for the picobirnaviruses, I evaluated two aspects of the viruses from my samples. First, I evaluated the near-complete picobirnavirus genomes for the presence of ORFs in the *RdRp*, segment 2, of the samples from Uganda with the use of the SGC and the AGC. Secondly, I translated a trimmed section of the genogroup I *RdRp*, segment 2, from my samples from Uganda and New Zealand by either the standard genetic code (SGC; translation table_1) or an alternative genetic code (AGC; translation table_4). I then performed a co-phylogenetic analysis of these samples (Figure 44).

Alternative genetic codes other than the standard genetic code were considered, including the mould protozoan mitochondrial code (translation table_4) and invertebrate mitochondrial code (translation table_5) based on prior studies of AGCs in picobirnaviruses [93, 94, 195]. I evaluated both the mould protozoan (transl_table 4) and the invertebrate mitochondrial (transl_table 5) codes for ORFs in my picobirnaviruses. The mould AGC is altered from the SGC by the change of the termination code to the amino acid tryptophan and alternative start codons, depending on the organism (TTA, TTG, CTG, ATT, ATA, ATG, ATC, GTG and GTA). The invertebrate AGC is altered from the SGC by the changes of multiple amino acids (arginine to serine, isoleucine to methionine and the termination codon to tryptophan) and alternative start codons, also depending on the organism (ATA, ATT, ATC, GTG, TTG) [137-139]. An overall virome study [94], reported that one of the most common AGCs utilized is the invertebrate mitochondrial genetic code. Yinda et al. 2018 additionally utilized the invertebrate mitochondrial genetic code for picobirnaviruses while Kleymann et al. 2020 evaluated the use of both the mould and invertebrate mitochondrial genetic codes and used the mould protozoan mitochondrial genetic code. I evaluated both the mould and invertebrate mitochondrial codes for the ORFs in the picobirnavirus *RdRp* from my samples; no alterations were noted in the ORF sizes including start codons and stop codons between the two alternative genetic codes with only an exchange of arginine to serine and isoleucine to methionine. I, therefore, decided to use the mould protozoan genetic code (translation table 4) for the AGC.

In the first evaluation of the picobirnavirus ORFs with the SGC and the AGC, only picobirnavirus sequences from two gorilla samples were altered between the usage of the two codes while all of the *Picobirnavirus* sequences from the cattle samples with ORFs were different between the SGC and AGC. In the samples where both genetic codes identified an ORF, the AGC extended the length of the ORF by an average of 70 bp (36–108 bp range). One picobirnavirus ORF from a sequence in a gorilla sample (Ug07_M1) was extended with the AGC by 48 bp; one picobirnavirus from a sequence in another gorilla sample (Ug39_M1) was identified to have an ORF in the SGC and not in the AGC.

In the second evaluation with the co-phylogeny between the SGC and the AGC, the comparison of picobirnaviruses from Uganda and New Zealand was used to evaluate changes in clustering of the viruses with the use of different genetic codes. The use of the AGC did shift some of the clusters and branches though the majority of the branches remained grouped together and no obvious identification of clustering by host was noted with the use of the different genetic code, similar to the analysis in Chapter 5. Also, when the picobirnaviruses were translated into the amino acid sequences, stop codons appeared in the middle of the *RdRp* translated region that the sequences were trimmed to with the known conserved motifs and domains for the SGC, but not for the AGC. This does not appear unusual in picobirnaviruses as stop codons have been previously identified [116, 161]. Stop codons within the ORFs may indicate frame-shifting and generation of various protein sizes, read-through, artefacts of PCR or amplification and sequencing, and/or deletions or insertions in the nucleotides during replication [116, 161]. The three samples that these stop codons were identified in the SGC amino acid sequences included only the conventional PCR and Sanger sequenced contigs, not the metagenomic contigs. This may indicate that the stop codon presence in my samples may be more consistent with PCR artefacts though as this is not a novel finding in picobirnaviruses, it may be secondary to frame-shifting, read-through of the stop codons or mutations during normal replication for RNA viruses [116, 161, 283, 284]. None of the AGC translated sequences contained stop codons. Read-through of the stop codons is not uncommon in viral genomes and picobirnavirus segment 1 contains overlapping ORFs with read-through of the stop codons [116, 127, 284]. As noted previously, I did not obtain long enough segment 1 picobirnaviruses to identify this read-through capability and, to test for functional proteins, I did not attempt viral isolation as picobirnaviruses have not been successfully grown in cell cultures.

7.2.2 Bacterial host?

7.2.2.1 Identification from bacterial samples

I evaluated colonies of various enteric bacteria for picobirnaviruses and did not identify any viruses from these samples. The lack of identification of picobirnaviruses from the bacterial colonies is not surprising as this method of identification of picobirnaviruses has not been previously confirmed and

instead, the supposition of picobirnavirus as a bacterial virus was based on the presence of the RBS motifs, not the actual identification of the virus in bacterial samples.

7.2.2.2 Ribosomal Binding Motifs

With the extensive diversity of picobirnaviruses, the questionable pathogenicity and the similarity to the partitiviruses, researchers have suggested that picobirnaviruses may actually be prokaryotic rather than single cell eukaryotic viruses and have detected motifs in the picobirnavirus nucleotides that are typically associated with prokaryotic viruses [133, 195, 203]. I did identify RBS motifs in some of the picobirnaviruses from my samples. Based on the studies evaluating the RBS in picobirnaviruses [133, 203], I attempted to identify ORFs (as noted above) and, in those where ORFs were identified in the *RdRp*, I searched for the 4-mer, 5-mer and 6-mer RBS motifs upstream of the ORF from -18 nucleotides of the start codon. In the Boros et al. study, six complete picobirnaviruses were identified in a single cloacal sample from a chicken and all picobirnaviruses contained the 6-mer RBS in either the segment 1 or segment 2 [133]. In the Krishnamurthy et al. study, picobirnaviruses were evaluated from the NCBI database and metagenomic data from macaque faeces for a total of 41 segment 1 and 40 segment 2 picobirnaviruses in which all viruses had the 4-mer RBS and 75% of the picobirnaviruses had the 5-mer or the 6-mer RBS identified [203]. In the near-complete picobirnaviruses from my samples, I only identified two 6-mer RBS in the correct position and direction from the ORF in the *RdRp* of segment 2.

The reason for the lack of identification of the RBS in the picobirnaviruses from my samples may have occurred for a multitude of factors: 1) many of my near-complete picobirnavirus genomes were, in fact, near-complete, not complete and so may not have included the RBS; 2) many of my picobirnaviruses did not have identifiable ORFs and even though the RBS motifs were identified in the contig and without the presence of the ORF, they could not be evaluated further; 3) I only used the *RdRp* segment 2 of picobirnaviruses in my sample, while the other studies where the authors found 100% of the picobirnaviruses with the RBS were complete picobirnavirus segments sequences from both segment 1 and segment 2. Additionally, one of the studies that looked at NCBI database picobirnavirus sequences for the presence of the RBS utilized S1 segments (>1600 bp), as these are typically longer contigs and may be more likely to contain the RBS in the UTR upstream from the ORF [203].

7.2.3 Other host options?

The high mutation rate of RNA viruses and the high genetic diversity of certain RNA viruses, picobirnaviruses included, have led to the question of these viruses as a quasispecies [130, 179]. Quasispecies are described as “a form of mutation-selection balance that applies to genetic systems

characterized by very high mutation rates. As natural selection acts on the population as a whole, in quasispecies, it can also be considered a form of group selection" [130]. Though quasispecies arguments do not answer the question of the actual host of a virus such as picobirnavirus, the theory proposes that the high genetic diversity is the consequence of the average of the fitness of the viruses as a population rather than individual viral fitness [130]. Though picobirnaviruses have high genetic diversity that may support the quasispecies theory, the evidence for picobirnaviruses as a quasispecies is lacking and the debate of quasispecies theory persists.

In addition to the question of the actual host of picobirnavirus, debate remains as to the actual pathogenicity of the virus in the hosts it has been identified in. Picobirnaviruses have been thought to be associated with symptoms such as gastroenteritis and diarrhea in monkeys [118], humans [141, 173, 202, 285], pigs [169, 286], children and immunocompromised humans [162, 167, 170, 171, 180, 287, 288], cattle [165] turkeys and chickens [168, 191] and marmots [164]. Despite the many studies reporting a possible association with enteric disease in mammals, a clear association has not been confirmed. Picobirnaviruses do not grow in cell culture nor has virus isolation been successful. The inoculation of the virus in rabbits did result in the identification of the virus in the inoculated rabbits which could indicate infection; however, the identified viruses were sufficiently genetically diverse from the inoculated viruses, questioning whether they produced an infection, or not in the rabbits themselves [147]. This same study, though, did show an antibody response in one rabbit post-inoculation and in another rabbit at a later timepoint after exposure [147]. Furthermore, the capsid segment of picobirnaviruses have been found to have liposomal-perforating activity which is typically associated with vertebrate viruses rather than viruses of single eukaryotes or prokaryotes [127].

7.2.4 Additional limitations

My ability to further clarify the actual host of picobirnavirus was mainly limited by the small sample size. My power analysis to identify how many samples of each possible host I needed to test in order to identify the presence of the dsRNA virus led to a low number of samples (10) to determine with 95% confidence the presence or absence of picobirnaviruses. Given this low sample size, I only identified one picobirnavirus from a protozoan sample and none from the bacterial samples. More samples should be tested to confirm these findings in addition to further bacterial genera. Moreover, there are different *Cryptosporidium* species and though I tested three different known species of *Cryptosporidium*, other species may be indicated, and so studies should have sufficient power to detect the presence or absence in different species (Table 20). This is also true for the bacteria. My study included 11 species, so sample sizes across each should be increased. Studies of picobirnaviruses and their role, if any, in *Cryptosporidium* life cycles are also hampered by the lack of reliable *in vitro* culture methods to allow the completion of their life cycles, though these are being developed [232].

7.3 SUMMARY

Due to the high genetic diversity, the lack of clustering and the various hosts that the viruses have been found associated with, I attempted to identify further evidence for non-vertebrate or eukaryotic host of picobirnaviruses. I identified RNA from a single picobirnavirus from purified oocysts of a protozoa, *Cryptosporidium*, though no further picobirnaviruses were identified from the other protozoa or bacterial hosts sampled. I then investigated the use of an alternative genetic code for translation into amino acids and found that the use of the alternative genetic code did not identify more ORFs, but did identify longer length ORFs and the exclusion of stop codons, as compared to the standard genetic code. Furthermore, I did identify the RBS associated with prokaryotes in the genetic code of a selection of the picobirnaviruses investigated from my samples. The methods of analyses provide further evidence for a possible protozoal or bacterial host for picobirnaviruses.

CHAPTER 8 GENERAL DISCUSSION

A dsRNA virus genus called picobirnavirus was found in all of the species within my Ugandan study system. Picobirnaviruses were identified using multiple laboratory methods including the use of published primers, primers designed within this study and shotgun metagenomic sequences. Data derived from these methods, which also included sequencing of cloned PCR products, were combined using bioinformatic tools to maximize the information on picobirnavirus diversity in this study. To expand the geographic and host system range of my PhD research, I also tested and identified picobirnaviruses from New Zealand mammalian samples along with protozoan and bacterial samples. Phylogenetic analyses were performed with the aim of determining whether host species correlated with picobirnavirus diversity and to check for evidence of cross-species transmission. In addition, I sought to identify the presence of multiple within-host picobirnaviruses in my study. These analyses were performed both on my study samples alone, and together with a set of sequences representative of global picobirnavirus diversity, which also allowed me to assess relationships between picobirnavirus diversity and broad scale geography.

The identified picobirnavirus RNA were described, compared to known picobirnaviruses in the NCBI database and further characterized based on conserved regions in the genetic code. Additionally, the picobirnavirus RNA sequences were analysed using either nucleotide or amino acid sequences along with trimming for potential spurious sequence components within the known genogroup classifications. Based on these analyses, further phylogenetic analyses were performed with amino acid sequences without further sequence trimming beyond those performed for adequate multiple alignments and poor-quality regions. Phylogenetic analyses of genogroup I and genogroup II picobirnaviruses from the study together with other known NCBI genogroup I or genogroup II picobirnaviruses did show clades that were distinctly separate from the other genogroups when rooted, but still with high diversity even within the genogroups themselves and some *Picobirnavirus* sequences which did not cluster by genogroup. Furthermore, within- and between-genogroup diversity of the identified picobirnaviruses was high.

My phylogenetic analyses revealed clustering by host with 55% (11 out of 20 sequences) of the genogroup I picobirnaviruses shared within gorillas in the cohort Uganda study system. Once additional picobirnaviruses were evaluated both within- and between- genogroups, though, no further host clustering was recognized. Due to the majority of data coming from targeted PCR-based amplification of conserved regions on the *RdRp* gene, initial phylogenetic analyses were limited to short gene fragments. A subset of samples (n = 17) was able to be assembled into longer sequence contigs using metagenomic reads and these were analysed together with other near-complete *RdRp*

sequences from NCBI. Even longer picobirnavirus sequence lengths did not indicate host clustering. The phylogeographic analyses of the identified picobirnaviruses and picobirnaviruses from the NCBI database also did not reveal geographic clustering both within- and between- genogroups. In addition, the use of longer picobirnavirus sequence lengths did not indicate any phylogeographic association of the virus.

I did identify multiple picobirnaviruses within the same individual host from my samples with an average of 4–5 picobirnaviruses per sample (range 2–8). The within-host picobirnaviruses were categorized in various genogroups and some of unknown genogroup designation. The diversity of the within-host picobirnaviruses was high with an average amino acid percent identity of 34.3%, ranging from 16% to 100%. Despite high diversity between picobirnaviruses within and between samples, two sets of picobirnaviruses were 100% identical, one between two gorillas and one between a gorilla and a cattle host. In contrast, most of the study cohort sequences (87%, *n* = 98/113; Table 18) were not closely related (<90% pairwise identity) to the top hit sequences from the NCBI database with an average pairwise identity of 79.6%. A few picobirnavirus sequences, though, matched to NCBI sequences with high pairwise identity (>97%): Ug03_H_S3, 97.5% to a porcine picobirnavirus from China (KC841460); Ug55_C_C1, 97.3% also to a porcine picobirnavirus from China (KP984805); and NZCR03_S1, 100% to a human picobirnavirus from India (AB214978).

Due to this consistently high diversity of the picobirnaviruses and ubiquitous but inconsistent nature of the dsRNA virus in terms of host and geography, I investigated the actual host of picobirnaviruses. There is ongoing debate about whether picobirnaviruses have a vertebrate, protozoal or bacterial host [116, 133, 195, 203, 204, 216-218]. I sought to amplify picobirnavirus from each of these potential hosts. I did identify a *Picobirnavirus* from a purified oocyst *Cryptosporidium* sample but did not identify further picobirnaviruses from other protozoa (*Giardia*) or bacterial samples. Further analyses revealed that the use of an alternative genetic code for mould mitochondrial protozoa did extend the ORF lengths for the *RdRp* segment 2 of the picobirnaviruses and also eliminated the stop codons present in three amino acid picobirnavirus sequences with the standard genetic code. The use of the alternative genetic code, though, did not reveal stronger host clustering on phylogenetic analysis. Lastly, in the attempts to evaluate for the possibility of a prokaryotic host, ribosomal binding motifs present in some prokaryotic viruses but less in eukaryotic viruses were investigated. I found ribosomal binding motifs in the correct position within two of my picobirnaviruses. These additional analyses into the possible host of picobirnaviruses did not resolve the actual host though supported the possibility of a protozoan or bacterial host hypothesis. Further work is needed with additional bacterial and protozoan samples to further investigate this ongoing debate. Furthermore, typically

viruses would be grown in cultured cells for confirmation of pathogenicity and further research though to date, cell culture techniques for picobirnaviruses have been unsuccessful.

My findings support and add to the knowledge of this interesting dsRNA virus, *Picobirnavirus*, including that it is a ubiquitous virus, has now been identified in New Zealand livestock, has a high diversity and likely either is categorized into multiple genogroups and/or may be more than one virus type. It still does not appear to cluster by host or by geography and can either cause co-infections or manifests as multiple viruses within the same host. Despite the lack of clustering by host species, geographical location or inconsistent clustering by genogroup designation, phylogenetic analyses showed picobirnaviruses from my study with 98–100% similarity to each other within the same host species and between host species in the system of study. This is surprising, especially considering the high diversity of the virus and the rarity of host factor clustering. Other reports of highly similar to 100% similar picobirnaviruses have been published including in pigs [119, 154], between humans and pigs [150], between buffalo and humans [206] and between humans and horses [163]. I found 100% similar picobirnavirus sequences between different gorillas usually from the same gorilla family and also between cattle and gorilla and cattle and a human. It is possible that the increased contact between the gorillas in the same family group and between the cattle and human as reported more commonly in the survey of contact in the same region, may have resulted in sharing of picobirnaviruses or host jumping [44].

In the genogroup II picobirnaviruses, a human, cattle and gorilla picobirnavirus were >99% identical on multiple sequence alignment. Additionally, the contact potential between cattle and gorillas and gorillas and humans was documented, though less common than within the same species and between humans and domestic livestock which could result in cross-species transmission of the virus [44]. Habitat loss of wildlife and increasing contact between wildlife and other species are known to increase the potential for cross-species transmission of infectious diseases [107]. In the attempt to understand the contributions to cross-species transmission (spillover) and continued transmission that could result in epidemics, we understand that the more information we can gather and understand on cross-species transmission and possible pathogens usually through surveillance, will help us better deal with the outbreaks and hopefully someday prevent further outbreaks in the future [2, 106, 108]. In addition, further evidence and knowledge of how pathogens such as viruses evolve and host jump will be instrumental in understanding cross-species transmission [112]. Tools such as the use of co-phylogenetic analyses as I performed in this research, are more commonly used to investigate for co-evolution and specifically co-divergence of the virus and host phylogenies with additional statistical tools to assess the support for the co-phylogenies [112, 255].

Further research to estimate the frequency of cross-species transmission events between species can be performed with statistical models and ancestral state reconstruction that was not used in these analyses. Many of these models or analyses are dependent on good sequence data, preferably from deeper sequencing techniques that I did not perform due to time and financial constraints. I also did not have data for the viruses from different timescales in order to investigate co-evolution or co-divergence [111, 178]. However, these approaches are also dependent on reliable sequence alignments, and currently the great diversity of picobirnaviruses and mixed within-host infections with segmented genomes makes this problematic. The utility of picobirnaviruses for evaluation of cross-species transmission is less than ideal due to the high diversity, though the use of these viruses due to their ubiquitous nature and multiple hosts, small genome size and segmented genomes could result in a higher likelihood of cross-species transmission. Additionally, if the host of picobirnaviruses were prokaryotic or protozoal in origin, it would be more challenging to assess for cross-species transmission, unless the bacteria or protozoa were fairly host-specific or had high host clustering. Other possible considerations besides animal-to-animal transmissions include a common source of the virus, replication of the virus outside of the host and/or contamination in the processing of the samples. In addition, the patterns of pathogen transmission between animals and humans can be variable, involve a multitude of factors and cannot be necessarily extrapolated from one ecosystem to another.

Long-read sequencing techniques (also called third generation sequencing) can also be used to obtain complete or near-complete genomes with the use of single-molecule real-time sequencing (Pacific Biosciences®) or nanopore sequencing (Oxford Nanopore Technologies®) [289]. Complete to near-complete virus genomes would allow for further investigation into the actual host debate with picobirnaviruses, give further and more in-depth information concerning the diversity of the virus genomes and potentially allow for further analyses of clustering based on host and/or geography, though even based on my smaller analysis with near-complete genomes of picobirnaviruses, this may be unlikely to add more information or evidence. Though long-read sequencing may decrease the errors associated with assembly and mapping of the shorter reads in second-generation sequencing, long-read sequencing has lower per-read accuracy compared to short-read sequencing [289]. Additional techniques to obtain full or near-full length genomes such as with rapid amplification of cDNA ends (RACE) could have also been utilized though I attempted to produce longer sequence reads with the metagenomic reads and cloned sequences in my research on picobirnaviruses.

Additionally, more data from larger sample sizes to increase the number of viruses including picobirnavirus and more host species would allow for improved phylogenetic analyses and

comparisons. My samples consisted of only three humans (all clinically unwell), domestic and wild animals in Uganda and New Zealand (most clinically well) and other prokaryotic and eukaryotic hosts. Sampling bias due to the limitation on sample size for some hosts, as noted previously, was a potential limitation. In addition, I did not evaluate for PCR inhibitors in my samples or various filter sizes in the extraction process which could have impacted the quantity of viruses identified if inhibitors were present or if the viruses were filtered out. Age-related effects on the presence of picobirnaviruses was also considered though most hosts were within similar age ranges with the exception of the gorillas. I did not identify more picobirnaviruses in the younger or older gorillas from my study. Further work with additional samples from the study system is ongoing to investigate these questions with *network of contacts* and *health survey data* published in Muylaert et al. (2021) to complement the analyses. Approximately 600 total samples from clinical and non-clinical humans (within the community), habituated and unhabituated mountain gorillas and community livestock including cattle and goats will be utilized for this and other investigations. With the findings of this research and these further investigations, we may not be able to understand all of the complexities of cross-species transmission, but, hopefully, it will add more pieces to the puzzle.

APPENDIX A: SUPPLEMENTARY MATERIAL TO CHAPTER 3

SUPPLEMENTARY 1: NGS PILOT RESULTS AND DISCUSSION

OBJECTIVES

An assessment of various storage solutions and subsequent nucleic acid extraction was performed to evaluate the extraction potential of RNA viruses with different storage solutions and nucleic acid extraction options. Initial pilot samples were investigated for various pathogens (Chapter 3, Section 4.3.1).

METHODS

An additional set of cattle samples were collected and placed in sterile containers from a local farm near Massey University, Palmerston North, New Zealand. Faecal samples were then separated and placed into containers with RNAlater® and TRIzol®. Samples stored in RNAlater® were further evaluated for nucleic acid extraction with three protocols/kits; Roche High Pure Viral Nucleic Acid kit (DNA and RNA), Bioline Isolate Faecal DNA kit (DNA only) and TRIzol® RNA isolation (RNA only) (Bioline Reagents Ltd, London, UK; Roche Diagnostics GmbH, Mannheim, Germany; ThermoFisher Scientific (Invitrogen), Life Technologies Corporation, Carlsbad, California, USA) (Chapter 3, Section 3.3). Samples in TRIzol® were homogenized, rinsed in phosphate-buffered saline (PBS), phase separated with chloroform, and RNA was precipitated, washed and redissolved per the TRIzol® RNA Isolation protocol from the W.M. Keck Foundation Biotechnology Microarray Resource Laboratory at Yale University (Invitrogen, California, USA). Nucleic acids (DNA and RNA as indicated) from the three different methods were then quantified using the appropriate Qubit kit (ThermoFisher, Massachusetts, USA).

RESULTS

The nucleic acids storage and extraction experiment revealed that the Roche kit in RNAlater® performed better at selectively extracting RNA organisms from faeces as compared to the Bioline kit in RNAlater® or TRIzol®, similar to prior research [290]. Due to these results, I continued to store the samples in RNAlater® and perform nucleic acid extraction with the Roche High Pure Viral Nucleic Acid kit to better obtain data from RNA organisms without excluding relevant DNA organisms for other analyses.

Next-generation sequencing (NGS) of the initial pilot study samples, produced from extracted RNA (3 human, 5 cattle and 5 gorilla) identified 45 million reads. After removal of the non-coding motifs, subtraction of the no-template negative control and taking the average of the multiple runs, I identified 2,235,000 reads with 1514 distinct 30-mer (k-mer) motifs. NGS results from the subset of

RNA-amplified nucleic acids from the initial pilot samples indicated many types of organisms present. Possible pathogens or organisms identified on k-mer based NGS included but were not limited to *Coxiella burnetti*, Lassa virus, *Haemophilus parainfluenzae*, Hepatitis C virus, *Salmonella* species, *Escherichia coli*, *Plasmodium* species and *Picobirnavirus* (Table 11).

In order to verify the presence of nucleic acids from the pathogens above, both DNA and cDNA, created from the RNA extractions, were screened with pathogen-specific PCRs. RNA samples were screened by PCR for all the pathogens; DNA samples were screened by PCR for *C. burnetti*, *H. parainfluenzae, Salmonella* species/*E. coli*/*Shigella* species and *Plasmodium* species.

C. burnetti was identified in cattle and human samples with NGS, but identified in gorilla and cattle samples with PCR (Table 11); Lassa virus was identified in a human sample with NGS but not confirmed with PCR and Sanger sequencing; *H. parainfluenzae* was identified in human and gorilla samples with NGS and only confirmed in a single human sample by PCR; Hepatitis C virus was identified in a gorilla sample with NGS but not confirmed by PCR; *Salmonella* species/*E. coli*/*Shigella* species were identified in all species with NGS but not confirmed by PCR; *Plasmodium* species were identified in a human sample and confirmed by PCR for human, gorilla and cattle samples; and picobirnavirus was identified in a gorilla sample with NGS and further confirmed in human, gorilla and cattle samples by PCR (See Table 11 and Table 12)

TABLE 11 IDENTIFICATION OF ORGANISMS FROM SAMPLES WITH NGS, PATHOGEN-SPECIFIC PCR METHODS AND BIOINFORMATIC ANALYSES

Organism	Identified by NGS			Identified by PCR		
	Human	Gorilla	Cattle	Human	Gorilla	Cattle
C. burnetti	1	0	4	0	7	2
Lassa virus	1	0	0	0	0	0
H. parainfluenzae	2	1	0	1	0	0
Hep C virus	0	1	0	0	0	0
Enteric bacteria	2	2	1	0	0	0
Plasmodium	0	1	0	1	4	2
Picobirnavirus	0	1	0	1	13	11

NGS=next generation sequencing; PCR=polymerase chain reaction.

As mentioned previously, negative controls were used during all stages of the laboratory methods and testing. In NGS, “no template” negative controls with the storage solution of RNAlater® identified

many possible organisms (Chapter 3, Section 3.5). On pathogen-specific PCRs, all negative controls either with the RNAlater® no-template controls or PCR-grade water were negative on PCR.

My gBlocks® plasmid controls for picobirnavirus were positive by PCR and Sanger sequencing and sequence alignment confirmed them to be the correct synthetic sequence.

Picobirnavirus was identified in three human faecal samples, 29 gorilla faecal samples and 15 cattle faecal samples with PCR (see Table 12). All positive samples were then Sanger sequenced, with resulting data trimmed and assembled. Sufficient quality overlapping sequences (contigs) were available from two of three positive human samples, though only one sequence was identified as picobirnavirus through BLAST and this had highest similarity to a porcine picobirnavirus (see Table 13). Contigs were assembled from 16 of 29 gorilla samples (55%) and matches to picobirnaviruses occurred in 13 out of 16 (81%) contigs using BLAST (see Table A.2). One contig did not BLAST match to a picobirnavirus but to a gram-negative soil bacterium, *Cupriavidus basilensis* (NCBI accession number CP010537). The gorilla picobirnaviruses BLAST matches included horse, porcine, human and wastewater/sewage samples (see Table 13). Contigs were formed in 11 out of 16 (67%) cattle samples. Contigs were matched to organisms in 11 out of 11 (100%) cattle samples, all picobirnaviruses (see Table 12). The cattle picobirnaviruses similarity matches included horse, pig, camel and monkey reference sequences on the NCBI database (see Table 13). Duplicate analyses of cattle/gorilla picobirnaviruses from repeat PCRs confirmed the same picobirnavirus in selected positive samples.

TABLE 12 IDENTIFICATION OF PICOBIRNAVIRUSES FROM SAMPLES BY NGS, PATHOGEN-SPECIFIC PCR METHODS AND BIOINFORMATIC ANALYSES

Species	PCR band	Contig	BLAST
Human	3/3 (44–100%)	2/3 (21–94%)	1/3 (6–79%)
Gorilla	29/44 (51–78%)	16/44 (24–51%)	13/44 (18–44%)
Cattle	15/16 (72–99%)	11/16 (39–82%)	11/16 (39–82%)

BLAST= Basic Local Alignment Search Tool; 95% confidence intervals for the prevalence are given in parentheses.

TABLE 13 HIGHEST SIMILARITY MATCHES ON NCBI NUCLEOTIDE BLAST SEARCH FOR STUDY SAMPLES OF PICOBIRNAVIRUSES

Sample number	Species	Top hit	NCBI Accession #	E-value
3	Human	Porcine picobirnavirus group II isolate CHHN-A6	KC841460	1.60×10^{-164}
4	Gorilla	Hunan wastewater109 picobirnavirus	KJ135899	6.59×10^{-69}

10	Gorilla	Human picobirnavirus 78-SZ12	KT720487	1.91×10^{-39}
14	Gorilla	Washington Raw Sewage15 picobirnavirus	EU938913	3.65×10^{-37}
15	Gorilla	Hunan wastewater6 picobirnavirus	KJ135796	5.26×10^{-76}
16	Gorilla	Porcine picobirnavirus strain CYZ-II-1	KP984805	8.14×10^{-129}
17	Gorilla	Hunan wastewater112 picobirnavirus	KJ135902	5.56×10^{-36}
19	Gorilla	Porcine picobirnavirus CYZ-II-1	KP984805	
22	Gorilla	Hunan wastewater112 picobirnavirus	KJ135902	8.95×10^{-33}
23	Gorilla	Picobirnavirus horse/BRA-02/2009 clone	GU230508	1.45×10^{-55}
24	Gorilla	Human picobirnavirus 78-SZ12	KT720487	
27	Gorilla	Picobirnavirus horse/BRA-02/2009 clone	GU230508	4.05×10^{-51}
42	Gorilla	Porcine picobirnavirus isolate CHHN-B15	KC846794	7.12×10^{-14}
43	Gorilla	Picobirnavirus pig/BRA-02/1999 clone O	GU230540	6.87×10^{-39}
44	Gorilla	Hunan wastewater33 picobirnavirus	KJ135823	5.16×10^{-60}
47	Gorilla	Picobirnavirus horse/BRA-02/2009 clone II	GU230508	6.73×10^{-54}
49	Cattle	Dromedary picobirnavirus isolate c5128**	KM573807	4.26×10^{-56}
51	Cattle	Picobirnavirus horse/BRA-02/2009 clone II	GU230508	9.56×10^{-28}
52	Cattle	Porcine picobirnavirus strain PBV/NER-2/2015	KT380849	1.17×10^{-66}
54	Cattle	Porcine picobirnavirus isolate PBV/Por-wt/GG-II/NER-TRI/IND/Por-207/2014	KP868555	2.96×10^{-92}
55	Cattle	Picobirnavirus monkey/CHN-46/2002	JQ710483	1.25×10^{-14}
56	Cattle	Porcine picobirnavirus isolate PBV/Por-wt/GG-II/NER-TRI/IND/Por-207/2014	KP868555	10.01×10^{-117}
58	Cattle	Picobirnavirus horse/BRA-02/2009 clone II	GU230508	5.37×10^{-61}
59	Cattle	Picobirnavirus PREDICT_PbV-118	KT335254	1.59×10^{-30}
60	Cattle	Picobirnavirus horse/BRA-02/2009 clone II	GU230508	1.52×10^{-60}
60	Cattle	Porcine picobirnavirus PBV/NER-2/2015	KT380849	7.07×10^{-26}
61	Cattle	Porcine picobirnavirus strain PBV/India/2013/NER/19P	KJ650569	1.57×10^{-40}
61	Cattle	Human picobirnavirus isolate: GPBV6G2	AB517738	2.55×10^{-58}

62	Cattle	Picobirnavirus PREDICT_PbV-118	KT335253	9.97×10^{-128}

DISCUSSION

NGS results from the initial pilot samples indicated many types of organisms present, some expected and some unexpected within the system and species of study. Possible pathogens or organisms identified on k-mer based NGS included but were not limited to *Coxiella burnetti*, Lassa virus, *Haemophilus parainfluenzae*, Hepatitis C virus, *Salmonella* species, *Escherichia coli*, *Plasmodium* species and *Picobirnavirus*.

PCR results from gorilla, human and cattle samples indicated one pathogen that was identified on all species evaluated in more than one individual: picobirnavirus. My analysis confirmed the preliminary NGS findings and identified picobirnavirus through PCR in the faeces of three mammalian species in Africa. I have identified picobirnaviruses in faecal samples from humans, cattle and mountain gorillas, all residing in the same geographical region in south-western Uganda in and around the Bwindi Impenetrable Forest.

TABLE 14 NCBI PICOBIRNAVIRUS ACCESSIONS

AB186898.1	AB517731.1	AB517732.1	AB517734.1	AB517735.1	AB517736.1	AB517737.1	AB517738.1	AF246939.1	AF246940.1
GQ221268.1	GQ915026.1	GQ915029.1	GU968924.1	JF755419.1	JF755420.1	JQ710507.1	JX680467.1	KC692366.1	KF792838.1
KF823810.1	KF823811.1	KF861773.1	KJ206569.1	KJ495690.1	KJ650570.1	KJ663814.1	KJ663816.1	KM254161.1	KM254164.1
KM285234.1	KM573798.1	KM573799.1	KM573800.1	KM573801.1	KM573802.1	KM573803.1	KM573804.1	KM573805.1	KM573806.1
KM573807.1	KM573808.1	KP264973.1	KP941111.1	KR827415.1	KR827416.1	KR827417.1	KR827418.1	KR902503.1	KR902505.1
KR902507.1	KT934307.1	KT934308.1	KT984499.1	KU729755.1	KU729757.1	KU729758.1	KU729762.1	KU729763.1	KU729764.1
KU729767.1	KU729768.1	KU729769.1	KU892528.1	KU892529.1	KX374476.1	KX374478.1	KX374479.1	KY053141.1	KY120170.1
KY120171.1	KY120172.1	KY120173.1	KY120174.1	KY120175.1	KY120176.1	KY120177.1	KY120178.1	KY120179.1	KY120180.1
KY120182.1	KY120185.1	KY120188.1	KY174983.1	KY214430.1	KY214431.1	KY214432.1	KY399057.1	KY502850.1	KY502851.1
KY502853.1	KY502854.1	KY502855.1	KY502856.1	KY502863.1	KY502865.1	KY502867.1	KY502868.1	KY502870.1	KY502872.1
KY502874.1	KY502875.1	KY502876.1	KY502879.1	KY502963.1	KY502979.1	KY502980.1	KY502983.1	KY502988.1	KY502994.1
KY502999.1	KY503004.1	KY503020.1	KY928683.1	KY928684.1	KY928685.1	KY928686.1	KY928687.1	KY928688.1	KY928689.1
KY928690.1	KY928691.1	KY928692.1	KY928693.1	KY928694.1	KY928695.1	KY928696.1	KY928697.1	KY928698.1	KY928699.1
KY928700.1	KY928701.1	KY928702.1	KY928703.1	KY928704.1	KY928705.1	KY928706.1	KY928707.1	KY928708.1	KY928709.1
KY928710.1	KY928711.1	KY928712.1	KY928713.1	KY928714.1	KY928715.1	KY928716.2	KY928717.1	KY928718.1	KY928719.1
KY928721.1	KY928722.1	KY928723.1	KY928725.1	KY928726.1	KY928727.1	KY928728.1	KY928729.1	KY928731.1	KY928733.1
KY928734.1	KY928735.1	KY928736.1	KY928737.1	KY928738.1	LC338002.1	LC338003.1	LC338005.1	LC338006.1	LC338007.1
LC338008.1	MF071281.1	MF416389.1	MF416390.1	MF416391.1	MG003334.1	MG003339.1	MG003340.1	MG003341.1	MG010903.1
MG010904.1	MG010905.1	MG010906.1	MG010907.1	MG010908.1	MG010909.1	MG010910.1	MG010911.1	MG010912.1	MG010913.1
MG010914.1	MG010915.1	MG010916.1	MG010917.1	MG010918.1	MG010919.1	MG010920.1	MG190029.1	MG571903.1	MG571907.1
MG600064.1	MG821233.1	MG846401.1	MG846402.1	MG846403.1	MG846404.1	MG846405.1	MG846407.1	MG846408.1	MG846410.1
MG846411.1	MG846412.1	MH327934.1	MH412924.1	MH425584.1	MH425585.1	MH425586.1	MH425587.1	MH425588.1	MH425589.1
MH425590.1	MH453875.1	MH453878.1	MH933801.1	MH933802.1	MH933804.1	MH933806.1	MH933808.1	MH933809.1	MH933810.1
MH933811.1	MH933813.1	MH933815.1	MH933818.1	MH933819.1	MH933823.1	MH933825.1	MH933835.1	MH933839.1	MH933841.1
MK064212.1	MK064213.1	MK204395.1	MK204418.1	MK378834.1	MK378844.1	MK378845.1	MK378851.1	MK378856.1	MK378859.1
MK378865.1	MK378867.1	MK378868.1	MK521919.1	MK521920.1	MK521921.1	MK521922.1	MK521923.1	MK521924.1	MK521925.1
MK521926.1	MN563295.1	MN563296.1	MN563298.1	MN563300.1	MN563301.1	MN563302.1	MT129742.1	MT129743.1	MT129745.1
MT129746.1	MT129747.1	MT129748.1	MT129749.1	MT129750.1	MT129751.1	MT129752.1	MT129753.1		

TABLE 15 CODON TRANSLATION TABLES USED FOR PICOBIRNAVIRUSES

Code	Translation table	DNA codon	RNA codon	Translation with code	Standard translation
Standard	1				
Mould, protozoan coelenterate mitochondrial	4	TGA	UGA	Trp (W)	Stop (*)
Invertebrate mitochondrial	5	AGA	AGA	Ser (S)	Arg (R)
		AGG	AGG	Ser (S)	Arg (R)
		ATA	AUA	Met (M)	Ile (I)
		TGA	UGA	Trp (W)	Stop (*)

Refer to Table 4 for nucleotide and amino acid abbreviations [221]

APPENDIX B: SUPPLEMENTARY MATERIAL FOR CHAPTER 4

TABLE 16 SUMMARY OF PICOBIRNAVIRUS *RDRP* SEQUENCES IDENTIFIED FROM RT-PCR AND METAGENOMICS FROM UGANDA AND NEW ZEALAND

Sample	Species	Method	Primers	Seq name	%GC	Genogroup (if known)	Seq length	Conserved regions	Total PBVs per sample
Ug01	Human	S	25F	Ug01_H_S1	42.5	GI	181	D1	2
Ug01	Human	S	43R	Ug01_H_S2	43.3	GI	183		
Ug02	Human	S	25F	Ug02_H_S1	42.8	GI	182	PMA,D1	2
Ug02	Human	S	25F43R	Ug02_H_S2	42.6	GI	207	PMA,D1	
Ug03	Human	S	25F	Ug03_H_S1	44.7	GI	235		3
Ug03	Human	S	25F43R	Ug03_H_S2	48.3	GI	211	D1	
Ug03	Human	S	23F24R	Ug03_H_S3	44.1	GII	380		
Ug04	Gorilla	S	25F43R	Ug04_G_S1	45.1	GI	182		1
Ug06	Gorilla	S	25F43R	Ug06_G_S1	43.4	GI	378	PMF	2
Ug06	Gorilla	M	25F43R	Ug06_G_M1	46	GI	798	PMF,PMB,PMC, D2,D3	
Ug07	Gorilla	M	25F43R	Ug07_G_M1	40.5	GI	1009		1
Ug08	Gorilla	S	25F43R	Ug08_G_S1	47.5	GI	158		2
Ug08	Gorilla	S	25F43R	Ug08_G_S3	45.5	GI	157	D1	
Ug10	Gorilla	S	25F43R	Ug10_G_S1	40.3	GI	196	D1	1
Ug14	Gorilla	S	25F43R	Ug14_G_S1	44.1	GI	188	D1	1
Ug15	Gorilla	S	25F43R	Ug15_G_S1	45.3	GI	206		5
Ug15	Gorilla	S	NCF3R8	Ug15_G_S2	42.1		783	D3	
Ug15	Gorilla	M	25F43R	Ug15_G_M1	46.1	GI	893	PMG,PMF,PMA, PMB,D1,D2	
Ug15	Gorilla	C	NC	Ug15_G_C2.1	42.5		1887	D3	
Ug15	Gorilla	S	NCF5	Ug15_G_S3	43.7		309		
Ug16	Gorilla	S	25FNCF5	Ug16_G_S1	48.2	GI	473	D3	2
Ug16	Gorilla	M	25F43R	Ug16_G_M1	42.7	GI	447		
Ug17	Gorilla	S	25F	Ug17_G_S1	48.6	GI	148	D1	4
Ug17	Gorilla	S	NCF5	Ug17_G_S2	46.2		212		
Ug17	Gorilla	S	NCR8	Ug17_G_S3	43.5		254	D3	
Ug17	Gorilla	S	NCF3R5	Ug17_G_S4	45.2		654	D2,D3	
Ug17	Gorilla	M	25F43R	Ug17_G_M1	42.5	GI	393		
Ug22	Gorilla	C	NC	Ug22_G_C1.1	44.7		1427	D2,D3	4
Ug22	Gorilla	C	25F43R	Ug22_G_C2	47.5	GI	204	D1	
Ug22	Gorilla	C	25F43R	Ug22_G_C2.1	42.1	GI	589		
Ug22	Gorilla	M	25F43R	Ug22_G_M1	44.8	GI	1242	D2,D3	
Ug23	Gorilla	S	25F43R	Ug23_G_S1	41.1	GI	188	D1	1
Ug25	Gorilla	M	25F43R	Ug25_G_M1	43.8	GI	742		1
Ug27	Gorilla	S	25F43R	Ug27_G_S1	40.9	GI	179	D1	1
Ug28	Gorilla	M	23F24R	Ug28_G_M1	47.5	GII	1035	D3	1
Ug39	Gorilla	S	25F43R	Ug39_G_S1	36.8	GI	182		1
Ug41	Gorilla	S	25F43R	Ug41_G_S1	42.9	GI	185	D1	1
Ug42	Gorilla	S	25F43R	Ug42_G_S1	42.8	GI	153		1
Ug43	Gorilla	S	25F43R	Ug43_G_S1	43.2	GI	185	D1	6
Ug43	Gorilla	S	NCF3R8	Ug43_G_S2	43.7		675	D3	
Ug43	Gorilla	C	25F43R	Ug43_G_C1	44.1	GI	204	PMA,D1	
Ug43	Gorilla	C	25F43R	Ug43_G_C2	48.1	GI	208	D1	
Ug43	Gorilla	M		Ug43_G_M1	42.1		732	PMC, PMD, D2, D3	
Ug43	Gorilla	M		Ug43_G_M2	42.4		453	PMD,D3	

Sample	Species	Method	Primers	Seq name	%GC	Genogroup (if known)	Seq length	Conserved regions	Total PBVs per sample
Ug43	Gorilla	M		Ug43_G_M3	42.7		637	D3	
Ug44	Gorilla	S	25F43R	Ug44_G_S1	41.2	GI	187	D1	1
Ug45	Gorilla	M	25F43R	Ug45_G_M1	46.4	GI	1131	PMG,PMF,PMA, PMB,D1,D2,D3	2
Ug45	Gorilla	M	25F43R	Ug45_G_M2	44.4	GI	315	PMF,D1	
Ug47	Gorilla	S	25F43R	Ug47_G_S1	43.9	GI	1671	D1	1
Ug48	Cattle	M	25F43R	Ug48_C_M1	43.7	GI	1027	PMF,PMA,PMB, D1,D2,D3	1
Ug49	Cattle	S	25F43R	Ug49_C_S1	46.3	GI	291	D1	7
Ug49	Cattle	C	25F43RNC	Ug49_C_C2	51.1	GI	221	D1	
Ug49	Cattle	C/M	NC	Ug49_C_C3.1	43.1		1367	PMG,PMF,PMA, PMB,PMC,D1, D2,D3	
Ug49	Cattle	C/M	NC	Ug49_C_C4.1	45.9		1525	D3	
Ug49	Cattle	C	25F43R	Ug49_C_C5	46.1	GI	206	PMA,D1	
Ug49	Cattle	M	25F43R	Ug49_C_M1	44.2	GI	1291	PMG,PMF,PMA, PMB,PMC,D1, D2,D3	
Ug49	Cattle	M	25F43R	Ug49_C_M2	41.9	GI	954	PMF,PMA,PMB PMC,D1,D2,D3	
Ug49	Cattle	M	25F43R	Ug49_C_M3	44.8	GI	967	D1,D2	
Ug50	Cattle	M	25F43R	Ug50_C_M1	44	GI	1032	PMG,PMF,PMA, D1	3
Ug50	Cattle	M	25F43R	Ug50_C_M2	42	GI	1063	PMF,D1	
Ug50	Cattle	M	25F43R	Ug50_C_M3	43.4	GI	392	PMF,PMA,D1	
Ug51	Cattle	S	25F43R	Ug51_C_S1	45.1	GI	126	D1	1
Ug52	Cattle	S	25F43R	Ug52_C_S1	46.2	GI	197	D1	1
Ug53	Cattle	S	25F43R	Ug53_C_S1	41.7	GI	187	D1	2
Ug53	Cattle	M	25F43R	Ug53_C_M1	43.7	GI	1223	PMF,D1,D2,D3	
Ug54	Cattle	S	24R	Ug54_C_S1	43.5	GII	349		6
Ug54	Cattle	S	43R	Ug54_C_S2	42.7	GI	261		
Ug54	Cattle	S	25F	Ug54_C_S3	44.3	GI	173		
Ug54	Cattle	S	NCR8	Ug54_C_S4	45.9		362		
Ug54	Cattle	S	NCR5	Ug54_C_S5	48.8		226		
Ug54	Cattle	M	25F43R	Ug54_C_M1	51.2	GI	557	PMG,PMF,PMA, D1	
Ug54	Cattle	M	25F43R	Ug54_C_M2	45.1	GI	308	PMF,PMA,D1	
Ug55	Cattle	S	25F43R	Ug55_C_S1	38.3	GI	188	D1	6
Ug55	Cattle	S	DP1358R	Ug55_C_S2	49.4		158	PMF	
Ug55	Cattle	C	23F24R	Ug55_C_C1	48.5	GII	490		
Ug55	Cattle	C	NC	Ug55_C_C2	47.1		933	D3	
Ug55	Cattle	C	23F24RNC	Ug55_C_C3	47.6	GII	578	PMA,D1,D2	
Ug55	Cattle	M	25F43R	Ug55_C_M1	41.3	GI	1363	PMF,PMA,PMC	
Ug56	Cattle	S	25F	Ug56_C_S1	40.1	GI	180		2
Ug56	Cattle	S	43R	Ug56_C_S2	42.9	GI	187		
Ug56	Cattle	S	23F	Ug56_C_S3	43.7	GII	760		
Ug56	Cattle	S	NCR8	Ug56_C_S4	41.8		361		
Ug57	Cattle	S	25F43R	Ug57_C_S1	48.4	GI	188	D1	1
Ug57	Cattle	M	25F43R	Ug57_C_M1	45.9	GI	558	PMF,D1	
Ug58	Cattle	S	25F	Ug58_C_S1	41.8	GI	148		3
Ug58	Cattle	S	NCR8	Ug58_C_S2	39.7		230		
Ug58	Cattle	M	25F43R	Ug58_C_M1	42.7	GI	419	PMF,PMA,D1	
Ug58	Cattle	M	25F43R	Ug58_C_M2	40.5	GI	679	PMD.D1,D2,D3	
Ug59	Cattle	S	25F	Ug59_C_S1	42.6	GI	176		3
Ug59	Cattle	S	NCR8	Ug59_C_S2	44.3		317		

Sample	Species	Method	Primers	Seq name	%GC	Genogroup (if known)	Seq length	Conserved regions	Total PBVs per sample
Ug59	Cattle	C/M	23F24R	Ug59_C_C1.1	43.1	GII	704	D3	
Ug59	Cattle	C	23F24R	Ug59_C_C2	44	GII	205	PMA,D1	
Ug60	Cattle	S	25F43R	Ug60_C_S1	44	GI	211	PMA,D1	9
Ug60	Cattle	S	NCR8	Ug60_C_S2	48.2		323		
Ug60	Cattle	M	25F43R	Ug60_C_M1	40.8	GI	942	PMF,PMA,D1	9 (as above)
Ug60	Cattle	M	25F43R	Ug60_C_M2	46.1	GI	540	PMF,PMA,PMB, PMC,D1,D2,D3	
Ug60	Cattle	M	25F43R	Ug60_C_M3	44.7	GI	371	PMF,PMA,D1	
Ug60	Cattle	C	NC	Ug60_C_C1	44.9		236		
Ug60	Cattle	C	NC	Ug60_C_C2	47.9		335	D3	
Ug60	Cattle	C	NC	Ug60_C_C3	44.5		171		
Ug60	Cattle	C/M	NC	Ug60_C_C3.1	42		743		
Ug60	Cattle	C/M	NC	Ug60_C_C4.1	42		715	D3	
Ug60	Cattle	C/M	NC	Ug60_C_C6.1	44.7		885	D3	
Ug61	Cattle	S	25F43R	Ug61_C_S1	45.6	GI	196		2
Ug61	Cattle	S	23F24R	Ug61_C_S2	46.5	GII	344		
Ug62	Cattle	S	NCF3	Ug62_C_S1	47.6	GI	370		4
Ug62	Cattle	S	23F24R	Ug62_C_S2	45	GII	347		
Ug62	Cattle	M	25F43R	Ug62_C_M1	44.7	GI	1504	PMF,PMB,PMC, PMD,PME,D1, D2,D3	
Ug62	Cattle	M	25F43R	Ug62_C_M2	42.1	GI	1688	PMG,PMF,PMC, PME,D2,D3	
Ug63	Cattle	S	25F	Ug63_C_S1	45.5	GI	165		2
Ug63	Cattle	M	25F43R	Ug63_C_M1	42	GI	1472	PMG,PMF,PMA, PMB,PMC,D1, D2,D3	
NZC01	Cattle	C	25F43R	NZC01_C_C1	44.1	GI	203		1
NZCR03	Crypto	S	23F24R	NZCR03_S1	55.4	GII	148		1

Ug: Uganda; NZ: New Zealand. Species: Crypto=purified Cryptosporidium oocysts (see Chp 7). PBV: Picobirnavirus. Method: S=Sanger sequencing from RT-PCR; C=cloned and then Sanger sequencing from RT-PCR; M=Metagenomics. Primers: 25F/43R for *RdRp* genogroup I and 23F/24R for *RdRp* genogroup II from Rosen et al. 2000; F3/F5/R5/R8 primers uses all four primers for identification of any genogroup PBV from Anthony et al. 2015. %GC is percent guanine-cytosine content in sequence. Seq length is length of the sequence in base pairs or bp. Conserved regions: PMA=polymerase motif A; PMB=polymerase motif B; PMC=polymerase motif C; PMD=polymerase motif D; PME=polymerase motif E; PMF=polymerase motif F; PMG=polymerase motif G; D1=conserved domain 1; D2=conserved domain 2; D3=conserved domain 3.

TABLE 17 DATASET OF METAGENOMIC SEQUENCING READS

Sample	Host	Total reads	dsRNA virus reads	PBV reads
Ug01	Human	564,262	15	2
Ug02	Human	522,602	3	0
Ug03	Human	440,694	2	0
Ug04	Gorilla	293,174	2	0
Ug05	Gorilla	333,246	56	0
Ug06	Gorilla	327,014	2669	2587
Ug07	Gorilla	261,276	322	319
Ug08	Gorilla	434,752	6	0
Ug09	Gorilla	684,812	116	0
Ug10	Gorilla	724,874	59	1
Ug11	Gorilla	621,858	9	0
Ug12	Gorilla	672,300	12	8
Ug13	Gorilla	465,950	31	0
Ug14	Gorilla	410,956	12	0
Ug15	Gorilla	564,270	3419	3187

Ug16	Gorilla	478,780	1460	1432
Ug17	Gorilla	562,600	512	456
Ug18	Gorilla	426,420	44	40
Ug19	Gorilla	440,750	13	0
Ug20	Gorilla	393,456	155	155
Ug 21	Gorilla	316,180	10	10
Ug 22	Gorilla	300,424	1338	1325
Ug 23	Gorilla	201,362	31	31
Ug 24	Gorilla	404,090	4	0
Ug 25	Gorilla	656,562	923	677
Ug 26	Gorilla	502,684	115	4
Ug 27	Gorilla	415,296	18	16
Ug 28	Gorilla	484,832	296	194
Ug 29	Gorilla	315,790	7	2
Ug 30	Gorilla	365,150	338	113
Ug 31	Gorilla	427,138	13	3
Ug 32	Gorilla	380,558	81	0
Ug 33	Gorilla	703,630	49	2
Ug 34	Gorilla	439,430	24	2
Ug 35	Gorilla	540,084	25	0
Ug 36	Gorilla	251,086	0	0
Ug 37	Gorilla	274,290	276	0
Ug 38	Gorilla	196,082	596	569
Ug 39	Gorilla	348,502	417	413
Ug 40	Gorilla	350,958	0	0
Ug 41	Gorilla	834,326	39	15
Ug 42	Gorilla	666,438	121	97
Ug 43	Gorilla	473,232	281	279
Ug 44	Gorilla	570,726	196	188
Ug 45	Gorilla	576,848	419	381
Ug 46	Gorilla	448,712	209	121
Ug 47	Gorilla	612,666	28	0
Ug 48	Gorilla	471,806	3219	3219
Ug 49	Cattle	817,864	3968	3967
Ug 50	Cattle	370,178	2357	2356
Ug 51	Cattle	550,480	4630	4630
Ug 52	Cattle	46,212	21	21
Ug 53	Cattle	526,498	1561	1559
Ug 54	Cattle	525,614	1634	1632
Ug 55	Cattle	459,022	1775	1774
Ug 56	Cattle	343,540	1775	1769
Ug 57	Cattle	747,922	1293	1292
Ug 58	Cattle	762,592	8829	8823
Ug 59	Cattle	736,024	433	431
Ug 60	Cattle	512,756	3120	3114
Ug 61	Cattle	441,936	378	318
Ug 62	Cattle	419,030	5208	5204
Ug 63	Cattle	501,704	304	304
RNA_CONTROL	Control	469,730	8	0

TABLE 18 SUMMARY OF PICOBIRNAVIRUS *RDRP* SEQUENCES AND HIGHEST NCBI BLAST MATCH AND SPECIFICS

Seq Name	Accession number	Description of accession	Accession length	Bit score	E-value	Query cover	Pairwise ID
Ug01_H_S1	EU938815	Raw sewage PBV, Louisiana	201	114	1e-21	75	76.6
Ug01_H_S2	KT720491	Human PBV, 406-SZ12	202	104	2e-18	71	78
Ug02_H_S1	KX964659	Bovine PBV, isolate 06	162	159	3e-35	83	80
Ug02_H_S2	MG970262	Himalayan goral PBV, GH31	201	134	1e-27	97	74.1
Ug03_H_S1	MN729475	Mongoose PBV, M26	136	45.5	2.2	17	83.3
Ug03_H_S2	KC846784	Porcine PBV, CHHN-B5	201	150	2e-32	92	77.7
Ug03_H_S3	KC841460	Porcine PBV, group II, CHHN-A6	369	604	2e-168	93	97.5
Ug04_G_S1	KJ135796	Wastewater PBV, Hunan 6	201	275	6e-70	100	93.4
Ug06_G_S1	MH933835	Human PBV, CMRHP49B	1643	242	3e-59	100	82.5
Ug06_G_M1	MH933835	Human PBV, CMRHP49B	1643	842	0	91	85.6
Ug07_G_M1	KY928707	Marmot PBV, c332189	1530	198	93-46	77	66.7
Ug08_G_S1	KC846784	Porcine PBV, CHHN-B5	201	124	2e-24	84	81.5
Ug08_G_S3	KY120176	Bovine PBV, C343R 2-2	1695	156	3e-34	100	82.2
Ug10_G_S1	KT204491	Human PBV, 406-SZ12	202	221	1e-53	100	85.7
Ug14_G_S1	MG010911	Macaque PBV, 26	1627	174	1e-39	100	86.7
Ug15_G_S1	KJ135796	Wastewater PBV, Hunan 6	201	301	2e-77	96	93.4
Ug15_G_S2	MH425585	Chicken PBV, ChPBV-S2-Nov1	1677	682	0	95	80.5
Ug15_G_M1	KT334939	PREDICT PBV, PbV-36	544	432	2e-116	32	93.1
Ug15_G_C1.1	MH425585	Chicken PBV, ChPBV-S2-Nov1	1677	1468	0	99	81.3
Ug15_G_S3	KR902507	Equ3 PBV	1704	134	2e-27	74	70.7
Ug16_G_S1	MH933803	Human PBV, CMRHP20A	1648	328	2e-85	74	87
Ug16_G_M1	JF755420	Microtus PBV, V-111 USA	1245	190	5e-44	64	74.5
Ug17_G_S1	KJ135902	Wastewater PBV, Hunan 112	201	137	3e-28	100	80.4
Ug17_G_S2	MH933803	Human PBV, CMRHP20A	1648	225	1e-54	96	84.1
Ug17_G_S3	MH933822	Human PBV, CMRHP52	1387	271	1e-68	100	83.5
Ug17_G_S4	KY928708	Marmot PBV, c332280	1668	240	2e-58	92	68.9
Ug17_G_M1	JF755420	Microtus PBV, V-111 USA	1245	190	5e-44	73	74.5
Ug22_G_C1.1	KY928700	Marmot PBV, c327427	1747	390	1e-103	75	69.1
Ug22_G_C2	KJ135902	Wastewater PBV, Hunan 112	201	171	2e-38	98	82.3
Ug22_G_C2.1	JF755420	Microtus PBV, V-111 USA	1245	190	8e-44	49	74.5
Ug22_G_M1	JF755420	Microtus PBV, V-111 USA	1245	192	5e-44	23	74.6
Ug23_G_S1	KY120170	Bovine PBV, C343N 2-2	1671	244	9e-61	100	90.4
Ug25_G_M1	KT934307	Wolf PBV, 416 PRT	1667	180	2e-40	52	75.2
Ug27_G_S1	KY120170	Bovine PBV, C343N 2-2	1671	239	3e-59	100	90.5
Ug28_G_M1	MH933831	Human PBV, CMRHP21C	1163	902	0	98	82.3
Ug39_G_S1	MF693848	Ovine PBV, PBV_OVI29	201	131	2e-26	100	75.8
Ug41_G_S1	KJ135922	Wastewater PBV, Hunan 132	201	233	6e-57	100	87.6
Ug42_G_S1	KC865806	Avian PBV, AVE_57 BRA	201	124	2e-24	100	78.6
Ug43_G_S1	MK378853	Porcine PBV, JL01 C3	864	203	3e-48	100	85
Ug43_G_S2	MG010912	Macaque PBV, 27	1460	656	0	94	82.9
Ug43_G_C1	KJ135922	Wastewater PBV, Hunan 132	201	253	2e-63	98	88
Ug43_G_C2	KJ135902	Wastewater PBV, Hunan 112	201	190	2e-44	96	81
Ug43_G_M1	KY120175	Bovine PBV, C343R 2-1	1692	386	2e-102	92	73.5
Ug43_G_M2	KY502872	Gorilla PBV	1552	135	4e-27	40	76.9
Ug43_G_M3	MH933824	Human PBV, CMRHP63B	1680	593	7e-165	94	81.9
Ug44_G_S1	KJ135823	Wastewater PBV, Hunan 33	201	252	7e-63	100	89.8
Ug45_G_M1	KT334939	PREDICT PBV, PbV-36	544	509	1e-139	41	90.3
Ug45_G_M2	MH425590	Chicken PBV, ChPBV-S2-Nov6	1664	344	3e-90	97	85.3
Ug47_G_S1	KY120170	Bovine PBV, C345N 2-2	1671	236	3e-58	96	91.6

Seq Name	Accession number	Description of accession	Accession length	Bit score	E-value	Query cover	Pairwise ID
Ug48_C_M1	MK378865	Porcine PBV, NX02 C1	1309	910	0	90	85.5
Ug49_C_S1	MF693847	Ovine PBV, PBV_OVI05 BRA	201	260	2e-65	84	88.9
Ug49_C_C2	KJ135902	Wastewater PBV, Hunan 112	201	153	6e-33	90	77.5
Ug49_C_C3.1	KY120170	Bovine PBV, C345N 2-2	1671	626	3e-174	93	70.9
Ug49_C_C4.1	KR827415	Human PBV, ChXz-4	1585	696	0	91	71.8
Ug49_C_C5	MN196313	Bovine PBV, HAR-EQ-15/CI	201	255	6e-64	97	88.1
Ug49_C_M1	KY120170	Bovine PBV, C345N 2-2	1671	545	6e-150	86	70.8
Ug49_C_M2	LC338004	Dromedary PBV, 78C/GpI	1694	721	0	95	80.3
Ug49_C_M3	MG846408	Chicken PBV, RS/BR/15/5R-1	1336	242	2e-59	54	71.5
Ug50_C_M1	KY120176	Bovine PBV, C343R 2-2	1695	1030	0	92	86.2
Ug50_C_M2	KY120182	Simian PBV, 1R 2-2	1681	487	1e-132	91	71.6
Ug50_C_M3	MN871976	African green monkey PBV, S-3	1641	259	1e-64	68	81.3
Ug51_C_S1	GU230508	Horse PBV, BRA-02/2009 II	201	134	1e-27	99	84.8
Ug52_C_S1	KY120170	Bovine PBV, C345N 2-2	1671	288	1e-73	98	92.8
Ug53_C_S1	GU230496	Horse PBV, BRA-02 2009 O	198	171	2e-38	100	80.2
Ug53_C_M1	MN563301	Mongoose PBV, M17B	1715	618	3e-172	97	73.9
Ug54_C_S1	KP868555	Porcine PBV, Por-208 II	310	358	1e-94	76	89.2
Ug54_C_S2	GU230540	Porcine PBV, BRA-02 1999 O	201	123	1e-23	45	82.8
Ug54_C_S3	MK378865	Porcine PBV, NX02 C1	1309	123	8e-24	91	74.4
Ug54_C_S4	KY928729	Marmot PBV, c424142	1615	193	1e-44	80	74.3
Ug54_C_S5	KY928717	Marmot PBV, c299351	1718	159	4e-35	74	80
Ug54_C_M1	MH425590	Chicken PBV, ChPBV-S2-Nov6	1664	223	1e-53	47	79
Ug54_C_M2	MK378853	Porcine PBV, JL01 C3	864	297	3e-76	86	85.1
Ug55_C_S1	MF693849	Ovine PBV, PBV_OVI55	201	187	2e-43	100	81.9
Ug55_C_S2	KY120170	Bovine PBV, C345N 2-2	1671	153	4e-33	67	91.6
Ug55_C_C1	KP984805	Porcine PBV, CYZ-II-1	1591	623	3e-174	75	97.3
Ug55_C_C2	MK521925	Tasmanian devil PBV, Stoney 5	1642	269	2e-67	87	67.8
Ug55_C_C3	MN196315	Bovine PBV, HAR-EQ-15 CII	607	187	9e-43	55	73.3
Ug55_C_M1	KY120170	Bovine PBV, C345N 2-2	1671	682	0	98	74.2
Ug56_C_S1	GU230525	Porcine PBV, BRA-01 1998 S	207	80.6	2e-11	62	71.4
Ug56_C_S2	KP941111	Fox PBV, 55590	1650	90.6	5e-14	91	76
Ug56_C_S3	KT335253	PREDICT PBV, PbV-118	520	522	1e-143	68	82.3
Ug56_C_S4	KY120178	Bovine PBV, C372N 2	1622	338	1e-88	88	83.2
Ug57_C_S1	MH933806	Human PBV, CMRHP25A	1376	141	9e-30	100	76.6
Ug57_C_M1	MG846411	Chicken PBV, RS/BR/15/5S-3	1607	210	7e-50	75	72.3
Ug58_C_S1	KY120170	Bovine PBV, C345N 2-2	1671	200	3e-47	99	89.8
Ug58_C_S2	MK378854	Porcine PBV, JL01 C4	791	113	6e-21	56	78.7
Ug58_C_M1	KY120170	Bovine PBV, C345N 2-2	1671	456	8e-124	86	88.2
Ug58_C_M2	KY120177	Bovine PBV, C369R 2-1	1736	420	1e-112	100	75
Ug59_C_S1	KJ135919	Wastewater PBV, Hunan 129	201	101	3e-17	67	81.2
Ug59_C_S2	KY120178	Bovine PBV, C372N 2	1622	384	3e-102	97	89
Ug59_C_C1.1	KY120178	Bovine PBV, C372N 2	1622	886	0	99	88.3
Ug59_C_C2	KJ135922	Wastewater PBV, Hunan 132	201	252	8e-63	97	87.5
Ug60_C_S1	GU230508	Horse PBV, BRA-02 2009 II	201	280	2e-71	95	91
Ug60_C_S2	MK521924	Tas devil PBV, 5 Stoney Head	1655	185	1e-42	94	74.4
Ug60_C_M1	KF823812	Genet PBV, S13	811	204	2e-47	28	77
Ug60_C_M2	KY120176	Bovine PBV, C343R 2-2	1695	438	3e-118	99	78.4
Ug60_C_M3	KM573801	Dromedary PBV, c4180	1632	339	1e-88	78	85.9
Ug60_C_C1	KR827416	Human PBV, ChXz-5	1455	123	3e-24	90	73.2
Ug60_C_C2	MK521924	Tas devil PBV, 5 Stoney Head	1655	244	2e-60	97	77
Ug60_C_C3	KY928727	Marmot PBV, c393404	1612	91.5	1e-14	80	73

Seq Name	Accession number	Description of accession	Accession length	Bit score	E-value	Query cover	Pairwise ID
Ug60_C_C3.1	KM573809	Dromedary PBV, c6608	954	209	3e-49	51	78.5
Ug60_C_C4.1	KM573809	Dromedary PBV, c6608	954	149	3e-31	73	66.7
Ug60_C_C6.1	MH425586	Chicken PBV, ChPBV-S2-Nov2	1652	178	7e-40	39	72.4
Ug61_C_S1	MH835431	Goat PBV, C5	234	205	9e-49	98	83
Ug61_C_S2	MF787750	Deer PBV, ABT73 II	366	359	1e-94	100	83
Ug62_C_S1	MH933821	Human PBV, CMRHP49A	1584	420	5e-113	99	85
Ug62_C_S2	KP868555	Porcine PBV, Por-207 II	310	444	1e-120	86	92.7
Ug62_C_M1	KT335048	PREDICT PBV, PbV-65	544	860	0	42	95
Ug62_C_M2	KY120194	Simian PBV, 1R 2-1	1613	664	0	84	71
Ug63_C_S1	LC338006	Dromedary PBV, 103C/GpI	1611	161	8e-36	98	81.3
Ug63_C_M1	LC338004	Dromedary PBV, 78C/GpI	1694	508	5e-139	91	69.9
NZC01_C_C1	MH835431	Goat PBV, strain C5	234	257	2e-64	100	88.3
NZCR03_S1	AB214978	Human PBV, pseudogene	193	268	6e-68	100	100

PBV: Picobirnavirus. Accession numbers are reported from NCBI database based on the highest BLAST match. Accession length is length of the accession in base pairs or bp. Bit score is “derived from the raw alignment score, taking the statistical properties of the scoring system into account.” E-value is representative of “the number of different alignments with scores equivalent to or better than is expected to occur in a database search by chance.” Query cover is the amount of nucleotides or amino acids that the subject shares with the input sequence, expressed as a percentage of the total number of nucleotides or amino acids in the query. Pairwise ID is pairwise identity and is the “extent to which two (nucleotide or amino acid) sequences have the same residues at the same positions in an alignment, expressed as a percentage.” [280]

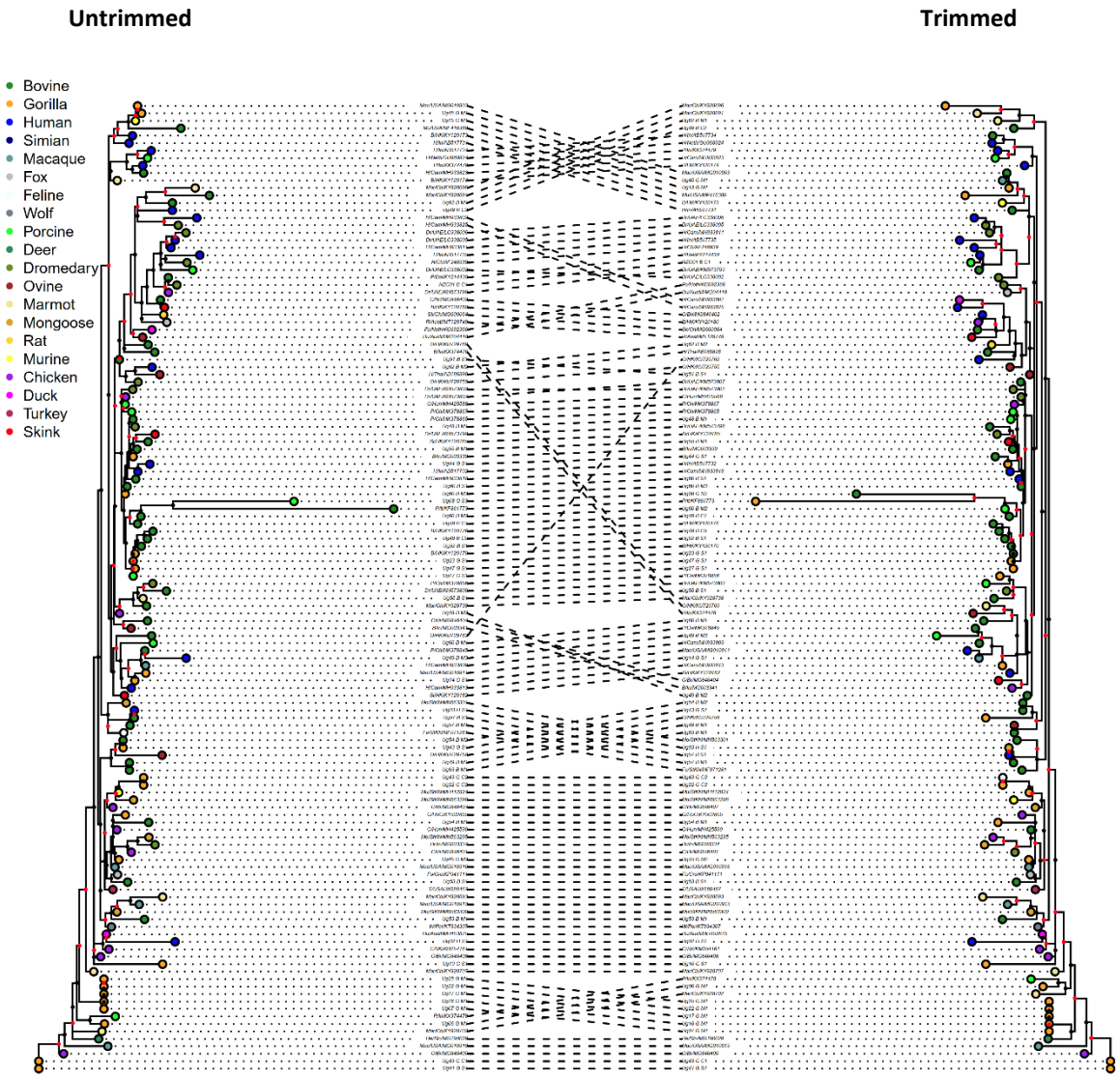


FIGURE 48 CO-PHYLOGENY OF UNTRIMMED TO TRIMMED GENOGROUP I PICOBIRNAVIRUS SEQUENCES WITH NAMES

Picobirnavirus *RdRp* genogroup I sequences from this study and highest BLAST match comparing untrimmed to trimmed sequences. Untrimmed sequences are on the left-hand side of the co-phylogeny trees and trimmed sequences are on the right-hand side of the co-phylogeny trees. Total of 130 sequences with the untrimmed sequences of 80–88aa in length and the trimmed sequences of 60–72aa in length. Percent identity of the untrimmed genogroup I was 65.2% and of the trimmed genogroup I was 73.8%. Best tree model was selected from PhyML (Section 3.7.5.3) as LG + G + I for untrimmed and trimmed trees. Fast-likelihood-based methods of aLRT SH-like were used for branch support values (Section 3.7.5.3). Black nodes denote branch support values >0.8 and red nodes <0.8. Coloured branch labels show different PBV hosts; dark green=B=cattle/bovine, bright green=P=porcine/pig, green=Dr=deer, orange=G=gorilla, purple=C=chicken, blue=H=human, dark blue=Si=simian, olive=Dr=dromedary, red=SK=skink, burgundy=O=ovine, khaki=Mar=marmot, magenta=Du=duck, dark purple=T=turkey, yellow=Mu=murine, cadet blue=Mac=macaque, goldenrod=Mo=mongoose, light azure=Fe=feline, light grey=Fo=fox, dark grey=W=wolf, gold=R=rabbit. Black nodes denote branch support values >0.8 and red nodes <0.8. Labels from the NCBI PBVs are host species as noted above/country/accession number. Labels for the samples from the study are: Ug# refers to the sample number from Uganda and NZ# from the sample number from New Zealand. G after the Ug# or NZ# refers to

gorilla samples, H refers to human samples and C refers to cattle samples. S#, C# or M# at the third site on the labels designates S#=Sanger sequence, C#=cloned sequence, M#=metagenomic sequence.

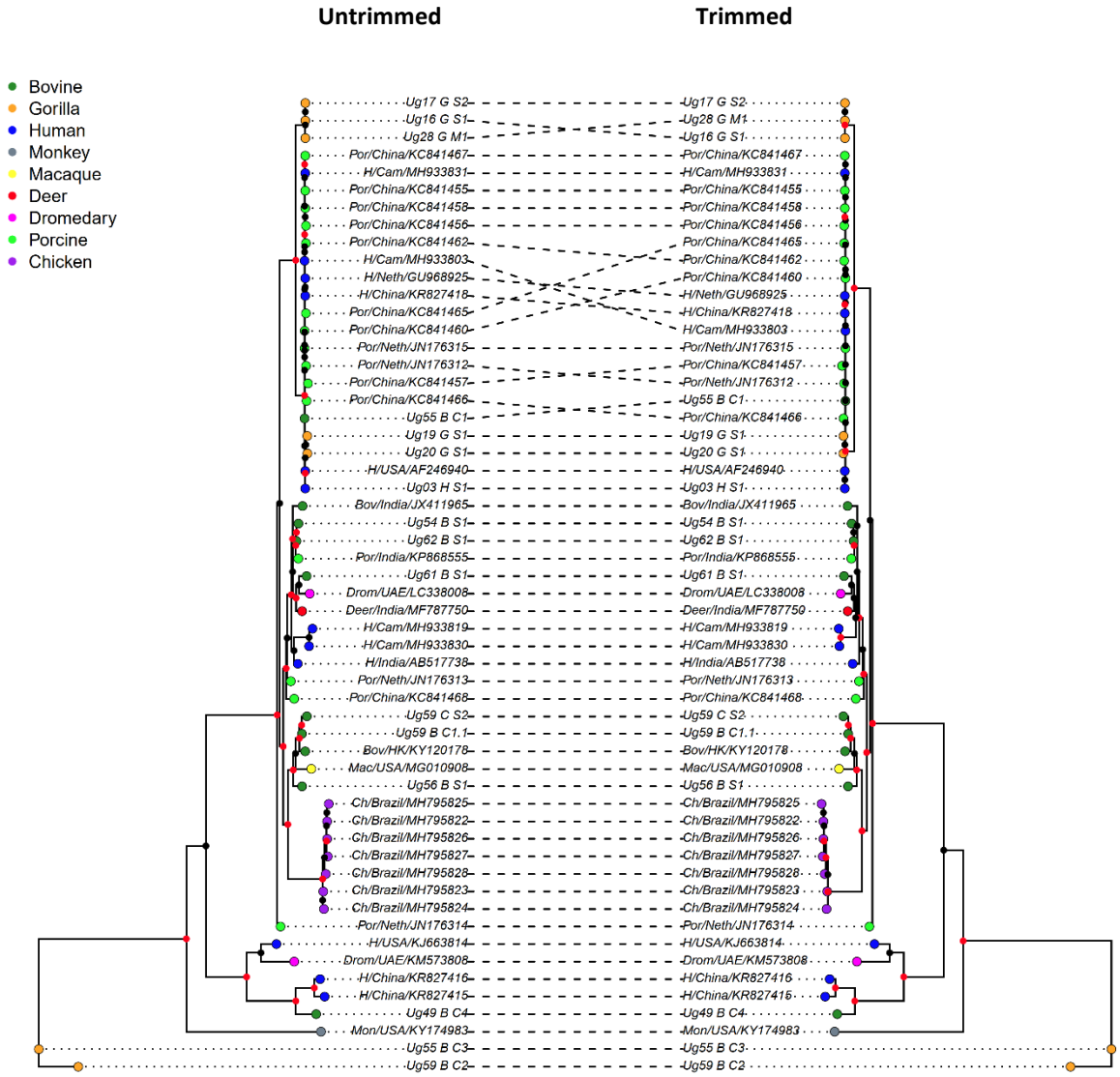


FIGURE 49 CO-PHYLOGENY OF UNTRIMMED TO TRIMMED GENOGROUP II PICOBIRNAVIRUS SEQUENCES

Picobirnavirus *RdRp* genogroup II sequences from this study and highest BLAST match comparing untrimmed to trimmed sequences. Untrimmed sequences are on the left-hand side of the co-phylogeny trees and trimmed sequences are on the right-hand side of the co-phylogeny trees. Total of 56 sequences with the untrimmed of 137aa in length and the trimmed of 67–121aa in length. Percent identify of the untrimmed genogroup II was 61.7% and of the trimmed genogroup II was 60.1%. Best tree model was selected from PhyML (Section 3.7.5.3) as LG + G for untrimmed and trimmed trees. Fast-likelihood-based methods of aLRT SH-like were used for branch support values (Section 3.7.5.3). Black nodes denote branch support values >0.8 and red nodes <0.8. Coloured branch labels show different PBV hosts; dark green=B or Bov=cattle/bovine, bright green=deer, orange=G=gorilla, blue=H=human, light blue=Mon=monkey, red=Drom=dromedary, grey=Mac=macaque, pink=Por=pig/porcine, yellow=Ch=chicken. Black nodes denote branch support values >0.8 and red nodes <0.8. Labels from the NCBI PBVs are host species/country/accession number. Labels for the samples from the study are: Ug# refers to the sample number from Uganda. G after the Ug# or NZ# refers to gorilla samples, H refers to human samples and C refers to cattle samples. S#, C# or M# at the third site on the labels designates S#=Sanger

sequence, C#=cloned sequence, M#=metagenomic sequence. Labels for the NCBI picobirnaviruses are: species as noted above with colours/Country of origin: USA=United States of America, HK=Hong Kong, UAE=United Arab Emirates, Cam=Cameroon, Neth=Netherlands,/NCBI Accession number.

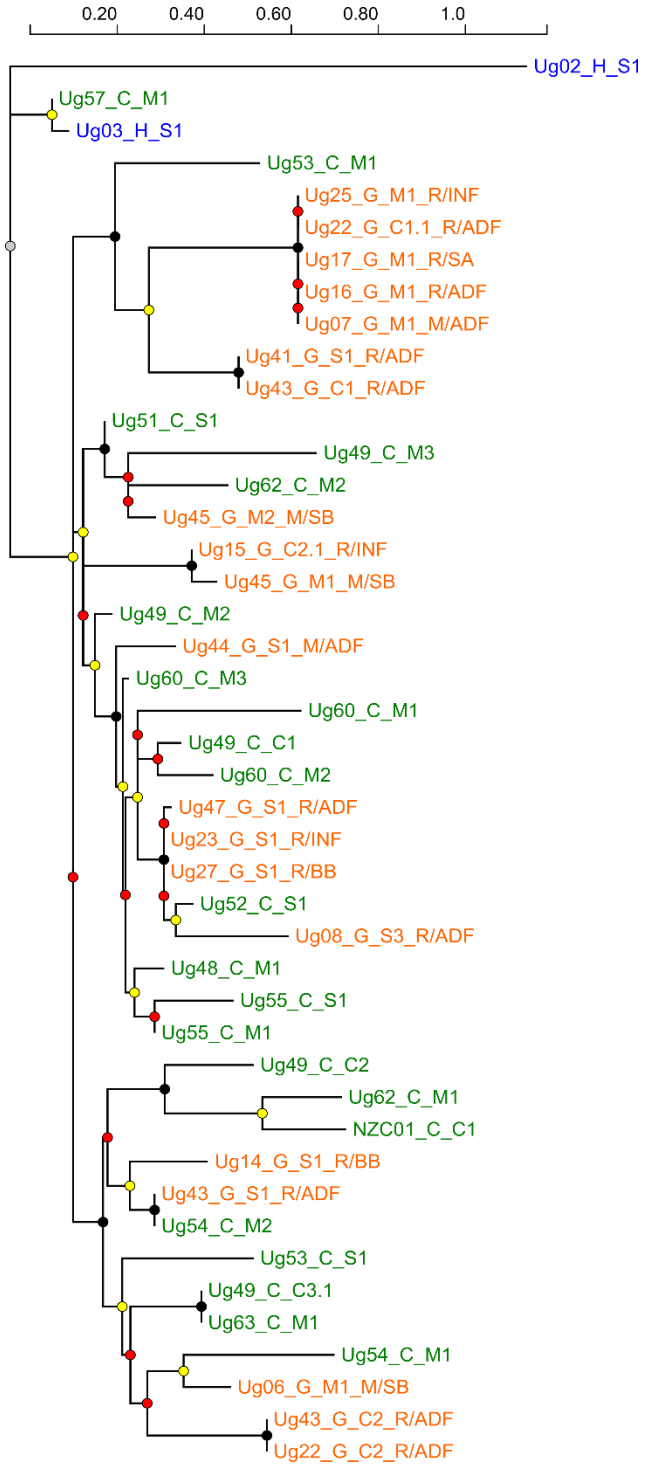


FIGURE 50 PHYLOGENETIC TREE OF THE GENOGROUP I PICOBIRNAVIRUSES FROM THIS STUDY UNROOTED

***RdRp* genogroup I picobirnavirus 52 sequences from Uganda and New Zealand with color-coding of hosts; orange: gorilla, green: cattle, blue: human. The tree includes duplicates from the same sample and is**

unrooted. Amino acid sequences trimmed to between 65-66 amino acids (Section 3.7.5.1). Percent identity for the multiple alignment of the tree was 85.2%. Best tree model was selected from PhyML (Section 3.7.5.3) as LG + G. Fast-likelihood-based methods of aLRT SH-like were used for branch support values (Section 3.7.5.3). Black nodes denote branch support values >0.8, yellow nodes between 0.5 and 0.8 and red nodes <0.5. Labels for the samples from the study are: Ug# refers to the sample number from Uganda and NZ# from the sample number from New Zealand. H after the Ug# or NZ# refers to human samples, G refers to gorilla samples and C refers to cattle samples. S#, C# or M# at the third site on the labels designates S#=Sanger sequence, C#=cloned sequence, M#=metagenomic sequence. The fourth site for the Gorilla samples designates the gorilla family and gender and age if applicable: R=Rushegura gorilla family, M=Mubare gorilla family/INF=infant, SA=sub-adult, BB=blackback male, SB=silverback male, ADF=adult female [225](per Mwebe et al 1998).

APPENDIX C: SUPPLEMENTARY MATERIAL FOR CHAPTER 5

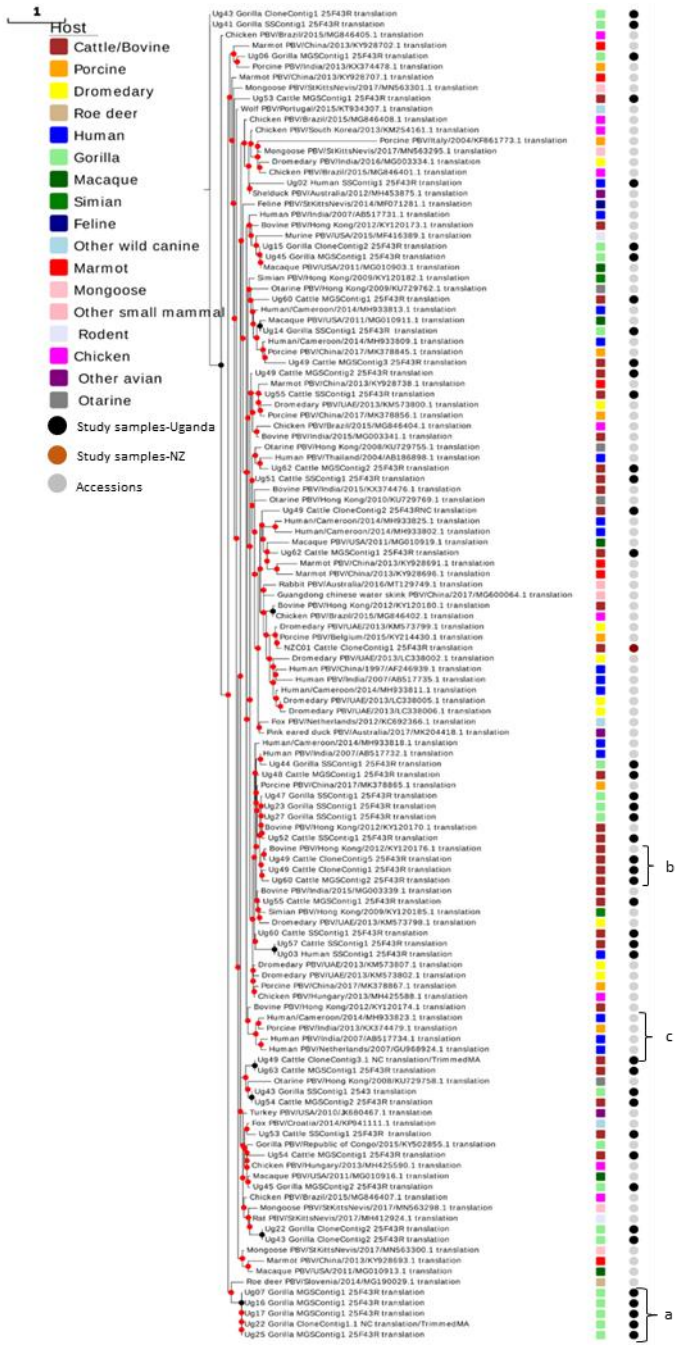


FIGURE 51 PHYLOGENETIC TREE OF THE GENOGROUP I PICOBIRNAVIRUSES BY HOST

Tree of host associations from the multiple alignment of 126 sequences of *RdRp* genogroup I picobirnaviruses from study and accessions from the NCBI database. Amino acid sequences trimmed to between 65-66 amino acids (Section 3.7.5.1). Percent identity for the host tree was 71.2%. Best tree model was selected from PhyML (Section 3.7.5.3) as LG + G + I. Fast-likelihood-based methods of aLRT SH-like were used for branch support values (Section 3.7.5.3). Black nodes denote branch support values >0.8 and red nodes <0.5. Bracket designated with a is a cluster of gorilla samples from this study; bracket b is a cluster of cattle samples from this study and NCBI picobirnaviruses; bracket c is a cluster of NCBI human and porcine picobirnaviruses. Color-coding in first set of squares for the host species are noted in the legend with additional clarification: other wild canine: wolf,

fox; other small mammal: rabbit; rodent: rat, murine or vole; other avian: shelduck, duck, turkey. Color-coding for the second set of circles: black: study samples from Uganda; brown: study sample from New Zealand; light grey: NCBI accession picobirnaviruses. Labels from the NCBI picobirnaviruses are host species/country/accession number. Labels for the samples from the study are: Ug# refers to the sample number from Uganda and NZ# from the sample number from New Zealand. G after the Ug# or NZ# refers to gorilla samples and C refers to cattle samples. S#, C# or M# at the third site on the labels designates S#=Sanger sequence, C#=cloned sequence, M#=metagenomic sequence. Labels for the NCBI PBVs are: Country of origin: USA=United States of America, UAE=United Arab Emirates/NCBI Accession number.

FIGURE 52 PHYLOGENETIC TREE OF THE GENOGROUP II PICOBIRNAVIRUSES BY HOST

Tree of host associations for the multiple alignment of 59 nucleotide sequences of *RdRp* genogroup II picobirnaviruses from study and a selection of genogroup II picobirnaviruses from the NCBI database. Amino acid sequences trimmed to between 120-130 amino acids (Section 3.7.5.1). Percent identity for the host tree was 72.4%. Best tree model was selected from PhyML (Section 3.7.5.3) as WAG + G. Fast-likelihood-based methods of aLRT SH-like were used for branch support values (Section 3.7.5.3). Black nodes denote branch support values >0.8 and red nodes <0.8. Bracket designated with a include a human sample from this study with a human NCBI *Picobirnavirus* sequence; bracket b is a cluster of gorilla samples from this study; bracket c is a cluster of NCBI human and porcine picobirnaviruses. Color-coding in first set of squares for the host species are designated by the legend. Color-coding for the second set of circles: black: study samples from Uganda; light grey: NCBI accession top hits picobirnaviruses. Labels from the NCBI picobirnaviruses are host species/country/year collected/accession number. Labels for the samples from the study are: Ug# refers to the sample number from Uganda. G after the Ug# or NZ# refers to gorilla samples, H refers to human samples and C refers to cattle samples. S#, C# or M# at the third site on the labels designates S#=Sanger sequence, C#=cloned sequence, M#=metagenomic sequence. Labels for the NCBI PBVs are: Country of origin: USA=United States of America, UAE=United Arab Emirates/NCBI Accession number.

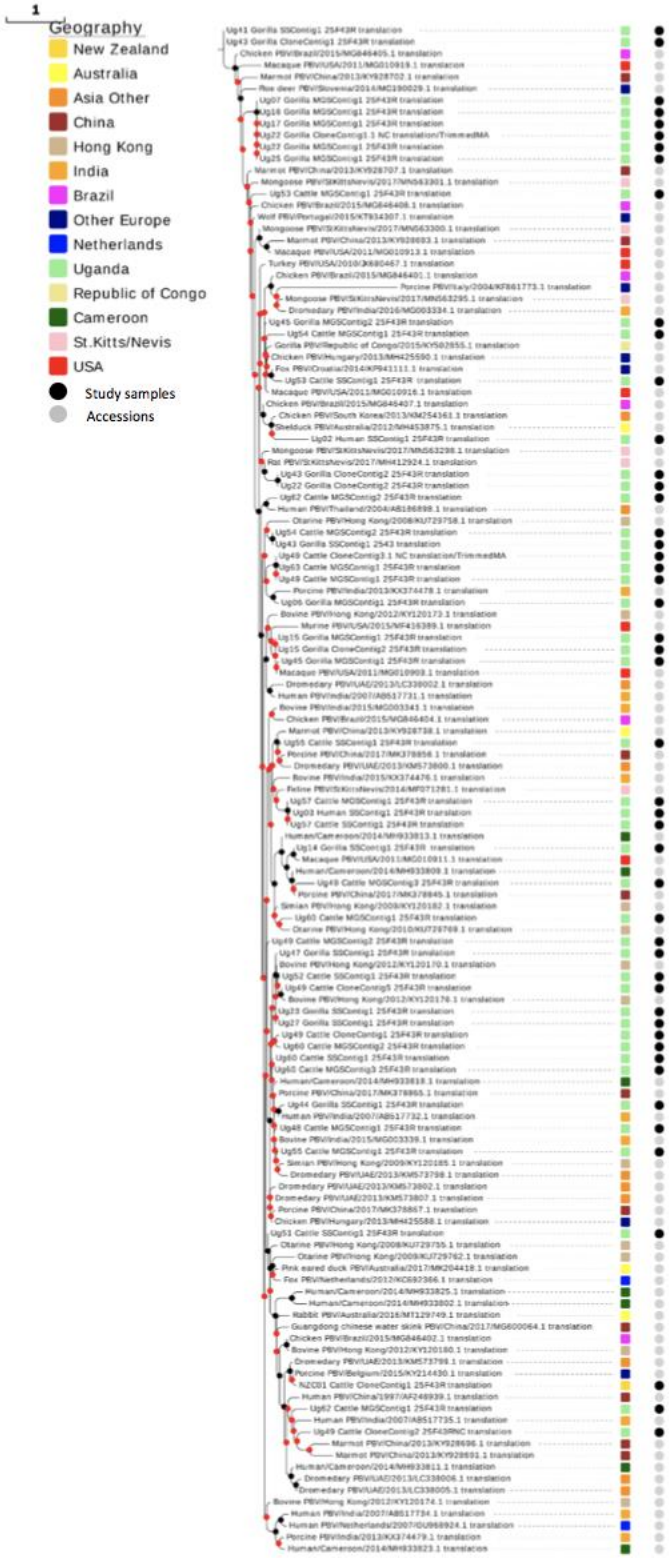


FIGURE 53 PHYLOGENETIC TREE OF THE GENOGROUP I PICOBIRNAVIRUSES BY GEOGRAPHY

Tree of geographic associations from the multiple alignment of 126 sequences of *RdRp* genogroup I picobirnaviruses from study and accessions from the NCBI database. Amino acid sequences trimmed to

between 65-66 amino acids (Section 3.7.5.1). Best tree model was selected from PhyML (Section 3.7.5.3) as LG + G + I. Fast-likelihood-based methods of aLRT SH-like were used for branch support values (Section 3.7.5.3). Black nodes denote branch support values >0.8, yellow nodes 0.8 to 0.5 and red nodes <0.5. Color-coding in first set of squares for the geographical regions are: yellow/brown for Asia/Oceania Pacific region (United Arab Emirates: UAE); blue/purple for European countries; green for Africa; pink/red for Americas: (United States of America: USA). Color-coding for the second set of circles: black: study samples from Uganda and New Zealand; light grey: NCBI accession picobirnaviruses. Labels from the NCBI picobirnaviruses are host species/country/year collected/accession number. Labels for the samples from the study are: Ug# refers to the sample number from Uganda. G after the Ug# or NZ# refers to gorilla samples, H refers to human samples and C refers to cattle samples. S#, C# or M# at the third site on the labels designates S#=Sanger sequence, C#=cloned sequence, M#=metagenomic sequence. Labels for the NCBI PBVs are: Country of origin: USA=United States of America, UAE=United Arab Emirates/NCBI Accession number.

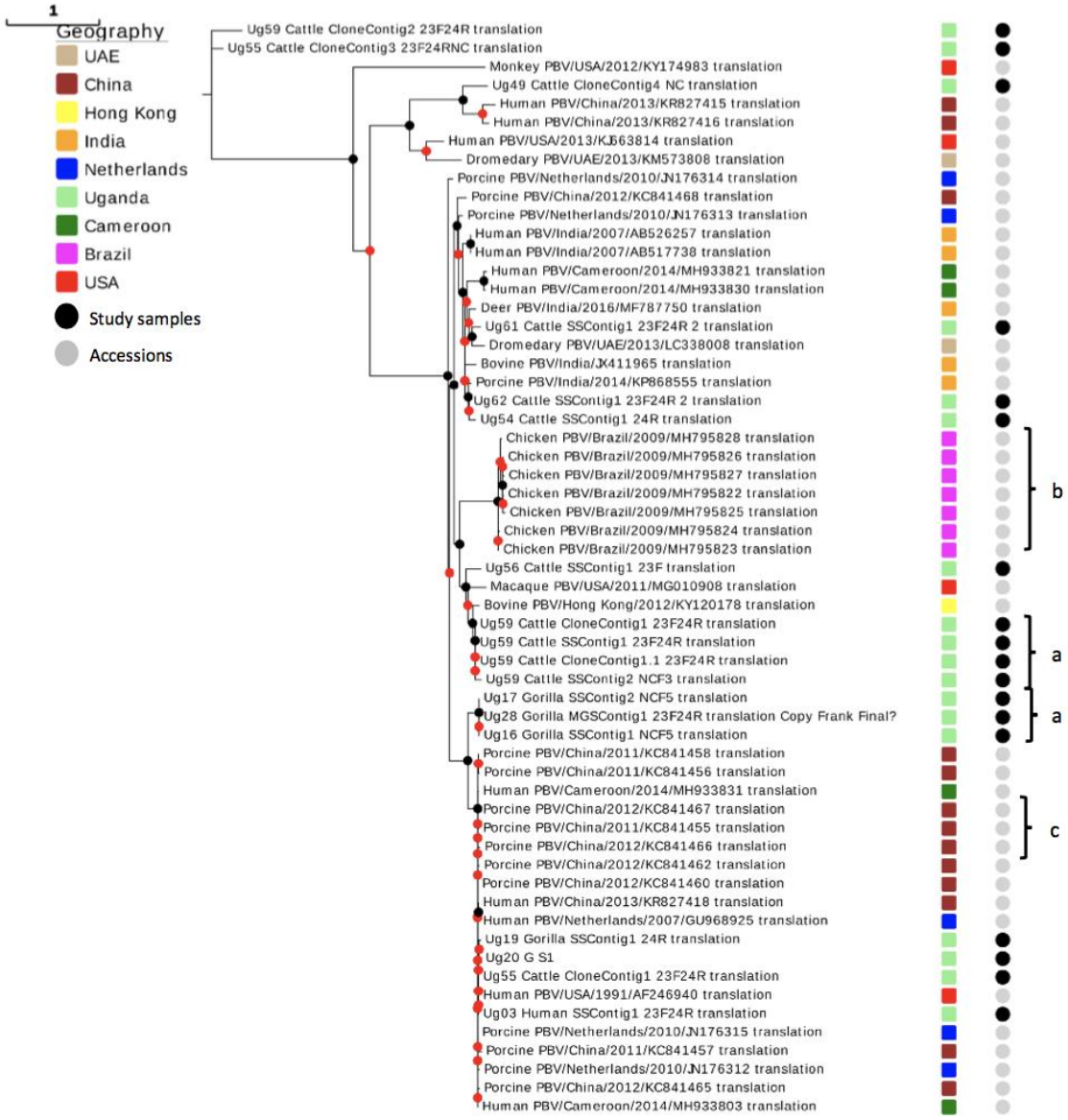


FIGURE 54 PHYLOGENETIC TREE OF THE GENOGROUP II PICOBIRNAVIRUSES BY GEOGRAPHY

Tree of geographic associations for the multiple alignment of 59 nucleotide sequences of *RdRp* genogroup II picobirnaviruses from study and a selection of genogroup II picobirnaviruses from the NCBI database. Amino acid sequences trimmed to between 120-130 amino acids (Section 3.7.5.1). Best tree model was selected from PhyML (Section 3.7.5.3) as WAG + G. Fast-likelihood-based methods of aLRT SH-like were used for branch support values (Section 3.7.5.3). Black nodes denote branch support values >0.8, yellow nodes 0.8 to 0.5 and red nodes <0.5. Brackets designated with a include two small clusters from Uganda; bracket b is a larger cluster from Brazil; bracket c is a smaller cluster from China. PID for the host tree was 72.4%. Color-coding in first set of squares for the geographical regions are: yellow/brown for Asia/Oceania Pacific region (United Arab Emirates: UAE); blue for Europe; green for Africa; pink/red for Americas (United States of America: USA). Color-coding for the second set of circles: black: study samples from Uganda and New Zealand; light grey: NCBI accession picobirnaviruses. Labels from the NCBI picobirnaviruses are host species/country/year collected/accession number. Labels for the samples from the study are: Ug# refers to the sample number from Uganda. G after the Ug# or NZ# refers to gorilla samples, H refers to human samples and C refers to cattle samples. S#, C# or M# at the third site on the labels designates S#=Sanger sequence, C#=cloned sequence, M#=metagenomic sequence. Labels for the NCBI PBVs are: Country of origin: USA=United States of America, UAE=United Arab Emirates/NCBI Accession number.

APPENDIX D: SUPPLEMENTARY MATERIAL FOR CHAPTER 6

TABLE 19 HEATMAPS OF THE MULTIPLE PICOBIRNAVIRUSES WITHIN THE INDIVIDUAL SAMPLES

	G1.1	G1.2
G1.1	100%	40%
G1.2	40%	100%
	G2.1	**G2.2**
G2.1	100%	69%
G2.2	69%	100%
	C3.1	**C3.2**
C3.1	100%	26%
C3.2	26%	100%

	G3.1	G3.2	G3.3	G3.4
G3.1	100%	72%	68%	18%
G3.2	72%	100%	76%	19%
G3.3	68%	76%	100%	16%
G3.4	18%	19%	16%	100%

	C2.1	C2.2	C2.3	C2.4	C2.5
C2.1	100%	40%	25%	24%	23%
C2.2	40%	100%	25%	29%	25%
C2.3	25%	25%	100%	93%	82%
C2.4	24%	29%	83%	100%	87%
C2.5	23%	25%	82%	87%	100%

	C1.1	C1.2	C1.3	C1.4	C1.5	C1.6	C1.7
C1.1	100%	73%	76%	29%	90%	85%	72%
C1.2	73%	100%	73%	29%	69%	74%	71%
C1.3	76%	73%	100%	30%	72%	78%	70%
C1.4	29%	29%	30%	100%	27%	30%	29%
C1.5	90%	69%	72%	27%	100%	82%	75%
C1.6	85%	74%	78%	30%	82%	100%	74%
C1.7	72%	71%	70%	29%	72%	74%	100%

	C4.1	C4.3	C4.4	C4.6	C4.7	C4.8
C4.1	100%	35%	56%	25%	26%	25%
C4.3	35%	100%	31%	30%	26%	27%
C4.4	56%	31%	100%	29%	30%	29%
C4.6	25%	30%	29%	100%	78%	79%
C4.7	26%	26%	30%	78%	100%	87%
C4.8	25%	27%	29%	79%	87%	100%

Percent identity (%) from the multiple alignment of the amino acid sequences from 50-79aa in length within each individual and distances allocated. G=Gorilla, C=Cattle samples from Uganda. G1=Ug15 (2 PBVs), G2=Ug22 (2 PBVs), G3=Ug43 (4 PBVs); C1=Ug49 (7 PBVs), C2=Ug55 (6 PBVs), C3=Ug59 (2 PBVs), C4=Ug60 (8 PBVs). Color-coding for percent identity between two sequences: blue=0–19%, green=20–39%, yellow=40–59%, orange=60–79%, red=80–100% (not including duplicates in tables which are not color-coded).

APPENDIX E: SUPPLEMENTARY MATERIAL FOR CHAPTER 7

TABLE 20 POSITIVE CONVENTIONAL PCR RESULTS OF THE BACTERIA COLONIES AND PROTOZOAL-PURIFIED PRODUCTS TESTED FOR THE IDENTIFICATION OF PICOBIRNAVIRUS

Sample	Host	Primer set	PCR	Contig	Length	BLAST match	Bit sc	E-value
B1	AHSC 1	25F/43R	-	-				
		23F/24R	-	-				
		F3/F5/R5/R8	-	-				
B2	YEP 75	25F/43R	+	+	564	Corynebacterium falsenii; CP007156	577	1.7e-160
		23F/24R	-	-				
		F3/F5/R5/R8	+	+	462	Corynebacterium kutscheri; CP011312	457	4e-124
B3	NZRM 3681	25F/43R	-	-				
		23F/24R	-	-				
		F3/F5/R5/R8	-	-				
B4	FS 296	25F/43R	-	-				
		23F/24R	-	-				
		F3/F5/R5/R8	+	-				
B5	FS/2 295	25F/43R	-	-				
		23F/24R	-	-				
		F3/F5/R5/R8	-	-				
B6	FS 294	25F/43R	-	-				
		23F/24R	-	-				
		F3/F5/R5/R8	-	-				
B7	FS 180	25F/43R	-	-				
		23F/24R	-	-				
		F3/F5/R5/R8	-	-				
B8	CMB 13	25F/43R	-	-				
		23F/24R	-	-				
		F3/F5/R5/R8	-	-				
B9	CMB 12	25F/43R	-	-				
		23F/24R	-	-				
		F3/F5/R5/R8	-	-				
B10	CMB 11	25F/43R	-	-				
		23F/24R	-	-				
		F3/F5/R5/R8	-	-				
B11	*C. zealandensis*	25F/43R	-	-				
		23F/24R	-	-				
		F3/F5/R5/R8	-	-				
B12	*C. pseudodip.*	25F/43R	-	-				
		23F/24R	-	-				
		F3/F5/R5/R8	-	-				
B13	*C. petrophilum*	25F/43R	+	+	238	*Staphylococcus aureus*; LS483309	228	2.6e-55
		23F/24R	-	-				
		F3/F5/R5/R8	-	-				
B14	FS 181	25F/43R	-	-				
		23F/24R	-	-				
		F3/F5/R5/R8	+					
B15	SA 200 a	25F/43R	-	-				
		23F/24R	-	-				
		F3/F5/R5/R8	+	+	165	*Shigella sonnei*; LR213458	250	1.1e-62
B16	SA 130 a	25F/43R	-	-				
		23F/24R	-	-				
		F3/F5/R5/R8	-	-				
B17	*Escherichia coli* (E. coli 5171)	25F/43R	-	-				
		23F/24R	-	-				
		F3/F5/R5/R8	+	+	441	*Shigella sonnei*; LR 213458	324	1.9e-84
B18	Serotype A2	25F/43R	-	-				
		23F/24R	-	-				
		F3/F5/R5/R8	+	+	250	*Klebsiella pneumonia*e; CP033242	243	3e-60
B19	AHSC 2	25F/43R	-	-				
		23F/24R	-	-				
		F3/F5/R5/R8	-	-				

Sample	Host	Primer set	PCR +	Contig +	Length	BLAST match	Bit sc	E-value
B20	ESR 916	25F/43R	-	-				
		23F/24R	-	-				
		F3/F5/R5/R8	+	+	621	*Shigella sonnei*; LR213458	308	2.7e-79
B21	FS 209	25F/43R	-	-				
		23F/24R	-	-				
		F3/F5/R5/R8	+	-				
B22	FS 208	25F/43R	-	-				
		23F/24R	-	-				
		F3/F5/R5/R8	+	+	283	*Klebsiella pneumoniae*; CP033242	254	1.6e-63
B23	FS 207	25F/43R	-	-				
		23F/24R	-	-				
		F3/F5/R5/R8	+	+	317	*E. coli* Es_ST2350; CP031321	93	4e-15
B24	YP 509 a	25F/43R	-	-				
		23F/24R	-	-				
		F3/F5/R5/R8	-	-				
B25	YP 510 b	25F/43R	-	-				
		23F/24R	-	-				
		F3/F5/R5/R8	-	-				
B26	YP 511 d	25F/43R	-	-				
		23F/24R	-	-				
		F3/F5/R5/R8	-	-				
B27	FS 182	25F/43R	-	-				
		23F/24R	-	-				
		F3/F5/R5/R8	-	-				
B28	SA 145 a	25F/43R	-	-				
		23F/24R	-	-				
		F3/F5/R5/R8	-	-				
B29	YEP 77	25F/43R	-	-				
		23F/24R	-	-				
		F3/F5/R5/R8	-	-				
B30	YEP 76	25F/43R	+	+	397	*Corynebacterium peudotuberculosis*; CP026501	401	2.3e-107
		23F/24R	-	-				
		F3/F5/R5/R8	+	+	397	No matches on sequence		
CR1	*Cryptosporidium hominis*	25F/43R	-	-				
		23F/24R	-	-				
		F3/F5/R5/R8	+	+	276	*Cryptosporidium hominis*; KY882333	156	2.2e-34
CR2	*Cryptosporidium hominis*	25F/43R	-	-				
		23F/24R	-	-				
		F3/F5/R5/R8	-	-				
CR3	*Cryptosporidium hominis*	25F/43R	-	-				
		23F/24R	+	+	148	Human picobirnavirus; AB214978	268	5.1e-68
		F3/F5/R5/R8	-	-				
CR4	*Cryptosporidium hominis*	25F/43R	-	-				
		23F/24R	-	-				
		F3/F5/R5/R8	-	-				
CR5	*Cryptosporidium parvum*	25F/43R	-	-				
		23F/24R	+	+	211	*Penicillium polonicum*; KU530219	339	5e-89
		F3/F5/R5/R8	-	-				
CR6	*Cryptosporidium parvum*	25F/43R	+	-				
		23F/24R	-	-				
		F3/F5/R5/R8	-	-				
CR7	*Cryptosporidium unknown*	25F/43R	-	-				
		23F/24R	+	-				
		F3/F5/R5/R8	-	-				
CR8	*Cryptosporidium unknown*	25F/43R	-	-				
		23F/24R	-	-				
		F3/F5/R5/R8	-	-				
CR9	*Cryptosporidium hominis*	25F/43R	+	+	378	*Cryptosporidium parvum*; AF040725	579	6.3e-161
		23F/24R	-	-				
		F3/F5/R5/R8	-	-				
CR10	*Cryptosporidium cuniculus*	25F/43R	-	-				
		23F/24R	-	-				
		F3/F5/R5/R8	-	-				
GD1	*Giardia intestinalis*	25F/43R	-	-				

		23F/24R	-	-				
		F3/F5/R5/R8	-	-				
Sample	**Host**	**Primer set**	**PCR +**	**Contig +**	**Length**	**BLAST match**	**Bit sc**	**E-value**
GD2	*Giardia intestinalis*	25F/43R	-	-				
		23F/24R	-	-				
		F3/F5/R5/R8	-	-				
GD3	*Giardia intestinalis*	25F/43R	-	-				
		23F/24R	-	-				
		F3/F5/R5/R8	-	-				
GD4	*Giardia intestinalis*	25F/43R	-	-				
		23F/24R	+	+	476	*Escherichia coli*; AP019189	731	0
		F3/F5/R5/R8	-	-				
GD5	*Giardia intestinalis*	25F/43R	+	+	299	*Cupriavidus basilensis*; CP010537	139	8.5e-29
		23F/24R	-	-				
		F3/F5/R5/R8	-	-				
GD6	*Giardia intestinalis*	25F/43R	-	-				
		23F/24R	-	-				
		F3/F5/R5/R8	-	-				
GD7	*Giardia intestinalis*	25F/43R	+	+	64	No matches on sequence		
		23F/24R	+	-				
		F3/F5/R5/R8	+	-				
GD8	*Giardia intestinalis*	25F/43R	-	-				
		23F/24R	+	-				
		F3/F5/R5/R8	-	-				
GD9	*Giardia intestinalis*	25F/43R	-	-				
		23F/24R	-	-				
		F3/F5/R5/R8	-	-				
GD10	*Giardia intestinalis*	25F/43R	-	-				
		23F/24R	-	-				
		F3/F5/R5/R8	+	-				

Bacteria were named based on the sample name B for bacteria and # of the samples 1–30; *Cryptosporidium* were named based on the sample name CR for *Cryptosporidium* and # of the samples 1–10; and *Giardia* were named based on the sample name GD for *Giardia* and # of the samples 1–10. Host is the genus and species of the "host" bacteria or protozoa if known; study is the study that the samples originated from with abbreviations designated based on possible host (YEP=Yellow-eyed penguins) or contract. All three primer sets were evaluated on all samples (see Chapter 3, 3.6.1.1) with 25F/43R for the genogroup I picobirnaviruses, 23F/24R for the genogroup II picobirnaviruses [117](Rosen et al. 2000) and the four primers [153](Anthony et al. 2015) for any genogroup picobirnavirus but only the positive results and primers are listed in the table. Contig length is the length of the sequence in base pairs or nucleotides from the conventional PCR and Sanger sequencing. Accession number and description of the accession are based upon BLASTn on the NCBI database done through Geneious BLAST search; maximum bit score and e-value are also based on the highest BLAST hit for each sequence.

REFERENCES

1. Taylor, L.H., S.M. Latham, and M.E.J. Woolhouse, *Risk factors for human disease emergence.* Philosophical Transactions of the Royal Society B-Biological Sciences, 2001. **356**(1411): p. 983-989.
2. Lloyd-Smith, J.O., et al., *Epidemic Dynamics at the Human-Animal Interface.* Science, 2009. **326**(5958): p. 1362-1367.
3. World Health Organization, *Global Health Observatory data: Top 10 causes of death.* 2015.
4. Study, J.W.-C., *World Health Organization (WHO)-convened Global Study of Origins of SARS-CoV-2: China Part*. 2021. p. 1-120.
5. Petersen, E., et al., *Comparing SARS-CoV-2 with SARS-CoV and influenza pandemics.* Lancet Infectious Diseases, 2020. **20**(9): p. E238-E244.
6. Wu, F., et al., *A new coronavirus associated with human respiratory disease in China (vol 579, pg 265, 2020).* Nature, 2020. **580**(7803): p. E7-E7.
7. Kahn, L.H., *Confronting zoonoses, linking human and veterinary medicine.* Emerging Infectious Diseases, 2006. **12**(4): p. 556-561.
8. Messenger, A.M., A.N. Barnes, and G.C. Gray, *Reverse Zoonotic Disease Transmission (Zooanthroponosis): A Systematic Review of Seldom-Documented Human Biological Threats to Animals.* PLOS ONE, 2014. **9**(2): p. e89055.
9. Haydon, D.T., et al., *Identifying reservoirs of infection: A conceptual and practical challenge.* Emerging Infectious Diseases, 2002. **8**(12): p. 1468-1473.
10. Dobson, A. and J. Foufopoulos, *Emerging infectious pathogens of wildlife.* Philosophical Transactions of the Royal Society of London Series B-Biological Sciences, 2001. **356**(1411): p. 1001-1012.
11. Alison G. Power and Charles E. Mitchell, *Pathogen Spillover in Disease Epidemics.* The American Naturalist, 2004. **164**(S5): p. S79-S89.
12. Wolfe, N.D., C.P. Dunavan, and J. Diamond, *Origins of major human infectious diseases.* Nature, 2007. **447**(7142): p. 279-283.
13. Cleaveland, S., M.K. Laurenson, and L.H. Taylor, *Diseases of humans and their domestic mammals: pathogen characteristics, host range and the risk of emergence.* Philosophical Transactions of the Royal Society B-Biological Sciences, 2001. **356**(1411): p. 991-999.
14. Heymann, D.L. and O.A. Dar, *Prevention is better than cure for emerging infectious diseases.* Bmj-British Medical Journal, 2014. **348**.
15. Azhar, E.I., et al., *Evidence for Camel-to-Human Transmission of MERS Coronavirus.* New England Journal of Medicine, 2014. **370**(26): p. 2499-2505.
16. Hayman, D.T.S., *As the bat flies.* Science, 2016. **354**(6316): p. 1099-1100.
17. Plowright, R.K., et al., *Transmission or Within-Host Dynamics Driving Pulses of Zoonotic Viruses in Reservoir-Host Populations.* Plos Neglected Tropical Diseases, 2016. **10**(8).
18. Olival, K.J., et al., *Host and viral traits predict zoonotic spillover from mammals.* Nature, 2017.
19. Pedersen, A.B. and T.J. Davies, *Cross-Species Pathogen Transmission and Disease Emergence in Primates.* EcoHealth, 2009. **6**(4): p. 496-508.
20. Liu, W., et al., *Origin of the human malaria parasite Plasmodium falciparum in gorillas.* Nature, 2010. **467**(7314): p. 420-U67.
21. Li, Y., et al., *Eastern Chimpanzees, but Not Bonobos, Represent a Simian Immunodeficiency Virus Reservoir.* Journal of Virology, 2012. **86**(19): p. 10776-10791.
22. Nishiura, H. and G. Chowell, *Early transmission dynamics of Ebola virus disease (EVD), West Africa, March to August 2014.* Eurosurveillance, 2014. **19**(36): p. 5-10.
23. Ji, W., et al., *Cross-species transmission of the newly identified coronavirus 2019-nCoV.* Journal of Medical Virology, 2020. **92**(4): p. 433-440.
24. Restif, O., et al., *Model-guided fieldwork: practical guidelines for multidisciplinary research on wildlife ecological and epidemiological dynamics.* Ecology Letters, 2012. **15**(10): p. 1083-1094.
25. Plowright, R.K., et al., *Pathways to zoonotic spillover.* Nat Rev Micro, 2017. **advance online publication**.
26. Wolfe, N.D., et al., *Bushmeat hunting deforestation, and prediction of zoonoses emergence.* Emerging Infectious Diseases, 2005. **11**(12): p. 1822-1827.
27. Woolhouse, M.E.J. and S. Gowtage-Sequeria, *Host Range and Emerging and Reemerging Pathogens.* Emerging Infectious Diseases, 2005. **11**(12): p. 1842-1847.
28. Pulliam, J.R.C., *Viral host jumps: Moving toward a predictive framework.* Ecohealth, 2008. **5**(1): p. 80-91.

29. Pulliam, J.R.C. and J. Dushoff, *Ability to Replicate in the Cytoplasm Predicts Zoonotic Transmission of Livestock Viruses.* Journal of Infectious Diseases, 2009. **199**(4): p. 565-568.
30. Rulli, M.C., et al., *The nexus between forest fragmentation in Africa and Ebola virus disease outbreaks.* Scientific Reports, 2017. **7**.
31. Kruse, H., A.-M. Kirkemo, and K. Handeland, *Wildlife as Source of Zoonotic Infections.* Emerging Infectious Diseases, 2004. **10**(12): p. 2067-2072.
32. Lloyd-Smith, J.O., et al., *Superspreading and the effect of individual variation on disease emergence.* Nature, 2005. **438**(7066): p. 355-359.
33. Nie, Q., et al., *Phylogenetic and phylodynamic analyses of SARS-CoV-2.* Virus Research, 2020. **287**.
34. Jones, K.E., et al., *Global trends in emerging infectious diseases.* Nature, 2008. **451**(7181): p. 990-U4.
35. Daszak, P., A.A. Cunningham, and A.D. Hyatt, *Emerging Infectious Diseases of Wildlife-- Threats to Biodiversity and Human Health.* Science, 2000. **287**(5452): p. 443-449.
36. Cranfield, M.R., *Mountain Gorilla research: The risk of disease transmission relative to the benefit from the perspective of ecosystem health.* American Journal of Primatology, 2008. **70**(8): p. 751-754.
37. Butynski, T.M. and J. Kalina, *Three new mountain national parks for Uganda.* Oryx, 1993. **27**(4): p. 214-224.
38. Dick, G.W.A., S.F. Kitchen, and A.J. Haddow, *Zika virus 1. Isolations and serological specificity.* Transactions of the Royal Society of Tropical Medicine and Hygiene, 1952. **46**(5): p. 509-520.
39. Dick, G.W.A., *Zika virus 2. pathogenicity and physical properties.* Transactions of the Royal Society of Tropical Medicine and Hygiene, 1952. **46**(5): p. 521-534.
40. Okware, S.I., et al., *An outbreak of Ebola in Uganda.* Tropical Medicine & International Health, 2002. **7**(12): p. 1068-1075.
41. Robbins, M.M., et al., *Population dynamics of the Bwindi mountain gorillas.* Biological Conservation, 2009. **142**(12): p. 2886-2895.
42. Plumptre, A., Robbins, M. & Williamson, E.A., *Gorilla beringei. (errata version published in 2016) The IUCN Red List of Threatened Species 2016*. 2016, International Union for Conservation of Nature (IUCN).
43. Plumptre, A.J., et al., *The biodiversity of the Albertine Rift.* Biological Conservation, 2007. **134**(2): p. 178-194.
44. Muylaert R.L., D.B., Ngabirano A., Kalema-Zikusoka G., MacGregor H., Lloyd-Smith J.O., Fayaz A., Knox M.A., Hayman D.T.S., *Community health and human-animal contacts on the edges of Bwindi Impenetrable National Park, Uganda.* PLoS Neglected Tropical Diseases, 2021.
45. Daszak, P., A.A. Cunningham, and A.D. Hyatt, *Anthropogenic environmental change and the emergence of infectious diseases in wildlife.* Acta Tropica, 2001. **78**(2): p. 103-116.
46. Woodford, M.H., T.M. Butynski, and W.B. Karesh, *Habituating the great apes: the disease risks.* Oryx, 2002. **36**(2): p. 153-160.
47. Nicole, W., *Seeing the Forest for the Trees How "One Health" Connects Humans, Animals, and Ecosystems.* Environmental Health Perspectives, 2014. **122**(5): p. A122-A129.
48. Robbins, M.M., *A demographic-analysis of male life-history and social-structure of mountain gorillas.* Behaviour, 1995. **132**: p. 21-47.
49. Hanson, T., *The Impenetrable Forest: Gorilla Years in Uganda*. 2001, United States of America: Curtis Brown Unlimited.
50. Doran-Sheehy, D.M., et al., *Habituation of western gorillas: The process and factors that influence it.* American Journal of Primatology, 2007. **69**(12): p. 1354-1369.
51. Rubanga, S.V., D. Bact, and G. Kalema-Zikusoka, *The establishment and use of field laboratories: Lessons from the Conservation Through Public Health gorilla research clinic, Uganda.* Journal of Exotic Pet Medicine, 2013. **22**(1): p. 34-38.
52. United Nations Educational, S.a.C.O. *Bwindi Impenetrable National Park*. World Heritage List, 2017.
53. Organization, U.N.E.S.a.C. *Uganda*. 2019.
54. Kamugisha, S.R., et al., *A Retrospective Cross Sectional Study of the Effectiveness of a Project in Improving Infant Health in Bwindi, South Western Uganda.* Frontiers in Public Health, 2018. **6**.
55. Mukasa, N., *The Batwa Indigenous People of Uganda and their Traditional Forest Land; Eviction, Non-Collaboration and Unfulfilled Needs.* Indigenous Policy Journal, 2014. **24**(4): p. 1-16.
56. Baker, J., E.J. Milner-Gulland, and N. Leader-Williams, *Park Gazettement and Integrated Conservation and Development as Factors in Community Conflict at Bwindi Impenetrable Forest, Uganda.* Conservation Biology, 2012. **26**(1): p. 160-170.

57. Graczyk, T.K., et al., *Hyperkeratotic mange caused by Sarcoptes scabiei (Acariformes : Sarcoptidae) in juvenile human-habituated mountain gorillas (Gorilla gorilla beringei).* Parasitology Research, 2001. **87**(12): p. 1024-1028.
58. Kalema-Zikusoka, G., R.A. Kock, and E.J. Macfie, *Scabies in free-ranging mountain gorillas (Gorilla beringei beringei) in Bwindi Impenetrable National Park, Uganda.* Veterinary Record, 2002. **150**(1): p. 12-15.
59. Guerrera, W., et al., *Medical survey of the local human population to determine possible health risks to the mountain gorillas of Bwindi Impenetrable Forest National Park, Uganda.* International Journal of Primatology, 2003. **24**(1): p. 197-207.
60. Harrison, M., et al., *Profiling unauthorized natural resource users for better targeting of conservation interventions.* Conservation Biology, 2015. **29**(6): p. 1636-1646.
61. Millan, J., et al., *Serosurvey of Dogs for Human, Livestock, and Wildlife Pathogens, Uganda.* Emerging Infectious Diseases, 2013. **19**(4): p. 680-682.
62. Salyer, S.J., et al., *Epidemiology and Molecular Relationships of Cryptosporidium spp. in People, Primates, and Livestock from Western Uganda.* Plos Neglected Tropical Diseases, 2012. **6**(4).
63. Johnston, A.R., et al., *Molecular Epidemiology of Cross-Species <italic>Giardia duodenalis</italic> Transmission in Western Uganda.* PLoS Negl Trop Dis, 2010. **4**(5): p. e683.
64. Rwego, I.B., et al., *Gastrointestinal Bacterial Transmission among Humans, Mountain Gorillas, and Livestock in Bwindi Impenetrable National Park, Uganda.* Conservation Biology, 2008. **22**(6): p. 1600-1607.
65. Tumusiime, D.M., *Protected areas and people in Uganda: costs, benefits, livelihoods and narratives around Bwindi Impenetrable National Park*. Protected areas and people in Uganda: costs, benefits, livelihoods and narratives around Bwindi Impenetrable National Park. 2012. 172 pp.-172 pp.
66. Twongyirwe, R., et al., *Dynamics of forest cover conversion in and around Bwindi impenetrable forest, Southwestern Uganda.* Journal of Applied Sciences and Environmental Management, 2011. **15**(1): p. 189-195.
67. Langhout, M.v.Z., P. Reed, and M. Fox, *Validation of multiple diagnostic techniques to detect Cryptosporidium sp and Giardia sp in free-ranging western lowland gorillas (Gorilla gorilla gorilla) and observations on the prevalence of these protozoan infections in two populations in Gabon.* Journal of Zoo and Wildlife Medicine, 2010. **41**(2): p. 210-217.
68. Hogan, J.N., et al., *Giardia in mountain gorillas (Gorilla beringei beringei), forest buffalo (Syncerus caffer), and domestic cattle in Volcanoes National Park, Rwanda.* Journal of Wildlife Diseases, 2014. **50**(1): p. 21-30.
69. Sak, B., et al., *Long-Term Monitoring of Microsporidia, Cryptosporidium and Giardia Infections in Western Lowland Gorillas (Gorilla gorilla gorilla) at Different Stages of Habituation in Dzanga Sangha Protected Areas, Central African Republic.* Plos One, 2013. **8**(8).
70. Graczyk, T.K., L.J. Lowenstine, and M.R. Cranfield, *Capillaria hepatica (Nematoda) infections in human-habituated mountain gorillas (Gorilla gorilla beringei) of the Parc National de Volcans, Rwanda.* Journal of Parasitology, 1999. **85**(6): p. 1168-1170.
71. Whittier, C.A., M.R. Cranfield, and M.K. Stoskopf, *Real-time PCR detection of Campylobacter spp. in free-ranging mountain gorillas (Gorilla beringei beringei).* Journal of Wildlife Diseases, 2010. **46**(3): p. 791-802.
72. Rwego, I.B., et al., *High rates of Escherichia coli transmission between livestock and humans in rural Uganda.* Journal of Clinical Microbiology, 2008. **46**(10): p. 3187-3191.
73. Melin, A.D., et al., *Comparative ACE2 variation and primate COVID-19 risk.* bioRxiv : the preprint server for biology, 2020.
74. Townsend, A.K., et al., *Emerging infectious disease and the challenges of social distancing in human and non-human animals.* Proceedings of the Royal Society B-Biological Sciences, 2020. **287**(1932).
75. Mattson, K., *ZOO GORILLAS RECOVERING FROM COVID-19.* Javma-Journal of the American Veterinary Medical Association, 2021. **258**(5): p. 441-441.
76. Palacios, G., et al., *Human Metapneumovirus Infection in Wild Mountain Gorillas, Rwanda.* Emerging Infectious Diseases, 2011. **17**(4): p. 711-713.
77. Kading, R.C., et al., *Prevalence of antibodies to alphaviruses and flaviviruses in free-ranging game animals and nonhuman primates in the greater congo basin.* Journal of Wildlife Diseases, 2013. **49**(3): p. 587-599.

78. Nizeyi, J.B., et al., *Cryptosporidium sp and Giardia sp infections in mountain gorillas (Gorilla gorilla beringei) of the Bwindi Impenetrable National Park, Uganda.* Journal of Parasitology, 1999. **85**(6): p. 1084-1088.
79. Graczyk, T.K., et al., *Cryptosporidium parvum Genotype 2 infections in free-ranging mountain gorillas (Gorilla gorilla beringei) of the Bwindi Impenetrable National Park, Uganda.* Parasitology Research, 2001. **87**(5): p. 368-370.
80. Nizeyi, J.B., et al., *Cryptosporidiosis in people sharing habitats with free-ranging mountain gorillas (Gorilla gorilla beringei), Uganda.* American Journal of Tropical Medicine and Hygiene, 2002. **66**(4): p. 442-444.
81. Graczyk, T.K., et al., *Anthropozoonotic Giardia duodenalis genotype (assemblage) A infections in habitats of free-ranging human-habituated gorillas, Uganda.* Journal of Parasitology, 2002. **88**(5): p. 905-909.
82. Nizeyi, J.B., et al., *Campylobacteriosis, salmonellosis, and shigellosis in free-ranging human-habituated mountain gorillas of Uganda.* Journal of Wildlife Diseases, 2001. **37**(2): p. 239-244.
83. Kalema-Zikusoka, G., J.M. Rothman, and M.T. Fox, *Intestinal parasites and bacteria of mountain gorillas (Gorilla beringei beringei) in Bwindi Impenetrable National Park, Uganda.* Primates, 2005. **46**(1): p. 59-63.
84. Rothman, J.M., A.N. Pell, and D.D. Bowman, *Host-parasite ecology of the helminths in mountain gorillas.* Journal of Parasitology, 2008. **94**(4): p. 834-840.
85. Bodewes, R., et al., *Viral metagenomic analysis of feces of wild small carnivores.* Virology Journal, 2014. **11**.
86. Aw, T.G., S. Wengert, and J.B. Rose, *Metagenomic analysis of viruses associated with field-grown and retail lettuce identifies human and animal viruses.* International Journal of Food Microbiology, 2016. **223**: p. 50-56.
87. Creer, S., et al., *The ecologist's field guide to sequence-based identification of biodiversity.* Methods in Ecology and Evolution, 2016. **7**(9): p. 1008-1018.
88. National Research Council Board on, B., *The National Academies Collection: Reports funded by National Institutes of Health*, in *Bioinformatics: Converting Data to Knowledge: Workshop Summary*, R. Pool and J. Esnayra, Editors. 2000, National Academies Press (US) Copyright ©2000, National Academy of Sciences: Washington (DC).
89. Krishnamurthy, S.R. and D. Wang, *Origins and challenges of viral dark matter.* Virus research, 2017.
90. Chong, R., et al., *Fecal Viral Diversity of Captive and Wild Tasmanian Devils Characterized Using Virion-Enriched Metagenomics and Metatranscriptomics.* Journal of Virology, 2019. **93**(11).
91. Cummings, M.J., et al., *Precision Surveillance for Viral Respiratory Pathogens: Virome Capture Sequencing for the Detection and Genomic Characterization of Severe Acute Respiratory Infection in Uganda.* Clinical Infectious Diseases, 2019. **68**(7): p. 1118-1125.
92. Finkbeiner, S.R., et al., *Metagenomic analysis of human diarrhea: Viral detection and discovery.* Plos Pathogens, 2008. **4**(2).
93. Yinda, C.K., et al., *Cameroonian fruit bats harbor divergent viruses, including rotavirus H, bastroviruses, and picobirnaviruses using an alternative genetic code.* Virus Evolution, 2018. **4**(1).
94. Shi, M., et al., *Redefining the invertebrate RNA virosphere.* Nature, 2016. **540**(7634): p. 539-+.
95. Ghurye, J.S., V. Cepeda-Espinoza, and M. Pop, *Metagenomic Assembly: Overview, Challenges and Applications.* The Yale journal of biology and medicine, 2016. **89**(3): p. 353-362.
96. Breitwieser, F.P., J. Lu, and S.L. Salzberg, *A review of methods and databases for metagenomic classification and assembly.* Briefings in Bioinformatics, 2019. **20**(4): p. 1125-1139.
97. Altschul, S.F., et al., *BASIC LOCAL ALIGNMENT SEARCH TOOL.* Journal of Molecular Biology, 1990. **215**(3): p. 403-410.
98. Duraisamy, R., et al., *Detection of novel RNA viruses from free-living gorillas, Republic of the Congo: genetic diversity of picobirnaviruses.* Virus Genes, 2018. **54**(2): p. 256-271.
99. Baker, K.S., et al., *Metagenomic study of the viruses of African straw-coloured fruit bats: Detection of a chiropteran poxvirus and isolation of a novel adenovirus.* Virology, 2013. **441**(2): p. 95-106.
100. Nyaga, M.M., et al., *Complete genome analyses of the first porcine rotavirus group H identified from a South African pig does not provide evidence for recent interspecies transmission events.* Infection Genetics and Evolution, 2016. **38**: p. 1-7.
101. McMullan, L.K., et al., *Using next generation sequencing to identify yellow fever virus in Uganda.* Virology, 2012. **422**(1): p. 1-5.

102. Albarino, C.G., et al., *Novel Paramyxovirus Associated with Severe Acute Febrile Disease, South Sudan and Uganda, 2012.* Emerging Infectious Diseases, 2014. **20**(2): p. 211-216.
103. Masembe, C., et al., *Viral metagenomics demonstrates that domestic pigs are a potential reservoir for Ndumu virus.* Virology Journal, 2012. **9**(218): p. (24 September 2012)-(24 September 2012).
104. Blomstrom, A.L., et al., *Viral metagenomic analysis of bushpigs (Potamochoerus larvatus) in Uganda identifies novel variants of Porcine parvovirus 4 and torque teno sus virus 1 and 2.* Virology Journal, 2012. **9**(192): p. (11 September 2012)-(11 September 2012).
105. Bennett, A.J., et al., *Naturally Circulating Hepatitis A Virus in Olive Baboons, Uganda.* Emerging Infectious Diseases, 2016. **22**(7): p. 1308-1310.
106. Iacono, G.L., et al., *A unified framework for the infection dynamics of zoonotic spillover and spread.* PLoS Neglected Tropical Diseases, 2016. **10**(9): p. e0004957-e0004957.
107. Wood, J.L.N., et al., *A framework for the study of zoonotic disease emergence and its drivers: spillover of bat pathogens as a case study.* Philosophical Transactions of the Royal Society B-Biological Sciences, 2012. **367**(1604): p. 2881-2892.
108. Grange, Z.L., et al., *Ranking the risk of animal-to-human spillover for newly discovered viruses.* Proceedings of the National Academy of Sciences, 2021. **118**(15): p. e2002324118.
109. Morse, S.S., et al., *Zoonoses 3 Prediction and prevention of the next pandemic zoonosis.* Lancet, 2012. **380**(9857): p. 1956-1965.
110. Woolhouse, M.E.J., L.H. Taylor, and D.T. Haydon, *Population biology of multihost pathogens.* Science, 2001. **292**(5519): p. 1109-1112.
111. Holmes, E.C., *Evolutionary History and Phylogeography of Human Viruses.* Annual Review of Microbiology, 2008. **62**: p. 307-328.
112. Geoghegan, J.L., S. Duchene, and E.C. Holmes, *Comparative analysis estimates the relative frequencies of co-divergence and cross-species transmission within viral families.* Plos Pathogens, 2017. **13**(2).
113. Knox, M.A., K.R. Gedye, and D.T.S. Hayman, *The Challenges of Analysing Highly Diverse Picobirnavirus Sequence Data.* Viruses-Basel, 2018. **10**(12).
114. Pereira, H.G., et al., *Novel viruses in human feces.* Lancet, 1988. **2**(8602): p. 103-104.
115. Pereira, H.G., et al., *A virus with a bisegmented double-stranded-RNA genome in rat (Oryzomys-nigripes) intestines.* Journal of General Virology, 1988. **69**: p. 2749-2754.
116. Green, J., et al., *Genomic characterisation of the large segment of a rabbit picobirnavirus* and comparison with the atypical picobirnavirus of Cryptosporidium parvum.* Archives of Virology, 1999. **144**(12): p. 2457-2465.
117. Rosen, B.I., et al., *Cloning of human picobirnavirus genomic segments and development of an RT-PCR detection assay.* Virology, 2000. **277**(2): p. 316-329.
118. Wang, Y., et al., *Detection of viral agents in fecal specimens of monkeys with diarrhea.* Journal of Medical Primatology, 2007. **36**(2): p. 101-107.
119. Banyai, K., et al., *Genogroup I picobirnaviruses in pigs: evidence for genetic diversity and relatedness to human strains.* Journal of General Virology, 2008. **89**: p. 534-539.
120. Duarte Fregolente, M.C., et al., *Molecular characterization of picobirnaviruses from new hosts.* Virus Research, 2009. **143**(1): p. 134-136.
121. Symonds, E.M., D.W. Griffin, and M. Breitbart, *Eukaryotic Viruses in Wastewater Samples from the United States.* Applied and Environmental Microbiology, 2009. **75**(5): p. 1402-1409.
122. Woo, P.C.Y., et al., *Complete Genome Sequence of a Novel Picobirnavirus, Otarine Picobirnavirus, Discovered in California Sea Lions.* Journal of Virology, 2012. **86**(11): p. 6377-6378.
123. Woo, P.C.Y., et al., *Metagenomic analysis of viromes of dromedary camel fecal samples reveals large number and high diversity of circoviruses and picobirnaviruses.* Virology, 2014. **471**: p. 117-125.
124. Ganesh, B., G. Masachessi, and Z. Mladenova, *Animal Picobirnavirus.* VirusDisease, 2014. **25**(2, Sp. Iss. SI): p. 223-238.
125. Conceicao-Neto, N., et al., *Reassortment among picobirnaviruses found in wolves.* Archives of Virology, 2016. **161**(10): p. 2859-2862.
126. Collier, A.M., et al., *Initiation of RNA Polymerization and Polymerase Encapsidation by a Small dsRNA Virus.* Plos Pathogens, 2016. **12**(4).
127. Duquerroy, S., et al., *The picobirnavirus crystal structure provides functional insights into virion assembly and cell entry.* Embo Journal, 2009. **28**(11): p. 1655-1665.
128. Mertens, P., *The dsRNA viruses.* Virus Research, 2004. **101**(1): p. 3-13.
129. Simmonds, P., et al., *Virus taxonomy in the age of metagenomics.* Nature Reviews Microbiology, 2017. **15**(3): p. 161-168.

130. Holmes, E.C., *The Evolution and Emergence of RNA Viruses*. Oxford Series in Ecology and Evolution, ed. O.U.P. Inc. 2009, New York, USA: Oxford University Press Inc.
131. Wakuda, M., Y. Pongsuwanna, and K. Taniguchi, *Complete nucleotide sequences of two RNA segments of human picobirnavirus.* Journal of Virological Methods, 2005. **126**(1-2): p. 165-169.
132. Ganesh, B., et al., *Picobirnavirus infections: viral persistence and zoonotic potential.* Reviews in Medical Virology, 2012. **22**(4): p. 245-256.
133. Boros, A., et al., *Multiple divergent picobirnaviruses with functional prokaryotic Shine-Dalgarno ribosome binding sites present in cloacal sample of a diarrheic chicken.* Virology, 2018. **525**: p. 62-72.
134. Ghosh, S. and Y.S. Malik, *The True Host/s of Picobirnaviruses.* Frontiers in Veterinary Science, 2021. **7**.
135. Da Costa, B., et al., *Picobirnaviruses encode a protein with repeats of the ExxRxNxxxE motif.* Virus Research, 2011. **158**(1-2): p. 251-256.
136. Malik, Y.S., et al., *Epidemiology, phylogeny, and evolution of emerging enteric Picobirnaviruses of animal origin and their relationship to human strains.* BioMed Research International, 2014. **2014**: p. 780752-Article ID 780752.
137. (ENA)/GenBank, D.D.B.o.J.D.E.N.A., *Feature Table Definition*. November 2019, EMBL-EBI; NCBI: Mishima, Japan; Cambridge, UK; Bethesda, MD, USA.
138. Cornishbowden, A., *NOMENCLATURE FOR INCOMPLETELY SPECIFIED BASES IN NUCLEIC-ACID SEQUENCES - RECOMMENDATIONS 1984.* Nucleic Acids Research, 1985. **13**(9): p. 3021-3030.
139. Nomenclature, I.U.o.P.a.A.C.-I.U.o.B.I.-I.J.C.o.B., *Nomenclature and Symbolism for Amino Acids and Peptides.* Eur J Biochem, 1984. **138**: p. 9-37.
140. Ng, K.K.S., J.J. Arnold, and C.E. Cameron, *Structure-function relationships among RNA-dependent RNA polymerases.* Rna Interference, 2008. **320**: p. 137-156.
141. Smits, S.L., et al., *New Viruses in Idiopathic Human Diarrhea Cases, the Netherlands.* Emerging Infectious Diseases, 2014. **20**(7): p. 1218-1222.
142. Zhang, S., et al., *Detection and evolutionary analysis of picobirnaviruses in treated wastewater.* Microbial Biotechnology, 2015. **8**(3): p. 474-482.
143. Day, J.M. and L. Zsak, *Molecular and Phylogenetic Analysis of a Novel Turkey-Origin Picobirnavirus.* Avian Diseases, 2014. **58**(1): p. 137-142.
144. Li, L., et al., *Exploring the virome of diseased horses.* Journal of General Virology, 2015. **96**: p. 2721-2733.
145. Perez, L.J., G.A. Cloherty, and M.G. Berg, *Understanding the Genetic Diversity of Picobirnavirus: A Classification Update Based on Phylogenetic and Pairwise Sequence Comparison Approaches.* Viruses, 2021. **13**(8): p. 1476.
146. Gillman, L., A. Maria Sanchez, and J. Arbiza, *Picobirnavirus in Captive Animals from Uruguay: Identification of New Hosts.* Intervirology, 2013. **56**(1): p. 46-49.
147. Gallimore, C., D. Lewis, and D. Brown, *DETECTION AND CHARACTERIZATION OF A NOVEL BISEGMENTED DOUBLE-STRANDED-RNA VIRUS (PICOBIRNAVIRUS) FROM RABBIT FECES.* Archives of Virology, 1993. **133**(1-2): p. 63-73.
148. Ghosh, S., et al., *Molecular characterization of full-length genomic segment 2 of a bovine picobirnavirus (PBV) strain: evidence for high genetic diversity with genogroup I PBVs.* Journal of General Virology, 2009. **90**: p. 2519-2524.
149. Silva, R.R., et al., *Genogroup I avian picobirnavirus detected in Brazilian broiler chickens: a molecular epidemiology study.* Journal of General Virology, 2014. **95**: p. 117-122.
150. Giordano, M.O., et al., *Evidence of closely related picobirnavirus strains circulating in humans and pigs in Argentina.* Journal of Infection, 2011. **62**(1): p. 45-51.
151. Kunz, A.F., et al., *High detection rate and genetic diversity of picobirnavirus in a sheep flock in Brazil.* Virus Research, 2018. **255**: p. 10-13.
152. Masachessi, G., et al., *Establishment and maintenance of persistent infection by picobirnavirus in greater rhea (Rhea Americana).* Archives of Virology, 2012. **157**(11): p. 2075-2082.
153. Anthony, S.J., et al., *Non-random patterns in viral diversity.* Nature Communications, 2015. **6**.
154. Carruyo, G.M., et al., *Molecular characterization of porcine picobirnaviruses and development of a specific reverse transcription-PCR assay.* Journal of Clinical Microbiology, 2008. **46**(7): p. 2402-2405.
155. Woo, P.C.Y., et al., *High Diversity of Genogroup I Picobirnaviruses in Mammals.* Frontiers in Microbiology, 2016. **7**.
156. Gonzalez, G., et al., *An optimistic protein assembly from sequence reads salvaged an uncharacterized segment of mouse picobirnavirus.* Scientific Reports, 2017. **7**.

157. Duarte Fregolente, M.C. and M.S. Viccari Gatti, *Nomenclature proposal for picobirnavirus.* Archives of Virology, 2009. **154**(12): p. 1953-1954.
158. Lambden, P.R., et al., *CLONING OF NONCULTIVATABLE HUMAN ROTAVIRUS BY SINGLE PRIMER AMPLIFICATION.* Journal of Virology, 1992. **66**(3): p. 1817-1822.
159. Attoui, H., et al., *Strategies for the sequence determination of viral dsRNA genomes.* Journal of Virological Methods, 2000. **89**(1-2): p. 147-158.
160. Woo, P.C.Y., et al., *Novel Picobirnaviruses in Respiratory and Alimentary Tracts of Cattle and Monkeys with Large Intra- and Inter-Host Diversity.* Viruses-Basel, 2019. **11**(6).
161. Ng, T.F.F., et al., *Divergent picobirnaviruses in human feces.* Genome announcements, 2014. **2**(3).
162. Sun, G., et al., *Viral metagenomics analysis of picobirnavirus-positive feces from children with sporadic diarrhea in China.* Archives of Virology, 2016. **161**(4): p. 971-975.
163. Ganesh, B., et al., *Genogroup I picobirnavirus in diarrhoeic foals: Can the horse serve as a natural reservoir for human infection?* Veterinary Research, 2011. **42**.
164. Luo, X., et al., *Investigation and sequence analysis of diarrhea related viruses in the feces of Marmota himalayana on the Qinghai-Tibet Plateau, China.* Zhongguo Meijie Shengwuxue ji Kongzhi Zazhi = Chinese Journal of Vector Biology and Control, 2016. **27**(4): p. 333-340.
165. Malik, Y.S., et al., *Picobirnavirus detection in bovine and buffalo calves from foothills of Himalaya and Central India.* Tropical Animal Health and Production, 2011. **43**(8): p. 1475-1478.
166. Masachessi, G., et al., *Picobirnavirus (PBV) natural hosts in captivity and virus excretion pattern in infected animals.* Archives of Virology, 2007. **152**(5): p. 989-998.
167. Zhao, D., et al., *The relationship between picobirnavirus infection and diarrhea: meta-analysis.* Journal of Tropical Medicine (Guangzhou), 2014. **14**(5): p. 651-655.
168. Tamehiro, C.Y., et al., *Segmented double-stranded genomic RNA viruses in fecal samples from broiler chicken.* Brazilian Journal of Microbiology, 2003. **34**(4): p. 349-353.
169. Wilburn, L., et al., *Molecular detection and characterization of picobirnaviruses in piglets with diarrhea in Thailand.* Archives of Virology, 2017. **162**(4): p. 1061-1066.
170. Bhattacharya, R., et al., *Detection of genogroup I and II human picobirnaviruses showing small genomic RNA profile causing acute watery diarrhoea among children in Kolkata, India.* Infection Genetics and Evolution, 2007. **7**(2): p. 229-238.
171. Giordano, M.O., et al., *Diarrhea and enteric emerging viruses in HIV-infected patients.* Aids Research and Human Retroviruses, 1999. **15**(16): p. 1427-1432.
172. Telengech, P., et al., *Diverse Partitiviruses From the Phytopathogenic Fungus,Rosellinia necatrix.* Frontiers in Microbiology, 2020. **11**.
173. van Leeuwen, M., et al., *Human Picobirnaviruses Identified by Molecular Screening of Diarrhea Samples.* Journal of Clinical Microbiology, 2010. **48**(5): p. 1787-1794.
174. Luo, X.-l., et al., *Marmota himalayana in the Qinghai-Tibetan plateau as a special host for bi-segmented and unsegmented picobirnaviruses.* Emerging Microbes & Infections, 2018. **7**.
175. Smits, S.L., et al., *Genogroup I and II Picobirnaviruses in Respiratory Tracts of Pigs.* Emerging Infectious Diseases, 2011. **17**(12): p. 2328-2330.
176. Smits, S.L., et al., *Picobirnaviruses in the Human Respiratory Tract.* Emerging Infectious Diseases, 2012. **18**(9): p. 1539-1540.
177. Anthony, S.J., et al., *A Strategy To Estimate Unknown Viral Diversity in Mammals.* Mbio, 2013. **4**(5).
178. Lam, T.T.Y., et al., *Genomic Analysis of the Emergence, Evolution, and Spread of Human Respiratory RNA Viruses.* Annual Review of Genomics and Human Genetics, Vol 17, 2016. **17**: p. 193-218.
179. Lauring, A.S., *Within-Host Viral Diversity: AWindow into Viral Evolution.* Annual Review of Virology, Vol 7, 2020, 2020. **7**: p. 63-81.
180. Chao, R., X. Ma, and X. Liao, *Infantile diarrhea with Picobimavirus, Chengdu.* Modern Preventive Medicine, 2015. **42**(23): p. 4250-4251.
181. Davies, T.J. and A.B. Pedersen, *Phylogeny and geography predict pathogen community similarity in wild primates and humans.* Proceedings of the Royal Society B-Biological Sciences, 2008. **275**(1643): p. 1695-1701.
182. Pedersen, A.B., et al., *Patterns of host specificity and transmission among parasites of wild primates.* International Journal for Parasitology, 2005. **35**(6): p. 647-657.
183. Cooper, N., et al., *Phylogenetic host specificity and understanding parasite sharing in primates.* Ecology Letters, 2012. **15**(12): p. 1370-1377.
184. Lyons, S., et al., *Species Association of Hepatitis B Virus (HBV) in Non-Human Apes; Evidence for Recombination between Gorilla and Chimpanzee Variants.* Plos One, 2012. **7**(3).

185. Wevers, D., et al., *Novel Adenoviruses in Wild Primates: a High Level of Genetic Diversity and Evidence of Zoonotic Transmissions.* Journal of Virology, 2011. **85**(20): p. 10774-10784.
186. Peeters, M. and E. Delaporte, *Simian retroviruses in African apes.* Clinical Microbiology and Infection, 2012. **18**(6): p. 514-520.
187. Yang, X., et al., *Geographic Distribution and Genetic Diversity of Rice Stripe Mosaic Virus in Southern China.* Frontiers in Microbiology, 2018. **9**.
188. Sardar, R., et al., *Integrative analyses of SARS-CoV-2 genomes from different geographical locations reveal unique features potentially consequential to host-virus interaction, pathogenesis and clues for novel therapies.* Heliyon, 2020. **6**(9).
189. Starkman, S.E., et al., *Geographic and species association of hepatitis B virus genotypes in non-human primates.* Virology, 2003. **314**(1): p. 381-393.
190. Masachessi, G., et al., *Maintenance of picobirnavirus (PBV) infection in an adult orangutan (Pongo pygmaeus) and genetic diversity of excreted viral strains during a three-year period.* Infection Genetics and Evolution, 2015. **29**: p. 196-202.
191. Verma, H., et al., *Prevalence and complete genome characterization of turkey picobirnaviruses.* Infection Genetics and Evolution, 2015. **30**: p. 134-139.
192. Kluge, M., et al., *Metagenomic Survey of Viral Diversity Obtained from Feces of Subantarctic and South American Fur Seals.* Plos One, 2016. **11**(3).
193. Hamza, I.A., et al., *Evaluation of pepper mild mottle virus, human picobirnavirus and Torque teno virus as indicators of fecal contamination in river water.* Water Research, 2011. **45**(3): p. 1358-1368.
194. Kuhar, U. and U. Jamnikar-Ciglenecki, *High detection rate and high genetic diversity of genogroup I Picobirnaviruses from roe deer.* Infection Genetics and Evolution, 2019. **73**: p. 210-213.
195. Kleymann, A., et al., *Detection and Molecular Characterization of Picobirnaviruses (PBVs) in the Mongoose: Identification of a Novel PBV Using an Alternative Genetic Code.* Viruses, 2020. **12**(1).
196. Ganesh, B., et al., *Detection and Molecular Characterization of Porcine Picobirnavirus in Feces of Domestic Pigs from Kolkata, India.* Indian Journal of Virology, 2012. **23**(3): p. 387-391.
197. Chen, M., et al., *Molecular detection of genogroup I and II picobirnaviruses in pigs in China.* Virus Genes, 2014. **48**(3): p. 553-556.
198. Martinez, L.C., et al., *Picobirnavirus causes persistent infection in pigs.* Infection Genetics and Evolution, 2010. **10**(7): p. 984-988.
199. Grange, et al. *SpillOver: A new tool for ranking the risk of viral spillover to humans using big data.* in *GeoVET 2019*. 2019. University of California, Davis; Davis, California, USA.
200. Amimo, J.O., et al., *Metagenomic analysis demonstrates the diversity of the fecal virome in asymptomatic pigs in East Africa.* Archives of Virology, 2016. **161**(4): p. 887-897.
201. Banyai, K., et al., *Genome sequencing identifies genetic and antigenic divergence of porcine picobirnaviruses.* Journal of General Virology, 2014. **95**: p. 2233-2239.
202. Banyai, K., et al., *Sequence heterogeneity among human picobirnaviruses detected in a gastroenteritis outbreak.* Archives of Virology, 2003. **148**(12): p. 2281-2291.
203. Krishnamurthy, S.R. and D. Wang, *Extensive conservation of prokaryotic ribosomal binding sites in known and novel picobirnaviruses.* Virology, 2018. **516**: p. 108-114.
204. Gallimore, C.I., et al., *DETECTION OF A PICOBIRNAVIRUS ASSOCIATED WITH CRYPTOSPORIDIUM POSITIVE STOOLS FROM HUMANS.* Archives of Virology, 1995. **140**(7): p. 1275-1278.
205. Lojkic, I., et al., *Faecal virome of red foxes from peri-urban areas.* Comparative Immunology Microbiology and Infectious Diseases, 2016. **45**: p. 10-15.
206. Malik, Y.S., et al., *Identification and characterisation of a novel genogroup II picobirnavirus in a calf in India.* Veterinary Record, 2014. **174**(11).
207. Taniguchi, K. and M. Wakuda, *Picobirnavirus.* Virus (Nagoya), 2005. **55**(2): p. 297-302.
208. Kim, H.-R., et al., *Viral metagenomic analysis of chickens with runting-stunting syndrome in the Republic of Korea.* Virology Journal, 2020. **17**(1).
209. Wille, M., et al., *Virus-virus interactions and host ecology are associated with RNA virome structure in wild birds.* Molecular Ecology, 2018. **27**(24): p. 5263-5278.
210. Moore, N.E., et al., *Metagenomic analysis of viruses in feces from unsolved outbreaks of gastroenteritis in humans.* Journal of Clinical Microbiology, 2014.
211. Martinez, L.C., et al., *Molecular diversity of partial-length genomic segment 2 of human picobirnavirus.* Intervirology, 2003. **46**(4): p. 207-213.

212. Khramtsov, N.V. and S.J. Upton, *Association of RNA polymerase complexes of the parasitic protozoan Cryptosporidium parvum with virus-like particles: Heterogeneous system.* Journal of Virology, 2000. **74**(13): p. 5788-5795.
213. Khramtsov, N.V., et al., *Presence of double-stranded RNAs in human and calf isolates of Cryptosporidium parvum.* Journal of Parasitology, 2000. **86**(2): p. 275-282.
214. Nibert, M.L., et al., *Cryspovirus: a new genus of protozoan viruses in the family Partitiviridae.* Archives of Virology, 2009. **154**(12): p. 1959-1965.
215. Mercado, R., et al., *Cryptosporidium hominis infection of the human respiratory tract.* Emerging Infectious Diseases, 2007. **13**(3): p. 462-464.
216. Volotao, E.M., et al., *First evidence of a trisegmented double-stranded RNA virus in canine faeces.* Veterinary Journal, 2001. **161**(2): p. 205-207.
217. Leite, J.P.G., et al., *A NOVEL AVIAN VIRUS WITH TRISEGMENTED DOUBLE-STRANDED-RNA AND FURTHER OBSERVATIONS ON PREVIOUSLY DESCRIBED SIMILAR VIRUSES WITH BISEGMENTED GENOME.* Virus Research, 1990. **16**(2): p. 119-126.
218. Ludert, J.E. and F. Liprandi, *IDENTIFICATION OF VIRUSES WITH BISEGMENTED AND TRISEGMENTED DOUBLE-STRANDED-RNA GENOME IN FECES OF CHILDREN WITH GASTROENTERITIS.* Research in Virology, 1993. **144**(3): p. 219-224.
219. Wolf, Y.I., et al., *Origins and Evolution of the Global RNA Virome.* Mbio, 2018. **9**(6).
220. Shine, J. and L. Dalgarno, *DETERMINANT OF CISTRON SPECIFICITY IN BACTERIAL RIBOSOMES.* Nature, 1975. **254**(5495): p. 34-38.
221. Elzanowski A, O.J., *The Genetic Codes*. 7 January 2019, National Center for Biotechnology Information (NCBI): Bethesda, Maryland, U.S.A.
222. Keen, E.C., *Phage therapy: concept to cure.* Frontiers in Microbiology, 2012. **3**.
223. Drulis-Kawa, Z., G. Majkowska-Skrobek, and B. Maciejewska, *Bacteriophages and Phage-Derived Proteins - Application Approaches.* Current Medicinal Chemistry, 2015. **22**(14): p. 1757-1773.
224. Chibani, C.M., et al., *Classifying the Unclassified: A Phage Classification Method.* Viruses-Basel, 2019. **11**(2).
225. Mwebe, R.A., *Survey on Cryptosporidium prevalence within the mountain gorilla population (Gorilla gorilla beringei) the Bwindi Impenetrable National Park in South-western Uganda.*, in *Department of Wildlife and Animal Resource Management*. 1998, Makarere University: Kampala, Uganda. p. 132.
226. Crump, J.A., D.R. Murdoch, and M.G. Baker, *Emerging infectious diseases in an island ecosystem: The New Zealand perspective.* Emerging Infectious Diseases, 2001. **7**(5): p. 767-772.
227. Ltd., T.I.o.E.S.a.R., *New Zealand public health surveillance report*. August 2021.
228. Browne, A.S., et al., *Use of Genomics to Investigate Historical Importation of Shiga Toxin-Producing Escherichia coli Serogroup O26 and Nontoxigenic Variants into New Zealand.* Emerging Infectious Diseases, 2019. **25**(3): p. 489-500.
229. Holland, B.R., *Quantification of historical livestock importation into New Zealand 1860-1979 (vol 62, pg 309, 2014).* New Zealand Veterinary Journal, 2015. **63**(6): p. 347-347.
230. Hayman, D.T.S. and M.A. Knox, *Estimating the age of the subfamily Orthocoronavirinae using host divergence times as calibration ages at two internal nodes.* Virology, 2021. **563**: p. 20-27.
231. Meloni, B.P. and R.C.A. Thompson, *Simplified methods for obtaining purified oocysts from mice and for growing Cryptosporidium parvum in vitro.* Journal of Parasitology, 1996. **82**(5): p. 757-762.
232. Ogbuigwe, P., et al., *High-Yield Purification of Giardia intestinalis Cysts from Fecal Samples.* Current protocols in microbiology, 2020. **59**(1): p. e117-e117.
233. Hall, R.J., et al., *Evaluation of rapid and simple techniques for the enrichment of viruses prior to metagenomic virus discovery.* Journal of Virological Methods, 2014. **195**(0): p. 194-204.
234. Jesser, K.J., et al., *Clustering of Vibrio parahaemolyticus Isolates Using MLST and Whole-Genome Phylogenetics and Protein Motif Fingerprinting.* Frontiers in Public Health, 2019. **7**.
235. de Bruin, A., et al., *Detection of Coxiella burnetii in Complex Matrices by Using Multiplex Quantitative PCR during a Major Q Fever Outbreak in The Netherlands.* Applied and Environmental Microbiology, 2011. **77**(18): p. 6516-6523.
236. Tian Guo, Z., et al., *Rapid Detection of Haemophilus influenzae and Haemophilus parainfluenzae in Nasopharyngeal Swabs by Multiplex PCR.* Biomedical and Environmental Sciences, 2012. **25**(3): p. 367-371.
237. Sandres-Saune, K., et al., *Determining hepatitis C genotype by analyzing the sequence of the NS5b region.* Journal of Virological Methods, 2003. **109**(2): p. 187-193.

238. Margall, N., et al., *Two unusual hepatitis C virus subtypes, 2j and 2q, in Spain: Identification by nested-PCR and sequencing of a NS5B region.* Journal of Virological Methods, 2015. **223**: p. 105-108.
239. Asogun, D.A., et al., *Molecular Diagnostics for Lassa Fever at Irrua Specialist Teaching Hospital, Nigeria: Lessons Learnt from Two Years of Laboratory Operation.* Plos Neglected Tropical Diseases, 2012. **6**(9).
240. Singh, B., et al., *A genus- and species-specific nested polymerase chain reaction malaria detection assay for epidemiologic studies.* American Journal of Tropical Medicine and Hygiene, 1999. **60**(4): p. 687-692.
241. Kearse, M., et al., *Geneious Basic: An integrated and extendable desktop software platform for the organization and analysis of sequence data.* Bioinformatics, 2012. **28**(12): p. 1647-1649.
242. Untergasser, A., et al., *Primer3-new capabilities and interfaces.* Nucleic Acids Research, 2012. **40**(15).
243. Koressaar, T. and M. Remm, *Enhancements and modifications of primer design program Primer3.* Bioinformatics, 2007. **23**(10): p. 1289-1291.
244. Roux, K.H. and K.H. Hecker, *One-step optimization using touchdown and stepdown PCR.* Methods in molecular biology (Clifton, N.J.), 1997. **67**: p. 39-45.
245. Don, R.H., et al., *TOUCHDOWN PCR TO CIRCUMVENT SPURIOUS PRIMING DURING GENE AMPLIFICATION.* Nucleic Acids Research, 1991. **19**(14): p. 4008-4008.
246. Federhen, S., *The NCBI Taxonomy database.* Nucleic Acids Research, 2012. **40**(D1): p. D136-D143.
247. Pruitt, K.D., et al., *NCBI Reference Sequences (RefSeq): current status, new features and genome annotation policy.* Nucleic Acids Res, 2012. **40**.
248. Guindon, S., et al., *New Algorithms and Methods to Estimate Maximum-Likelihood Phylogenies: Assessing the Performance of PhyML 3.0.* Systematic Biology, 2010. **59**(3): p. 307-321.
249. Gascuel, O., *BIONJ: An improved version of the NJ algorithm based on a simple model of sequence data.* Molecular Biology and Evolution, 1997. **14**(7): p. 685-695.
250. Lefort, V., J.-E. Longueville, and O. Gascuel, *SMS: Smart Model Selection in PhyML.* Molecular Biology and Evolution, 2017. **34**(9): p. 2422-2424.
251. Anisimova, M. and O. Gascuel, *Approximate likelihood-ratio test for branches: A fast, accurate, and powerful alternative.* Systematic Biology, 2006. **55**(4): p. 539-552.
252. Subramanian, B., et al., *Evolview v3: a webserver for visualization, annotation, and management of phylogenetic trees.* Nucleic Acids Research, 2019. **47**(W1): p. W270-W275.
253. He, Z., et al., *Evolview v2: an online visualization and management tool for customized and annotated phylogenetic trees.* Nucleic Acids Research, 2016. **44**(W1): p. W236-W241.
254. Zhang, H., et al., *EvolView, an online tool for visualizing, annotating and managing phylogenetic trees.* Nucleic Acids Research, 2012. **40**(W1): p. W569-W572.
255. Charleston, N.A. and S.L. Perkins, *Traversing the tangle: Algorithms and applications for cophylogenetic studies.* Journal of Biomedical Informatics, 2006. **39**(1): p. 62-71.
256. Team, R.C., *R: A language and environment for statistical computing*, in *R Foundation for Statistical Computing*. 2020: Vienna, Austria.
257. Antonio Balbuena, J., R. Miguez-Lozano, and I. Blasco-Costa, *PACo: A Novel Procrustes Application to Cophylogenetic Analysis.* Plos One, 2013. **8**(4).
258. Guangchuang Yu, D.S., Huachen Zhu, Yi Guan, Tommy Tsan-Yuk Lam. , *ggtree: an R package for visualization and annotation of phylogenetic trees with their covariates and other associated data.* Methods in Ecology and Evolution, 2017. **8**(1): p. 28-36.
259. Guangchuang Yu, T.T.-Y.L., Huachen Zhu, Yi Guan. , *Two methods for mapping and visualizing associated data on phylogeny using ggtree.* Molecular Biology and Evolution, 2018. **35**(2): p. 3041-3043.
260. K., P.E.S., *ape 5.0: an environment for modern phylogenetics and evolutionary analyses in R.* Bioinformatics, 2019. 35: p. 526-528.
261. Juan Antonio Balbuena, T.P., Matthew Hutchinson and Fernando Cagua, *paco: Procrustes Application to Cophylogenetic Analysis.* 2019.
262. Morgan, M., *BiocManager: Access the Bioconductor Project Package Repository.* 2019.
263. Revell, L.J., *phytools: An R package for phylogenetic comparative biology (and other things).* Methods Ecol. Evol, 2012. **3**: p. 217-223.
264. Slowikowski, K., *ggrepel: Automatically Position Non-Overlapping Text Labels with 'ggplot2'*. 2020.
265. Wickham, H., *ggplot2: Elegant Graphics for Data Analysis*. 2016, New York: Springer-Verlag.
266. Wickham, H., *stringr: Simple, Consistent Wrappers for Common String Operations*. 2019.
267. Yu, G., *Using ggtree to visualize data on tree-like structures.* Current Protocols in Bioinformatics, 2020. **69**: p. e96.

268. Hutchinson, M.C., et al., *paco: implementing Procrustean Approach to Cophylogeny in R.* Methods in Ecology and Evolution, 2017. **8**(8): p. 932-940.
269. Katoh, K. and D.M. Standley, *MAFFT Multiple Sequence Alignment Software Version 7: Improvements in Performance and Usability.* Molecular Biology and Evolution, 2013. **30**(4): p. 772-780.
270. Sanchez, R., et al., *Phylemon 2.0: a suite of web-tools for molecular evolution, phylogenetics, phylogenomics and hypotheses testing.* Nucleic Acids Research, 2011. **39**: p. W470-W474.
271. Capella-Gutierrez, S., J.M. Silla-Martinez, and T. Gabaldon, *trimAl: a tool for automated alignment trimming in large-scale phylogenetic analyses.* Bioinformatics, 2009. **25**(15): p. 1972-1973.
272. Altschul, S.F., et al., *Gapped BLAST and PSI-BLAST: a new generation of protein database search programs.* Nucleic Acids Research, 1997. **25**(17): p. 3389-3402.
273. Pfennig, D.W., *Effect of Predator‐Prey Phylogenetic Similarity on the Fitness Consequences of Predation: A Trade‐off between Nutrition and Disease?* The American Naturalist, 2000. **155**(3): p. 335-345.
274. Ricklefs, R.E. and S.M. Fallon, *Diversification and host switching in avian malaria parasites.* Proceedings of the Royal Society B-Biological Sciences, 2002. **269**(1494): p. 885-892.
275. Perlman, S.J. and J. Jaenike, *Evolution of multiple components of virulence in Drosophila-nematode associations.* Evolution, 2003. **57**(7): p. 1543-1551.
276. Hanada, K., Y. Suzuki, and T. Gojobori, *A large variation in the rates of synonymous substitution for RNA viruses and its relationship to a diversity of viral infection and transmission modes (vol 21, pg 1074, 2004).* Molecular Biology and Evolution, 2004. **21**(7): p. 1462-1462.
277. Parrish, C.R., P.R. Murcia, and E.C. Holmes, *Influenza Virus Reservoirs and Intermediate Hosts: Dogs, Horses, and New Possibilities for Influenza Virus Exposure of Humans.* Journal of Virology, 2015. **89**(6): p. 2990-2994.
278. Joycelyn, S.J., et al., *High detection rates and genetic diversity of picobirnaviruses (PBVs) in pigs on St. Kitts Island: Identification of a porcine PBV strain closely related to simian and human PBVs.* Infection Genetics and Evolution, 2020. **84**.
279. Yinda, C.K., et al., *Gut Virome Analysis of Cameroonians Reveals High Diversity of Enteric Viruses, Including Potential Interspecies Transmitted Viruses.* Msphere, 2019. **4**(1).
280. Fassler J, C.P., *BLAST Glossary In: BLAST® Help [Internet].* 2008, Bethesda, Maryland, USA: National Center for Biotechnology Information (US).
281. Goldman, D. and K. Domschke, *Making sense of deep sequencing.* International Journal of Neuropsychopharmacology, 2014. **17**(10): p. 1717-1725.
282. Bhattacharya, R., et al., *Molecular epidemiology of human picobirnaviruses among children of a slum community in Kolkata, India.* Infection Genetics and Evolution, 2006. **6**(6): p. 453-458.
283. Honigman, A., et al., *Cis Acting Rna Sequences Control the Gag Pol Translation Readthrough in Murine Leukemia-Virus.* Virology, 1991. **183**(1): p. 313-319.
284. Loughran, G., et al., *Evidence of efficient stop codon readthrough in four mammalian genes.* Nucleic Acids Research, 2014. **42**(14): p. 8928-8938.
285. Yan, Y., et al., *Analysis on the pathogeny of viral diarrhea in summer and fall of 2010 in Guiyang districts.* Modern Preventive Medicine, 2013. **40**(7): p. 1201-1203.
286. Kylla, H., et al., *Coinfection of diarrheagenic bacterial and viral pathogens in piglets of Northeast region of India.* Veterinary World, 2019. **12**(2): p. 224-230.
287. Ganesh, B., et al., *Detection of closely related Picobirnaviruses among diarrhoeic children in Kolkata: Evidence of zoonoses?* Infection Genetics and Evolution, 2010. **10**(4): p. 511-516.
288. Valle, M.C., et al., *Diarrheic syndrome related to viral agents in kidney transplanted patients.* Medicina-Buenos Aires, 2001. **61**(2): p. 179-182.
289. Amarasinghe, S.L., et al., *Opportunities and challenges in long-read sequencing data analysis.* Genome Biology, 2020. **21**(1).
290. Li, L., et al., *Comparing viral metagenomics methods using a highly multiplexed human viral pathogens reagent.* J Virol Methods, 2015. **213**: p. 139-46.