University of Nottingham

UK | CHINA | MALAYSIA

A three-pronged approach to virus discovery

Jack Douglas Hill, MSc. (hons), BSc. (hons)

Thesis submitted to the University of Nottingham
for the degree of Doctor of Philosophy

November 2023

Abstract

Rodents and bats are two key groups of host species for viral diversity and zoonotic transmission to humans. Whilst the use of NGS and PCR based virus discovery is improving our collective knowledge of the virome, very little is known about the virome overall, and even less is known about the specific virome of these mammalian hosts. Similarly, information is lacking regarding the prevalence of viruses within UK rodents. As climate change progresses and rodents are driven into closer proximity to humans and livestock, the risk of rodent zoonotic transmission is increasing, increasing the risk of transmission of viruses with pandemic potential. Evolutionary information about key viruses may help to improve pandemic preparedness, but the field of historic virology is still relatively underdeveloped, particularly regarding historic RNA virus investigations. Therefore, this project was designed to take a three-pronged approach to investigating virus diversity within UK rodents via degenerate PCR and NGS screening, and to advance the field of historic RNA virus discovery.

140 modern rodents from 5 species were screened by degenerate PCR for the presence of adenoviruses, hantaviruses, coronaviruses, *Rotaviruses* and *Rubiviruses*. 2 adenoviruses and 1 hantavirus were identified, and all other viruses were not found in. This screening was complemented by unbiased NGS sequencing of these rodents and analysis of the NGS data. The metagenomics screening yielded a total of 216 viral hits across 19 viral genera, including potentially important viruses such as *Cardioviruses*, *Hepaciviruses* and hantaviruses, amongst others. These viruses were PCR confirmed, and expanded the known viral diversity of bank voles, field voles, wood mice and yellow-necked mice to include *Picobirnaviruses, Kunsaigiviruses, Rosaviruse, Pegiviruses, Bocaparvoviruses* and polyomaviruses. *Hepacivirus F, Rosaviruses* and *Orbiviruses* were also found in the UK for the first time, and accurate abundance estimates for 19 viral genera were quantified, ranging from 0.7%-67.1%. Collectively, this drastically improves our collective knowledge of the UK virome.

This project also advanced the techniques used for historic RNA virus discovery and manipulation. A protocol for historic RNA extraction from preserved animals ranging from at least 35-156 years old was developed with over 90% extraction success. NGS library preparation processes were improved to yielded functional NGS libraries, albeit with substantial adaptor dimer contamination. qPCR screening methods for historic coronaviruses were also developed, and historic coronaviruses were found in 3 samples. Collectively, this advances the methodology and techniques of historic RNA virus discovery, and shows as a proof of concept that conventional screening methods can be used for the discovery of historic viruses in well preserved samples.

Acknowledgments

I would like to begin by thanking my supervisors Prof. Jonathan Ball and Dr. Patrick McClure, without whom this project would not have been possible. Their guidance and expertise has been invaluable throughout this project. I would also like to thank Dr. Joseph Chappell for his help at every stage in the project, and for allowing me to follow in his footsteps with this project, Dr.Theocharis Tsoleridis for his help with phylogenetics, and Sophie Wartnaby for being a good friend in the lab and helping out in the day to day of this project.

Outside of the Virus Research Group, I would like to thank Dr. Andrea Sartorius and Roberto Portela-Miguez for providing the modern and ancient samples respectively- there literally would not be a project without them. I would also like to thank the DeepSeq team for their assistance with the tapestation system and Dr. Stuart Astbury for operating the high performance computer throughout this project.

On a more personal side, I would like to thank my 4 best friends Adela Boboc, Luke Mudie, Abbie Tomes and Joseph Hinds for their friendship and for keeping me sane throughout the project (and an extra thank you to Joe for being the best office mate I could ask for!). I would like to thank my Mum and Dad, brothers Charlie and Murphy and sister Bonnie for all their love and support throughout these 4 years, without them there's simply no way I would have made it to the end of this.

Finally, and most importantly, I would like to thank my fiancée Emma most of all. Thank you for all your help over these years, for being my teammate, proofreader and best friend throughout it all. I love you, and couldn't have done this without you, your support and the joy you bring each day.

Table of contents

Contents

List of figures

List of tables

Abbreviations

AAV	*Adeno-associated virus*
aDNA	Ancient DNA
aeDNA	Ancient and environmental DNA facility
AHSV	*African horse sickness virus*
aRNA	Ancient RNA
B19V	Parvovirus B19
BLAST	Basic local alignment search tool
BLASTX	Translated protein BLAST
bp	Base pairs
BTV	*Bluetongue virus*
BV	Bank vole
cDNA	Complementary DNA
CMV	*Cytomegalovirus*
CNS	Central nervous system
CoV	Coronavirus
COVID-19	Coronavirus disease 2019
CPV	Canine parvovirus
CRS	Congenital rubella syndrome
CSF	Cerebrospinal fluid
Ct	Cycle threshold
CZID	Chan-Zuckerberg IDSeq
dH_2O	Deionised water
DNA	Deoxyribonucleic acid
dNTP	Dioxynucleotide triphosphate
dsDNA	Double-stranded DNA
dsRNA	Double-stranded RNA
EAV	*Equine arterivirus*
ECMV	*Encephalomyocarditis virus*
EDTA	Ethylenediaminetetraacetic acid
EHDV	Epizootic haemorrhagic virus
ELISA	Enzyme-linked immunosorbent assay
FDA	Food and Drug Administration
FFPE	Formalin-fixed paraffin embedded
FFPET	Formalin-fixed paraffin embedded tissue
FPV	Feline panleukopenia virus
FV	Field vole
GAPDH	Glyceraldehyde-3-phosphate dehydrogenase
GSNAP	Genomic Short-read Nucleotide Alignment Program
GVP	Global virome project
HAV	Hepatitis A virus
HCMV	*Human cytomegalovirus*
HCPS	Hantavirus cardiopulmonary syndrome

HCV	*Hepatitis C virus*
HCV	Hepacivirus C/ hepatitis C virus
HFMD	Hand, foot and mouth disease
HFRS	Haemorrhagic fever with renal syndrome
HIV	Human immunodeficiency virus
HPC	High performance computer
HPeV	*Human parechovirus*
HPgV	*Human pegivirus*
HPS	Hantavirus pulmonary syndrome
HTS	High throughput sequencing
ICTV	International Committee on Taxonomy of Viruses
IgG	Immunglobulin G
IgM	Immunglobulin M
IPA	Isopropanol
IUPAC	International Union of Pure and Applied Chemistry
kb	Kilobases
KSHV	*Kaposi's sarcoma-associated herpesvirus*
LASV	*Lassa virus*
LCMV	Lymphocytic Choriomeningitis Virus
LEDC	Less economically developed country
LZW	Lempel-Ziv-Welch
MCD	Multicentric Castleman disease
MEDC	More economically developed country
MERS-CoV	Middle-Eastern respiratory syndrome coronavirus
MLV	*Murine leukemia virus*
MMR	Measles, Mumps, Rubella
N/A	Not available
NCBI	National Centre for Biotechnology Information
NE	Nepropathia epidemica
NEB	New England Biolabs
NGS	Next generation sequencing
NHM	Natural History Museum
NTBLAST	Nucleotide BLAST
NW	New-World
OPV	Oral *poliovirus* vaccine
ORF	Open reading frame
OW	Old-World
PACS	Post-acute COVID-19 syndrome
PAGE	Polyacrylamide gel electrophoresis
PARV4	*Parvovirus 4*
PBV	*Picobirnavirus*
PCR	Polymerase chain reaction
PEL	Primary effusion lymphoma
PRRSV	Porcine reproductive and respiratory syndrome virus
qPCR	Quantitative polymerase chain reaction

RAM	Random Access Memory
RdRp	RNA-dependent RNA polymerase
RIN	RNA intergrity number
RNA	Ribonucleic acid
RT-PCR	Reverse transcription PCR
RT-qPCR	Reverse transcription quantitative PCR
RubV	*Rubella virus*
RuhV	*Ruhugu virus*
RusV	*Rustrela virus*
SAFV	Saffold virus
SARS-CoV	Severe acute respiratory syndrome coronavirus
SARS-CoV-2	Severe acute respiratory syndrome coronavirus 2
SFTP	Secure file transport protocol
SPRI	Solid-phase reversible immobilisation
ssDNA	Single-stranded DNA
ssRNA	Single-stranded RNA
STAR aligner	Spliced Transcripts Alignment to a Reference
SV40	*Simian virus 40*
TAE	Tris-Acetate-EDTA
TATV	Tatenale virus
TDAV	*Theiler' disease associated virus*
TE	Tris-EDTA
TMEV	*Theiler's murine encephalomyocarditis virus*
TS	Trichodysplasia spinulosa
UTR	Untranslated region
UV	Ultraviolet
VDPV	Vaccine-derived poliovirus
VHF	Viral haemorrhagic fever
WHO	World Health Organisation
WM	Wood mouse
WPV	Wild *poliovirus* species
YM	Yellow-necked mouse

Chapter 1- Introduction

1. Understanding the virome

a. What is the virome?

The virome can be described as the collective term for all viruses on Earth, including an estimated 10^{31} virions[1,2,3,4]. This project is more concerned with the mammalian virome, which can be described as including all viruses that infect eukaryotic organisms or cells, virus derived genetic elements within mammalian eukaryotic chromosomes such as endogenous retroviruses that can modulate host gene or protein expression, archaeal viruses that can infect archaea, and bacteriophage viruses that can infect bacteria[5]. Prions may also be considered a part of the mammalian virome, although there is some debate regarding their inclusion, and they are not relevant to this project and will not be considered further[5]. Bacteriophages, archaeal viruses and viruses of protozoa were also not considered to be within the scope of this project, as they are not directly relevant for mammalian infection, although it should be noted that their indirect effects on human and animal health can be significant[1,6].

b. How much do we understand?

i. Estimated understanding of the virome

Whilst our understanding of the virome in some more common host species such as *Mus musculus* is relatively deep (largely due to their use as laboratory model species), as of 2014, it was estimated that approximately 1% of the virome had been discovered[4,5,7]. Whilst the majority of the virome is comprised of bacteriophage, limited further virome composition information is available, in part due to the lack of an analogue for the bacterial 16S rRNA gene as a consistent marker limiting sequence information availability[1]. With recent advancements in metagenomics technologies, more of the virome has been discovered across both small mammals and humans[6,8]. Despite these advancements, it is still sometimes estimated that only 1-2% of the total virome has been discovered, although more optimistic updates suggest that up to 6% of the viral diversity in vertebrates has now been identified[5,9]. Recently, the Global Virome Project (GVP) has begun, and aims to use modern technologies to identify approximately 70% of the virome within the next 10 years[10].

ii. Rapid evolution limiting our understanding

The rapid evolution of viruses limits our understanding of the overall virome. Viral evolution can lead to cross-species transmission, which can in turn lead to rapid viral evolutionary adaptation to the new host species, potentially leading to a closely related yet distinct viral species[5,11]. Hantaviruses are a good example of this concept, where the Tula and Tatenale viruses are phylogenetically very similar and are derived from similar host species but are divergent enough to be classified as distinct species, largely due to their coevolution within their respective hosts[11,12]. This virus-host coevolution can lead to increased transmission, leading to further evolution in a continuous cycle[13]. Because of this, viruses with rapid transmission potential such as

influenza viruses evolve quickly and divergently relative to eukaryotes, bacteria and other kingdoms of life, and even relative to other less rapidly evolving viruses[14,15]. There is also evidence of horizontal gene exchange between RNA viruses of different Baltimore classes, which adds further diversity and evolutionary potential[16]. RNA viruses tend to have RNA-dependent RNA polymerases (RdRps) that replicate extremely quickly with little (if any) proofreading capacity, resulting in a very high mutation rate that in turn drives rapid evolution, with an average mutation rate of 10^{-2}- 10^{-4} mutations per site per year[17,18,19]. The large population sizes of RNA viruses also drives genetic variability and evolution, as there are many viruses reproducing within a given population at any given moment, any of which may mutate and evolve[18,20]. Segmented viruses may also undergo reassortment, where two different viruses of the same species or family co-infect a cell and segments are exchanged during the packaging of progeny viruses. This is commonly observed phenomenon in Rotaviruses, where 3-5% of viruses show reassortment, and influenza, where in swine shows in 2018 at least 7 different reassortment events were observed[20,21]. Between cross-species transmission and adaptation, rapid evolution and large population sizes, unreliable replication, recombination and reassortment, the virome is constantly evolving and fluctuating, limiting our overall understanding[5,17,18, 20].

iii. Sampling bias and other issues of understanding

Our understanding of the virome is also subject to sampling bias. As expected, there have been significant investigations and studies into the human virome and its composition, and some mammalian reservoir species have also undergone extensive virome investigations, whilst other species have been undersampled or not sampled at all[6,7,22,23]. However, whilst some rodent and bat species have been sampled to some degree, this kind of extensive metagenomics analysis has not been performed on the majority of species, including key livestock species such as goats, or even the majority of bat or rodent species[8,24,25]. Rare and endangered species are also undersampled, and host spatiotemporal dynamics are often not considered when designing virome investigation studies[23]. Also, metagenomics virology study results are often poorly annotated or the data itself is not publicly available, which limits their value and the virome information provided to the wider scientific community[24].

Sampling methods, including metagenomics library preparation procedures, RNA extraction and data analysis pipelines, can lead to skewing of results and may result in low copy number viruses being missed, particularly when there is a high bacteriophage background or multiple animal pooling is performed[6,24,26,27]. Sampling methods often also bias towards double-stranded DNA (dsDNA) viruses rather than RNA viruses and viral investigations tend to steer understandably towards known zoonotic diseases, causing sections of the virome to be inadequately investigated[1,23]. The type of sample can also bias virome investigations. For example, gut or faecal samples may contain viruses from the food of the animal sampled, whereas viral tissue tropism may result in viruses present in the animal but not the sampled tissue being missed[23]. Similarly, many virome investigations rely on convenience sampling

rather than probability sampling, which may introduce bias due to characteristics often associated with convenience samples, such as premature death and illness (often from persistent infections, in turn leading to bias against the detection of non-lethal acute infections)[28]. Additionally, there is bias towards sampling locations- for example, rodents and bats in China have been sampled extensively alongside rodents in some major US cities such as New York, whereas other regions have been poorly sampled such as some hyperdiverse regions of Brazil and the Americas[8,23,29-32]. Not all environments have been adequately investigated either- for example, farms and urban areas have been relatively well investigated due to their proximity and impact on human health, whilst other areas have been undersampled[24,30,32,33].

The above factors contribute to the presence of significant biases in investigations into the virome, and the overall understanding of the virome. However, as metagenomics technologies advance, these biases are likely to become less significant, provided that the appropriate consideration is given to the host species targeted and habitats tested[25,34].

c. Why does it matter?

i. Human zoonosis

Zoonotic viruses are defined as viruses that can be transmitted from vertebrates to humans and cause an infectious disease, and which may be transmitted either via direct contact or indirect contact[35]. Whilst this definition is imperfect- for example, it fails to distinguish between viruses that are directly transmitted from animals in every case such as rabies and those that have jumped species and are now near exclusively transmitted amongst humans such as SARS-CoV-2- it is broadly accepted for its simplicity and broad applicability[36]. Whilst further breakdowns of the term zoonosis have been proposed, for the purposes of this work the above definition offered by Mollentze and Streicker will be considered sufficient[35,36]. One definition of reservoir hosts is that they are animals that can be infected by a virus and then proceed to transmit the virus to another susceptible host- such as another animal of the species or a human via zoonotic transmission- without suffering symptoms of disease from the virus, and one definition of a maintenance host is a host species in which a pathogen can persist without requiring zoonotic transmission from another host[37,38]. However, the precise definition of reservoir host is difficult both conceptually and practically, and will vary via context and by who is defining the term, to the extent that some consider a reservoir to be a single species and others consider a reservoir to be a complex set of species amongst which a pathogen may be transmitted[39]. In practice, the terms zoonotic host and reservoir host are often used somewhat interchangeably. Dead-end hosts are those that can be infected by a virus but are unable to transmit the virus further, although in some rare cases dead-end hosts may be involved in infrequent and circumstantial transmission[38]. Many important viruses have zoonotic potential, and most viruses that are important for human public health have close relatives in bats, rodents or other reservoir species[31].

Each spillover event increases the risk that a virus that can transmit laterally between humans enters a human host, potentially leading to major international pandemics[21].

Whilst all kingdoms of life can be infected by viruses, it is largely mammalian species that are likely to act as reservoir species for zoonotic viruses, with bats, rodents and to a lesser extent primates and birds acting as important hosts largely due to their vast species diversity and associated large zoonotic virus pool, although invertebrates and other kingdoms of life can also act as reservoirs under some circumstances[7,25,32,35]. Most newly emerging infectious viral diseases of humans are likely caused by spillover events[31]. Broadly, increased host species phylogenetic relatedness to humans is associated with increased zoonotic spillover potential, as is species diversity and viral richness, although other factors such as opportunity to interact with humans, reproductive rate and availability of key resources such as food for the host species are also important when considering spillover risk[32,40]. Both virus traits and host traits contribute towards spillover risk- for example, RNA viruses are more likely to spillover and cause a zoonotic infection than DNA viruses, and host sympatry (having a geographical and ecological range that overlaps with other potential hosts) also increases zoonotic virus host potential[7]. Some clinical features also increase transmissibility and zoonotic potential, such as symptoms that exacerbate viral shedding[32]. Trying to elucidate the factors that are involved in causing or enabling zoonotic spillover events is difficult and often inconsistent, as the variety of host and viral factors, the role of reservoirs, temporal and spatial variability and other ecological factors all interact in a complex web that varies on a case-by-case basis[38].

Zoonotic transmission rates are linked to many biotic and abiotic factors, but usually the most important factor is the frequency of contact between humans and the animal reservoir species[30]. As global urbanisation increases, humans and wildlife are driven into closer proximity, allowing for more cross-species interactions and a greater risk of zoonotic infection[31,32]. For example, there has been an increase in bats and rodents in urban areas due to habitat change, which increases the risk of zoonotic infections[22]. Seasonal virus prevalence is also a variable risk factor for zoonotic transmission and is likely to become more important as climate change advances[41,42]. As climate change progresses, a variety of factors are likely to influence zoonotic risk. One such example is the expansion of host ranges for insect vectors of arboviruses, as global temperature increases expand the regions at which temperatures are suitable for mosquitoes and other insects, as well as increased rainfall providing more insect breeding sites. Some rodent populations will likely increase due to the favourable breeding conditions of warmer winters and spring leading to an increase of food, followed by heat waves which drive rodents towards human dwellings as shelter, in turn increasing the risk of rodent-human interaction and zoonotic transmission[42]. Whilst this effect on rodent populations is largely theoretical at this point, there is evidence from Germany that favourable rodent breeding conditions is associated with increased rodent populations, which are

associated with an increase in tick-vector population increases, which in turn is associated with an increased transmission risk of tick-borne diseases such as tick-borne encephalitis virus[43]. Increased rainfall and flooding may lead to outbreaks of pathogens such as *Norovirus* due to water contamination, and increased winds may lead to increased airborne transmission of viruses, such as strains of avian influenza (although in the case of avian influenza this is likely to act as a secondary transmission route relative to enteric transmission)[42]. Overpopulation and rural to urban human migration will also increase the risk of transmission of zoonotic viruses, and human global migration may lead to zoonotic diseases being transmitted globally into naïve human populations[42]. These factors and others will likely lead to an increased rate of viral spillover events, which may then lead to outbreaks and viral host-adaptation, in turn possibly driving some zoonotic viruses to be maintained in human populations and become endemic[42].

It is difficult to accurately estimate the incidence of zoonotic infections, as they are often not routinely screened for and typically present with mild or indistinct symptoms[30]. For example, an estimated 80% of *Lassa virus* (LASV) infections are asymptomatic, and other diagnostic issues regarding LASV render it impossible to accurately estimate LASV spillover infections, therefore it is likely that the zoonotic risk of LASV is frequently undersetimated[44,45]. There are also likely many missing zoonotic viruses that have yet to be discovered, largely in the Americas but also worldwide, with significant variability by host taxonomic order[32]. Indeed, some estimates suggest that up to 99.9% of potentially zoonotic viruses have yet to be discovered, which represents approximately 750,000 individual viruses[46]. Whilst it is unlikely that all or even most of the undiscovered zoonoses will have the potential to cause a major pandemic, it is probable that at least some of these viruses will[10].

ii. Animal cross-species transmission

Whilst animal to human spillover is common, most cross-species viral transmission is between non-human mammalian species (for example, influenza spillover from birds to pigs and vice versa), including important livestock animals[21,40]. Viral spillovers into farm animals can have major health consequences for the livestock and economic consequences for the farmers- for example, a relatively recent CoV spillover and outbreak in pigs in China was fatal to the animals[47]. Approximately 1/3rd of all meat consumed worldwide is pork, and as pigs are well known reservoirs for a variety of potentially pathogenic viruses that can cause disease in humans, the increase in swine farming to meet food demands leads to both an increased risk of spillover into humans and more infections affecting animal welfare in pigs[48]. Zoonotic disease comes at a significant economic cost- as of 2010, an estimated $2 billion per year were lost to zoonotic disease (excluding SARS-CoV-2), with estimates including indirect costs of zoonoses being 10-fold higher[49]. If SARS-CoV-2 is also considered to be zoonotic (as discussed below), this figure increases into the trillions of dollars, with one estimate in December 2023 of total economic loss due to SARS-CoV-2 since 2020

reaching $14.7-21.8 trillion in total, equivalent to approximately $5-7 trillion per year[50].

It is possible for "reverse zoonotic infection" to occur, where humans transmit viruses to animals, one example of which is the reverse zoonotic spillover of astroviruses from humans to marine mammals[32,51]. There is also evidence of zoonotic and reverse zoonotic transmission on farms, through a pathway known as the wildlife-livestock-human interface where wildlife infects livestock via zoonotic transmission, which then infects humans, or humans infect livestock, which then infects wildlife[24]. For example, human swine handlers have been shown to transmit influenza viruses to swine via reverse zoonotic transmission, which have then in turn transmitted the influenza virus or a recombinant variant to another human via zoonotic transmission[21].

The farm environment can often lead to favourable conditions for rodent reservoir host species such as the brown rat (*Rattus norvegicus*) amongst others, where factors such as the availability of food and water for the livestock can lead to a population boom of the zoonotic host, in turn leading to increased zoonotic risk[15,52]. So far, metagenomics studies on farm animals are limited and of variable quality, leading to inadequate information about the virome and health of important livestock animals[24]. Intensive farming processes can also lead to rapid outbreaks of disease amongst farm animals due to cross-species transmission and increased host population sizes, and as farming demands increase to match an increasing population this is likely to become more commonplace[24,25].

Domestic and companion animals are also important to consider regarding zoonoses. Whilst there is some variability regarding the definition of a companion animal, domestic pets such as cats, dogs, small rodents and pet birds are usually considered to be companion animals[53]. The most common companion animals within the EU are cats and dogs, which have been shown to harbour and transmit a variety of viruses (such as astroviruses) via zoonotic transmission[51,53]. Whilst many zoonoses are still transmissible from domestic animals, due to effective vaccination and animal control others are becoming less common, such as *rabies lyssavirus*, which in many parts of the world (typically in MEDCs with effective vaccination programmes) is now significantly more likely to be transmitted by a wild animal than a domestic dog, although it should be noted that dog rabies is still a significant threat and public health burden in the majority of the world[23]. Pet rodents have also been shown to harbour and transmit zoonotic viruses- for example, in 2013 two people in the UK became infected with *Seoul orthohantavirus* from their pet rats, and consequently developed HFRS[11]. Pet parrots have been known as the cause the spillover of Newcastle disease virus into the poultry industry, with important animal welfare and economic consequences[53]. A variety of other viruses, including *West Nile virus*, *Bluetongue virus* (BTV) and *Crimean-Congo haemorrhagic fever orthonairovirus* have also been highlighted as important zoonoses in companion animals within the EU[53].

iii. Improving evolutionary understanding

Whilst zoonoses are important, there are other reasons for studying the virome. For example, information regarding zoonotic viruses and their transmission is significantly more useful in the wider context of the mammalian virome, as are zoonotic risk assessments[25]. Understanding the virome, hosts, environments and other such links in the disease network is essential for adequate pandemic preparedness, as a broader general understanding of these factors leads to a quicker response and counter effort when a pandemic emerges and can potentially allow for the mitigation of a pandemic before it reaches a critical point[10,32]. Improving our knowledge of the virome- both of currently known and currently unknown viral species- can improve the development and efficacy of therapeutics and vaccines, and lead to a better success rate in clinical trials for these medical interventions[5]. Gaining a greater insight into the virome and the evolutionary history of the viruses within may also grant new perspectives and clinical options regarding known clinically important viruses[31]. Monitoring the virome is also important to identify novel strains or species in rapidly evolving viruses and to identify and act against variants that may have evolved to circumvent existing antiviral therapies and preventative measures[17]. Finally, viruses and virus-like particles have major uses in a variety of fields, such as vaccine development, genetic engineering and cancer treatment, and in other non-medical fields such as plant farming and cosmetics[54]. As both virological discovery and manipulation technologies and our understanding of the virome develop, it stands to reason that these applications will be enhanced by the discovery of new and more useful species or strains of viruses, as well as finding applications for viruses that are inconceivable with current knowledge[5,54].

2. Methods of virus discovery

a. History of virus discovery

The tobacco mosaic virus was the first virus discovered in the 1890s and the term virus was first coined in 1898, although the first theories of viruses for agents of disease were recorded in the 9th century[4,55,56]. Shortly after, the first human viruses were isolated, including the *poliomyelitis virus* in 1912 and *influenza A virus* in 1918, as well as *rabies lyssavirus*, *Dengue virus* and measles virus amongst an estimated 40 others, largely due to the development of tissue culture techniques in 1907[4,56]. The evolution of virus discovery is fundamentally tied to the technology available, and improvements in techniques such as tissue culture allow for the advancement of virus discovery and understanding, even to date[4,55,56]. For example, the rapid increase of vaccine development to usher in the "golden age" of vaccination in the 1950s was largely possible due to the development and uptake of cell culture techniques[57]. Early virus discovery methods such as the use of Chamberland filters allowed for the illustration of the presence of the virus, but no further information about viral properties[55].

Virus particles were first able to be visualised using X-ray crystallography and electron microscopes in the 1940s, and whilst the first isolation of poliovirus via cell culture techniques in 1950 and subsequent use of cell culture technology has been revolutionary and fundamental to improving our understanding of viruses, the major leap forward in virus discovery came in the early 1980s with the back to back inventions of Sanger sequencing and the polymerase chain reaction (PCR)[55,58]. For the first time viruses could be amplified and sequenced, allowing for significant insights into viral genomes and their associated proteins[55]. The 21st century then led to the development of high-throughput NGS (next generation sequencing) technology and the development of bioinformatics, increasing the efficiency and frequency of virus discovery near exponentially[59]. Finally, due to the recent mass generation of bioinformatics data, data-driven virus discovery is becoming more and more common, where bioinformaticians and virologists scour data available in public repositories in an effort to find the viruses that will likely have been sequenced and disregarded or missed within the data. Whilst there are mixed feelings amongst the virology community regarding these researchers with some declaring them as "data parasites", this is an efficient, cheap and risk-free (due to the lack of wet lab steps) method of virus discovery which is likely to become more common in the future[55]. Whilst most historical virus discovery methods have fallen out of common use and been superseded by modern molecular methods, it is likely that for the foreseeable future a combination of PCR and NGS approaches will provide the best results for virologists for both virus discovery and diagnostics[38].

b. PCR- based virus discovery

PCR is still one of the most important techniques for virus discovery and sequencing used today- indeed, PCR is often still used as the confirmatory gold standard when assessing NGS based virus discovery[47,60]. Invented in the 1980s by Kary Mullis, PCR

works by using specific primers to bind to target DNA (or RNA for reverse-transcription PCR (RT-PCR)), and through multiple cycles of denaturation at high temperature, primer annealing at relatively low temperature and extension at an intermediate temperature, a polymerase is used to amplify the target nucleic acids in an efficient and accurate manner[61]. PCR is typically used as the method of choice when screening many samples for a specific virus, due to the ability to design specific primers against a conserved region of the target virus, in turn allowing for efficient amplification and identification of even low-copy number viruses[,61]. Alternatively, more degenerate pan-family or pan-genus primers can be designed, allowing for the identification of a broad range of related viruses and more divergent viruses within these groups, or even the identification of new viral families[12,59]. PCR can also be used for viral prevalence assessment through mass PCR screening, or for virus full-genome retrieval by "primer-walking" (essentially sequencing the entire genome via a series of smaller PCR reactions), although this is labour intensive and inadvisable for large genomes[47,62]. Metadata regarding animal samples can also be gained or confirmed via PCR screening. For example, PCR screening for the Y chromosome can be used to identify the sex of animals, and cytochrome B PCR and sequencing can be used for species identification[63,64]. PCR amplification steps are also fundamental for the generation of NGS libraries and NGS screening technology[59].

One major advantage of PCR screening is the extreme sensitivity, where a well-designed PCR reaction can theoretically amplify and identify even a single copy of a target virus, far surpassing the sensitivity of current NGS methods[26,37,65]. Another advantage of PCR is that it requires very little input material, in turn preserving precious and limited samples whilst still generating useful data[66]. Quantitative PCR (qPCR) can be used to quantify the amount of target genome present in the sample, allowing for accurate estimations of viral copy number and viral load, and qPCR reactions are often used as the "gold-standard" for diagnostic screening for many viruses and other pathogens, with immense value in rapid diagnostics and outbreak situations[30,34,60]. PCR is also relatively cheap, easy and quick to perform, allowing for rapid diagnostic testing and research advancement[67].

However, there are disadvantages to using PCR for virus detection. One is that prior sequence knowledge is required for primer design, meaning that it is necessary to target a specific virus. Primers are highly specific for the genome region that they were designed using, so a more divergent virus may be missed due to a lack of primer binding[60,61,65]. Similarly, PCR reagents and cycle conditions must be optimised to be efficient, and the presence of contaminants such as PCR inhibitors can reduce the efficiency of or prevent the PCR reaction entirely[11,61]. Whilst PCR is effective for identifying the presence of a virus and qPCR is useful for quantifying the virus, PCR is unable to determine whether the individual screened is still infective and likely to transmit the virus, although current NGS technologies also cannot do this[38]. However, in a well-designed PCR reaction, the advantages far outweigh the

disadvantages, and PCR is likely to remain a key tool for virus discovery and screening for the foreseeable future[38].

c. NGS-based virus discovery

NGS technology has been the most recent major innovation in virus discovery, first coming into routine use in the mid-2000s with the Roche 454 sequencing and early Illumina Solexa sequencing systems[55,68]. NGS technology is highly parallel, high-throughput technology that results in the generation of gigabytes or terabytes of data, theoretically sequencing every section of DNA or RNA in the target sample (although in practice the sensitivity of NGS is not sufficient for this)[65,69]. This can lead to the delivery and analysis of billions of reads and hundreds of billions of bases of sequence, leading to the generation of an extreme amount of data[69]. NGS sequencing has become a key approach to virus discovery and has facilitated the rapid expansion of our understanding of the virome and viral phylogeny[5,59].

The most prominent NGS technology is Illumina technology, which generates short read fragments ragning from 50-500bp (base pairs), although typically of 150bp[69]. At the time of writing, the most advanced Illumina sequencer is the Illumina NovaSeq 6000, which has the capacity to generate up to 20 billion reads within 48 hours [69]. However, the Illumina library preparation process remains largely the same as that for previous Illumina sequencers, albeit with some optimisation (such as the removal of a gel-based size selection step) and with increased availability of kit-based library preparation methods such as the NEBNext Ultra II Library prep kit[70].

The initial input into the kit is double stranded DNA, therefore RNA to cDNA synthesis steps are required to be performed if using RNA samples[69,70]. The DNA is then fragmented to the approximate target size either chemically or mechanically, before proceeding to the end repair and polishing step where DNA overhangs are removed. This is followed by the A-tailing step, where a poly-A adenine tail is enzymatically added to the target DNA[70]. The next step is adaptor ligation, where a double-stranded DNA adaptor binds to the double-stranded target DNA, before being enzymatically cleaved into two separate smaller adaptors- one at the 5' end and one at the 3' end of the target DNA[69]. The 3' adaptor contains a recognition site for a P7 index primer and a "barcode", a 6 base recognition sequence that is later used to identify sequences pertaining to a specific input sample, and the 5' adaptor contains a recognition site for a universal P5 primer used for amplification and during the library synthesis reaction. These primers are added sequentially, and the library is amplified by PCR after primer ligation to generate sufficient material to sequence[69]. Between the adaptor ligation and the library amplification, purification is performed using reversible magnetic SPRI (Solid-phase reversible immobilisation) beads, which help to remove unbound contaminants and to allow for size selection and filtration[70]. This is also performed at the end of most protocols. Finally, before loading onto the sequencer, it is typical to perform quality control and quantification to ensure a successful library generation, often by using a Tapestation or Bioanalyzer instrument

to assess fragment size[69]. Figure 1 shows an example of Illumina library preparation using the NEBNext Ultra II kit.

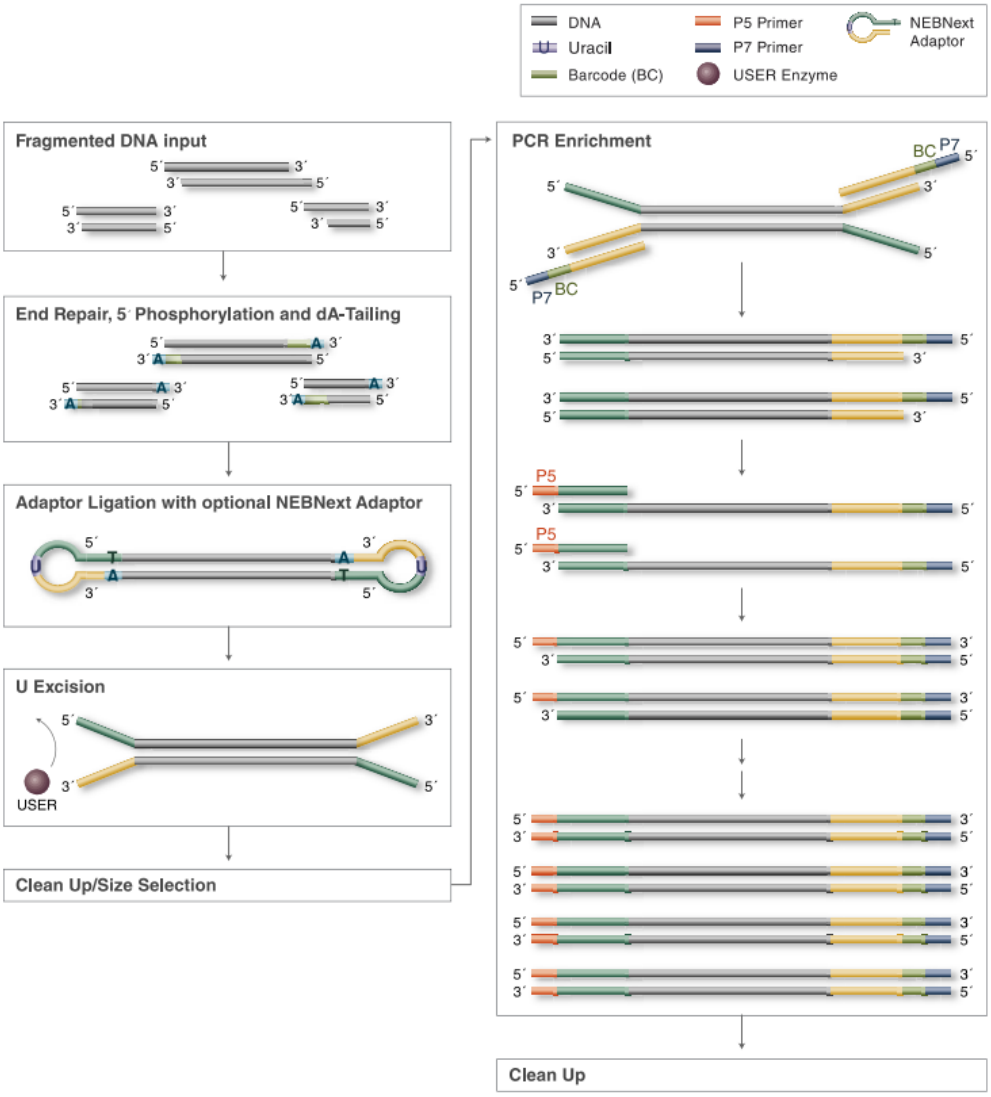


Figure 1- Library preparation process for the NEBNext Ultra II Library Prep kit. Figure is taken from the kit insert (NEB). Left hand side shows end repair and poly-A tailing, double stranded adaptor ligation, uracil cleavage to produce 2 single stranded adaptors and a SPRI bead clean up step. Right hand side shows the annealing of the P7 and P5 primers and the amplification of the library, followed by another SPRI bead clean up step.

After library preparation, the library must be amplified and sequenced to generate data, and the Illumina platform does this via the proprietary "sequencing by synthesis" process[69]. Here, the library is loaded onto a flow cell, which is a glass slide containing millions of oligonucleotides that are bound to the surface and are complementary to the adaptor sequence of the library, all contained within millions of individual fluidic channels. The adaptors contained within the library bind to the surface cells and are immobilised, before buffers, dNTPS and polymerases are added to allow for an extension process similar to that used in PCR [69]. The bound fragment then undergoes bridge PCR, where it is repeatedly amplified by PCR to generate an individual cluster of identical molecules that can then be sequenced[69]. The sequencing process is relatively similar to Sanger sequencing, where a fluorescently labelled dNTP is reversibly bound and imaged at each site of the growing strand, with the fluorescent wavelength corresponding to an individual nucleotide. The labelled dNTP is then removed and replaced with a standard dNTP, and the next sequencing cycle continues in the same way to continue sequencing the cluster of molecules[69]. Most Illumina sequencing is performed using paired-end sequencing to improve read quality, where the same cluster is sequenced in the forward direction, and then again in the reverse direction[69].

The libraries must then be analysed by bioinformatics to be of any use, ideally using a high-performance computer (HPC) to minimise analysis time[27]. Initially, samples must be demultiplexed using their index primer barcode to identify which sequences were produced by each library- this is performed as standard by most commercial sequencing companies[69]. The libraries then undergo various steps to become suitable for analysis, including quality assessment, adaptor trimming, background sequence removal (and host sequence removal for pathogen analysis), often *de novo* assembly, and alignment to reference sequences[6,27]. There are many standard pipelines available for this such as the CZID pipeline and Genome Detective, all following broadly the same process and giving similar results but often using alternative specific software[27,71,72]. It is beyond the scope of this work to undertake an in-depth review of alternative bioinformatics tools- for an overview, see the mini-review by Chappell and colleagues as a starting point[71]. Regardless, following bioinformatic analysis, it is possible to match sequences to reference genomes, which shows exactly which organism is most similar to the reads that have been sequenced, the region of the genome that has been sequenced, and the number of reads that have been sequenced for this region[59,71]. This is extremely useful for virus discovery[5,59].

Other metagenomics technologies are also available, such as the Oxford Nanopore technology and MGI DNA nanoball technology[68,73]. The MGI DNA nanoball technology is an alternative approach to providing short read sequences via rolling circle replication, and is a relatively new and somewhat unproven technology which gives comparable results to Illumina sequencing (both being over 99.9% accurate in most cases) but at a lower cost[69,73]. Oxford Nanopore technology is typically more

useful for long-read analysis and is therefore less directly comparable to Illumina technology. This is performed by passing a single molecule through a nanopore, where each nucleotide is detected based on the change in charge relative to the previous nucleotide and timescale[68]. The Oxford Nanopore system has many advantages such as being more portable, quicker and often able to detect a full genome of a pathogen, leading to a potential future in diagnostic applications, although it is currently hampered by a 10-15% error rate limiting its utility and preventing it from replacing Illumina technology[68]. Hybrid approaches have been tested where Oxford Nanopore sequencing is used in parallel with Illumina sequencing, in which the nanopore sequence acts as a reference "backbone" and the Illumina reads are used to identify mismatches and to "polish" the genome to increase accuracy, although this is both expensive and slow[74].

There are many advantages to using NGS technology for virus discovery. The largest and most obvious advantage is that the sheer amount of data generated is almost overwhelmingly large- often numbering billions of reads- which is frequently able to produce well supported, high copy number full genomes for phylogenetic analysis[65]. Metagenomics is also unbiased as it gives reads for all microbes (excluding those with very low copy numbers) and host cells within the sample, which allows for a true snapshot of the microbial community and allows for the detection of pathogens in diagnostic situations where there are no obvious candidates for testing[27,65]. There is also no requirement for prior knowledge of the sequence of the target organism, as specific primers or probes are not required, which is particularly useful in rare or understudied species'[24,25,61]. NGS can also be used to analyse virome changes over time in some circumstances, such as when monitoring the same patient or in livestock management, although this does require some previous knowledge of what would be considered a "normal baseline" virome and is not currently routinely performed[24]. Additionally, NGS experiments are logistically easier and safer than other culture-based virus discovery techniques[25].

There are also many disadvantages of NGS. One is that when screening pooled samples for virus discovery, it is impossible to know which individual sample(s) contained the virus in question, and thus any hits must then be validated by PCR[66]. Additionally, the computing power and bioinformatics expertise required to process a NGS dataset is still considerable, despite the increased availability of simplified pipelines and cloud-based high performance computing capacity, and even with significant bioinformatics skill it is still possible to mischaracterise an endogenous retrovirus as a novel virus in animals with poor genome annotation[25,72,75]. Another disadvantage is the relatively high cost of NGS, with a cost of approximately £5000 per commercial flow cell, although the cost per sample can be decreased by pooling samples into one reaction with unique barcodes at the expense of decreased reads per sample, and it should also be noted that NGS experiments are becoming more affordable over time[30,76]. Additionally, due to the requirement to match reads to reference sequences, there is the risk that truly novel or highly divergent viruses may

not match any reference sequences and may be lost as “dark matter”, although with the increase in data-driven virus discovery these viruses may be identified later[6,55,75]. Database and reference sequence quality may also be variable[59,75]. There is also an element of unavoidable skewing of a library, as some viruses may amplify preferentially, and others may not be purified or sequenced adequately[6]. When considered in comparison to PCR this further hampers the already lower sensitivity of NGS investigations, although this issue is often able to be overcome with sufficient reads per sample and NGS sensitivity is increasing over time[26,65]. Despite all of these disadvantages, the sheer volume of data provided by NGS and metagenomics studies and their ability to provide an in-depth analysis of the virome far outweigh their associated issues, provided a researcher or laboratory has the resources and expertise to handle the data[25,75].

3. Ancient RNA viruses and their discovery

a. Introduction to ancient and historic genomics

Whilst the word "ancient" is commonly considered in a prehistoric context, in a molecular context ancient DNA and RNA can simply be defined as DNA or RNA that persists and is informative for a prolonged time after organismal death[77]. Whilst this definition is simplistic and understandable, it fails to adequately distinguish between freshly killed animals and preserved specimens, although this is admittedly a difficult line to draw[77,78]. Whilst each group will draw their own line for specimen age required to be considered as "ancient" vs historic, our group considers any specimen that was collected prior to the year 2000 and has been preserved in some manner (i.e through formalin fixation, ethanol preservation, or extreme freezing) to contain historic nucleic acids, and any specimen which has a fossil origin or is likely to contain nucleic acids that are over 1000 years old as containing ancient nucleic acids[79]. Most of the following discussion is applicable to both ancient and historic genomics, and therefore distinctions will not be drawn in many cases.

Ancient DNA and ancient RNA (aDNA and aRNA, respectively) and their historic counterparts are found in a variety of specimens, including animals frozen in permafrost and their faeces, formalin fixed and frozen human lung tissue, fossils and extinct animals such as the Tasmanian tiger[80-83]. aRNA has been discovered in exceedingly old animal samples, with the oldest sequenced to date being approximately 14,000 years old, although the very oldest aRNA found was in ice cores estimated to be up to 140,000 years old[80,81]. An important source of aRNA is natural history collections, such as those housed at the Natural History Museum in the UK (NHM), the Smithsonian National Museum of Natural History in the USA and the Berlin Museum of Medical History in Germany[78,84]. Whilst aDNA is now relatively well understood and working with it is not too unusual, far less is known about aRNA, and working with aRNA is still considered to be both difficult and unusual[77,78]. Initial aDNA research was limited to mitochondrial DNA investigations, and only in the last two decades have ancient genomic investigations become viable due to advances in technology[85].

b. Technology and challenges

Most frequently, ancient genomics studies are performed using metagenomics technologies with modifications for ancient nucleic acids, such as additional bead clean up steps or alternative filtration criteria to retrieve smaller fragments[80,81]. Most investigations focus on aDNA as the processes for working on aDNA are significantly more developed than for aRNA, due in part to the perception of RNA as much more fragile than DNA[80,86]. Whilst this may be broadly true, there are some factors that can increase the preservation of RNA- for example, circular RNAs and RNA preserved in encapsidated viruses tend to be relatively well preserved[86]. For both aDNA and aRNA, the nucleic acids tend to be highly fragmented with fragment lengths generally around and often below approximately 100bp, which many metagenomics protocols tend to filter out by default[77,78,84,87]. Additionally, nucleic acid preservation quality

and fragmentation may vary both between specimens and tissues, and even within the same tissue within an individual specimen, although it has previously been found that liver tends to be a (relatively) well preserved tissue in most small animal specimens[83,87].

The nature of the sample itself and its preservation can also affect ancient genomic investigations. One major factor that can affect nucleic acid preservation is the method of preservation of the sample, whether chemical or temperature based[77,80]. Chemical fixatives such as formalin and ethanol effectively preserve nucleic acids, although formalin is known to damage the tissue and nucleic acid contained within during the preservation process[77,88]. However, working with formalin in particular presents safety risks and makes processing the tissue more difficult, whereas ethanol readily evaporates if not carefully monitored and topped up, potentially leading to drying out and damage to the specimen[77,83]. Additionally, some museum samples predate formalin fixation, and others were collected shortly after the beginning of the usage of formalin but before widespread use, leading to questions regarding the level of fixation and how best to process the tissue[83,84]. Another issue is that for natural history collections, whilst the date of entry to the collection is usually known, the exact date of animal death, time from death to fixation, transportation information including whether the specimen has ever dried out and other important metadata such as animal age at time of death is often missing, potentially complicating downstream analysis of this censored data[83].

Another major issue of ancient genomics work is the risk of contamination with modern nucleic acids. The primary route of contamination is from the user or the environment throughout every step of the process[82]. To try and avoid this, efforts such as deep cleaning laboratories before and after ancient genomics work and ideally working in a specific ancient genomics laboratory should be taken, although due to a collective lack of experience and specific and robust protocols for ancient genomics work across the scientific community this is often insufficient to prevent contamination entirely[83]. Another issue is potential contamination of reagents- for example, even a previously unopened kit can be contaminated, as is known to occur with standard filtration steps using silica matrices within some column filtration DNA extraction protocols[81]. Even with the most stringent control measures, it is never possible to rule out contamination with 100% certainty throughout ancient genomics work, so steps must be taken to assess contamination when analysing results[82]. A common method for this is to assess nucleic acid damage using software such as mapDamage 2.0, which uses the relatively predictable accumulation of nucleotide damage to assess whether a read is from an ancient sample or modern contamination[89].

Finally, the analysis of ancient sequences presents its own challenges. One of the most important is that most reference sequences and reference databases are based on modern genomes, and as a result, any significantly divergent ancient sequence data may not be identified as the animal or microbe in question, leading to the loss

of likely novel sample reads[82]. This is particularly true for rapidly evolving organisms such as many RNA viruses, or when screening extinct animal hosts[5,83]. Another is that any form of PCR screening of ancient samples is difficult, as the primers will likely be designed based on modern genomes, which are unreliable as described above[82]. This presents difficulties when attempting to validate the findings of ancient genomics studies, which are understandably subjected to a high level of skepticism and scrutiny as standard[84].

c. Virus discovery

Despite the technical challeneges, there have been examples of both DNA and RNA virus sequences isolated from ancient samples within the last 10 years[81,90]. One study found plant DNA and RNA viruses in frozen faeces of an estimated age of 700 years[81]. Ancient plant viruses have also been found in leaves preserved within herbarium sheets, where an RNA virus of the genus *Tobamovirus* was identified in an estimated 100-year-old specimen[86]. Fungal viruses have also been isolated from the skin and muscle tissue of an extinct Tasmanian tiger[83]. Recently, Rustrela virus was isolated from preserved lion tissue from the 1980s[91].

Human infecting viruses have also been found, primarily in frozen human remains and formalin fixed tissues[81,86]. The first human virus identified in this way was found in human remains frozen in permafrost and contained the 1918 pandemic strain of *influenza A virus*, which suggested that other viruses could be found in preserved human tissue[81,86,92]. The 1918 strain of *influenza A virus* has also been identified in human lung FFPE (formalin-fixed paraffin embedded) tissue in museum collections in German samples, as has the full genome of a measles virus isolate from German museum collection sample FFPE lung tissue from 1912[78,92]. Human parvovirus B19 has been isolated from Neolithic skeletal remains with an estimated age of between 500 and 7000 years old, with 63.9%-99.7% genome coverage, suggesting that complete or near complete virus genomes can be found in extremely old samples under the correct conditions[90]. Picornavirus or picornavirus-like sequences have been recently identified in the extinct Tasmanian tiger, although further research is required to provide more specific information on these sequences[83]. Finally, 2 different bacteriophages have also been found in archival bat samples from the 1800s-1900s within the Smithsonian Museum archives in New York[84]. Other non-viral pathogens have also been identified through ancient genomics, such as *Borellia burgoferi* in an ancient human sample[85].

d. Importance and potential application

There are a variety of applications for ancient genomics, ranging from virus discovery and characterisation through to using aRNA quantities to assess gene expression levels in tissues[83,90]. In both pathogens and mammals ancient genomics can be used to analyse genome changes over time, and to observe and characterise evolutionary history, including genomic sequencing of extinct species and their closest living relatives today[77,85]. From a virology perspective, this can allow for the rolling back of the molecular clock and the identification of common ancestors to modern viruses,

providing key information about their origin, evolution, and if there are any previously unnoticed evolutionary hotspots in the genome that may evolve and mutate in the future[77,81]. This also allows for improved phylogenetic analysis, including potential reclassification of viruses that share a newly discovered common ancestor, and potentially the identification of recombinant viral strains[81,90]. Museum collections in particular can be useful for providing a "snapshot in time" of the virome at different points throughout history, and for tracking changes within the virome, potentially including the generation of divergent strains, geographic expansion and even host switching events[77,90]. This can help to provide insight into key pathogens and their evolution over time, in turn enhancing our collective knowledge of these pathogens and improving pandemic preparedness[83].

4. Rodents and bats as virus host species

This project primarily considers the virome of rodents and bats, and an introduction to these host species is provided here.

a. Rodents

Rodents belong to the order *Rodentia*, which contains approximately 2300 species across 33 families and 43% of all mammalian species, making it the largest mammalian order[8,62]. Excluding Antarctica, rodents are found on all continents and in most environments in all countries, and are often widespread and present in large numbers in any given area[13]. Rodents possess many characteristics that make them effective virus hosts, including extremely high species diversity, shared and often sympatric habitats amongst large groups of rodents, and rapid generation times and short lifespans, as well as being evolutionarily ancient animals[7,93,94]. Accordingly, they are reservoir hosts for viruses of a minimum of 22 families and hundreds of species (over 500 of which were found as recently as 2023), including viruses such as LASV and CoVs, and are known to transmit some of these viruses to other species, including humans[7,29,40,45,95]. Indeed, other than bats, rodents host the most viruses that may be zoonotically transmitted per species of all mammals, although this may be in part due to research bias and disproportionate research effort into rodent viruses[7]. Often, infections by viruses transmitted from rodents to humans are mild and undetected, making it difficult to truly estimate the burden of zoonotic viruses from rodents due in part to a limited understanding of the rodent virome, with a large number of predicted "missing" rodent viruses remaining to be discovered worldwide[30,32]. Zoonotic transmission from rodents can occur in a variety of ways, including via the consumption of infected urine or meat, inhalation of aerosolised waste, directly through bites or scratches, or indirectly through a vector such as an insect, tick or flea[13,95,96]. The rodent species considered below are those that are of interest to this project, although it should be noted that a variety of other rodent species are relevant in this context.

Many viruses have previously been isolated from bank voles (*Myodes glareolus*), including CoVs, picornaviruses (including *Parechovirus B*), and *Anelloviruses*, amongst others[29,41,97,98]. Bank voles are also reservoirs for hantaviruses, including *Puumala virus*, which can be transmitted to humans[13,99]. Bank voles are widespread within Great Britain and Europe, suggesting a broad application to virome investigations into these animals[41,100]. Similarly, field voles (*Microtus agrestis*) have also been shown to be reservoirs for many viruses, including CoVs, TATV, *rotavirus A*, astroviruses, and picornaviruses, amongst others[12,29,62,97]. These viruses have been detected within the UK and across Europe, suggesting similar research value as for bank voles[12,62,64,101].

Yellow-necked mice (*Apodemus flavicollis*) are also found in the UK and are less well studied, but have been associated with CoVs, picornaviruses and paramyxoviruses, amongst others[41]. They are also believed to act as incidental hosts for *Tula orthohantavirus*, although further research is required into the viral profile of yellow-

necked mice and their association with hantaviruses[102]. Wood mice (*Apodemus sylvaticus*) are similarly found in the UK and understudied, but have been associated with *Anelloviruses* and picornaviruses, although it is likely that other viruses will be present within wood mice[41,98]. Finally, the African multimammate mouse (*Mastomys natalensis*) is the primary reservoir for LASV and is known to be responsible for the majority of zoonotic transmission to humans[45,96]. *M. natalensis* is frequently found in and around domestic dwellings, and is the most common rodent in Africa, which somewhat explains the frequency of zoonotic spillover into humans and high incidence of LASV[45].

b. Bats

The second most diverse group of mammals is bats of the order *Chiroptera*, which consist of approximately 1400 species and represent approximately 20% of mammalian species[22,103]. Much like rodents, bats are found on all continents bar Antarctica, but some regions such as Central and South America are home to significantly more bats[23]. Bats are also frequent and effective viral hosts, due to many traits shared with rodents including high species diversity and interspecies habitat sharing, but also due to characteristics not found in rodents such as long life spans for their body size allowing for chronic infections and living in close proximity in large and often sympatric groups, with up to 1 million bats sharing a single roost[7,35,93]. Indeed, bats host the most viruses per species of any mammalian order with a minimum of 58 viral families detected in bats[7,67]. Whilst bats are often considered to be effective viral hosts, some families of bats are hosts to a greater diversity of viruses than other families of bats, and this is often associated with increased host species diversity within the family and a broad distribution as demonstrated by the *Rhinolophidae*[104].

Spillover events from bats to humans cause significant disease outbreaks annually, and much like rodents it is believed that there are still many "missing viruses" circulating within bats which are yet to be discovered, particularly in South and Central America[7,29,32]. Bats are perhaps most importantly associated with CoVs, where they have been found to act as reservoirs for both α and β CoVs, including those with significant zoonotic potential and those closely related to SARS-CoV-2[13,25,29,67]. Another major virus associated with bats is *rabies lyssavirus*, where in some countries with effective dog vaccination programs bats are the primary cause of zoonotic rabies infections, although as previously discussed this is not the case throughout much of the world[26,67]. However, bats are also associated with a variety of other important viruses and virus families, including *Nipah virus*, hantaviruses, *Hepaciviruses* and *Reoviruses*, amongst others[12,23,105,106].

Whilst some viruses such as *rabies lyssavirus* are transmitted to humans directly from bats via bites, more commonly an intermediate host or insect vector is responsible for the zoonotic transmission, and larger animals such as primates can act as intermediate hosts for key viruses such as *Ebolaviruses*[23,25,38,67]. As is the case with rodents, land use changes and climate change are driving bats and humans into

closer proximity, increasing the zoonotic transmission risk. For example, the range of bat species *Pipistrellus kuhlii* has increased almost 4-fold in the last 40 years, and the frequency of these bats roosting near to humans has increased over the same time period, in a response that is believed to be linked to climate change and urbanisation via multi-temporal epidemiological modelling[107]. *Pipistrellus kuhlii* is known to be infected with viruses including the *Rhabdovirus Vaprio virus* and an α-CoV, suggesting that there may be a potential zoonotic transmission risk to humans and demonstrating an example of the importance of climate change when considering bat to human zoonotic transmission[108].

5. Key viruses

This section is designed to introduce viruses of relevance to this project. These are virus species that were found screened for by degenerate PCR (chapter 4), found in the metagenomics data produced (chapter 5) or screened for in historic specimens (chapter 6). Whilst each virus considered here is discussed to a reasonable depth, this section is not intended to act as an extensive literature review of the viruses considered here, nor to cover all potentially important viruses with zoonotic potential.

a. Adenovirus

Adenoviruses are members of the family *Adenoviridae* and are non-enveloped dsDNA viruses (Baltimore classification group I) with genomes ranging from 25-48kb, although typically around 36kb[109,110]. Traditionally, the adenovirus genome was believed to be stable and unlikely to recombine, although further research has shown that for some species of adenovirus this is not the case, with the majority of recently identified pathogenic adenoviruses being recombinant[111-113]. The adenovirus genome is complex, and typically contains approximately one ORF for each genomic kilobase, encoding several multi-functional structural proteins and some non-structural proteins[112]. Most adenoviruses use the coxsackie-adenovirus receptor, but some serotypes use other receptors such as CD46[114]. Adenoviruses infect a variety of hosts, including humans, rodents such as wood mice and shrews, bats, pigs, cats, dogs, non-human primates, reptiles such as lizards, fish, and birds, amongst others[115-119]. There are over 80 species of adenovirus, distributed amongst 6 genera, the largest and arguably most important of which is the *Mastadenovirus* genus which infects mammals including humans[109]. Adenoviruses are distributed globally and are typically transmitted via the respiratory or faecal-oral routes, although due to their ability to persist as fomites they can also be transmitted via touching contaminated materials[119,120]. Adenoviruses are typically considered to be relatively host-specific[110,115,121]. However, there is now significant evidence supporting zoonotic transmission from non-human primates and cats to humans, and reverse zoonotic transmission from humans to non-human primates and potentially to bats, as well as interspecies transmission amongst non-human mammalian species[119].

There are 7 species of adenovirus that infect humans, named *Adenovirus A-G*, and over 90 serotypes, most of which fall under the species *Adenovirus D*[112,120]. Some serotypes of *Adenovirus A, F* and *G* are known to cause gastroenteritis, some serotypes of *Adenovirus B, C* and *E* are known to cause respiratory infections (typically upper respiratory tract infections), and some serotypes of *Adenovirus B, D* and *E* are known to cause conjunctivitis[96,113,120]. Adenovirus infections in humans occur worldwide with no obvious seasonal pattern, and infections tend to be asymptomatic in otherwise healthy individuals but can cause severe respiratory or gastrointestinal disease in immunocompromised people and children, particularly when co-infection occurs with bacterial pathogens or other respiratory viruses[109,110,119,122,123]. Global prevalence across all serotypes is high, with an

estimated 20 million annual cases in the USA alone, but serotype prevalence and disease burden varies by location[112,120].

In terms of gastroenteritis, adenoviruses are the second most important viral pathogen worldwide (after *Rotaviruses*) and cause significant illness and mortality, and whilst adenovirus prevalence varies globally, it can reach up to 31% in children with gastroenteritis in LEDCs[122,124]. Whilst typically an estimated 3% of respiratory disease is caused by adenovirus infections, in outbreak situations the prevalence can reach up to 16% and transmission rates to close contacts can reach almost 50%, and whilst most respiratory adenovirus infections are mild and self-limiting, severe cases can lead to symptoms such as adenoviral pneumonia which can have a mortality rate of up to 30%[113,120,125]. Up to 90% of conjunctivitis cases can be attributed to adenovirus infections and most cases are mild and self-limiting, but severe cases can lead to significant complications including loss of vision[120]. Adenovirus infections are typically diagnosed by PCR (except for conjunctivitis, which is usually diagnosed clinically), but can also be diagnosed via enzyme immunoassays or viral culture, although differential diagnosis is often not performed due to the typically mild nature of the disease[120,122]. There are no specific antiviral adenovirus treatments available, and whilst vaccines against human adenovirus 4 and 7 are available for military personnel they are not typically administered to the general public[112,120].

Adenoviruses are promising tools for vaccine development, largely due to their large and well-understood genome allowing for a significant delivery capacity (up to 36kb in third generation vectors), which are also traits that may be useful for gene therapy[114,126]. The broad cell tropism of adenoviruses allows for a variety of target tissues, and they have been shown to elicit an effective innate and humoral immune response in clinical trials[114]. A variety of adenovirus-based vaccines are in clinical trials including those for HIV-1 and for SARS-CoV-2, and an adenovirus-based *Ebolavirus* vaccine has been approved for clinical use. There are issues with adenovirus-based vaccines, such as a seroprevalence of up to 80% for some serotypes limiting effectiveness, although attempts are being made to mitigate these issues via using recombinant non-human primate adenoviruses as vectors[114]. Adenoviruses are also under investigation as anti-cancer agents, both as anti-cancer vaccines and as therapeutic agents in combination with existing therapies[114,126].

b. Arenavirus

The family *Arenaviridae* consists of enveloped negative sense ssRNA viruses (Baltimore classification group V)[127,128]. Arenavirus genomes are approximately 10.5kb, and typically consists of a small (S) segment of approximately 3.5kb that encodes the nucleoprotein and glycoprotein precursor, and a large (L) segment of approximately 7.2kb that encodes the RdRp and matrix protein, and some (although not many) arenaviruses also encode a medium (M) segment[45,127]. Unusually for segmented viruses, there is very little evidence of arenavirus reassortment, although an error prone RdRp does lead to a high mutation rate and rapid strain generation[129,130]. There are over 40 species within four genera of arenavirus, namely

the *Hartminivirus, Antennavirus, Reptarenavirus* (which infects reptiles and snakes) and *Mammarenavirus*, which infects mammals, and unless otherwise stated will be the genus discussed here[127,131]. Arenaviruses can be broadly split into old world (OW) viruses, which circulate within Africa and Europe and include (LASV), *lymphocytic choriomeningitis virus* (LCMV) and lujo virus, and new world (NW) viruses which circulate within the Americas and include junin virus, machupo virus, *guanarito virus* and *chapare virus*, amongst others, although there are some arenaviruses found in fish that do not fit into either category[59,130]. NW viruses are then further divided into clades A-D, with viruses that infect and cause disease in humans found in clade B and very rarely in clade D[130]. Whilst *tacaribe virus* naturally infects bats, rodents act as the natural host for most arenaviruses, and zoonotic transmission to humans via inhalation of aerosolised droppings or the consumption of food contaminated by infected urine is common, and transmission from direct contact with rodents can occur but is much less common[44,99,130,132]. Human to human transmission occurs naturally in some New World *Arenaviruses* such as *machupo virus* and *chapare virus* and can occur rarely via contaminated organ transmission in LCMV[99,133,134].

LASV is the most clinically important arenavirus, and whilst the incidence of LASV is likely underreported, it is believed to infect up to 3 million individuals with approximately 10,000 deaths annually[44,132]. The largest burden of LASV is in Nigeria where it is endemic, but other West African countries such as Sierra Leone, Guinea and Ghana amongst others also report LASV infections, with up to 7 geographically distinct LASV lineages circulating in this region[44,45]. The main reservoir species for LASV is *M. natalensis*. Rodent infection appears to be persistent and asymptomatic, and transmission is primarily zoonotic although human to human transmission via direct contact of infected bodily fluids can occur. LASV diagnosis is difficult, due to symptomatic similarity to other diseases within the area, a long incubation period of up to 10 days, a lack of any form of reliable standardised diagnostic assay, and up to 80% of infections being asymptomatic or mild[45]. LASV causes Lassa fever, which tends to cause mild febrile illness, but can progress to severe symptoms including seizures, coma and death. Whilst Lassa fever generally has a low mortality rate of < 1%, in severe outbreaks a mortality rate of up to 70% has been recorded, and approximately 20% of cases in hospitalised patients are fatal[45]. There is currently no vaccine available for LASV prevention, and no specific antiviral LASV treatments are available, although intravenous Ribavirin administered shortly after contracting LASV has shown modest benefits[44,45,128].

LCMV is the only *Mammarenavirus* reported to be native to Europe to date, where the natural reservoir of *M. musculus* is widespread across mainland Europe[99]. The incidence and prevalence of LCMV is unknown, and most infections are either asymptomatic or cause a mild and self-resolving febrile illness, although fatalities can occur at a rate of < 1%[99,135]. LCMV is believed to be under-diagnosed, but when performed diagnosis is typically made serologically or via PCR[99,135]. Whilst naturally acquired LCMV is generally mild, LCMV acquired via organ transplant is usually fatal,

with a 90% mortality rate. Congenital LCMV is also severe and can cause a variety of neurological symptoms ranging from mild learning disabilities to severe mental retardation, epilepsy and ocular disorders[135]. Much like LASV, there is no vaccine approved for LCMV prevention, and no specific antiviral therapy is available, although off-label use of Ribavirin may show some limited benefit[128,135].

Lujo virus is unique amongst OW arenaviruses that can infect humans as it causes viral haemorrhagic fever (VHF) symptoms including vomiting, diarrhoea and haemodynamic collapse amongst others, with a very high mortality rate of up to 80% in some settings[129]. The rodent host of lujo virus is unknown, as is the incidence due to its rarity, and there are no specific treatments or vaccines available. To date, lujo virus has only been reported in Zambia and South Africa, and the rarity of lujo virus may lead to under diagnosis or misdiagnosis, although an accurate PCR assay is available for lujo virus diagnosis[129].

NW arenaviruses typically cause fewer infections than OW arenaviruses, but with much higher mortality[130]. NW arenaviruses tend to be relatively geographically constrained, so are typically only tested for in their endemic area by PCR[130,134]. Junin Virus, which causes Argentine haemorrhagic fever, is endemic within Argentina where it is transmitted zoonotically from its rodent host (primarily *Calomys musculinus*) to humans[131]. Prior to vaccine development, an estimated 500 cases per year were observed, which has since reduced to approximately 13 cases per year, with mortality due to VHF symptoms of up to 20%. There are no specific antiviral drugs against junin virus, but treatment with convalescent serum reduces mortality to approximately 1%, and an effective vaccine has been available since the 1990s[131]. Machupo virus causes Bolivian haemorrhagic fever, another VHF with approximately 25-35% mortality[133]. Machupo virus infections occur sporadically within Bolivia, and are transmitted via the rodent *Calomys callosus*, although human to human transmission has also been reported[130,133]. There are no specific antiviral therapies or vaccines available for machupo virus, although vaccine research is underway[133].

Guanarito virus is another VHF causing virus that causes Venezuelan haemorrhagic fever, with a mortality rate of approximately 23%[136]. It is transmitted zoonotically by the rodent reservoir *Zygodontomys brevicauda* primarily to agricultural workers, but this is an infrequent occurrence, and annual cases are estimated to be in the low hundreds. There are no specific antiviral treatments or vaccine available for *guanarito virus*, and treatment is supportive only[136]. Finally, *chapare virus* is an extremely rare arenavirus which is the causative agent of chapare haemorrhagic fever, with a mortality rate approaching 50%[134]. Cases occur sporadically within Bolivia, and there is evidence that *chapare virus* transmission is both zoonotic (possibly from the rodent *Oligoryzomys microtus*, but this has yet to be confirmed) and human to human, particularly within nosocomial settings. There are no specific antiviral treatments or vaccines against *chapare virus*, and misdiagnosis as dengue fever is common[134].

c. Arterivirus

Arteriviruses are positive sense ssRNA viruses (Baltimore classification group IV) within the family *Nidovirales*, and the so called "Picornavirus supergroup"[16,137,138]. They are small, enveloped viruses, with a monopartite genome of ~12.5-16kb[138,139]. The genome encodes 10-15 ORFs, including 2 major ORFs (ORF1a and ORF1b) and 8-13 minor ORFs, encoding a minimum of 10 proteins including 5 glycoproteins and one large polyprotein which is later cleaved to give the RdRp protein[138,140]. Macrophages are believed to be the primary target cell, although arteriviruses have also been detected in liver, lung, brain, spleen, and heart, and specific receptors for entry remain unknown[138,140].

The rapidly expanding *Arteriviridae* are currently split into 6 subfamilies, 13 genera (largely demarcated by host species), 11 subgenera and 23 species, and are believed to be globally distributed, including within the UK[137,140,141]. Prevalence is highly variable, and differs according to geographical location, virus and animal host[138,142]. A variety of host species can be infected with arteriviruses, including pigs, horses, rodents, non-human primates, hedgehogs, shrews and turtles[8,137,140,141,143]. Arteriviruses are currently believed to be the only RNA viruses that infect mammals that are not known to infect humans, either directly or indirectly[143]. Cross species transmission is believed to occur between small mammals and non-human primates, yet despite the taxonomic relatedness between non-human primates and humans and the opportunities for spillover, no zoonotic infection has ever been reported[137,141,143]. Some arteriviruses, such as *porcine reproductive and respiratory syndrome virus* (PRRSV) and equine arteritis virus (EAV) are considered important pathogens of livestock, whereas others such as *wobbly possum disease virus*, *hedgehog arterivirus 1* and simian haemorrhagic fever virus are considered to be pathogens of wildlife[142-144].

Arterivirus transmission is believed to be primarily sexual, although vertical transmission and aerosol transmission also occurs[142-144]. Arterivirus infection may be asymptomatic or may lead to severe and potentially fatal respiratory disease or haemorrhagic fever, and is associated with late-term abortion in pregnant livestock[138,142]. Infection may also become persistent, which can in turn cause animals to act as reservoirs for transmission- this is particularly an issue with stud stallions used in horse breeding and EAV transmission[139,144]. Arterivirus infections can be diagnosed through RT-PCR or through ELISA, although ELISA diagnosis gives no information on whether an animal is acutely infected, chronically infected, or has been transiently infected in the past[138,142]. Arterivirus diagnostics are often unreliable, and testing is often not performed at all in many settings[138,144]. No specific treatments are available, and most arteriviruses are not currently vaccine preventable, although vaccines are available for PRRSV and are in development for EAV[138].

PRRSV infects pigs, and is the most important livestock arterivirus, estimated to cost over $700 million annually within the USA alone[142]. Outbreaks can cause severe

disease and variant strain outbreaks can reach mortality rates approaching 100% in infected herds, particularly when co-infection with other pathogens occurs[138]. Prevalence is highly variable ranging from 18.5-53.8% seroprevalence in Nepal and Nigeria respectively, and appears to be more common in female pigs and older pigs[142]. Whilst limited vaccination options are available for PRRSV these tend to be strain specific and of limited use in outbreak situations, although improved vaccines are in development[138]. EAV is also an economically important Arterivirus that can cause severe disease and mortality in horses, and often causes chronic infections in stallions which enter a "carrier state" and spread EAV when mating[138,140,144]. EAV is distributed globally (excluding New Zealand, and seroprevalence varies significantly by breed of horse and geographic region, although tends to be ~20% across Europe[144]. No vaccines are available for EAV[138].

d. Astrovirus

Astroviruses are small, positive sense ssRNA viruses (Baltimore classification group IV) within the family *Astroviridae* that were first identified in 1975 and lack an envelope, with a genome of ~6.8-7.9kb which encodes a 5' UTR, 3 open reading frames, a 3'UTR and a poly-A tail[19,76,124]. Astroviruses have a highly error prone polymerase leading to a high mutation rate can readily recombine (including across species barriers) due to the circulation of multiple strains, further increasing diversity and evolution[19,48,51]. Astroviruses are globally distributed and infect a variety of species, including rodents such as field voles, bats, companion animals and humans, with prevalence estimates in rodents of ~15%[31,51,62]. Due to their high mutation rate and variety of host species astrovirus spillover is believed to be relatively frequent, largely due to their variety of hosts and host species proximity[19,51]. There is no clear major reservoir species for astroviruses, making evolutionary and spillover patterns difficult to elucidate[19]. *Astroviridae* are split into the genera *Mamastroviruses* and *Avastroviruses* traditionally believed to infect mammals and birds respectively, although recently spillover between the two genera has been observed in both directions[51]. Until recently, only 8 human astroviruses were known, although recent studies have identified more species and strains of human and animal astroviruses[19,145].

Astrovirus prevalence and incidence worldwide is difficult to estimate. An outdated estimate of global incidence is ~11% although this varies significantly by region, population density, season, and urban or rural environment, with estimates ranging from 2-42% and seroprevalence studies suggesting a much higher rate of exposure of up to 65%[76,124,145-147]. Astroviruses are transmitted via the faecal-oral route and are estimated to cause 2-9% of non-bacterial acute gastrointestinal disease annually, with co-infection between astroviruses and other viruses of gastrointestinal significance being relatively common[124,146,148]. Astrovirus infections are commonly asymptomatic and tend to only be of clinical significance in children, the elderly and the immunocompromised, with children under 2 years of age bearing most of the disease burden[146,147]. Astrovirus infections are usually either asymptomatic or cause

mild gastroenteritis with diarrhea that self-resolves following an incubation period of 4-5 days, but in severe cases symptoms can include wasting, encephalitis, meningitis and rarely death[76,148]. Traditional astrovirus diagnosis involved electron microscopy or immunoassays, although these methods are difficult and have largely been superseded by PCR and qPCR based molecular methods or metagenomics analysis in severe cases[76,145]. There are no specific treatments for astrovirus infection and management is generally limited to rehydration via the provision of fluids and electrolytes to combat the effects of diarrhea. There is some evidence that the antiviral Nitazoxanide may have some benefit if provided quickly, although further studies and clinical trials will be needed to confirm this and to elucidate its mechanism of action as an antiviral compound[148].

There are still large knowledge gaps regarding astroviruses. Due to the difficulty of using cell-culture systems to test astrovirus infections and characteristics and the lack of any easily usable animal system other than turkey poults, studies are difficult to perform[76,148]. As a result little is known about the evolutionary history and genome composition of astroviruses, and the astrovirus receptor remains unknown[19,51,148]. With the advent of new technologies and increased usage of NGS, some of the knowledge gaps are likely to close as research continues[145].

e. Coronavirus

Coronaviruses (CoVs) are enveloped viruses with positive-sense ssRNA genomes (Baltimore classification group IV) within the family *Coronaviridae* and order *Nidovirales*[149,150]. *Coronaviridae* are split into two subfamilies, namely *Orthocoronavirinae* which contains all coronaviruses that infect mammals, and *Letovirinae*, which contains one genus that infects frogs[13]. *Orthocoronavirinae* are then split into 4 genera, including *alpha-coronavirus* and *beta-coronavirus*, which infect mammals, and *gamma-coronavirus* and *delta-coronavirus*, which primarily infect birds but can also rarely infect mammals[150,151]. The genus *beta-coronavirus* contains 4 subgenera, including *Merbecovirus, Nobecovirus, Sarbecovirus* and *Embecovirus*, and across all genera over 50 CoV species are known[149,152]. Genomes range from 22-36kb and tend to be of approximately similar sizes within a genus, with an average size of approximately 30kb[149,151,153]. The CoV genome is monocistronic and contains two major open reading frames and several smaller minor open reading frames, encoding 3 structural and typically 5 non-structural proteins, and replication is cytoplasmic[149,154,155]. Due to a relatively inaccurate polymerase CoVs undergo rapid recombination, leading to the rapid generation of new strains[47,156,157].

Many species are natural hosts for CoVs, including humans, rodents including bank voles, bats, cats, birds, camels, and marine animals amongst others, and cross-species transmission has been reported, with bats and rats believed to act as the reservoir species for most α and β CoV species[14,15,29,103,150,157]. CoVs are ubiquitous and globally distributed in humans and likely in rodents, although limited research is available on the prevalence and distribution of rodent CoVs[13,152,157,158]. 7 CoVs are

known to be pathogenic in humans- namely the four seasonal CoVs (two α-CoVs 229E and NL63 and two β-CoVs HKU1 and OC43), the three pandemic coronaviruses SARS-CoV (severe acute respiratory syndrome coronavirus), MERS-CoV (Middle East respiratory syndrome coronavirus) and SARS-CoV-2, but rarely canine coronavirus and porcine coronavirus HKU15 can also cause infections in humans[149,158]. It is believed that CoVs 229E and NL63 were originally transmitted to humans from bats, and that CoVs OC43 and HKU-1 were transmitted to humans from rodents[159]. Whilst the impact of the SARS-CoV-2 pandemic cannot be overstated, it is beyond the scope of this work to provide any more than a surface level overview of the unprecedented impact of this virus.

All four seasonal CoVs are globally distributed, and typically cause either asymptomatic infections or mild cold and flu-like symptoms in humans via an upper respiratory tract infection[158]. Whether there is a zoonotic element to seasonal CoV infections is unknown, but it is possible that bats may act as potential reservoir hosts, as CoVs with approximately 80% nucleotide similarity to the seasonal coronaviruses NL63 and 229E have been identified in bats via PCR[159]. Despite spanning two genera and using different receptors, the pathogenesis and progression of all four seasonal CoVs is similar, and all four are transmitted via the respiratory route and are more prevalent in children[123]. When diagnosed, seasonal CoVs are diagnosed via RT-PCR, but due to the mild nature of the symptoms of these diseases they are often not diagnosed, in turn limiting epidemiological understanding of these viruses. However, it is estimated that 15-30% of human respiratory infections are caused by seasonal CoVs, with OC43 causing the most infections[123]. These viruses show a seasonal pattern within the northern hemisphere, with infection peaks in the winter and particularly February (although this seasonality is not observed within the tropics), and demonstrate an alternating biennial pattern, where the seasonal α-CoVs (NL63 and 229E) peak one year, and the seasonal β-CoVs (OC43 and HKU1) peak the following year[158,161,162]. The reason for the alternating pattern is not fully understood and may have been disrupted by the lockdowns and health interventions associated with the COVID-19 pandemic- further research is required to discern whether this pattern rematerialises in the coming years[161,163].

The first pandemic CoV was SARS-CoV, a β-CoV first identified in 2002 in China[154]. Approximately 8000-8500 people were infected with SARS-CoV, and whilst most had self-resolving flu-like symptoms, approximately 25% of those infected suffered from respiratory failure with an approximately 10% mortality rate. SARS-CoV was transmitted via the respiratory and possibly the faecal-oral route and was infamous for rapid transmission[154]. SARS-CoV was zoonotic, and whilst it is unknown whether palm civets were the reservoir host, they as a minimum acted as an intermediate host for transmission to humans[151]. There are no specific antiviral treatments or vaccines available for SARS-CoV, and no cases have been reported since 2003[154].

MERS-CoV was the second pandemic CoV and is a β -CoV that was first identified in 2012[15]. Whilst the peak of infections occurred directly after the discovery of MERS-

CoV in the Middle East some cases are still reported today[15]. It is believed that zoonotic spillover occurred from camels, although it is unclear whether camels are reservoir hosts or whether other species such as rodents or bats act as the true reservoir[15,164]. Approximately 1400 cases of MERS-CoV infection have been confirmed of which approximately 80% occurred in Saudi Arabia, and transmission occurs via the respiratory route allowing for rapid viral spread[154]. Whilst symptomatically similar to SARS-CoV, MERS-CoV infection has a much higher mortality rate of approximately 35%, although in some settings this reached up to 60%, particularly in people with underlying health conditions. There are currently no effective antiviral treatments or vaccines for MERS-CoV, although many vaccines are in development and clinical trials[154].

SARS-CoV-2, the causative agent of COVID-19, is responsible for the largest viral pandemic of the 21st century. Whilst it is difficult to truly estimate the number of SARS-CoV-2 infections due to the frequent asymptomatic transmission, lack of testing since the peak of the pandemic and deliberate misinformation campaigns, at the time of writing this is estimated to be over 700 million with almost 7 million deaths worldwide, and these values are constantly increasing[165]. The SARS-CoV-2 pandemic has also modified the epidemiology of other respiratory viruses such as influenza and seasonal CoVs, largely due to barrier interventions and lockdowns reducing their transmission. Whether this modified epidemiology persists remains to be seen[163]. Whilst the natural host is of SARS-CoV-2 is unknown, it is believed to most likely be a zoonotic disease of bat origin due to the genomic similarity of SARS-CoV-2 and other CoVs found in bats (such as the beta-CoV RaTG13 found in *Rhinolophus affinis*), possibly spilling over to humans via an intermediate host[14,166].

SARS-CoV-2 is transmitted via the respiratory route and typically causes mild or asymptomatic disease, although approximately 1.3% of cases are fatal with an increased mortality rate in the immunocompromised and elderly[154,165]. Amongst the CoVs, SARS-CoV-2 is unique in that following symptomatic infection post-acute COVID-19 syndrome or "PACS" (also known as long COVID) occurs in an estimated 35% of patients, reaching up to 87% in hospitalised patients[165]. PACS is defined by the WHO as having at least one persistent symptom following probable or confirmed SARS-CoV-2 infection presenting within 4 months of infection and persisting for 2 or more months that cannot be otherwise explained by an alternative diagnosis. Common symptoms include fatigue, memory problems, dyspnea, joint pain and sleep problems[165]. SARS-CoV-2 diagnosis is typically made via RT-PCR, and accurate assays can reliably detect SARS-CoV-2 infections in saliva samples[167]. Initial treatment for SARS-CoV-2 infection was purely supportive, but recently a variety of antiviral drugs have been shown to be effective for SARS-CoV-2 treatment [154].

A variety of safe and effective vaccinations are now available for SARS-CoV-2, and whilst efficacy varies slightly by vaccine brand and demographic they all provide effective prevention in all demographics against the progression to severe COVID-19 disease and death, including against variants of concern[168]. Whilst these vaccines are

effective after one dose, subsequent doses increase protection and reduce the rate of waning immunity[168]. Although there are some side-effects associated with SARS-CoV-2 vaccines such as a low risk of Guillain-Barre syndrome and myocarditis, these side-effects are less likely following vaccination than from natural SARS-CoV-2 infection, and the vaccines are safe for pregnant people[168]. It is currently unclear whether the vaccines provide any protection from the development of PACS[165].

f. *Cytomegalovirus*

The genus *Cytomegalovirus* (CMV) belongs to the subfamily *betaherpesvirinae* within the family *Herpesviridae* and consists of enveloped viruses with extremely large dsDNA genomes (Baltimore classification group I) of 230-240kb[169,170]. CMV infections are largely considered in the context of humans (HCMV) but can also infect non-human primates and rodents including rats, mice and guinea pigs, with seemingly no capacity for cross-species transmission[170]. HCMV infection is distributed globally, with approximately 50% of individuals in higher income countries and 90% of individuals in LEDCs proving seropositive and harbouring a latent infection[171]. Infection is typically asymptomatic but can cause disease in the immunocompromised including transplant recipients, and is particularly harmful to newborns and neonates that are infected during pregnancy[171,172]. Transmission occurs directly via saliva and urine and sexually, and vertical transmission can also occur[172]. Diagnosis is typically made via RT-PCR of saliva or urine but is often not performed except for when newborns are symptomatic[172]. Congenital CMV infections are the most common and important CMV infections and are the leading cause of non-genetic hearing loss in the world, although other potentially serious symptoms can also occur in babies including seizures and microcephaly[171,172]. Children can be treated with a 6-month course of ganciclovir which shows moderate benefit in children with hearing loss, but no effective antiviral treatments are available for adults[170-172]. Despite decades of vaccine development no vaccines are currently available for prevention of CMV infection[170,171].

g. Hantavirus

Hantaviruses, of the family *Hantaviridae* within the order *Bunyavirales*, are enveloped viruses with a tripartite negative sense ssRNA genome (Baltimore classification group V)[11,12,160]. There are four genera of hantavirus, specifically *Orthohantavirus*, *Mobatvirus*, *Thottimvirus* and *Loanvirus*, and over 50 known hantavirus species of which at least 22 infect humans and are typically found in the *Orthohantavirus* genus[13,173]. Unless otherwise stated this work will be considering the *Orthohantavirus* genus. The hantavirus genome consists of three segments: a small (S) segment, which encodes the nucleocapsid protein and is typically ~1.8kb, a medium (M) segment, which encodes two glycoproteins and is typically ~3.5kb, and a large (L) segment, which encodes the RdRp and is typically ~6kb but can reach up to 12kb[13,63]. Hantaviruses are distributed globally but functionally they can be further divided into OW hantaviruses that circulate within Asia and Europe (including the UK), and NW hantaviruses which circulate within the Americas, although

hantaviruses have also rarely been detected in Africa[12,64,173-175]. Rodents are the primary hosts for hantaviruses, including field voles and common voles for OW hantaviruses and pygmy rice rats for NW hantaviruses, although they have also been detected in bats and can spillover into humans where they can cause severe disease[12,33,63,176]. Zoonotic transmission to humans is believed to primarily occur through the inhalation of aerosolised virus via rodent faeces or urine, although in the specific case of the NW hantavirus *Andes orthohantavirus* horizontal transmission amongst humans has been observed[176-179].

OW hantaviruses are found in Europe and Asia and can cause haemorrhagic fever with renal syndrome (HFRS) or a milder form entitled nepropathia epidemica (NE) in humans[33,160]. It is difficult to accurately predict OW hantavirus incidence due to the relative frequency of asymptomatic infection, but some estimates reach up to 200,000 cases per year with a mortality rate ranging from 1-15% with up to 90% of infections and fatalities occurring in China[30,173,180]. Important OW hantaviruses include *Hantaan orthohantavirus* and *Seoul orthohantavirus*, which primarily circulate in Asia and can cause HFRS, and *Puumala orthohantavirus* and *Dobrava-Belgrade orthohantavirus*, which typically circulate in Europe and are more likely to cause NE[180]. Another OW hantavirus is *Tula orthohantavirus* which circulates in Europe and is typically considered to be non-pathogenic, although there is a single report of severe infection in an immunocompromised host[101,181].

HFRS symptoms include febrile and gastrointestinal symptoms, dizziness and rarely potentially fatal renal failure[13]. Diagnosis is made by RT-PCR, but due to the rare and often asymptomatic or mild nature of OW hantavirus infections are often underdiagnosed[13,64]. There is no specific antiviral treatment for HFRS and vaccines against OW hantavirus infections are only available in China and South Korea, where their efficacy is unclear[125,180].

NW hantaviruses are found in the Americas and have rarely been reported elsewhere such as in Europe due to imported cases[176]. Human infections with NW hantaviruses are rare but severe, with an estimated 4000 cases reported in South America and likely fewer in North America, and a seroprevalence typically ranging from 0.5-6% with mortality rates reaching up to 50%[33,176]. NW hantavirus infection is heavily linked to rodent density and factors such as deforestation that drive rodents into closer proximity with humans such as extreme weather events are NW hantavirus risk factors associated with increased infections[42,177]. NW hantavirus infections can cause a clinical syndrome known as hantavirus cardiopulmonary syndrome (HCPS, also known as Hantavirus pulmonary syndrome, HPS), and key NW hantavirus species include *Sin Nombre orthohantavirus*, which primarily causes HCPS in North America, and *Andes orthohantavirus*, which primarily causes HCPS in South America, which combined cause most HCPS infections[33,63,173,180]. HCPS infection can have a long incubation period of up to 54 days followed by mild febrile disease that often progresses to severe disease including symptoms including haemorrhage, sepsis and organ failure potentially leading to death[176]. Diagnosis is typically made through RT-

PCR, and there are no specific antiviral treatments available for HCPS. Treatment is purely supportive and palliative, although a variety of antiviral treatments are under investigation[176,177,180]. Similarly, there are no approved vaccines for the prevention of HCPS, but many candidates are undergoing clinical trials[180].

h. *Hepacivirus*

Hepaciviruses are enveloped positive sense ssRNA viruses (Baltimore classification group IV) belonging to the *Flaviviridae* family[100,182,183]. In 2011, 2 species belonged to the *Hepacivirus* genus, but since the increased adoption of NGS technologies and screening of other species, the genus has expanded significantly and currently contains 14 species which can then often be subdivided into several genotypes[100,184]. *Hepaciviruses* replicate within the cytoplasm and contain an 8.9-10.5kb genome which is translated as a single polyprotein and then cleaved into three structural and seven non-structural proteins[182,185,186]. The *Hepacivirus* genus and the *Pegivirus* genus are phylogenetically similar to the point that some unclassified viruses are described as "Hepegivirus", although further genomic sequencing and phylogenetic analysis will likely assign these viruses to either genus[105,185,187].

Hepaciviruses are found in a variety of hosts, including humans, horses, bats, rodents including bank voles, dogs, cattle and others, and tend to be highly species restricted with limited if any zoonotic transmission, although it is believed that *Hepacivirus C* (HCV) was transmitted originally into humans via zoonotic transmission from a reservoir species suspected to be horses[105,143,183,187]. It is believed that *Hepaciviruses* have evolved with their hosts, which may be partially responsible for limiting zoonotic transmission[186]. Some *Hepaciviruses* such as HCV are globally distributed whereas other *Hepacivirus* species have a much less widespread distribution- for example, to date *Hepacivirus F* and *J* are found exclusively in mainland Europe[100,183,188]. The prevalence of most *Hepaciviruses* are difficult to estimate due to their wild animal hosts and limited sampling[187]. The route of transmission for most *Hepaciviruses* is unknown, although *equine hepacivirus* is known to be transmitted vertically in some cases, and other *Hepaciviruses* are suspected to possibly be transmitted via insect vectors[100,186]. *Hepacivirus* viral loads are highest in the liver in most infections, although virion can also be found in blood and other organs[183].

HCV is the species of causative agent of hepatitis C and is also known as *Hepatitis C virus*, and is one of the most important pathogens worldwide[105]. The incidence of HCV infection is high with estimates of up to 3 million new acute infections and 1 million deaths annually excluding those from hepatocellular carcinoma and up to 160 million people living with chronic hepatitis C infection, although accurate incidence and prevalence data is often lacking in many environments[187,188]. With 8 different genotypes and 86 subtypes that tend to largely be confined to specific geographic regions, HCV is globally but unevenly distributed, with Pakistan, Nigeria, Egypt, Russia, China and India accounting for > 50% of all deaths, and a larger burden in LEDCs than MEDCs[187,188]. HCV transmission is primarily through percutaneous exposure, largely via needle sharing, blood transfusion and blood products, and

intravenous injections of drugs, although vertical transmission and sexual transmission can also occur[188].

Hepatitis C infection is often asymptomatic leading to significant underdiagnosis, but may also cause acute or chronic disease, which can in turn cause cirrhosis, liver failure and frequently fatal hepatocellular carcinoma. Diagnosis is typically made through HCV specific RT-PCR, although serological methods can also be used for disease monitoring during treatment[188]. There are many anti-HCV treatments available and whilst the specific combination of drugs used to treat patients varies by location, HCV genotype and patient demographic, treatment can be curative and is near 100% effective in most cases, although the severe side effects and cost can be prohibitive in some settings[183,188]. No HCV is vaccine is currently available although promising vaccine candidates are in development, with one candidate in phase II clinical trials[183,188].

Rodent *Hepaciviruses* include *Hepacivirus* species *E-J* with a variety of rodent hosts including bank voles, South African four-striped mice and brush-tailed possums[105]. Of these rodent *Hepaciviruses*, *Hepacivirus F* and *Hepacivirus J* are particularly interesting as these may be beneficial as an effective small animal model for HCV pathogenesis[100,183]. Previously known as *bank vole hepacivirus 1* and *2* respectively, *Hepacivirus F* and *J* are common in bank voles within mainland Europe with prevalence estimates of 10.99% and 17.79% respectively[100,105]. In laboratory infected animals, liver pathogenesis and progression in *Hepacivirus F* and *J* infected bank voles was similar to that of HCV pathogenesis in humans, and since bank voles are effective research animals they have been suggested as HCV small animal model species[183].

i. *Murine leukemia virus*

Murine leukemia virus (MLV) is a member of the genus *gamma-retrovirus* within the family *Retroviridae*[189,190]. MLV is an enveloped virus with a positive sense ssRNA genome of approximately 8-9kb which forms a DNA intermediate as part of its replication cycle (Baltimore classification group VI) and is often considered to be a "simple" retrovirus as the genome encodes only the *gag*, *pol* and *env* genes without accessory proteins[19,191]. Like all retroviruses, MLV is known to integrate into the host genome at random sites with (usually) no negative impacts on the infected cell, at which point it is called a provirus and becomes heritable if integrated into germline cells[189,191-193]. Retroviruses and integrated proviruses (also known as endogenous retroviruses) evolve over time due to transcription errors from an inaccurate polymerase and can be used to track changes in the host genome and for molecular dating in paleogenomics[59,194]. However, this integration renders it difficult to distinguish between endogenous retroviruses and infectious retroviruses when analysing metagenomics data[75].

Whilst retroviruses have been identified in all vertebrate classes, MLV has been shown to be effective at host-switching within the same host order but not at any more phylogenetically divergent levels such as host class[193]. MLV strains can be

considered ecotropic (only infects rodent cells), xenotropic (only infects cells from hosts other than mice or rats) or amphotropic (infects both mice and other host cells)[194,195]. MLV has also been shown to undergo recombination- indeed, the accidental recombination of two laboratory strains of MLV led to the identification of xenotropic murine leukemia virus-related virus (XMRV)[194]. XMRV was falsely linked to prostate cancer and chronic fatigue syndrome due to its ability to infect human cells under cell culture conditions, prior to approximately 5 years of research attempting to identify the origin and association of XMRV disproving this link[194]. MLV has been found globally in many species of both laboratory and wild rodents, where some strains can cause lymphoma and neurological deficits[189,195,196]. Similarly, individual strains of MLV tend to be relatively geographically isolated and some are more common than others, with MLV-E strains appearing to be the most common[196]. Like other retroviruses, MLV is transmitted vertically and horizontally via several routes, including salivary transmission, via bloodborne transmission and via sexual transmission[190].

MLV has been an effective tool in cell culture investigations due to its simplicity, capacity to infect human cells, and ability to integrate into the host genome and still cause a productive infection[189,191]. MLV genomes are also amenable to pseudotyping, where the envelope glycoprotein of MLV is substituted for an envelope protein of a different virus of interest, in turn allowing for cellular entry and expression of this glycoprotein. Whilst an effective research tool, MLV is not an effective basis for prophylactic or therapeutic vaccines in humans due to a potential oncogenic effect upon genomic integration and the need to infect actively replicating cells to produce new virions[191].

j. Orbivirus

Orbiviruses are a genus of non-enveloped dsRNA viruses (Baltimore classification group III) within the family *Reoviridae* and subfamily *Sedoroviriae*, with a genome consisting of 10-12 genomic segments with sizes ranging from ~3kb to ~0.8kb, resulting in an overall genome size of ~18-20kb[106,197-199]. These genomes are often split into large, medium and small segments, and typically encode 7 structural and 3-5 non-structural proteins, although the exact number varies by *Orbivirus* species[197]. There is significant variability within some genomic segments, and this variability along with frequent reassortment by nature of the multi-partite genome has led to many circulating species and subtypes of *Orbivirus*. To date, at least 22 distinct species and 160 subspecies of *Orbivirus* have been identified[197-199].

Orbiviruses are arboviruses that replicate in and are transmitted by a variety of vector species, primarily biting midges of the *Culicodies* genus and ticks[106,198]. *Orbiviruses* are broadly associated with tropical regions, although they have also been reported in temperate regions and desert regions and have been reported in most of the inhabited world excluding northern Europe, Asia and Canada[199-201]. Prevalence varies significantly by season due to vector populations and activity, which also varies geographically by region and largely follows vector range, and is expected to increase

as climate change expands vector ranges[200-202]. *Orbiviruses* infect a variety of host species, including deer, cattle, rodents, bats, horses, dogs and very rarely humans as incidental hosts[106,197,202]. Within ruminants, vertical transmission is relatively common, and rarely horizontal transmission via sexual transmission, sharing food substances or blood products can occur[202]. Important *Orbiviruses* include BTV, *epizootic haemorrhagic disease virus* (EHDV), *African horse sickness virus* (AHSV), as well as the rarely human infecting viruses of the Kemorovo complex, *lebombo virus, changuinola virus* and *orungo virus*[106,198].

Human *Orbiviruses* are extremely rare with less than 200 confirmed cases worldwide, and no human fatalities have ever been associated with *Orbiviruses*[106]. Symptoms for all types of human *Orbivirus* infection are febrile in nature, and Kemerovo complex viruses can also lead to vomiting and abdominal pain, *orungo virus* can cause headache and encephalitis, and *lebombo virus* and the single reported case of *changuinola virus* produced no specific symptoms[106,198]. There has also been one incident of AHSV in humans in 1989, where veterinary workers were accidentally exposed to virus aerosols and developed uveochoriretinitis and three of four individuals developed frontotemporal encephalitis, although there is no record of natural or zoonotic AHSV infection in humans[106]. Diagnosis is made serologically, and population studies in the Czech Republic and Sub-Saharan Africa have showed seroprevalence of 18% and up to 35% for Kemorovo complex viruses and *orungo virus* respectively, suggesting a possible underreporting of human *Orbivirus* infections, possibly due to non-specific symptoms and a lack of diagnostic investigations. No human *Orbivirus* treatments or vaccines are available[106].

The most important *Orbivirus* is BTV which infects many wild and livestock animals and is estimated to cause an economic loss of at least $3 billion annually, as well as causing severe disease and suffering in animals[197,202]. Sheep are the most commonly and severely infected animals, although cattle and deer can also become infected with BTV[198]. BTV in primarily transmitted by bites of *Culicoides* flies, although vertical transmission occurs with variable frequency based on isolate, and tick bite transmission can also occur[198,202]. Infections may be asymptomatic, but symptoms can include oedema, haemhorrage and a characteristic cyanosis of the tongue, and in severe cases can lead to respiratory distress and death[198,202]. Even non-fatal infections tend to lead toward reduced productivity within the animal and can cause spontaneous abortion in pregnant ruminants[202]. Locally enzootic strains of BTV tend to cause mild disease whilst incursive strains tend to cause severe disease in a given area, although the reasons for this difference remain unknown[198]. BTV vaccines are available but tend to be serotype specific live-attenuated vaccines with limited cross-protection[202]. The multipartite nature of the live attenuated vaccines also presents a risk of recombination with circulating natural strains, potentially resulting in the generation of new infectious strains. New recombinant and vector-based vaccines are currently under development to attempt to mitigate these risks and aim to prevent vertical transmission[202].

EHDV is another important *Orbivirus* that infects key livestock species such as cattle and white-tailed deer, although it is difficult to accurately predict the economic impact of EHDV due to frequent misdiagnosis of EDHV as BTV due to similar clinical presentation[200,201]. EHDV has a similar distribution to BTV, largely due to sharing vector species in the *Culicoides* flies, although EHDV has a larger impact in the North American deer farming industry[200,202]. EHDV infections in cattle have become increasingly common throughout the 21st century, and whilst infections are often asymptomatic or mild they can lead to haemorrhagic disease and death in approximately 20% of wild animals or up to 80% of livestock in severe outbreaks. Diagnosis of EHDV is usually made through qPCR, and serotype specific vaccines are available, although vector management strategies such as removing breeding sites are often considered preferable to vaccination[200].

Whilst the importance of *Orbiviruses* is clear, there are significant issues in the investigation of and knowledge gaps regarding *Orbvivirus* infection. For example, the role of specific genome segments and protein functions are often still unknown, as is the full spectrum of vector species and mechanism of vertical transmission[198]. Various cell lines are used in studying *Orbiviruses in vitro*, and small animal models are available for *Orbivirus* investigations[198].

k. Paramyxovirus

The *Paramyxoviridae* is the second largest family of negative sense ssRNA viruses (Baltimore classification group V), and is comprised of four subfamilies, 17 genera and 86 species, although this family is rapidly expanding[203,204]. Paramyxoviruses are enveloped monopartite viruses with genomes of approximately 15-21kb that encode 6 major proteins[93,203,205]. Due to the highly specific nature of paramyxovirus replication, all paramyxovirus genomes adhere to the "rule of six", which states that the genome length must be exactly a multiple of six to allow for efficient viral replication[206]. Most paramyxoviruses use one of three receptors for cell entry, specifically ephrin B2 and ephrin B3 for *Henipaviruses* and glycoproteins such as sialic acid for other paramyxoviruses, which allows them to enter into target cells which are generally lung epithelial, endothelial and neuronal cells[93,207].

Paramyxoviridae can be split into four subfamilies of which the *Orthoparamyxovirinae* and *Avulavirinae* are the most significant due to their capacity for causing diseases in humans and key animal species[93,203,208]. *Orthoparamyxovirinae* can infect a variety of host species, including but not limited to humans, bats, pigs, horses, rodents including bank voles and field voles, dogs and cats[203,205,207,207]. Prevalence estimates and host geographic range are difficult to quantify for most of these viruses, although it is believed that due to frequent asymptomatic transmission, limited surveillance and globalisation that prevalence and host range are significantly underestimated for the majority of paramyxovirus[93,204]. Paramyxoviruses frequently spillover from reservoir hosts (generally bats or rodents) into other hosts, likely due their use of highly conserved receptors, often causing severe disease in humans and key animal species[93,209].

Important human and animal pathogens within the *Paramyxoviridae* include *hendra* and *nipah viruses* of the *Henipah* genus, measles, mumps, and Newcastle Disease virus, amongst others[203,208].

Measles virus is believed to be the most contagious pathogen on Earth, with a basic reproduction number of 12-18, and prior to vaccination programs caused approximately 135 million cases and 6 million deaths annually[210]. Measles infection is usually mild and self-limiting, typically causing a rash and febrile symptoms, but in severe cases could lead to pneumonia and fatal subacute sclerosing panencephalitis[210]. Despite only being able to infect humans measles was distributed globally, but successful eradication of measles has been achieved in 83 countries, changing the distribution of measles to be more prevalent in LEDCs and those with less successful vaccination campaigns[210]. Since the introduction of the measles mumps rubella (MMR) vaccine and other measles vaccines, measles infections and deaths have decreased by an estimated 73%, and whilst issues such as vaccine cost and distribution and vaccine misinformation have hampered the success of vaccination campaigns the elimination of measles is possible[210,211].

Mumps virus is a paramyxovirus that causes inflammation and swelling of glands, and in severe cases can lead to meningitis, encephalitis, oophoritis and orchitis amongst other symptoms[211]. Mumps virus is considered a pathogen of humans (although there is some evidence that bats may also be infected), and prior to vaccination an estimated 0.1-1% of the global population contracted mumps annually following a global distribution pattern. As of 2016 121 countries had introduced mumps virus vaccinations resulting in up to 97% decrease in incidence, although this varied by location, and much of Africa and Asia still has limited vaccination coverage and severe Mumps outbreaks[211].

Nipah virus is a zoonotic Paramyxovirus with important health impacts on both humans and livestock, and is most common in South-East Asia[93,207]. The natural reservoir for *nipah virus* is *Pteropus* bats, where the virus is shed in the faeces, saliva and is contained in blood and other tissue[207]. Pigs may consume infected bat or pig faeces or contaminated fruit and may develop febrile and potentially fatal illness which may be horizontally transmitted between the herd of pigs, and horses can be similarly infected with a fatality rate exceeding 50%[207]. Transmission to humans then occurs via human contact with infected pig saliva, urine or faeces, although direct transmission to humans via the consumption of date palm sap and infected horse meat can also occur[93,207]. In humans, *nipah virus* can be asymptomatic, but symptoms can include severe respiratory symptoms, encephalitis, flu like symptoms, encephalitis and vomiting often leading to a fatality rate exceeding 50%. There is also evidence of limited human to human transmission via close contact with patient saliva[207]. Symptomatic diagnosis of *nipah virus* is difficult, so serological methods including ELISAs, PCR testing and cell culture techniques are used diagnostically[207]. There are no specific treatments for *nipah virus* and no vaccines for humans or

livestock are available, although one human vaccine is currently in phase I clinical trials[93,207].

The other major *henipah virus* is *hendra virus*, which tends to cause infections in horses and has been known to spill over into humans causing fatal disease in over 50% of cases[209]. Pteropid bats are believed to be reservoir hosts for *hendra virus*, which has so far only been observed in South-East Asia and Australia where it causes annual spillover infections[93,209]. Spillover into horses via the faecal-oral route may lead to a subclinical infection or a potentially lethal symptomatic infection, which may then be transmitted to humans via close contact where febrile symptoms may develop which in turn may be fatal, although no human to human transmission has been observed[93,207,209]. *Hendra virus* diagnosis is generally made via serological ELISAs, and no specific treatment is available[93,209]. Whilst there are no *hendra virus* vaccines available in humans an effective vaccine is available in horses[93,209].

l. *Parvoviridae*

The *Parvoviridae* are a family of highly diverse non-enveloped negative sense ssDNA viruses (Baltimore classification group II)[212]. The *Parvoviridae* family is rapidly expanding, with the discovery of new viruses via metagenomics investigations occurring relatively frequently[212,213]. The *Parvoviridae* are split into three subfamilies, namely the *Parvovirinae, Densovirinae,* and *Hamaparvovirinae*, which tend to infect vertebrates, invertebrates, and both, respectively[214,215]. As of 2021, 23 genera and over 100 species constitute the *Parvovirinae*- for all genera assigned to each subfamily at the time of writing, see Table 1[212-214]. *Parvoviridae* genomes range from 3.9-6.3kb in size, and all *Parvoviridae* encode up to four major non-structural proteins and at least two capsid proteins, although other accessory proteins are often also encoded and are highly variable due to their tendency to recombine[212,214,216]. Due to the lack of a viral polymerase, parvovirus replication requires either the use of host enzymes or the presence of a "helper virus" that co-infects the same cell and provides the necessary polymerase to replicate. Replication is also tied to the host cell cycle, with increased viral replication in rapidly dividing host cells[212]. Some *Parvovirinae* such as some members of the *Dependoparvovirus* and *Protoparvovirus* genera can rarely integrate into the host genome during replication, in turn giving rise to heritable endogenous virus elements (EVEs) that can be replicated with the host genome[217,218].

Table 1- Organisation of the Parvoviridae family.
The genera comprising each subfamily of the Parvovirinae, and the number of species accurate as of 2021. Genera of importance to this project have been highlighted in **bold.** Adapted from[214].

Subfamily	Genus	Number of species
Densovirinae	*Aquambidensovirus*	2
	Blattambidensovirus	1
	Hemiambidensovirus	2
	Iteradensovirus	5
	Miniambidensovirus	1
	Pefuambidensovirus	1
	Protoambidensovirus	2
	Scindoambidensovirus	3
Hamaparvovirinae	*Brevihamaparvovirus*	2
	Chaphamaparvovirus	8
	Hepanhamaparvovirus	1
	Ichthamaparvovirus	1
	Penstylhamaparvovirus	1
Parvovirinae	*Amdoparvovirus*	5
	Artiparvovirus	1
	Aveparvovirus	2
	Bocaparvovirus	25
	Copiparvovirus	7
	Dependoparvovirus	8
	Erythroparvovirus	7
	Loriparvovirus	1
	Protoparvovirus	13
	Tetraparvovirus	6

Parvoviridae as a family are globally distributed, although it is difficult to estimate overall incidence and prevalence due to the variability and frequent animal tropism of these viruses, and prevalence varies significantly by geography even when considering the same parvovirus species[219-221]. Parvovirus transmission is poorly understood and whilst it is believed that droplet and aerosol transmission is likely the major route of transmission, iatrogenic transmission and vertical transmission may also occur[212,222,223]. Many parvoviruses are non-pathogenic although others are pathogens of humans and animals often with severe consequences, and some may have zoonotic potential, although this has yet to be confirmed[212,224].

The genus *Dependoparvovirus* belongs to the *Parvovirinae* subfamily and contains 9 species, which are known to infect a variety of species including humans, rodents, bats, lagomorphs such as rabbits, waterfowl, cats, birds, sea lions and others[212,214,217,225]. Many *Dependoparvoviruses* can replicate autonomously within the host cell without using cellular machinery to do so, although the presence of helper viruses is still required[212]. *Dependoparvoviruses* can infrequently integrate into the host cell genome, and analysis of *Dependoparvovirus* EVEs show that they are amongst the most ancient *Parvoviridae* and are likely to have diverged from the common parvovirus ancestor between 23 and 79 million years ago[217]. First identified in adenovirus preparations in the 1960s, the most important *Dependoparvoviruses* belong to the *Adeno-associated virus* (AAV) species, and AAVs can be found both in humans and other animals such as bats[212,217]. AAVs are considered non-pathogenic in both humans and animals, although other *Dependoparvoviruses* such as *anseriform dependoparvovirus 1* can cause potentially severe disease within waterfowl[217,226]. AAVs are distributed globally with an estimated seroprevalence of approximately 70% in humans and 18.6% in bats[212,217].

Since the 1980s, AAVs have been considered to be the best vector available for human gene therapy, and the only FDA or European Medicines Agency approved gene therapy treatments (Luxturna to treat retinal disease and Glybera to treat lipase deficiency, respectively) utilise AAV vectors[212,217,226]. AAV therapy is currently only used to treat monogenic recessive diseases, although over 100 phase I and II clinical trials are underway using AAVs as vectors to treat a variety of diseases[226]. Most diseases amenable to AAV mediated gene therapy are those that affect the central nervous system (CNS), liver and striated muscle due to the natural tropism of AAVs for these cell types[226]. There are issues with AAV mediated gene therapy, such as the small genome size limiting the size of the gene that can be supplied via AAV infection, rare and unpredictable genomic intergation, and prohibitive costs. High AAV seroprevalence can lead to immunological prevention of effective therapies, although methods of capsid alteration such as using historical capsids, *in silico* design or using rare AAV strains can help to minimise immunological interference[226]. Despite these limitations AAV mediated gene therapy is an extremely useful tool, and with further development may usher in a new age of treatments for genetic disorders[212,217,226].

The genus *Protoparvovirus* consists of 15 species, and is known to infect a variety of hosts, including humans, rodents, cats, dogs, pigs, foxes and bats, amongst other species[212,213,215,225,227]. The pathogenicity of *Protoparvoviruses* varies significantly by species- for example, feline panleukopenia virus (FPV) and canine parvovirus (CPV) within the species *Protoparvovirus carnivoran 1* can cause illness and significant mortality in cats and dogs respectively, whereas human *Protoparvovirus* infections cause mild, if any, disease[212,213,215]. Similarly to *Dependoparvoviruses*, *Protoparvoviruses* can integrate into the host genome via EVEs, and due to their non-autonomous replication they have a tropism for fast dividing cells, rendering them useful for cancer therapy[218,228].

FPV is a pathogen of cats primarily (although rarely FPV can spillover into dogs) which can cause severe disease in kittens over 6 weeks of age if untreated, although the availability of an effective vaccine has significantly reduced the disease burden of FPV[212,215]. CPV is believed to have emerged in the 1970s due to a cross-species transmission of FPV and subsequent host adaptation, and quickly caused a global panzootic in dogs[212,215]. CPV is transmitted via the faecal-oral route and is highly contagious, with symptomatic onset following a 3-7 day incubation period, including diarrhoea and vomiting, loss of appetite, fever and dehydration which may be fatal if left untreated[215]. Vaccines have been developed for CPV and have reduced the disease burden, although these are strain specific and do not guarantee absolute protection from all strains[215]. *Protoparvoviruses* can also infect rodents including rats and mice, where disease appears to be asymptomatic[212]. Similarly, the newly identified *Protoparvovirus* called *Newlavirus* appears to be asymptomatic in at least 2 different species of fox in Canada, although further investigations will need to be conducted to confirm this[227]. Rodent *Protoparvoviruses*, and in particular rat H-1 parvovirus, may act as oncolytic viruses with a potential role in cancer therapy[228].

To date, three species of *Protoparvovirus* have been known to infect humans- bufavirus, discovered in 2012, tusavirus, discovered in 2014, and cutavirus, discovered in 2016[213]. Bufavirus is the most studied of these and appears to cause mild and self-limiting gastrointestinal disease in children with a global distribution, although the prevalence of bufavirus infection varies from approximately 1% - 5% by location. Due to relative clinical insignificance of bufavirus infection no specific treatments or vaccines are available[213]. Tusavirus has been identified in a single stool sample via metagenomics and in one individual in seroprevalence investigations, giving an overall prevalence of approximately 0.5% in these investigations- accordingly, very little is known about tusavirus infections[213]. Cutavirus was identified in faecal samples from Brazillian children with diarrhoea and has since been isolated in skin biopsies from cutaneous T-cell lymphoma patients, at a prevalence of approximately 1% and 23.5% in their studies respectively, although little else is known about cutavirus[219]. Transmission methods are unknown for these three viruses although their presence in stool does to some extent imply that the

faecal-oral route may be important, and it is unclear whether cutavirus may be transmitted via direct skin to skin contact[213,219].

The genus *Bocaparvovirus* (formerly known as *Bocavirus*) consists of 29 species and is the largest genus of *Parvoviridae*[212]. *Bocaparvoviruses* have most frequently been discovered via metagenomics analysis and have been found in a variety of host species including humans, rodents, cows, pigs, dogs, cats, non-human primates, camels and other mammals[212,225]. Some of the animal *Bocaparvoviruses* such as the canine minute virus and feline bocavirus have been shown to cause mild and self-limiting disease and typically mild but rarely fatal disease respectively in their animal hosts, whereas others are asymptomatic[212,215,225]. Whilst it is difficult to estimate prevalence and distribution for many of these viruses, rodent *Bocaparvoviruses* are globally distributed and may infect a variety of rodent species, illustrating a potential zoonotic transmission risk[224].

Human *Bocaparvoviruses* are globally distributed, and are considered to be primarily respiratory but also gastrointestinal pathogens that are often identified in co-infections with other respiratory pathogens[229]. There are four serotypes of human *Bocaparvovirus* (human bocavirus 1-4), at least some of which are believed to have arisen through recombination events[216,221]. These viruses most commonly infect children and usually cause mild and self-limiting common cold symptoms in infected individuals, although they have also been rarely associated with meningitis and meningoencephalitis[221,230]. Human *Bocaparvovirus* infections are often not diagnosed due to their mild nature and are dismissed as a common cold, but when they are diagnosed it is usually via specific PCR assays[230]. These viruses are transmitted via aerosols, and no specific treatments or vaccines are available for human *Bocaparvovirus* infection[229,230]. Human *Bocaparvovirus* prevalence varies by location but in Japan are estimated to cause approximately 16% of respiratory infections, and seroprevalence is believed to reach up to 96% in adults worldwide[221,230].

Chapparvoviruses include the two genera *Icthamaparvovirus* and *Chaphamaparvovirus* within the subfamily *Hamaparvovirinae*[212,214,231]. As the genus *Icthamaparvovirus* contains only 1 species that infects the gulf pipefish, this project will only consider the genus *Chaphamaparvovirus*[214]. This genus contains 15 species including those identified in bats, birds, pigs, rodents, fish, dogs, cats, chickens and other species[212,214]. *Chaphamaparvoviruses* are globally distributed and usually do not cause disease in their hosts, although there are some exceptions such as mouse kidney parvovirus, fechavirus and tilapia parvovirus, each of which can be fatal[212,218,225]. EVEs have been observed in *Chaphamaparvoviruses* suggesting a potential for genomic recombination and evolutionary studies[218]. Whilst it is possible that a human *Chaphamaparvovirus* exists and could even cause kidney disease, this has yet to be observed, and whilst there is a zoonotic risk due to the relatively high prevalence of rodent *Chaphamaparvoviruses* in major cities such as New York City, zoonotic spillover has never been confirmed[212,231]. No transmission route has been

confirmed, although the ability of *Chaphamaparvoviruses* to persist outside of the body suggests a possible fomite or respiratory route, and the presence of *Chapamaparvovirus* DNA in urine may also suggest a transmission route involving consumption of contaminated food[231]. Regardless, some *Chaphamaparvoviruses* such as fechavirus appear to be highly contagious and can cause rapidly progressing outbreaks in confined areas[225].

Other *Parvoviridae* are also important, such as the *Erythroparvovirus* parvovirus B19 (B19V) and the *Tetraparvovirus* parvovirus 4 (PARV4)[223,232]. B19V is a pathogen of children that is globally distributed and follows a seasonal pattern with outbreaks every 3-5 years[220]. Global prevalence varies significantly with areas including China, the Indian subcontinent and Eastern Europe having lower prevalence than in other areas of the world, and seroprevalence is highly variable ranging from approximately 30-70%[220,223]. B19V infection typically causes relatively mild febrile disease and a characteristic rash known as "fifth disease" in childhood and febrile disease without a rash in adulthood, although there have been reports of rare and serious neurological symptoms including encephalitis, meningitis and peripheral neuropathy, amongst others[220,233]. B19V is transmitted via aerosol droplets and can be transmitted vertically via cross-placental transmission in up to 33% of infected pregnant people, which can lead to abortion or fetal hydrops in the infant with a decreasing risk as the pregnancy progresses toward term[220,222,223]. B19V diagnosis is typically made serologically using commercial immunoassays against IgM and IgG and no specific treatment or vaccine is available, although there has been a reported use of high-dose intravenous immunoglobulin which may have improved the patient outcome[222,223].

PARV4 comprises 3 genotypes- genotypes 1 and 2 are found in Europe, Asia and North America, and genotype 3 is found in Sub-Saharan Africa, leading to a broad and potentially global distribution[232]. Seroprevalence for PARV4 is geographically variable ranging from 0-50%, and whilst PARV4 is not regularly screened for it can be diagnosed serologically[221,232]. PARV4 transmission routes are currently unknown, although previously PARV4 infection has been associated with intravenous drug users suggesting that subcutaneous injury may be a viable, yet likely secondary, transmission route[221]. The link between PARV4 and disease is unclear- some studies suggest that PARV4 causes mild influenza-like febrile illness, others suggest B19V symptoms include a rash and fetal hydrops, and yet others suggest that B19V causes gastrointestinal symptoms or alternatively an asymptomatic infection[221,232]. Alternatively, PARV4 may not be intrinsically pathogenic, but may act as a helper virus for co-infections, potentially exacerbating the symptoms of the other pathogen[232]. Further research is required to determine PARV4 pathogenicity or lack thereof[221,232].

m. *Pegivirus*

Pegivirus is a genus of the *Flaviviridae* family, and *Pegiviruses* are small positive sense ssRNA viruses (Baltimore classification group IV) with genomes of 8.9-11.3 kb[234,235].

The *Pegivirus* genome consists of one ORF which translates into a single polyprotein that is cleaved to give rise to two structural and six non-structural proteins, as well an internal ribosome entry site (IRES) that directs polyprotein translation and a 5' non-translated region[235,236].The *Pegivirus* genus consists of 11 species with further genotypic subdivisions and can infect a variety of mammalian hosts including humans, rodents, bats, primates, pigs, horses and birds[234,235,237,238]. Primates are believed to be the natural hosts of *Pegiviruses*, which are believed to be evolutionarily ancient viruses[234,235]. To date there is no evidence of *Pegivirus* spillover events or zoonotic transmission, possibly due to relatively slow evolution by RNA virus standards[235,237].

In most hosts *Pegiviruses* do not cause disease, although there is evidence that *pegivirus equi* (also known as Theiler's disease associated virus, TDAV) may cause or contribute to the development of Theiler's disease in horses, which can cause asymptomatic to severe hepatitis following the administration of blood products and is fatal in up to 18% of infected horses[185,238]. *Pegivirus* infections are similarly considered non-pathogenic in humans, although *Pegivirus* RNA has been isolated from brain biopsies of encephalitis patients and may rarely be associated with the development of lymphoma[185,235,239]. As it is generally considered non-pathogenic diagnostic screens and blood donation screens do not routinely screen for *Pegivirus* infection, and no specific treatments or vaccines are available[185,235].

Pegiviruses are distributed globally with substantial variation in prevalence by area[235,240]. Whilst 1-20% of the global population is estimated to be seropositive for *Pegivirus* exposure, in MEDCs seroprevalence estimates tend to range from 1.1-6% whilst in LEDCs they can reach as high as 75.3%[235,239]. Only 2 species of *Pegivirus* are known to infect humans- HPgV-1 (previously known as GB virus C) consisting of 7 geographically isolated genotypic strains, and HPgV-2[235,236,240]. *Pegiviruses* can cause both chronic and acute infections, and may spontaneously clear within 2 years of infection or may persist for decades for unknown reasons that are suspected to link to host genetics[235,236].

Pegivirus transmission is primarily through percutaneous needlestick injuries, but can also be transmitted sexually, iatrogenically via blood products or via vertical transmission- accordingly, intravenous drug users are the most likely to contract *Pegivirus* infection and *Pegiviruses* are often seen in coinfections with other bloodborne viruses such as HIV-1 and HCV[235]. Active *Pegivirus* infection has been associated with improved outcomes and prolonged survival in patients suffering with HIV-1, hepatitis C, Ebola and malaria co-infections[235,236,240]. Whilst the mechanisms through which *Pegivirus* infection improves the prognosis of these diseases is unknown, it is generally believed that immune system modulation is involved[235,236,240]. There are suggestions that *Pegivirus* infection may be used as an effective therapy or as a potential "bio-vaccine" in resource limited settings where the availability of HIV treatments such as HAART are limited, although the potential

for *Pegivirus* associated disease requires further research and investigation prior to this becoming feasible[235,239].

n. *Picobirnavirus*

The *Picobirnavirus* genus is the only member of the family *Picobirnaviridae*, and *Picobirnaviruses* (PBVs) are usually bipartite viruses with a dsRNA genome (Baltimore classification group III)[241,242]. Typically, the total genome size of PBVs is approximately 4.2kb, consisting of a large segment (segment 1) of approximately 2.3-2.6kb and a small segment (segment 2) of approximately 1.5-1.9kb, although some PBVs have a small genome profile where segments 1 and 2 are approximately 1.75 and 1.55kb respectively[242,243]. Whilst most PBVs are bipartite, monopartite genomes resulting from the fusion of genomic segments 1 and 2 have been reported, and tripartite genomes have also been reported, although further research is required regarding whether these tripartite genomes were detected due to the presence of a mixed infection or a genuine tripartite genome[243,244]. Segment 1 contains 2-3 ORFs, leading to 2-3 proteins including a capsid protein and another protein of unknown function, whilst segment 2 contains 1 ORF encoding the RdRp[245]. PBV genomes are highly variable and can be separated into three distinct genogroups showing < 40% intergenogroup similarity, where genogroups I and II are found in mammals and genogroup III is found in invertebrates, although it has also been suggested that 2 further genogroups (IV and V) are circulating in humans[241,242,245]. There are currently only two ICTV recognised species within the *Picobirnavirus* genome (*Human Picobirnavirus* and *Rabbit Picobirnavirus*), and it has been suggested that due to their variability that PBVs may actually exist as a quasispecies rather than as individual species[16,243,244].

PBVs have been isolated from a variety of host species including humans, rodents, bats, pigs, cattle, horses, wolves, dogs, birds and invertebrates, and appear to be globally distributed[8,41,245-247]. The broad host range of PBVs is believed to be due to their ability to undergo recombination and the natural host of PBVs is currently unknown, although frequent interspecies transmission between mammalian species and zoonotic transmission via an unknown but likely faecal-oral or waterborne transmission route has been observed[241,243,246,247]. There have also been suggestions that PBVs may in fact be bacteriophages and that their broad host range may be explained as infecting bacteria that are ubiquitous throughout mammals[9,241,246]. Whilst PBVs have never been successfully propagated in a mammalian cell system, they have also never been successfully propagated in a prokaryotic culture system either, lending support to neither the prokaryotic nor eukaryotic origin theories[243].

PBVs are usually found in the stool of infected individuals, although they have also been isolated from respiratory tracts of humans, pigs and cattle and found in serum samples from mammals and could likely be found in other tissues following further investigations[243,246,248]. PBV infection can persist for a prolonged time and often goes through cycles of silence followed by periods of high viral activity, possibly due to immunological responses[242,243]. There is significant debate regarding the

pathogenicity of PBV infection in humans and animals. Some studies indicate that PBVs are non-pathogenic in mammalian hosts, others suggest that they can cause opportunistic infections in immunocompromised hosts, and yet others suggest that they can directly cause symptomatic infections as the primary infectious agent[6,241,248]. PBV infection either as the sole pathogen or as a co-infection is often associated with diarrhoea, although recently PBVs have also been implicated in causing acute respiratory infection as the sole pathogen identified in hospitalised patients[242,248]. No specific PBV treatments or vaccine are available[241,243].

PBV identification was traditionally performed via PAGE electrophoresis and size discrimination, and despite how insensitive this method is it is still performed today[242,243]. Recently, RT-qPCR has become a more frequent method of diagnosis as it is significantly easier and more sensitive, although the extreme variability of the PBV genome can lead to issues with successful priming[242,243]. PBVs are often also incidentally found during metagenomics studies[243,247]. Due to the diagnostic difficulties and the lack of concordance regarding the pathogenicity of PBVs leading to limited investigations, PBV prevalence and incidence is difficult to accurately assess[6,241,243,247]. A recent study in hospitalised patients suggested a prevalence of 19.2%, although this is likely to be lower in the non-hospitalised general population and is likely to be highly variable according to geographic area and demographic groups[242,248]. In pigs, a prevalence of 20.9% has been reported in Argentina and of 75.8% in the Caribbean, as has a prevalence of 14.28% in horses, 23.4% in Brazillian cattle, and from 0-42.35% in goats in Turkey, illustrating the variability in PBV prevalence[243,245]. Finally, PBVs have been identified in wastewater in multiple countries suggesting a possible waterborne transmission route[242,248].

o. Picornavirus

Picornaviruses, of the family *Picornaviridae*, are RNA viruses with genomes ranging from ~6.5-10kb and are non-enveloped viruses[249]. Whilst *Dicipiviruses* have a dsRNA genome, all other *Picornaviridae* have positive sense ssRNA genomes (Baltimore classification groups III and IV, respectively)[67,250]. Picornaviruses are extremely diverse and to date 68 genera, 158 species and approximately 700 genotypes have been identified, in turn identifying picornaviruses as the largest and most diverse Baltimore classification group IV viruses. In the age of metagenomics, picornavirus discovery is increasing in frequency with new picornaviruses often identified as incidental findings[59,250,251]. Picornavirus genomes typically encode a single polyprotein that is cleaved by viral proteases into at least 7 non-structural proteins and 3-4 capsid proteins, although this is not always the case and multiple polyproteins may rarely be encoded[249,250,252]. The picornavirus polymerase is highly error prone, leading to a high mutation rate and significant recombination, in turn leading to the extreme diversity of picornaviruses[20,111]. Picornaviruses are ubiquitous, both in terms of their global distribution and host range, where they have been found on all continents excluding Antarctica and infect hosts of all major vertebrate lineages including humans, rodents, bats, livestock animals such as pigs,

and birds, amongst others[48,111,251,253-255]. Some picornaviruses have shown zoonotic potential, with some rodent picornaviruses such as some *Enteroviruses* having previously been shown to undergo zoonotic transmission[67,256]. Of the picornavirus species identified to date, an estimated 24 species from 9 genera are known to infect humans, with disease profiles ranging from asymptomatic infection through to potentially lethal disease as demonstrated by polio[256-258]. Due to the diversity of the *Picornaviridae*, it is difficult to identify general traits for this family and it is beyond the scope of this work to examine the whole family in-depth[251].

Enteroviruses are one of the earliest described genera of picornavirus and are arguably the most important genus[250]. There are 15 species of *Enterovirus*, of which 7 infect humans (human *Enterovirus A-D* and *Rhinovirus A-C*)[259,260]. The species *Enterovirus C* is then further divided into three genetic subgroups with a variety of serotypes within each subgroup, and due to rapid recombination between *Enterovirus* species and strains new serotypes emerge frequently, leading to over 300 serotypes reported to date[111,260,261]. Animal hosts of *Enteroviruses* include cattle and pigs for *Enteroviruses E-G*, non-human primates for *Enteroviruses H* and *J*, dromedary camels for *Enterovirus I*, and currently unclassified *Enterovirus* species in rodents and goats[18,41,262]. The zoonotic potential of these viruses is unknown, although it is possible that *Enterovirus E and H* may infect humans via contact with infected livestock faeces, and there is evidence that swine vesicular disease virus may have evolved from a strain of human coxsackievirus B5 that spilled over from humans into pigs[256,262]. Some *Enteroviruses* are relatively geographically constrained, whilst others such as coxsackievirus A are globally distributed, in part due to their effective transmission via the faecal-oral and respiratory routes and their prolonged survival as fomites[258,263,264]. Many *Enterovirus* infections are asymptomatic, whereas others cause hand, foot and mouth disease (HFMD), which is most common in children, and some cases may cause potentially fatal CNS symptoms such as meningoencephalitis, although the specific *Enterovirus* and immune state of the individual may affect disease outcome[259,264].

Historically, the most important picornavirus was the poliovirus, which used to cause an estimated 350,000 cases and up to 35,000 deaths worldwide prior to a concentrated global vaccination campaign[111,258]. Poliovirus consists of three wild poliovirus strains (WPV1-3), all of which can cause both acute poliomyelitis and post-polio syndrome[258]. WPV 2-3 have been eradicated worldwide, and whilst WPV 1 has not successfully been eradicated, its prevalence and incidence has dropped significantly and its endemic range has reduced to only include Pakistan and Afghanistan[258]. Like other *Enteroviruses*, poliovirus is transmitted via the respiratory and faecal-oral routes and has been known to be transmitted via contaminated food and water[258].

Many poliovirus cases are asymptomatic but following an incubation period of up to 10 days, acute poliomyelitis may develop, which primarily causes flu-like symptoms. Up to 1% of cases may progress to temporary or acute flaccid paralysis limiting

movement without sensory loss, and up to 10% of paralytic cases may be fatal[111,258,265]. 40-50% of acute poliomyelitis patients develop post-polio syndrome, where debilitating neuromuscular symptoms may develop up to 40 years post infection. Previously, ELISA serological assays were used for poliomyelitis diagnosis, whereas more recently RT-PCR and cell culture techniques have become more commonplace[258]. There is no specific antiviral treatment for either acute poliomyelitis or for post-polio paralysis, and no way to reverse the paralytic effects of poliovirus infection[265]. However, there are effective anti-polio vaccines, including a live attenuated oral poliovirus vaccine (OPV) and an inactivated vaccine[258]. The OPV vaccine is commonly considered to be more effective and is preferentially used, although the high recombination rate of poliovirus can lead to reversion to a pathogenic strain and the development of vaccine-derived poliovirus (VDPV) infection, which can in turn cause paralysis[258,265]. Outside of Pakistan and Afghanistan, all non-imported poliovirus cases are VDPV cases, and since 2021 VDPV has been reported in 36 countries and ~2000 cases worldwide. Improved vaccines with reduced risks of VDPV development are currently in development[258].

Many other *Enteroviruses* are of clinical importance. For example, enterovirus 71 causes HFMD in children, and whilst most infections are mild and self-limiting, potentially fatal neurological symptoms and breathing difficulties have been reported[264]. Whilst enterovirus 71 consists of many genotypes and is globally distributed, the majority of the disease burden is Asia, where China alone reported ~7.2 million cases and ~2500 deaths from 2008-2012, and whilst the death rate is usually low in outbreak settings mortality can spike to over 50%[264]. Diagnosis is primarily made via RT-PCR, and whilst no specific treatments are available for enterovirus 71 infections vaccines are available, albeit at potentially prohibitive costs[259,264]. Coxsackievirus A also causes HFMD primarily in children, and some serotypes such as coxsackievirus-A6 can cause significant mortality in adults[263,266]. Much like enterovirus 71, coxsackievirus A is globally distributed but most of the disease burden is in Asia, and there is no specific antiviral therapy for coxsackie virus A infetion or prophylactic vaccine available[264,266]. Other human *Enteroviruses* such as Echovirus 30 and Enterovirus D68 can also cause disease in humans[259]. Livestock *Enteroviruses* have a relatively mild disease burden and a minimal effect on the farming industry but can still cause disease in animals- for example, some can cause severe diarrhoea in goats, and *Enterovirus G* strains can infect pigs[261].

The genus *Cardiovirus* contains 6 species (*Cardiovirus A-F*) and approximately 32 genotypes[256]. *Cardiovirus A* consists of one genotype (encephalomyocarditis virus, EMCV) that primarily infects rodents, *Cardiovirus B* consists of the rodent viruses Theiler's murine encephalomyelitis virus (TMEV) and thera virus, as well as Vilyusk human encephalomyocarditis virus[267,268]. *Cardiovirus C* consists of 2 serotypes of Boone cardiovirus that infects rats, *Cardiovirus D* consists of 11 genotypes of Saffold virus (SAFV) which infects humans, *Cardiovirus E* consists of one genotype that infects red-backed voles and *Cardiovirus* F consists of one genotype that infects grey-backed

voles[267]. *Cardioviruses* as a genus are globally distributed, although the exact distribution varies by species and even genotype- for example, whilst Boone cardioviruses are globally distributed, others such as *Cardiovirus E* and *Cardiovirus F* are not[253,267]. Prior to 2007 *Cardioviruses* were not believed to infect humans, but retrospective studies showed that a SAFV strain had been present in the faecal sample of a child in 1982, illustrating at least 41 years of human *Cardiovirus* infections[111].

EMCV is a virus of rodents, which is readily transmitted to pigs via faecal-oral transmission, and then can be onwardly transmitted zoonotically to humans, or to non-human primates, dogs, boars and elephants[48,95,268]. To date there has been no reports of a direct transmission route from rodents to humans, although this is suspected to occur[268]. EMCV consists of two strains (although it has been proposed that 4 species and 7 serotypes may be more appropriate) and is endemic in rodents of a variety of species with a seroprevalence of up to ~38%[268]. EMCV infections cause disease ranging from asymptomatic to fatal myocarditis, and disease severity and progression appears to vary according to serotype and host species and age[48,268]. EMCV seroprevalence in humans ranges from approximately 3-50% depending on location, and usually causes an asymptomatic or mild and self-limiting febrile disease, although in severe cases can cause reproductive failure, encephalitis, diabetes mellitus and infant death[48,268]. TMEV infects mice and in a natural infection tends to cause an asymptomatic enteric infection, although TMEV genotype 2 strains can be more virulent and may cause fatal myocarditis[95,257,269,270]. Very little is known about *Cardiovirus C* species except that they were identified in *Rattus norvegicus* and laboratory rats[267].

SAFV is a ubiquitous virus of humans that consists of 11 genotypes, some of which are globally distributed such as SAFV1-3 and others of which are not, such as SAFV4-8 which are primarily found in Afghanistan and Pakistan[257,269]. The majority of SAFV infections appear to be asymptomatic and seroprevalence in children is typically approximately 80%, although it can reach up to 100% in some areas[257,271]. SAFV is believed to be transmitted via the faecal-oral route and can cause mild and self-limiting febrile respiratory disease or gastroenteric disease that is typically diagnosed by RT-PCR[257,272]. In rare cases SAFV can cause more severe disease, including acute flaccid paralysis, meningitis, pancreatitis, and rarely sudden death, particularly in children[271,272]. Due to the rapid mutation rate and broad distribution of SAFV, there are concerns about a more dangerous strain evolving and potentially causing a global pandemic[257].

The genus *Parechovirus* consists of 6 species (*Parechovirus A-F*) and 29 genotypes to date[256]. *Parechovirus A* primarily infects humans, whilst *Parechovirus B* and *C* primarily infect rodents[102,256]. *Parechovirus D* primarily infects bats but also infects ferrets, *Parechovirus E* primarily infects birds, and *Parechovirus* F primarily infects lizards[256]. *Parechoviruses* are globally distributed and can cause infections ranging from asymptomatic to potentially fatal[111,273]. *Parechovirus B* can infect a variety of

rodent species across mainland Europe and the USA including bank voles and yellow-necked mice, with a prevalence of up to ~26% by RT-PCR[102]. Zoonotic transmission of *Parechovirus B* into humans is considered unlikely, although there have previously been reports of intrauterine foetal death and sudden infant death syndrome being associated with *Parechovirus B* infection in humans and in rodents[95,99,102].

To date, 19 genotypes of human parechovirus (HPeV) have been identified, and HPeVs have been isolated worldwide[273,274]. HPEV 1 and HPEV 3 appear to be the most clinically important and most commonly isolated HPEVs, with HPeV 3 incidence reaching up to 57% in Malawi and causing an estimated 2% of viral infections in many countries[273]. HPeVs are transmitted via the faecal-oral route and appear to be seasonal, and are usually diagnosed via RT-PCR[273,275]. Most HPeV genotypes tend to cause relatively mild and self-limiting disease, but HPeV 3 has been associated with a variety of neurological sequelae including sepsis and meningoencephalitis alongside severe symptoms including cardiac arrest and premature death [273,275]. HPeVs are primarily pathogens of children, and in particular premature babies are at risk of developing severe symptoms upon HPeV infection, although immunocompromised and elderly adults are also at risk of severe HPeV[273,274]. The severity and outcome of HPeV infection is linked to both the viral genotype and the demographics of the patient, and those with neurological sequelae are approximately 3.5-fold more likely to require ventilator support in severe cases[274]. There are no specific treatments or vaccines available for HPeVs and treatment is largely supportive[273,275].

Hepatitis A virus (HAV) is a member of the *Hepatovirus* genus that is ubiquitous and distributed worldwide[41,276]. Initially HAV was only isolated in humans and non-human primates, but recently *Hepatoviruses* have been isolated from bats and rodent species too, although cross-species transmission has yet to be observed[276]. Similarly, there was initially only one member of the *Hepatovirus* genus, but in 2015 Drexler and colleagues identified 13 novel species in a variety of hosts, dramatically expanding the genus[276]. HAV is unusual in that it exists as a non-enveloped virus and is shed in the faeces as such but derives an envelope from host-cell membranes when circulating in blood[276]. HAV is primarily transmitted via the faecal-oral route but is also the leading cause of water-borne hepatitis, and a conservative estimate of HAV incidence is given as 1.4 million by the WHO[111,277]. The HAV burden is significantly higher in LEDCs, as poor sanitation, poor hygiene and crowded conditions are all associated with increased transmission and can lead to severe outbreaks[277]. HAV is known to cause acute hepatitis which becomes more severe with advanced age, and can rarely cause fulminant hepatitis, liver failure and death. A safe and effective vaccine that provides long-term immunity is available, but vaccination rates are variable and generally too low to adequately prevent widespread transmission[277].

There are also less clinically important genera of *Picornaviridae*. The genus *Rosavirus* is a relatively new genus of *Picornavirus* that was discovered in wild canyon mouse stool in 2011 prior to full genome capture in 2013[96,278]. Retrospectively named

Rosavirus A, *Rosavirus B* has since been discovered in the Street rat and Norway rat in China, and *Rosavirus C* has been discovered in five rat species in China[95]. A second genotype of *Rosavirus A* was found in a faecal sample from a child with diarrhoea in The Gambia at a prevalence of 0.55%, but very little is known about the potential disease profile, transmissibility or overall prevalence for this or any other *Rosavirus*[252]. The genus *Kunsagivirus* is another new picornavirus genus first discovered in 2013[255]. *Kunsagivirus A* was found in a faecal sample from a European roller bird (*Coracias garrulus*) by metagenomics, although it was also suggested that the natural host of this virus could be a shrew that was eaten by the bird due to the relatively close relation of *Kunsagivirus A* to other rodent picornaviruses[255]. *Kunsagivirus B* has since been discovered in Cameroonian fruit bats, and *Kunsagivirus C* was discovered in archival blood of a Tanzanian yellow baboon from 1986, suggesting that *Kunsagiviruses* have been circulating since at least the mid-1980s[254,279]. In 2020, a new genotype of *Kunsagivirus C* was identified in wild vervet monkeys in Uganda suggesting that it is possible that non-human primates are the true reservoir hosts of *Kunsagiviruses*, or at least an effective intermediate host[251]. Whilst very little is known about the life cycle, disease profile or spillover ability of *Kunsagiviruses*, it has been speculated that they likely have a broader host range than currently known and may have a relatively broad geographic range including Africa and Europe[251].

p. Polyomavirus

Polyomaviruses are non-enveloped dsDNA viruses with a circular genome (Baltimore classification group I) of the family *Polyomaviridae*[280,281]. Polyomavirus genomes are approximately 5kb and contain early, late and non-control regions, which in turn encode at least 3 structural proteins in all polyomaviruses alongside an extra structural protein in avian polyomaviruses and some mammalian polyomaviruses[281]. The *Polyomaviridae* consists of four genera- *Alphapolyomavirus* which infects mammals and humans, *Betapolyomavirus* which also infects mammals and humans, *Gammapolyomavirus* which infects and can cause fatal disease in birds, and *Deltapolyomavirus* which infects humans[281]. Polyomaviruses are known to infect a variety of species including humans, rodents, bats, cattle, goats, pigs, non-human primates and fish, amongst others[103,282,283]. Whilst it is believed that some newly discovered species of polyomavirus may be zoonotic, transmission between host species is rare even when both hosts are genetically similar and further research is required to confirm the zoonotic potential of new polyomaviruses[103,280,282]. There are over 100 species of polyomavirus, some of which are globally distributed such as BK polyomavirus, and new species are being discovered frequently in a variety of hosts[282,284,285].

14 polyomaviruses are associated with infections in humans, some of which are potentially pathogenic[281,284]. BK polyomavirus and JC polyomavirus were the first polyomaviruses discovered in humans and are broadly distributed globally with a seroprevalence of up to 80%, with a suspected but unconfirmed respiratory or faecal-

oral transmission route[284]. Like most human polyomaviruses, they tend only to cause disease in immunosuppressed individuals and cause asymptomatic but persistent infections in immunocompetent people. In immunosuppressed patients, BK polyomavirus can cause kidney failure or haemorrhagic cystitis, whereas JC polyomavirus can cause potentially fatal progressive multifocal leukoencephalopathy and may be linked to the development of tumours[284]. Merkel cell polyomavirus is asymptomatic in immunocompetent humans and is believed to be present as a persistent infection in up to 80% of the global adult population. Other than a raccoon polyomavirus, Merkel cell polyomavirus is the only polyomavirus with compelling evidence for causing cancer in the host, and 80-85% of Merkel cell carcinomas contain Merkel cell polyomavirus DNA[284].

Trichodysplasia spinulosa polyomavirus is the cause of trichodysplasia spinulosa (TS), a skin condition characterised by follicular pauples, keratonic protrusions and eyebrow alopecia in immunocompromised individuals[283,284]. TS is a very rare condition and can be treated with cidofovir, and since being isolated from respiratory samples trichodysplasia spinulosa polyomavirus is believed to very rarely cause respiratory disease and to be transmitted via a possible respiratory route[283]. *Human polyomaviruses 6* and *7* have high global seroprevalences of 83-93% and 63-83% respectively and can cause puritic rash symptoms in immunosuppressed patients. *Human polyomavirus 6* may also be involved in the development of a variety of tumours, although further research is required to confirm this[284]. Many other human polyomaviruses, such as *human polyomaviruses 10* and *12*, are not believed to cause disease[284]. Finally, humans have previously been infected with the polyomavirus simian virus 40 (SV40), which was discovered as a contaminant of some batches of poliovirus vaccine in the 1960s which were grown in SV40 cells and have since been associated with a variety of human tumours[284].

q. Rhadinovirus

The genus *Rhadinovirus* is a member of the sub-family *gammaherpesvirinae* within the family *Herpesviridae* and is an enveloped virus with a dsDNA genome (Baltimore classification group I) of approximately 130kb[169,286]. *Rhadinoviruses* are globally distributed and establish a persistent infection within the host due to a biphasic lytic and latent life cycle[286]. Whilst primarily considered in the context of human infection, *Rhadinoviruses* can also infect rats, mice, deer and cattle, and in cattle they can cause post-partum metritis[287,288]. Like most *Herpesviridae*, *Rhadinoviruses* are generally well adapted to the host and can cause disease in immunosuppressed individuals or in an alternative host following cross-species transmission[169].

The most important *Rhadinovirus* is *human herpesvirus 8*, also known as the Kaposi's sarcoma-associated herpesvirus (KSHV)[286]. KSHV is globally but unevenly distributed, and whilst incidence and prevalence are difficult to estimate, it is believed to be most prevalent in endemic regions within the Middle East and Africa where seroprevalence can reach up to 85%[289]. KSHV is suspected to cause 40,000 new cases of cancer annually and up to 20,000 deaths, although this is likely an underestimate

due to a lack of effective diagnostic assays and underreporting[289]. KSHV has been implicated in many diseases but has only been clearly epidemiologically linked with Kaposi's sarcoma, primary effusion lymphoma (PEL) and multicentric Castleman disease (MCD), all of which are cancerous conditions[286,289]. These diseases don't usually present in the immunocompetent host but are typically found in the immunocompromised and specifically in AIDS patients, where Kaposi's sarcoma presents in 20% of individuals[289]. Average survival time following diagnosis with one these conditions is low- 10 months for PEL and approximately 1 year for MCD[289]. There is no specific antiviral treatment for KSHV, and treatment for diseases of KSHV is symptomatic and variable, with Kaposi's sarcoma treatment being based on which form of disease presents and no generalised treatment plans for PEL or MCD due to their rarity[286,289]. No vaccine is currently available for KHSV although vaccine development research is ongoing[289].

r. *Rotavirus*

Rotaviruses are multi-partite dsRNA viruses (Baltimore classification group III), with an approximately 18.5kb genome split across 11 genomic segments which in turn range from ~3kb to ~0.6kb[13,290,291]. *Rotaviruses* belong to the *Reoviridae* family and the *Sedoreovirinae* subfamily[13,292]. The *Rotavirus* genome encodes 6 structural and 6 non-structural proteins, and the majority of segments are monocistronic, although segment 11 has 2 ORFs encoding 2 different non-structural proteins[291]. *Rotaviruses* are classified according to their VP7 glycoprotein and VP4 protease sensitive protein segments, although where possible a full 11 segment "barcode" should be used for identification[293].

Rotavirus infections in humans are primarily caused by group A *Rotaviruses*, although there is also evidence of group B and C *Rotavirus* infections in humans[291]. *Rotavirus* infection is the leading cause of severe gastroenteritis in children under 5 globally, causing millions of cases and deaths with estimates ranging from 125,000-600,000 deaths annually[291,294,295]. Approximately half of infections (50.2%) are in children under 5 years of age, although *Rotavirus* infection can occur at any age and can be severe in the immunocompromised and the elderly[291,293,296]. LEDCs bear the majority of the global *Rotavirus* burden, with approximately half of all *Rotavirus* deaths occuring in Nigeria, Pakistan, India and the Democratic Republic of Congo[293]. *Rotaviruses* can infect a variety of species including humans, rodents such as bank voles, pigs, cows, horses, dogs, cats and birds[62,293,297]. Different *Rotavirus* species tend to infect different host species, although due to significant genetic reassortment zoonotic transmission and interspecies transmission is relatively common[293,296,297]. An estimated 2.7-5.4% of *Rotaviruses* are reassortant rendering it difficult to perform evolutionary investigations on these *Rotaviruses*[20,62]. The risk of zoonotic transmission is increased in LEDCs, where an often-increased proximity to livestock and rodents promotes zoonotic transmission[293,297]. As well as being potential reservoir species and allowing for zoonotic transmission, pigs infected with *Rotaviruses* may suffer from a variety of symptoms including reduction in weight and

potentially death, leading to a large economic and animal welfare burden. This disproportionately affects LEDCs, where *Rotavirus* prevalence in pigs may reach up to 67% with an even higher seroprevalence[48].

Rotavirus transmission typically occurs via the faecal-oral route, and in humans generally causes mild and self-limiting gastroenteritis with symptoms including diarrhoea, vomiting, nausea and stomach pain, although infections can be fatal, particularly in young children[294,298]. Symptomatic diagnosis is difficult due to the non-specific nature of *Rotavirus* symptoms and is often not performed in mild cases, potentially leading to an underestimation of *Rotavirus* incidence[298]. Since 1990 *Rotavirus* diagnosis has been performed via PCR, and whilst the original Gouvea primers are still used today multiplex PCR techniques such as a semi-nested *Rotavirus* specific multiplex PCR panel are now the standard, although serological and traditional culture methods are also infrequently used[290,292,295,299]. *Rotavirus* treatment entails oral rehydration and intravenous rehydration in severe cases and no specific *Rotavirus* antiviral treatments are available[298].

Vaccination has been relatively successful for *Rotavirus* prevention. At least 118 countries have approved a *Rotavirus A* vaccination for use with an estimated 40% global coverage, resulting in approximately 40% reduced hospitalization and 25% reduced *Rotavirus*-associated mortality globally[293]. Whilst vaccination has been effective worldwide, the benefit is more substantial in MEDCs, with LEDCs reporting reduced vaccine effectiveness comparatively. This is believed to be due to socioeconomic differences and increased zoonotic transmission, and to some extent vaccine cost[291,293]. Live-attenuated vaccines have been FDA approved for use including the monovalent Rotarix and the pentavalent RotaTeq vaccines, however, both provide protection against specific strains of Rotavirus with limited protection against new or less common strains[291]. These vaccines are not associated with complete sterile immunity but do significantly reduce the incidence of severe disease and death and have been shown to drive significant herd immunity-based protection from *Rotavirus* infection[293]. *Rotavirus* vaccines have been associated with a slightly increased risk of intussusception and are unsuitable for administration to immunocompromised and premature children[291]. As these are live attenuated vaccines there are also concerns regarding potential reassortment between vaccine strains and wild strains, potentially leading to immunological evasion and reduced vaccine effectiveness[291,293]. Vaccination programmes vary significantly by country and there is currently no individual vaccine suitable for global use[293]. Various new human vaccines using recombinant technology are in development[291]. *Rotavirus A* vaccines are also available for livestock[291].

s. Rubivirus

The genus *Rubivirus* is a member of the *Matonaviridae* family and contains positive sense ssRNA viruses (Baltimore classification group IV)[300]. The genus contains 3 species- *Rubivirus rubella*, also known as rubella virus (RubV), *Rubivirus ruteetense*, more commonly known as ruhugu virus (RuhV), and *Rubivirus strelense*, also known

as rustrela virus (RusV)[300]. The only known *Rubivirus* prior to 2020 was RubV and until 2018 this virus was classified as a member of the *Togaviridae*[301,302]. In 2020 RuhV and RusV were discovered, and since then other *Matonaviridae* viruses described as "Rubi-like" have been discovered from unexpected hosts such as the Pacific electric ray[301,303]. *Rubivirus* genomes are approximately 10kb long and contain 2 ORFS which encode 2 polyproteins, which are then in turn cleaved into 3 structural and 2 non-structural proteins[300,304].

RubV is a highly contagious virus that is endemic globally with a strict tropism for humans and an estimated 13 genotypes[300,305,306]. In 2019 an estimated 45,000 cases of rubella were reported worldwide, but the incidence of congenital rubella syndrome (CRS) is significantly higher at 100,000 confirmed and 236,000 estimated cases per year in LEDCs alone[301,307,308]. RubV is transmitted via aerosols, and RubV diagnosis is made in by a variety of methods but most commonly via serological assessment of IgG and IgM, or PCR testing[306]. Up to 50% of infections are asymptomatic or of sub-clinical significance and symptoms include low-grade fever, headache, and other cold like symptoms, often known as "German measles"[301,307,308]. There is no specific antiviral treatment for RubV infection and treatment is symptomatic[306].

CRS is a congenital birth syndrome that can cause heart defects, hearing loss, cataracts, and retinopathy and rarely neonatal death. CRS risk varies throughout pregnancy with a higher risk of CRS occurring if the infection is active during the first 12 weeks and a reduced risk following 20 weeks of pregnancy, although if the mother has an active rubella infection at the time of delivery the risk of CRS increases to 90%[308]. The exact mechanism of CRS mediated damage is unknown, although it is currently believed that the syndrome may develop due to apoptosis of key organs and immune system modulation by the virus[304]. Effective vaccines are available for RubV and are included in the routine vaccination of at least 150 countries, leading to eradication in an estimated 80 countries[303,307,308]. Vaccination is often delivered as part of the MMR vaccine and whilst one dose is considered protective, two doses are commonly delivered, with an effectiveness of 95% protection after one dose and 99% protection after two doses[307]. As the RubV vaccine is a live attenuated virus of the 1a vaccine strain, it can very rarely lead to reactivation and symptomatic rubella infection and should not be given to pregnant people[306,308].

RusV is a newly discovered *Rubivirus* with a broad host range in zoo animals including red-necked wallabies, South American coati and lions, and other animals including domestic cats, otters, donkeys, capybaras, wood mice and yellow-necked mice[91,301,302,305]. Whilst RusV was first discovered in 2020, studies on archival tissues of zoo lions have shown that RusV has been circulating since the 1980s, if not earlier[91,301]. Initially identified in Germany, RusV has since been found in Austria and Sweden, although there have been no reports of RusV beyond the Germanic region of mainland Europe[91,301]. RusV has been found in cases of fatal non-suppurative meningoencephalitis in all infected animals except for *Apodemus* mice, which implies

that *Apodemus* mice are potentially the reservoir hosts of RusV[91,301,302,305]. Due to the widespread range of *Apodemus* rodents and their proximity to humans the zoonotic potential of RusV should be assessed, although there is no evidence of zoonotic spillover into humans reported to date[301,302]. RuhV was also first identified in 2020 in healthy leaf-nosed bats (*Hipposideros cyclops)* in Uganda. The lack of disease in these bats suggests that these are potential reservoir hosts, and to date no RuhV spillover has been observed into any other species[301]. There is currently no information available on the transmission route of either RusV or RuhV and further research into these viruses is required to adequately assess spillover risk[301,302].

6. Project aims and overall approach

As wildlife (and particularly UK wildlife) is undersampled in terms of virus discovery, the underlying hypothesis for this project was that an in-depth investigation into both modern and historic wildlife would yield new information regarding viruses of key families, and potentially allow for the discovery of modern viruses. By identifying viruses in UK wildlife, our collective understanding of viral zoonotic risk would be improved in a manner that may aid with agricultural and medical decision-making process. The potential identification of novel viruses would also be important, as it would not only potentially shed further light on the evolution and phylogenetic relationships of key virus families and the relationships between them, but would also allow for the assessment of their pathogenic potential and allow for mitigation of the threat (if any) posed by these viruses before they caused significant disease and mortality.

To meet these aims, this project utilised two different sample types via three total approaches- the three "prongs" described in the title. The first group of samples is the liver and gut tissue of 140 freshly collected and well-preserved Welsh rodents of five species. The first prong was to screen these samples via degenerate PCR for viruses with zoonotic potential. The second prong involved using NGS to sequence these animals, in turn gaining a snapshot of the virome and possibly identifying novel or unexpected viruses. This prong also aimed to provide an estimate of viral positive proportion within these animals without pre-existing section bias. The third and final prong was to attempt to perform historic virus discovery, using preserved animals from the collection at the NHM. This would in turn involve advancing historic RNA extraction and library preparation techniques, followed by screening using both conventional PCR and metagenomics. Overall, these three prongs aimed to provide a greater insight into both the virome and viral prevalence of community rodents in the UK, and of historic viruses within Africa, with a hope to identify novel viruses and provide phylogenetic and evolutionary information of key viral species such as CoVs, hantaviruses and LASV.

Chapter 2- Materials and methods

1. Sample types and collection

a. Modern samples

Modern samples (i.e. fresh samples that have only been subjected to short-term preservation processes) were collected from 3 sites in Fongoch and Nant Y Mwyn regions of Wales across 3 expeditions by Dr. Andrea Sartorius (Figure 2). A total of 108 animals were sampled in May, September and October 2019, and a further 32 animals were sampled in September 2021, resulting in a total of 140 animals sampled. These included 13 bank voles (*Myodes glareolus*), 7 field voles (*Microtus agrestis*), 17 yellow-necked mice (*Apodemus flavicollis*), 100 wood mice (*Apodemus sylvaticus*) and 1 least weasel (*Mustela nivalis*). Animals were live trapped and euthanised, followed immediately by dissection, where a section of liver and a section of gut were taken from each animal (excluding one bank vole, where only the liver was taken). Only animals that appeared healthy upon basic inspection were sampled to prevent sample bias, and all animal experiments were approved by, and carried out in strict accordance with, thr University of Nottingham's experiments and ethics comittees and complied with the Home Office of Great Britain and Northern Ireland's Animals (Scientific Procedures) Act 1986. Samples were stored in RNAlater (Thermofisher, USA) to stabilise and preserve RNA. Samples were stored on dry ice during transport from the field to the collecting laboratory where they were stored at -20^{o}C until further transport to the Virology Research Group laboratory, where they were then stored at -70^{o}C.

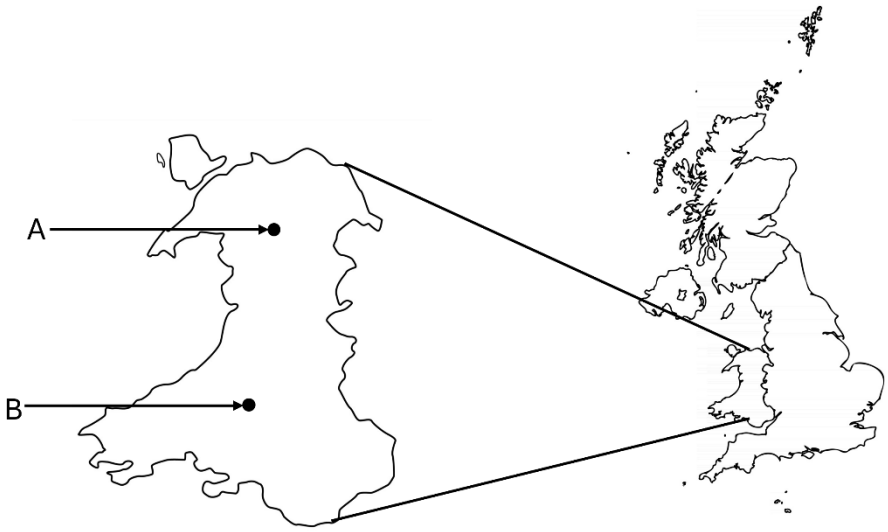


Figure 2- UK sampling sites.
A map of the UK and Wales showing the sampling sites for the UK rodents. A represents Frongoch, B represents Rhandrimwhyn, the closest village to the Nant Y Mwyn mine.

b. Historic samples

Animals were collected over many expeditions throughout the 19th and 20th centuries and were stored in sealed ethanol jars in the Natural History Museum archive at a cool room temperature. Animals sampled were selected from African specimens to represent a diverse range of sample ages (ranging from an estimated 40 years old to over 150 years old) species (Table 2). Animals (excluding sampling batch 2 and 3 specimens) were checked for pre-existing incisions into the chest, and if none were present a scalpel was used to open the chest cavity of the specimen. Small samples (approximately 1 cm^3) of liver and/or gut were then taken using a scalpel blade. A pilot collection of 6 samples was performed in 2019, where liver samples were collected from 2 Franquet's epauletted fruit bat (*Epomops franqueti*) specimens, 2 greater long-fingered bat (*Miniopterus inflatus*) specimens and 2 hammer-headed bat (*Hysignathus monstrosus*) specimens by Dr. Patrick McClure and Roberto Portela Miguez (NHM) using scalpels. A second collection was performed in early 2021, where gut samples were collected from the previously sampled *E. franqueti* by Roberto Portela Miguez and Darren Choonea (NHM) using a biopsy needle. A third collection was performed in October 2021, where gut samples were collected from 12 *E. franqueti* and 5 *M. natalensis* by Roberto Portela Miguez using a biopsy needle. A fourth collection was performed in December 2021, where gut and liver samples were taken from 5 previously sampled *E. franqueti* and 2 previously sampled *M. natalensis* specimens by Roberto Portela Miguez using a scalpel. Based on RNA extraction results and concerns about food contaminants in the gut, it was decided that all sampling from this point onwards would focus on liver samples. A fifth collection was performed in April 2022 where liver samples were taken from 4 previously sampled *E. franqueti* and 2 previously sampled *M. natalensis* specimens by Roberto Portela Miguez using a scalpel. A sixth and final (to date) collection was performed in May 2022, where liver samples from 3 *E. franqueti*, 2 *Epomops franqueti strepitans*, 9 Buettikofer's epauletted fruit bat (*Epomops buettikoferi*) specimens, 2 *M. africanus*, 4 *H. monstrosus*, 2 intermediate horseshoe bat (*Rhinolophus affinis superans*) specimens, 13 *M. natalensis*, 2 Guinea multimammate mouse (*Mastomys erythroleucus*) specimens, 1 *Mastomys coucha erythroleucus* (a specimen that may either be *M. erythroleucus* or *Mastomys coucha*, the Southern multimammate mouse), 1 *Mastomys* that could not be identified to the species level and 3 Tullberg's soft-furred mouse (*Praomys tullbergi*) specimens were collected by the author, Roberto Portela Miguez and Dr. Joseph Chappell using scalpels. All specimens from the sixth collection had not previously been sampled, and the aim was to collect as many samples as possible across a variety of specimen species. Samples were stored in O-ring tubes filled with ethanol taken from the sample jar and were double bagged in plastic storage bags. Samples were then transported at room temperature and stored in the Virology Research Laboratory at either room temperature or 4°C. All samples were collected within Africa, but due to limited

record keeping for historic samples, further geographic specificity is often unavailable.

Table 2- Historic specimen metadata.

List of specimens investigated and their species, minimum age, and sample tissue type. Sample species is as described on the jar for the specimen. Minimum age represents how long since the sample entered the NHM collection- it is not possible to provide specific ages for the samples due to archival information gaps resulting in censored data. Specimens collected prior to 1891 are considered likely not fixed, and those collected after 1891 are considered likely to have been fixed[309].

Sample name	Species	Tissue type	Sampling batch	Minimum age of sample (years)	Formalin fixation status
Epo1	*Epomops franqueti*	Liver	1	Unknown	Unknown
Epo2	*Epomops franqueti*	Liver	1	Unknown	Unknown
Min1	*Miniopterus inflatus*	Liver	1	Unknown	Unknown
Min2	*Miniopterus inflatus*	Liver	1	Unknown	Unknown
Hyp1	*Hypsignathus monstrosus*	Liver	1	Unknown	Unknown
Hyp2	*Hypsignathus monstrosus*	Liver	1	Unknown	Unknown
1966.3498G	*Epomops franqueti*	Gut	2	57	Likely fixed
1966.3498H	*Epomops franqueti*	Liver	2	57	Likely fixed
1966.3499G	*Epomops franqueti*	Gut	2	57	Likely fixed
1966.3499H	*Epomops franqueti*	Liver	2	57	Likely fixed
1984.1654	*Epomops franqueti*	Gut	3	39	Likely fixed
1984.1655	*Epomops franqueti Tomes*	Gut	3	39	Likely fixed
1966.3502	*Epomops franqueti Tomes*	Gut	3	57	Likely fixed
1966.3503	*Epomops franqueti Tomes*	Gut	3	57	Likely fixed
1966.3504	*Epomops franqueti Tomes*	Gut	3	57	Likely fixed
1966.3505	*Epomops franqueti Tomes*	Gut	3	57	Likely fixed
1966.3506	*Epomops franqueti Tomes*	Gut	3	57	Likely fixed
1880.7.21.3	*Epomops franqueti*	Gut	3	143	Not fixed
1880.7.21.1	*Epomops franqueti*	Gut	3	143	Not fixed
1880.7.21.4	*Epomops franqueti Tomes*	Gut	3	143	Not fixed
1948.598	*Epomops franqueti Tomes*	Gut	3	75	Likely fixed
1867.4.12.324	*Epomops franqueti Tomes*	Gut	3	156	Not fixed
1932.1.17.14	*Mastomys natalensis*	Gut	3	91	Likely fixed
1932.1.17.15	*Mastomys natalensis*	Gut	3	91	Likely fixed
1979.1229	*Mastomys natalensis*	Gut	3	44	Likely fixed

1979.1230	*Mastomys natalensis*	Gut	3	44	Likely fixed
1979.1240	*Mastomys natalensis*	Gut	3	44	Likely fixed
1966.3503G	*Epomops franqueti Tomes*	Gut	4	57	Likely fixed
1966.3503H	*Epomops franqueti Tomes*	Liver	4	57	Likely fixed
1966.3506G	*Epomops franqueti Tomes*	Gut	4	57	Likely fixed
1966.3506H	*Epomops franqueti Tomes*	Liver	4	57	Likely fixed
1880.7.21.1G	*Epomops franqueti*	Gut	4	143	Not fixed
1880.7.21.1H	*Epomops franqueti*	Liver	4	143	Not fixed
1880.7.21.4G	*Epomops franqueti Tomes*	Gut	4	143	Not fixed
1880.7.21.4H	*Epomops franqueti Tomes*	Liver	4	143	Not fixed
1979.1229G	*Mastomys natalensis*	Gut	4	44	Likely fixed
1979.1229H	*Mastomys natalensis*	Liver	4	44	Likely fixed
1979.1240G	*Mastomys natalensis*	Gut	4	44	Likely fixed
1979.1240H	*Mastomys natalensis*	Liver	4	44	Likely fixed
1867.4.12.324G	*Epomops franqueti Tomes*	Gut	4	156	Not fixed
1867.4.12.324H	*Epomops franqueti Tomes*	Liver	4	156	Not fixed
1966.3503H2	*Epomops franqueti Tomes*	Liver	5	57	Likely fixed
1966.3506H2	*Epomops franqueti Tomes*	Liver	5	57	Likely fixed
1979.1229H2	*Mastomys natalensis*	Liver	5	44	Likely fixed
1979.1240H2	*Mastomys natalensis*	Liver	5	44	Likely fixed
1880.7.21.4H2	*Epomops franqueti Tomes*	Liver	5	143	Not fixed
1984.1655H	*Epomops franqueti Tomes*	Liver	5	39	Likely fixed
66.3498	*Epomops franqueti*	Liver	6	57	Likely fixed
69.963	*Epomops buettikoferi*	Liver	6	54	Likely fixed
59.205	*Epomops buettikoferi*	Liver	6	64	Likely fixed
47.588/B62	*Epomops buettikoferi*	Liver	6	76	Likely fixed
79.428	*Epomops buettikoferi*	Liver	6	44	Likely fixed
62.1819	*Epomops buettikoferi*	Liver	6	61	Likely fixed
68.962	*Epomops buettikoferi*	Liver	6	55	Likely fixed
68.966	*Epomops franqueti strepitans*	Liver	6	55	Likely fixed
79.434	*Epomops buettikoferi*	Liver	6	44	Likely fixed
62.1818	*Epomops buettikoferi*	Liver	6	61	Likely fixed
68.357	*Epomops buettikoferi*	Liver	6	55	Likely fixed

66.3509	*Epomops franqueti*	Liver	6	57	Likely fixed
68.356	*Epomops franqueti strepitans*	Liver	6	55	Likely fixed
68.970	*Hypsignathus monstrosus*	Liver	6	55	Likely fixed
64.59	*Hypsignathus monstrosus*	Liver	6	59	Likely fixed
68.971	*Hypsignathus monstrosus*	Liver	6	55	Likely fixed
84.800	*Epomops franqueti*	Liver	6	39	Likely fixed
71.607	*Miniopterus inflatus africanus*	Liver	6	52	Likely fixed
79.424	*Hypsignathus monstrosus*	Liver	6	44	Likely fixed
72.285	*Miniopterus inflatus africanus*	Liver	6	51	Likely fixed
61.1642	*Rhinolophus affinis superans*	Liver	6	62	Likely fixed
61.16.43	*Rhinolophus affinis superans*	Liver	6	62	Likely fixed
76.1540	*Mastomys natalensis*	Liver	6	47	Likely fixed
79.1241	*Mastomys natalensis*	Liver	6	44	Likely fixed
76.1547	*Mastomys natalensis*	Liver	6	47	Likely fixed
48.1163	*Mastomys coucha/ Mastomys erythroleucus*	Liver	6	75	Likely fixed
79.1181	*Mastomys natalensis*	Liver	6	44	Likely fixed
79.1198	*Mastomys natalensis*	Liver	6	44	Likely fixed
1988.167	*Mastomys natalensis*	Liver	6	35	Likely fixed
79.1295	*Praomys tullbergi*	Liver	6	44	Likely fixed
79.1297	*Praomys tullbergi*	Liver	6	44	Likely fixed
79.1294	*Praomys tullbergi*	Liver	6	44	Likely fixed
79.1537	*Mastomys natalensis*	Liver	6	44	Likely fixed
1988.169	*Mastomys natalensis*	Liver	6	35	Likely fixed
77.3425	*Mastomys natalensis*	Liver	6	44	Likely fixed
77.3427	*Mastomys natalensis*	Liver	6	44	Likely fixed
68.652	*Mastomys erythroleucus*	Liver	6	55	Likely fixed
71.393	*Mastomys (Praomys) natalensis*	Liver	6	52	Likely fixed
71.402	*Mastomys (Praomys) natalensis*	Liver	6	52	Likely fixed
68.654	*Mastomys erythroleucus*	Liver	6	55	Likely fixed
12.11.25.16	*Mastomys natalensis*	Liver	6	111	Likely fixed
9.1.9.25	*Mastomys spp.*	Liver	6	114	Likely fixed

2. RNA extraction

a. Modern Rodents

For modern rodent gut and liver samples RNA extraction was performed using the GenElute^{TM} Mammalian Total RNA Miniprep kit (Sigma-Aldrich, USA) according to the protocol provided by the manufacturer with some modifications. The tissue in the lysis solution was homogenised using a ribolyser (Bio-Rad, USA) for 40 seconds at a speed of 6 and checked for remaining chunks of tissue. If tissue was present, the sample was ribolysed a second time as before. An extra spin at 14,000 rpm for 1 minute was added at the end of the protocol and the RNA was then transferred to a fresh collection tube to prevent fibres from the tube from interfering with RNA quantification. Samples were eluted in 50 µl of elution buffer, 1 µl of which was used for purity assessment and quantification. Quantification and purity assessment was then performed using a Nanodrop One spectrophotometer (Thermofisher, USA), where concentration was reported and the 260 nm/280 nm and 260 nm/230 nm absorbance ratios (detecting protein or phenol contamination respectively) were assessed, before RNA was stored at -70^{o}C. Liver samples were labelled as (sample number)H and gut samples as (sample number)G.

b. Historic samples

For historic samples, RNA extraction was performed using the Roche High Pure FFPET RNA Isolation kit (Roche, Switzerland) according to the protocol provided by the manufacturer with some modifications. Modifications included homogenisation following the addition of 100 µl of RNA tissue lysis buffer and 40 µl of proteinase K using a ribolyser for 40 seconds at a speed of 5. After homogenization samples were checked for chunks of tissue and the homogenisation was repeated if tissue chunks were present. This was performed up to 3 times as necessary, followed by the addition of the recommended 16 µl of 10% SDS and the continuation of the standard protocol. The second incubation (following the addition of extra proteinase K) was extended to 45 minutes to assist with tissue breakdown. Samples were eluted in 35 µl of elution buffer and stored at -70^{o}C. 1 µl was used for Nanodrop quantification and purity assessment as before. A further 2 µl was used for quantification and fragment size assessment using a TapeStation 4200 electrophoresis platform (Agilent, USA). Either a high sensitivity RNA screentape (Agilent, USA) was used according to the manufacturer instructions if the Nanodrop concentration was < 5 ng/µl, or a standard sensitivity screentape was used (Agilent, USA) if the Nanodrop concentration was ≥ 5 ng/µl.

3. PCR screening (modern samples)

a. cDNA synthesis

cDNA synthesis was performed using the RNA to cDNA EcoDry™ Premix with random hexamer primers (Takara, USA) according to the manufacturer's protocol with some modifications. Each EcoDry premix was reconstituted in 2 µl of Dep-C water (Sigma-Aldrich, USA) and split into 1 µl aliquots. For each sample, up to 19 µl of RNA was added to provide approximately 1500 ng of RNA where possible, or as close to 1500 ng as possible if the RNA concentration was too low. The reaction volume was increased to 20 µl using Dep-C water, and the reaction was performed according to the manufacturer instructions. Samples were stored at -20°C until further use.

b. Primer design

Primers were designed in Geneious Prime version 2022.1.1 (Dotmatics, USA). Reference sequences for human and animal viral species of the target genus and family were downloaded from NCBI Genbank (NCBI, USA) and were aligned in Geneious Prime using the Clustal Omega nucleotide aligner function. This alignment was then used to manually design primers targeting highly conserved regions of the target genomes to maximise virus capture. Primers were designed with an optimal GC content of 40-60% where possible, a length of approximately 17-23 nucleotides, high GC content at the 3' end known as a GC clamp, and an annealing temperature (Tm) of 55-58°C which matched the paired primers to within 2°C where possible, as this offers the highest chance of a successful PCR[61]. Primers were designed with the minimum required degeneracy to allow broad annealing to the target virus with the minimum possible hairpin likelihood, and a PCR product of between 300 and 1000bp where possible[61]. Figure 3 shows an example of a primer designed in this manner. All primers were synthesised by Sigma-Aldrich (Merck, USA), and degenerate screening primers were tested by using a gradient PCR- a PCR where a variety of annealing temperatures around the expected Tm are tested to find the optimal annealing temperature- using positive control DNA. All primer notation is according to standard IUPAC convention (Table 3).

Table 3- IUPAC base notation.

Standard IUPAC base notation was used throughout this project.

Single letter code	Corresponding bases
A	Adenine
C	Cytosine
G	Guanine
T	Thymine
R	A or G
M	A or C
W	A or T
S	C or G
Y	C or T
K	G or T
V	A, C or G
H	A, C or T
D	A, G or T
B	C, G or T
N	A, C, G or T

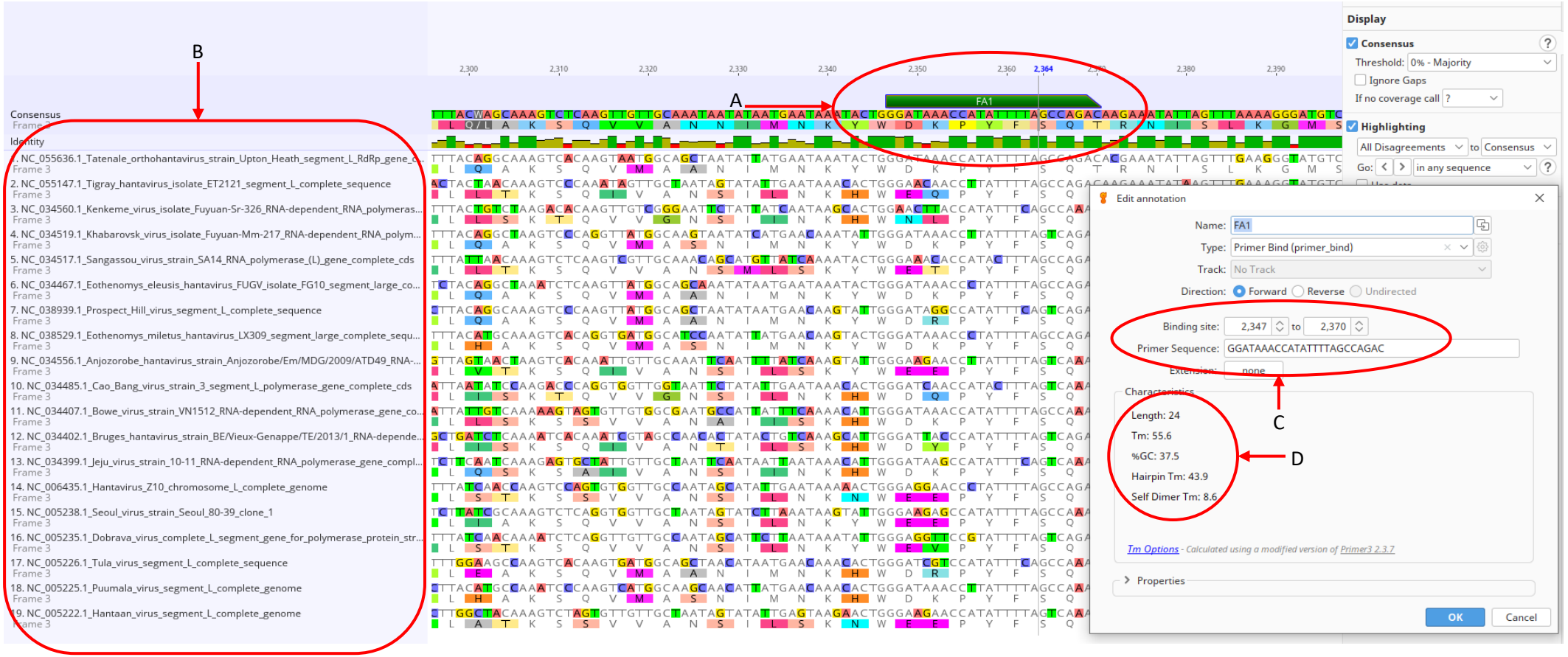


Figure 3- PCR primer design in Geneious prime.

A shows primer location within consensus sequence. Identity bar shows consensus across aligned samples (dark green represents ≥70% identity, pale green represents 30-70% identity, red represents ≤ 30% identity). B shows aligned sequences and their individual sequences. C shows primer sequence and location within alignment. D shows primer length, Tm, GC%, hairpin Tm and self-dimer Tm.

c. PCR result confirmation

All PCR results were visualised using agarose gel electrophoresis to view DNA band sizes. A 2% agarose gel with 5 µl of ethidium bromide was produced and submerged in a 10% TAE running buffer containing 30 µl of ethidium bromide. 2 µl of 10x FastDigest Green buffer loading dye (Theromofisher, USA) was added to each sample, and 7 µl of sample-buffer mix was added to the gel. 3.5µl of GeneRuler Mix DNA ladder (Thermofisher, USA) was added to assess band sizes, at no fewer than 1 ladder to every 10 samples. The gel then underwent electrophoresis at 90 volts for 36 minutes for expected products of < 500bp, or 42 minutes for expected products of ≥ 500bp, prior to visualisation using UV light and PCR band identification. Samples producing a product of the expected size were then sent to be sequenced by Sanger sequencing (Source Bioscience, UK), and samples were diluted either 1:5 or 1:10 in Dep-C water (for low and high visual intensity bands respectively) prior to postage as recommended by Source Bioscience. Sequence quality was then assessed using Geneious Prime, and sequence identity was confirmed using NCBI nucleotide BLAST (NTBLAST, NCBI, USA).

d. GAPDH screen

All samples underwent PCR screening for GAPDH, a ubiquitously expressed housekeeping gene which is often used as a positive control for RNA extractions which in this case doubles as a positive control for a successful cDNA synthesis reaction. For each reaction, 0.5 µl of cDNA was added to 0.05 pM of laboratory developed GAPDH forward primer (5'-CCATCTTCCAGGAGCGAGA), 0.05 pM of laboratory developed GAPDH reverse primer (5'-GCCTGCTTCACCACCTTCT), 0.25 mM of dNTPs (Sigma-Aldrich, USA), 1.24 µl of 10x buffer (Qiagen, Netherlands), 1 unit of hot-start Taq polymerase (Qiagen, Netherlands) and 9.7 µl of Dep-C water. The PCR cycles used were 95°C for 15 minutes, followed by 50 cycles of 94°C for 20 seconds, 55°C for 20 seconds and 72°C for 30 seconds, before a final step of 72°C for 10 minutes. Samples were then stored at 4°C until further use. A negative control reaction was included in all GAPDH PCRs.

e. Species identification

The liver sample for each animal underwent species identification via mitochondrial cytochrome B PCR, an established method for determining host species[29,64,100,103]. For each reaction, 0.5 µl of cDNA was added to 0.05 pM of CytoB forward primer (5'-TGAGGBGCYACAGTWATYACAAAC), 0.05 pM of CytoB reverse primer (5'-CGYAGGATDGCRTATGCRAATA)[310], 0.25 mM of dNTPs (Sigma-Aldrich, USA), 1.24 µl of 10x buffer (Qiagen, Netherlands), 1 unit of hot-start Taq polymerase (Qiagen, Netherlands), and 9.7 µl of Dep-C water. The PCR cycles used were 95°C for 15 minutes, followed by 50 cycles of 95°C for 20 seconds, 51°C for 20 seconds and 72°C for 1 minute, before a final step of 72°C for 10 minutes was performed. Samples were then stored at 4°C until further use. A negative control reaction was included in all CytoB PCRs.

f. α and β Coronavirus screen

3 sets of CoV primers were tested. The Woo_CoV_F (5'-GGTTGGGACTATCCTAAGTGTGA) and Woo_CoV_R (5'-CCATCATCAGATAGAATCATCATA), and Woo_Cov_2_F (5'-GHTGGGAYTAYCCTAARTGYGA) and Woo_CoV_2_R (5'-CCATCRTCAGWCARDATCATCATD) primer sets were initially tested[153]. The Woo_CoV primers were already validated but did not include CoVs identified since 2005, and the Woo_CoV_2 primers were designed *in silico* to include modern CoVs but were not yet validated. The third primer set was Alt_CoV_F (5'-GGWCCWCATGARTTTTGYTCNC) and Alt_CoV_R (5'-ACACAATARTAACTCADACCDCC), which was designed to attempt to confirm a potentially positive sample. A gradient PCR was performed for each primer set using an OC43 CoV as a template using reagent quantities of 1 µl of template cDNA, 0.05 pM of each primer, 0.25 mM of dNTPs, 1.24 µl of 10x buffer, 1 unit of hot-start Taq polymerase, and 9.7 µl of Dep-C water per reaction. These reagent quantities are the standard quantities for all PCR reactions throughout this project and were used for all PCRs unless otherwise stated. The PCR cycles used were 95°C for 15 minutes, followed by 50 cycles of 95°C for 20 seconds, annealing for 20 seconds and 72°C for 30 seconds, and a final step of 72°C for 1 minute. The annealing temperatures tested were 48.8°C, 49.8°C, 51.8°C, 54.6°C, 58°C, 62°C, 65.4°C or 68.1°C.

All rodent gut and liver samples underwent CoV screening using the Woo_CoV primers using the standard reagent quantities described above, a negative control and an OC43 CoV positive control. The PCR cycles used were 95°C for 15 minutes, followed by 50 cycles of 95°C for 20 seconds, 49°C (the optimal annealing temperature as shown by the gradient PCR) for 20 seconds and 72°C for 30 seconds, and a final step of 72°C for 1 minute.

One potentially positive wood mouse liver was screened with the Alt_CoV primers and Woo_CoV_2 primers, using standard reagent quantities and the same controls as used for the Woo_CoV screen. The cycles were as described above, except for annealing at 51°C and 49°C, respectively.

g. Old-World Hantavirus screen

All rodent gut and liver samples underwent Old-World Hantavirus screening using the established and validated HAN_L_F (5'- ATGTAYGTBAGTGCWGATGC) and HAN_L_R (5'- AACCADTCWGTYCCRTCATC) primers using the standard reagent quantities described above, a negative control and a known *Tatenale orthohantavirus* (TATV) positive sample as a positive control[175]. The PCR cycles used were 95°C for 15 minutes, followed by 50 cycles of 95°C for 20 seconds, 59°C for 20 seconds and 72°C for 30 seconds, and a final step of 72°C for 1 minute. Samples were also screened with the OW_HAN_F (5'- TGCWGATGCHACWAARTGGTC) and OW_HAN_R (5'- GGNAAYTGGCTRCARGGWAA) primers under the same conditions, except for annealing at 55°C and extending for 30 seconds. Potentially positive

samples were stored at 4°C for short term use or frozen at -20°C for long term storage.

h. Rubivirus screen

Rubivirus screening primers Rub_1_F (5'- AAYAACTGCCGCACBCC) and Rub_1_R (5'- TCCTCVAGYTCGTSGAG) were designed *in silico*. Very few reference sequences were available for *Rubivirus* genomes, thereofre the alignments were supplemented with other reported sequences to ensure appropriate primer target breadth[301]. A gradient PCR was performed for each primer set using a negative control and a rubellavirus vaccine extract sample as a positive control, with standard reagent quantities. The PCR cycles used were 95°C for 15 minutes, followed by 50 cycles of 95°C for 20 seconds, annealing for 20 seconds and 72°C for 30 seconds, and a final step of 72°C for 1 minute. The annealing temperatures tested were 54.4°C, 56.4°C, 58.8°C, 61.8°C, 65.4°C, 68.3°C, 70.7°C or 72.3°C. All rodent gut and liver samples underwent rubivirus screening using the Rub_1 primers and standard reagent quantities, with the same controls as above. The PCR cycles used were 95°C for 15 minutes, followed by 50 cycles of 95°C for 20 seconds, 59°C (the optimal annealing temperature as shown by the gradient PCR) for 20 seconds and 72°C for 30 seconds, and a final step of 72°C for 1 minute.

i. Adenovirus screen

4 sets of adenovirus primers were tested: AdHex_365_F (5'- GTGGAAYCMKGCDGTKGAC) and AdHex_365_R (5'- GTTCATGTASTCRTAGGTGTTB), Alt_Adeno_F (5'- CCMTTYAACCAYCMCCG) and Alt_Adeno_R (5'- GTRTTGYGNGCCATGGG), Alt_Adeno_2_F (5'- TWYACNCTGGCYGTGGG) and Alt_Adeno_2_R (5'- VCCRGCMARCACHCCCA), and Alt_Adeno_3_F (5'- TAYGCYAMYTTYTTCCCC) and Alt_Adeno_3_R (5'- CAGCTVACNGARGAGTC). A gradient PCR was performed for each primer set using a confirmed adenovirus positive clinical isolate as a positive control and standard reagent quantities. The PCR cycles used were 95°C for 15 minutes, followed by 50 cycles of 95°C for 20 seconds, annealing for 20 seconds and 72°C for 30 seconds, and a final step of 72°C for 1 minute for all primer sets. The annealing temperatures tested were 54.6°C, 56.4°C, 58.8°C, 61.8°C, 65.4°C, 68.3°C, 70.7°C or 72.3°C. Samples were visualised using agarose gel electrophoresis as previously described.

All rodent gut and liver samples underwent adenovirus screening using the Adhex_365 primers using the standard reagent quantities and positive control described above, as well as a negative control reaction. The PCR cycles used were 95°C for 15 minutes, followed by 50 cycles of 95°C for 20 seconds, 56°C (the optimal annealing temperature as shown by the gradient PCR) for 20 seconds and 72°C for 20 seconds, and a final step of 72°C for 1 minute. Potentially positive samples were stored at 4°C for short term use, or frozen at -20°C for long term storage. No samples were screened with the Alt_Adeno, Alt_Adeno_2 or Alt_Adeno_3 primers as they failed to produce a product at the gradient PCR stage at any temperature.

j. Rotavirus screen

2 sets of *Rotavirus* screening primers were tested- R_VP1_F (5'-CCAWTRGGAASAAGARATGTHCC) and R_VP1_R (5'- TRTGYTGMGABGMRTCCC) which were designed *in silico*, and the more established R_VP7_1_F (5'-CTCCTTTTAATGTATGGTATTGAATATACC) and R_VP7_1_R (5-RGTATAAAANACTTGCCACCATTTTTTCCA) primer set, which target the VP1 and VP7 segments of the *Rotavirus* genome respectively[290]. A gradient PCR was performed for each primer set using a *Rotavirus* positive clinical isolate sample as a positive control. cDNA was produced for the positive control using the same process as described for the samples, and a second batch of heat-inactivated control cDNA was produced by heating the RNA of the control to 95°C for 5 minutes before following the process as described for the samples. A gradient PCR was performed using both the heat-inactivated control and the normal control for each primer set, using standard reagent quantities. The PCR cycles used were 95°C for 15 minutes, followed by 50 cycles of 95°C for 20 seconds, annealing for 20 seconds and 72°C for 30 seconds, and a final step of 72°C for 1 minute. The annealing temperatures tested were 54.6°C, 56.4°C, 58.8°C, 61.8°C, 65.4°C, 68.3°C, 70.7°C or 72.3°C. No samples were screened with the R_VP1 or R_VP7 primer sets as they failed to produce a product at any temperature at the gradient PCR stage.

4. NGS investigations (modern samples)

a. Sample Pooling

Samples were pooled prior to NGS sequencing to reduce sequencing costs. Samples were stratified by tissue type (i.e. liver or gut), then stratified by sampling date, and then further stratified by species based on field identification and prior to species identification PCR. The only exception to this was Pool L, which included only the vole that could not be morphologically confirmed and the least weasel samples. The wood mouse samples and yellow-necked mouse samples were then further split into pools of 4-10 animals based on lead content in liver tissue, as this was of interest to a collaborating research group. Liver and gut pools were produced using the same constituent samples for each pool.

Once samples had been selected for a pool, sample cDNA was diluted in Dep-C water so that a total of 2 µg of RNA was contained within the pool. Each pool was then diluted in Dep-C water to give a total concentration of 67 ng/µl, and a total volume of 30 µl. This ensured that the sample was within the input range for the NGS library preparation kit (11 µl of sample, with a total RNA input of 5 ng-1 µg) whilst providing sufficient volume to perform 2 library preparations and all necessary quantification and quality control.

b. rRNA depletion and library preparation

rRNA depletion was performed to prevent excessive sequencing of sample rRNA, which would limit the sequencing of target viral RNA present within the sample[65]. rRNA depletion was performed using the NEBNext® Ultra™ II rRNA Depletion Kit v2 (New England Biolabs, USA) according to the manufacturer's protocol. NGS library preparation was performed using the NEBNext® Ultra™ II RNA Library Prep Kit for Illumina® (New England Biolabs, USA) according to the manufacturer's protocol and using SPRIselect beads (Beckmann Coulter, USA) during bead cleanup steps. Samples were then indexed using the NEBNext ® Multiplex Oligos for Illumina® (Index Primers set 3) (New England Biolabs, USA) for libraries A-L, or NEBNext® Multiplex Oligos for Illumina® (Index Primers set 4) (New England Biolabs, USA) for libraries M-R according to the manufacturer's instructions. Each library had a unique primer set assigned, and the liver and gut pools for each library had the same primer set assigned- for example, both the liver pool A samples and the gut pool A samples were assigned primer set 17.

c. Library quality control

To ensure the library was successfully synthesised, each library was quantified and sample purity was assessed using the Nanodrop One. The concentration of this was used to inform the operating parameters of the TapeStation. The TapeStation was used to assess library concentration and fragment size distribution prior to sequencing. Where the concentration given by the Nanodrop was ≥ 15 ng/µl, the standard sensitivity DNA screentape was used according to the manufacturer's instructions, and where the concentration given by the Nanodrop was < 15 ng/µl, the

high sensitivity DNA screentape was used according to the manufacturer's instructions. Library concentration and fragment sizes were assessed (a large, clean peak of between 240 and 300bp indicated the successful production of a standard 130-170bp insert library). The sample was also checked for the presence of excess of adaptor dimer, as indicated by a large peak of approximately 128-135bp. Where this was observed, TE buffer was added to the library to make the volume up to 50 µl and an additional 0.9x SPRISelect bead cleanup was performed as recommended by the manufacturer's instructions. Libraries that did not synthesise successfully in the first instance were repeated a second time to ensure successful library generation where reagent availability allowed. Before libraries were sent for sequencing, the gut and liver libraries for each sample set were then combined e.g. library A liver and library A gut libraries were combined to give a single library (library A). These combined libraries were then diluted to 15 nM as required by Genewiz. Samples were sent to Genewiz, where they underwent Illumina Nova-Seq sequencing using a single S4 sequencing lane.

d. Library analysis

Sequencing was downloaded from the Genewiz SFTP server using Cyberduck software and was imported into Geneious Prime[293,311]. Within Geneious Prime reads were paired using the default settings and were merged using the BBMerge plugin with default settings[312]. Reads also underwent quality control within Geneious to remove non-functional reads with a continuous repeated base. The merged reads and associated metadata were then uploaded to CZID using the command line interface.

e. Confirmatory PCR

The individual samples comprising each library were PCR screened for the presence of any likely viral hits from the CZID analysis. This made it possible to identify which sample(s) the hit came from and to provide some more genomic information for those viruses, and to confirm that the hit is a true positive for at least one sample from the library[22]. All screening was performed via endpoint PCR using standard reagent quantities unless otherwise stated (instances where qPCR screening was performed will be highlighted accordingly). Specific primers were designed and used for confirmatory screening according to the sequence information and contigs recovered from CZID. All endpoint PCR results were visualised using agarose gel electrophoresis as previously described. No positive controls were used throughout confirmatory PCR screening as all results of the correct band size were sequenced via Sanger sequencing as previously described, although negative control reactions were included in all screening PCRs. All PCR cycles were 95°C for 15 minutes, followed by 50 cycles of denaturation at 95°C for 20 seconds, annealing at the optimal primer temperature for 20 seconds, and extension at 72°C, followed by a final step of 72°C for 1 minute. qPCR reactions were performed using 20 µl of 2x SybrGreen (Thermofisher, USA), 0.08 pM of forward and reverse primer, 7.4 µl of dH_2O and 1 µl of cDNA template per sample. qPCR cycles were as follows: 95°C for 2 minutes,

followed by 40 cycles of 95°C for 5 seconds, annealing for 15 seconds at the optimal primer temperature and extension at 60°C for 10 seconds, followed by heating to 95°C for 10 seconds, cooling to 65°C for 5 seconds, and performing melt curve analysis by heating to 95°C in increasing 0.5°C increments at 1 second intervals. Sample quantification was performed after each cycle and after each temperature increase during the melt curve. All primers, primer sequences, annealing temperatures, extension times and product sizes are shown in Table 4.

Table 4- Primers used for NGS hit confirmation.

All primers, sequences, annealing temperatures, extension times and product sizes used in confirmatory PCR sequencing for modern sample analysis. Pan-Cardi_F and Pan-Cardi_R primers were designed and validated by Dr. Patrick McClure, all other primers were designed by the author for this project. Rot1 and Rot8 primers target *Rotavirus* segments 1 and 8, respectively. All *Picobirnavirus* primers were designed to target *Picobirnavirus* segment 2. TATVL primers targeted the L segment of TATV, and TATVM primers targeted the M segment of TATV.

Virus	Forward primer name	Forward primer sequence (5'-3')	Reverse primer name	Reverse primer sequence (5'-3')	Annealing temperature (°C)	Extension time (seconds)	Product size (bp)
Adenovirus	A-AdD-14113_F	GGGGTGAAGTTTGACACC	A-AdD-14351_R	AGAGCGGGTACGTTCC	55	30	238
	B-AdB-6969_F	GTTTTCAGCATACTCGGTGG	B-AdB-7281_R	GTTTAATGACATCACTTTTGGGC	57	30	312
	C-OAdC-19055_F	TCGGATCATCATCATCGTCC	C-OAdC-19620_R	AGAAGGAACTGTTGGGCG	57	45	565
	G-AdD-19965_F	TCATGTTCGACTCCAGCG	G-AdD-20301_R	TTCATGTAGGGGAAGCCG	57	30	336
	I-AdD-19347_F	CGACATCATGGACAACGTC	I-AdD-19925_R	GGAAGGTGTGGTTGAGG	57	45	578
	M-AdD-21334_F	GAGTCCGAGATCAAGCG	M-AdD-21689_R	GTCTGGGAGCTGACGG	55	30	355
	N-AdD-21082_F	GCAAATCAAACTCAAGTGCC	N-AdD-21689_R	GTCTGGGAGCTGACGG	56	45	607

	O-AdD-18404_F	TCGCCGCATCAGGAC	O-AdD-18777_R	TAGGGTTTGAAGGTGGGG	56	30	373
	O-AdD-12533_F	GACTTCCTGGCGTTCGT	O-AdD-12846_R	ACAGGTTCATCAAGTGCCC	57	30	313
Arterivirus	L-Art-14646_F	GAACGGGACACTGGTTC	L-Art-14843_R	CCTGGGTCTGGCGC	55	10	197
	N-Art-14586_F	CTACCCATACCAACAACTGAC	N-Art-14901_R	GCCAAGGGAAAATGAGGC	56	30	315
Astrovirus	B-Ast-3228_F	GATGCTGGTGACATAAGGC	B-Ast-3687_R	CAAATTCAAAACACTGCAGCC	57	30	459
	I-Ast-2638F	GAACAACGTGTCAGGCC	I-Ast-3241R	GCCTGATATCTGCCTCCTC	58	45	603
	K-Ast-768_F	GGAGATTGTTAGGTTCTTGGC	K-Ast-1394_R	GGAAATGTGAAGTAGTGCAGC	57	45	626
	M-Ast-4596_F	ACAGTCAGCAGTCTCTGG	M-Ast-4977_R	TAGATTGATCAGAGGAGCCC	56	30	381
	P-Ast-3912_F	CTATATGCTGGGCTAATAAAACC	P-Ast-4142_R	ATAAGCAGATTGACTTGGTCC	55	30	230

Bocaparvovirus	D-Boc-10_F	TCTGAACAACTTTCAAGCCG	D-Boc-312_R	CGAGTGCGAAGAAAGAGG	56	30	302
	P-Boc-470_F	GGAATGTGGAAAAGAATATCCTG	P-Boc-865_R	AATTCTGTGCTTCCTTCTTGT	56	30	395
Chapparvovirus	P-ChaP-689_F	CGATTCTTTACCCGAACTTGAG	P-ChaP-784_R	GTAGAGCGCTGTCAAAGG	56	10	95
Coronavirus	J-CoV-23225_F	AAGCAATCGTTATTATAATGACAGCCAG	J-CoV-23327_R	TCAAATCTACGCCCGTTGAAGACC	61	10	102
	J-CoV-14272_F	TTGCAGTGTTGTAAATGGGAAG	J-CoV-14394_R	GTACTTCAACGCTAAATGCTGT	57	10	122
Cytomegalovirus	I-CMV-174345F	TCAGACACATACCCTACCG	I-CMV-174701R	GCTATATGCCAATACACTGTCC	56	30	356
	I-CMV-174579_F	AAACTGCGGGCACTGG	I-CMV-174949_R	ATTCCCCGTGCCAAGAG	56	30	370
Dependoparvovirus	O-AAV-2937_F	GTCATTACCAAGTCCACCAG	O-AAV-3292_R	TTGCTGGCTACGTACGG	56	30	325
OW hantavirus	I-TATVL-4121F	GGTCTCGAGATAACTTGTTGG	I-TATVL-4342R	CCATGTCACTTTCTGTAGGC	56	30	221

	L-TATVL-3072_F	TCGATATATTGATGGAATGGAGG	L-TATVL-3514_R	GGAACTCAGCATTTGTAGGAG	56	30	442
	M-TATVM-1351_F	GTGATTTTGACTAAGACACTG	M-TATVM-1458_R	AACACAAAGCTCAACTGC	53	30	107
	N-TATVL-3623_F	CGCAGAGTAGGTGTGTC	N-TATVL-4019_R	CGCAGAGTAGGTGTGTC	55	30	396
	N-TATVL-5092_F	GTACAAGCAAACAGAAGGGC	N-TATVL-5624_R	CCAACAGCTTGTAAGGCC	57	45	532
Hepacivirus F	A-HepF-7583_F	ATGACCATCGCTGAGGC	A-HepF-8322_R	TTGTGACTCAAACACGCAC	57	45	739
	B-HepF-2127_F	TCTTACAAGGGCTCGTGG	B-HepF-2394_R	CAGCAACAATCTCAGGGG	56	30	267
	M-HepF-3761_F	AACTGGATCAGGCAAGACC	M-HepF-4251_R	ATGGAGTGAATATAAGATGGCG	56	30	490
	N-HepF-7761_F	CATCATGCCCAAAAGTGAGG	N-HepF-8157_R	CCTTGCATCACCATCGG	56	30	396
Murine Leukemia Virus	A-MLV-7041_F	CTTCAATGAATTAAGATTTGAGGG	A-MLV-7380_R	AGAAGCAAAATTAGGAGTGGTC	55	30	339
	C-MLV-4016_F	TGAACTGATAGCACTCACCC	C-MLV-4324_R	GTCTCTGTGATGGCTGCC	57	30	308

	I-MLV-6426_F	GCTCTCCCAGATCCGAC	I-MLV-6717_R	GGTCGTATTACAAAGGGCC	56	30	291
Orbivirus	E-Orb-2427_F	AGACGACATTTCTACCCGG	E-Orb-2641_R	CTTAAACAATAGTATCAGTTGGGC	56	30	214
Paramyxovirus	C-Jei-2660_F	TCATGCGGAGAATGCCC	C-Jei-2860_R	AGAGTCCTAGGCTAAGATTGG	55	30	200
	C-Para-16102_F	GATAAGTCATTAACCGGCGC	C-Para-16560_R	CGTTGTTGATCCAAAGTAAGGC	58	30	458
	D-Para-17276_F	GACAGAATTTAAAGTACTGGCC	D-Para-17584_R	TGTGTATTGATGTTAGGGCC	55	30	308
	N-Jei-8196_F	ATCTATATGACAGAATCTGGTGG	N-Jei-8317_R	AATCAGTTTTGGGTTTGAGGT	56	10	121
Pegivirus	L-Peg-4676_F	GTCAAACCCTGGGCCAC	L-Peg-4869_R	GTAGTGACCACGGAAGGG	57	10	193
	N-Peg-5219_F	CCTTTCTACGGCGTAACC	N-Peg-5523_R	TCTAGGGAAATGGAAAAGGTAG	55	30	304
Picobirnavirus	G-PiBiS2-811_F	GATGMGGTTGATAGGCGC	G-PiBiS2-1493_R	GCAAATTGATCACGTAGAGG	57	45	682

	H-PiBiS2-259_F	GGAAGACTCAAATGACACGC	H-PiBiS2-658_R	TGCGTCTTGAACTTCCAGC	57	30	399
	I-PiBiS2-1367_F	GCCTGAGATATCTCGAACGG	I-PiBiS2-1638_R	CTCGTTTATACCAGTGTTGGAC	57	30	316
	L-PiBiS2-566_F	ACTTCAAATAGGTGGGGAGC	L-PiBiS2-820_R	CCACAACATCGTCTTTCCC	57	30	254
	M-PiBiS2-788_F	CAGCATGGGTTAGCATGG	M-PiBiS2-892_R	GGTCAAACTTGGAGAAATCG	55	30	104
	O-PiBiS2-711_F	TCCATTCTCGGTAAACATAGC	O-PiBiS2-1475_R	TTACGTAGAGGATGATACTTCAC	55	50	764
	P-PiBiS2-1213_F	GTACAATCGTATACTAGCCATGG	P-PiBiS2-1666_R	ACCAATTTTCTATCCCAGACG	56	30	453
Picornavirus-*Cardiovirus*	A-Pic-19_F	TTTATATAAAGTAGATCTTCTACCCTC	A-Pic-697_R	TGGCAACTCTACAGAAGGG	55	45	678
	B-Pic-1980_F	ATTCCTCAGGCTCTTCCC	B-Pic-2432_R	TCATCATTGCAGTTCCTGTG	56	30	452
	D-Pic-4377_F	TCCACCACGGTTTCCC	D-Pic-4944_R	GCCATCAGGATTTTGCCC	56	45	567
	G-EMCV-810_F	TTTCCCTTTTGAAAACCACG	G-EMCV-1511_R	AATACATGGGGGAGGGG	55	45	701

	J-Pic-7528_F	GTTTTGACAGGGTTGACGG	J-Pic-7770_R	ATCTTCACCTAAATAGATTCAACC	56	30	242
	M-Pic-3046_F	TCCCAAGATGATGCCGC	M-Pic-3531_R	CCGTTGCGAAGAGTTGG	56	30	485
	N-Pic-248_F	GTTGTCCAGCGTTAATGGG	N-Pic-461_R	GTTGTCCAGCGTTAATGGG	57	10	213
	N-Pic-439_F	AAGGTGAAGAAAGTAAGTTTTGG	N-Pic-1543_R	GGCTTGCTGGGATTGTG	56	70	1104
	Q-Pic-565_F	TAGGATCCACTGCTGAAGG	Q-Pic-1104_R	TGTGGTTCCATTTCGATATCC	56	45	539
	R-Pic-2177_F	TTTGACTTCATGACAGGGGA	R-Pic-2578_R	GTGGGTAGGGGAAATAAAAGG	56	30	401
	Pan-Cardi_F	GGyCkAAGCCGCTTGGAATA	Pan-Cardi_R	GCTTTTGGCCSCAGAGG	60	30	341
Picornavirus-*Enterovirus C*	M-Ent-4599_F	CAGTTCTCGGTGATCATGG	M-Ent-4697_R	GTTGGGAAGGAAATGGGTAG	56	10	98
Picornavirus-*Kunsagivirus*	P-Pic-6405_F	TTTTCGCATGCAATTTCTATCC	P-Pic-6872_R	AATGCTGGAACATCAAGTGG	56	30	467

Picornavirus-*Parechovirus*	L-PEV-4140_F	GGTTCTGACTTTATGGATGGC	L-PEV-5100_R	CATTTTTTCCAGTGATGTAGGC	56	60	960
Picornavirus-*Rosavirus*	B-Ros-6869_F	TGAAAAAACAGCCTGCCG	B-Ros-7632_R	CAGCTTTCCTTTGTAGTGGTG	57	50	763
	M-Ros-7321_F	TCGCTCCCCCACATCA	M-Ros-7528_R	CAAAGCAGCACGAAGGC	57	30	207
Picornavirus-*Sapelovirus*	G-Sap-4607_F	CGATTGTGGTTGTGCTGC	G_Sap_4816R	AGTTGTAAAATCCTCACCATCC	57	30	209
Polyomavirus	F-Poly-1710_F	GATACAGTTAAAGCAAACCAGC	F-Poly-2145_R	CATTTTTTGCAGGATCTGGG	56	30	435
Protoparvovirus	F-PrPa-2536_F	CTCCTGCTGATCAACGC	F-PrPa-3336_R	TGTTAGAATCAAGAGCCACC	55	50	800
	I-PrPa-3850F	AGGTTCCAGTAGTACCAGC	I-PrPa-4292R	AGTCTAACTAAAAGTTGACCTGG	56	30	442
	I-PrPa-2375F	TCACTTGGTTCTAGGTTGGG	I-PrPa-2734R	GCTTGGTTAATGAAAATGTGAGC	57	30	359
	O-PrPa-3277_F	CCATTAAGGTTTACAACAATGACC	O-PrPa-3409_R	GGTACTGTTGGTTTCCATGGA	57	10	132
Rodent *Hepacivirus*	N-RHep-194_F	ACTGCGAGTGAAGCCG	N-RHep-471_R	CCCGAGTTCACAGCCTG	57	30	277

Rotavirus	B-Rot1-3142_F	CGATCAACTTCTACTTGACGC	B-Rot1-3700_R	CTCAGCTTCAAACTAATGGCG	57	45	558
	B-Rot8-385_F	GTGTTTGGTTTATATGAAGGGAG	B-Rot8-1023_R	CAGCTTCAAACTAATGGCG	55	45	638
Rhadinovirus	Q-Rhad-6777_F	GTTTGATTCTGCCACCACC	Q-Rhad-7093_R	TGTCTACCCACTCTTCTGGA	57	30	316

Library A adenovirus screening was performed with the primers A-AdD-14113_F and A-AdD_14351_R. Initial *Hepacivirus F* screening was performed with the A-HepF-7583_F and A-HepF-8322_R primers, and subsequent *Hepacivirus F* screening was performed with the B-HepF-2127_F and B-HepF-2394_R primers. MLV screening was performed with the primers A-MLV-7041_F and A-MLV-7380_R initially, and subsequently with the primers C-MLV-4016_F and C-MLV-4324_R. Picornavirus (*Cardiovirus*) screening was performed with the primers A-Pic-19_F and A-Pic-697_R.

Library B adenovirus screening was performed with the primers B-AdB-6969_F and B-AdB-7281_R. Initial astrovirus screening was performed using the B-Ast-3228_F and B-Ast-3687_R primers, and subsequent screening was performed with the K-Ast-768_F and K-Ast-1394_R primers. *Hepacivirus F* screening was performed using the primers B-HepF-2127_F and B-HepF-2394_R. Picornavirus (*Cardiovirus*) screening was performed with the B-Pic-1980_F and B-Pic-2432_R primers, and picornavirus (*Rosavirus*) screening was performed with the B-Ros-6869_F and B-Ros-7632_R primers. *Rotavirus* screening was performed by denaturing the target DNA at 95^{o}C for 5 minutes, followed by screening with the B-Rot1-3142_F and B-Rot1-3700_R primers alongside the B-Rot8-395_F and B-Rot8-1023_R primers.

Library C adenovirus screening was initially performed with the A-AdD-14113_F and A-AdD-14351_R primers, followed by subsequent screening with the C-oAdC-19055_F and C-oAdC-19620_R primers and the I-AdD-19347_F and I-AdD-19925_R primers. Initial astrovirus screening was performed with the K-Ast-768_F and K-Ast-1394_R primers, followed by subsequent screening using the B-Ast-3288_F and B-Ast-3687_R primers and the I-Ast-2638_F and I-Ast-3241_R primers. MLV screening was performed using the C-MLV-4016_F and C-MLV-4324_R primers. Initial paramyxovirus screening was performed using both the C-Para-16102_F and C-Para-16560_R primers and the C-Jei-2660_F and C-Jei-2860_R primer sets, followed by subsequent screening using the D-Para-17276_F and D-Para-17584_R primers.

Library D astrovirus screening was performed with the K-Ast-768_F and K-Ast-1394_R primers. *Bocaparvovirus* screening was performed with the D-Boc-10_F and D-Boc-312_R primers. MLV screening was initially performed with the A-MLV-7041_F and A-MLV-7380_R primers, and subsequent screening was performed with the I-MLV-6426_F and I-MLV-6717_R primers, as well as the C-MLV-4016_F and C-MLV-4324_R primers. Paramyxovirus screening was initially performed with the C-Para-16102_F and C-Para-16560_R primers, and subsequent screening was performed with the D-Para-17276_F and D-Para-17584_R primers. Picornavirus (*Cardiovirus*) screening was initially performed with the D-Pic-4377_F and D-Pic-4944_R primers, followed by the A-Pic-19F and A-Pic-607_R primers. Further *Cardiovirus* screening was performed using the B-Pic-1980_F and B-Pic-2432_R primers, as well as the G-EMCV-810_F and G-EMCV-1511_R primers.

Library E adenovirus screening was initially performed using the A-AdD-14113_F and A-AdD-14351_R primers, followed by subsequent screening using the B-AdB-6969_F

and B-AdB-7281_R primers as well as the I-AdD-19347_F and I-AdD-19925_R primers. Astrovirus screening was performed with the I-Ast-2638_F and I-Ast-3241_R primers. Initial hantavirus screening was performed with the I-TATVL-4121_F and I-TATVL-4342_R primers, followed by subsequent screening using the L-TATVL-3072_F and L-TATVL-3514_R primers and the N-TATVL-5092_F and N-TATVL-5624_R primers. *Hepacivirus F* screening was initially performed with the A-HepF-7583_F and A-HepF-8322_R primers, followed by subsequent screening using the B-HepF-2127_F and B-HepF-2394_R primers and the M-HepF-3761_F and M-HepF-4251_R primers. MLV screening was performed with the C-MLV-4016_F and C-MLV-4324_R primers. *Orbivirus* screening was performed with E-Orb-2427_F and E-Orb-2641_R primers. Initial *Picobirnavirus* screening was performed using both the I-PiBiS2-1367_F and I-PiBiS2-1638_R and G-PibiS2-811_F and G-PiBiS2-1493_R sets, and subsequent testing was performed using the L-PiBiS2-566_F and L-PiBiS2-820_R and O-PiBiS2-711_F and O-PiBiS2-1475_R primer sets. The following primer sets were using in order for Picornavirus (*Cardiovirus*) screening: A-Pic-19_F and A-Pic607_R, B-Pic-1980_F and B-Pic-2432_R, D-Pic-4377_F and D-Pic-4944_R, M-Pic-3046_F and M-Pic-3531_R.

Library F astrovirus screening was performed using the K-Ast-768_F and K-Ast-1394_R primers. Initial hantavirus screening was performed with the I-TATVL-4121_F and I-TATVL-4342_R primers, followed by subsequent screening using the L-TATVL-3072_F and L-TATVL-3514_R primers and the N-TATVL-5092_F and N-TATVL-5624_R primers. MLV screening was performed using the C-MLV-4016_F and C-MLV-4324_R primers. *Picobirnavirus* screening was performed using the G-PiBiS2-811_F and G-PiBiS2-1493_R primers. Polyomavirus screening was performed using the F-Poly-110_F and F-Poly-2145_R primers. *Protoparvovirus* screening was performed using both the F-PrPa-2536_F and F-PrPa-3336_R primer set and the I-Prpa-3850_F and I-PrPa-4292_R primer set.

Library G adenovirus screening was performed using the G-AdD-19965_F and G-AdD-20301_R primers. *Picobirnavirus* screening was performed using the G-PiBiS2-811_F and G-PiBiS2-1493_R primers. The following primer pairs were used in order for picornavirus (*Cardiovirus*) screening: A-Pic-19_F and A-Pic-607_R, B-Pic-1980_F and B-Pic-2432_R, D-Pic-4377_F and D-Pic-4944_R, G-EMCV-810_F and G-EMCV-1511_R, M-Pic-3046_F and M-Pic-3531_R. Picornavirus (*Sapelovirus*) screening was performed using the G-Sap-4607_F and G-Sap-4816_R primers.

Initial library H adenovirus screening was performed with the A-AdD-14113_F and A-AdD-14351_R primers, followed by subsequent screening with the B-AdB-6969_F and B-AdB-7281_R primer set and the I-AdD-19347_F and I-AdD-19925_R primer set. Astrovirus screening was performed with the I-Ast-2638_F and I-Ast-3241_R primers. MLV screening was initially performed with the I-MLV-6426_F and I-MLV-6717_R primers, followed with subsequent screening with the C-MLV-4016_F and C-MLV-4324_R primers. Initial *Picobirnavirus* screening was performed with the I-PiBiS2-1367_F and I-PiBiS2-1638_R primers, followed by subsequent screening with the G-

PiBiS2-811_F and G-PiBiS2-1493_R primer set and the H-PiBiS2-259_F and H-PiBiS2-658_R primers.

Initial library I adenovirus screening was performed using the I-AdD-19347_F and I-AdD-19925_R primers, followed by subsequent screening with A-AdD-14113_F and A-AdD-14351_R. Astrovirus screening was performed with the I-Ast-2638_F and I-Ast-3241_R primers. CMV screening was performed using the I-CMV-174345_F and I-CMV-174701_R primers, followed by subsequent screening using the I-CMV-174579_F and I-CMV-174949_R primers. Initial hantavirus screening was performed using the I-TATVL-4124_F and I-TATVL-4342_R primers, followed by subsequent screening using both the L-TATVL-3072_F and L-TATVL-3514_R primer set and the N-TATVL-5092_F and N-TATVL-5624_R primer set. Initial *Hepacivirus F* screening was performed with the A-HepF-7583_F and A-HepF-8322_R primers, followed by subsequent screening with both the B-HepF-2127_F and B-HepF-2394_R primer set, and the M-HepF-3761_F and M-HepF-4251_R primer set. Initial MLV screening was performed with the I-MLV-6426_F and I-MLV-6717_R primers, followed by subsequent screening with the C-MLV-4016_F and C-MLV-4324_R primers. Initial *Picobirnavirus* screening was performed with the I-PiBiS2-1367_F and I-PiBiS2-1638_R primers, followed by subsequent screening with the following primer sets: G-PiBiS2-811_F and G-PiBiS2-1493_R, L-PiBiS2-566_F and L-PiBiS2-820_R, O-PiBiS2-711_F and O-PiBiS2-1475_R. Initial picornavirus (*Cardiovirus*) screening was performed with the A-Pic-19_F and A-Pic-607_R primers, followed with subsequent screening by both the B-Pic-1980_F and B-Pic-2432_R primer set and the G-EMCV-810_F and G-EMCV-1511_R primer set, as well as the Pan-Cardi_F and Pan-Cardi_R primers. *Protoparvovirus* screening was performed using the I-PrPa-3850_F and I-PrPa-4292_R primers.

Library J adenovirus screening was performed with the A-AdD-14113_F and A_AdD-14351_R primers. Astrovirus screening was performed with the K-Ast-768_F and K-Ast-1394_R primers. Coronavirus screening was performed using the both J-CoV-23225_F and J-CoV-23227_R primer set and the J-Cov-14272 and J-CoV-14394_R primer set via qPCR. MLV screening was performed using the C-MLV-4016_F and C-MLV-4324_R primers. Initial *Picobirnavirus* screening was performed with the G-PiBiS2-811_F and G-PiBiS2-1493_R primers, followed by subsequent screening with both the L-PiBiS2-566_F and L-PiBiS2-820_R primer set and the H-PiBiS2-259_F and H-PiBiS2-658_R primer set. Picornavirus (*Cardiovirus*) screening was performed with the J-Pic-7528_F and J-Pic-7770_R primers.

Initial library K adenovirus screening was performed with the A-AdD-14113_F and A-AdD-14351_R primers, followed by subsequent screening with both the I-AdD-19347_F and I-AdD-19925_R primer set and the B-AdB-6969_F and B-AdB-7281_R primer set. Astrovirus screening was performed with the K-Ast-768-F and K-Ast-1394-R primers. Initial MLV screening was performed with the C-MLV-4016_F and C-MLV-4324_R primer set, followed with subsequent screening by both the A-MLV-7041_F and A-MLV-7380_R primer set and the I-MLV-6426_F and I-MLV-6717_R primer set.

Initial library L arterivirus screening was performed via qPCR using the L-Art-14646_F and L-Art-14843_R primers, followed by subsequent screening via endpoint PCR using the N-Art-14586_F and N-Art-14901_R primers. Astrovirus screening was performed using both the K-Ast-768_F and K-Ast-1394_R primer set and the I-Ast-2638_F and I-Ast-3241_R primer set. Hantavirus screening was performed using the L-TATVL-3072_F and L-TATVL-3514_R primers. Initial *Hepacivirus F* screening was performed using the B-HepF-2127_F and B-HepF-2394_R primers, followed by subsequent screening using the A-HepF-7583_R and A-HepF-8322_R primers. MLV screening was performed using the C-MLV-4016_F and C-MLV-4324_R primers. Picornavirus (*Parechovirus*) screening was performed with the L-PEV-4140_F and L-PEV-5100_R primers. *Picobirnavirus* screening was performed using the L-PiBiS2-566_F and L-PiBiS2-820_R primers. *Pegivirus* screening was performed via qPCR using the L-Peg-4676_F and L-Peg-4869_R primers.

Initial library M adenovirus screening was performed using the M-AdD-21334_F and M-AdD-21689_R primers, followed by subsequent screening using both the O-AdD-18404_F and O-AdD-18777_R primer set and the I-AdD-19347_F and I-AdD-19925_R primer set. Astrovirus screening was performed with the M-Ast-4596_F and M-Ast-4977_R primers. Initial hantavirus screening was performed using the M-TATVM-1351_F and M-TATVM-1458_R primers by qPCR, followed by subsequent screening using the N-TATVL-5092_F and N-TATVL-5624_R primers. *Hepacivirus F* screening was performed using the M-HepF-3761_F and M-HepF-4251_R primers. Initial *Picobirnavirus* screening was performed using the M-PiBiS2-788_F and M-PiBiS2-892_R primers, followed by subsequent screening using the O-PiBiS2-711_F and O-PiBiS2-1475_R primers. Picornavirus (*Cardiovirus*) screening was performed using the M-Pic-3046_F and M-Pic-3531_R primers. Picornavirus (*Enterovirus C*) screening was performed via qPCR using the M-Ent-4599_F and M-Ent-4697_R primers. Initial picornavirus (*Rosavirus*) screening was performed using the M-Ros-7321_F and M-Ros-7528_R primers, followed by subsequent screening with the B-Ros-6869_F and B-Ros-7632_R primers.

Initial library N adenovirus screening was performed using the N-AdD-21082_F and N-AdD-21689_R primers, followed by subsequent screening with the following primer pairs: M-AdD-21334_F and M-AdD-21689_R, O-AdD-18404_F and O-AdD-18777_R, I-AdD-19347_F and I-AdD-19925_R. Arterivirus screening was performed using the N-Art-14586_F and N-Art-14901_R primers. Initial astrovirus screening was performed using the M-Ast-4596_F and M-Ast-4977_R primers, followed by subsequent screening using the O-Ast-3194_F and O-Ast-3502_R primers. Initial hantavirus screening was performed using the N-TATVL-3623_F and N-TATVL-4019_R primers, followed by subsequent screening using the N-TATVL-5092_F and N-TATVL-5624_R primers. *Hepacivirus F* screening was performed using the N-HepF-7761_F and N-HepF-8157_R primers. Paramyxovirus screening was performed via qPCR using the N-Jei-8196_F and N-Jei-8317_R primers. *Pegivirus* screening was performed using the N-Peg-5219_F and N-Peg-5523_R primers. *Picobirnavirus*

screening was performed using the O-PiBiS2-711_F and O-PiBiS2-1475_R primers. Initial picornavirus (*Cardiovirus*) screening was performed using the N-Pic-248_F and N-Pic-461_R primers via qPCR, followed by subsequent screening with the Q-Pic-565_F and Q-Pic-1104_R primers via endpoint PCR. Rodent *Hepacivirus* screening was performed with the N-RHep-194_F and N-RHep-471_R primers.

Initial library O adenovirus screening was performed using the O-AdD_18404_F and O-AdD-18777_R primers, followed by subsequent screening using both the N-AdD-21082_F and N-AdD-21689_R primer set and the I-AdD-19347_F and I-AdD-19925_R primer set. Initial astrovirus screening was performed using the O-Ast-3194_F and O-Ast-3502_R primers, followed by further screening using the P-Ast-3912_F and P-Ast-4142_R primers. *Dependoparvovirus* screening was performed using the O-AAV-2937_F and O-AAV-3292_R primers. MLV screening was performed using the C-MLV-4016_F and C-MLV-4324_R primers. *Picobirnavirus* screening was performed using the O-PiBiS2-711_F and O-PiBiS2-1475_R primers. Initial picornavirus (*Cardiovirus*) screening was performed using the N-Pic-248_F and N-Pic-461_R primers via qPCR, followed by further screening using both the Q-Pic-565_F and Q-Pic-1104_R primer set and the M-Pic-3046_F and M-Pic-3531_R primer set via endpoint PCR. Initial *Protoparvovirus* screening was performed using the O-PrPa-3277_F and O-PrPa-3409_R primers, followed by subsequent screening using the I-PrPa-3850_F and I-PrPa-4292_R primers. *Rhadinovirus* screening was performed using the O-Rhad-6233_F and O-Rhad-6679_R primers.

Library P adenovirus screening was performed using the O-AdD-18404_F and O-AdD-18777_R primers. Astrovirus screening was initially performed using the P-Ast-3912_F and P-Ast-3912_R primers, followed by subsequent screening using both the M-Ast-4596_F and M-Ast-4977_R primer set and the O-Ast-3194_F and O-Ast-3502_R primer set. *Bocaparvovirus* screening was performed using the P-Boc-470_F and P-Boc-865_R primers. Chapparvovirus screening was performed via qPCR using the P-ChaP-689_F and P-ChaP-784_R primers. *Hepacivirus F* screening was performed using the M-HepF_3761_F and M-HepF-4251_R primers. *Picobirnavirus* screening was initially performed using the P-PiBiS2-1213_F and P-PiBiS2-1666_R primers, followed by subsequent screening using the O-PiBiS2-711_F and O-PiBiS2-1475_R primers. Picornavirus (*Cardiovirus*) screening was initially performed using the N-Pic-248_F and N-Pic_461-R primers via qPCR, followed by subsequent screening using both the Q-Pic-565_F and Q-Pic-1104_R primer set and the A-Pic-19_F and A-Pic-607_R primer set via endpoint PCR. Picornavirus (*Kunsagivirus*) screening was performed using the P-Pic-6405_F and P-Pic-6872_R primers.

Initial library Q adenovirus screening was performed using the M-AdD-21334_F and M-AdD-21689_R primers, followed by further screening using the I-AdD-19347_F and I-AdD-19925_R primer set. Initial astrovirus screening was performed using the M-Ast-4596_F and M-Ast-4977_R primers, followed by subsequent screening using both the O-Ast-3194_F and O-Ast-3502_R primer set and the P-Ast-3912_F and P-Ast-4142_R primer set. MLV screening was performed using the C-MLV-4016_F and C-

MLV-4324_R primer set. Picornavirus (*Cardiovirus*) screening was initially performed using the Q-Pic-565_F and Q-Pic-1104_R primers, followed by subsequent screening using both the B-Pic-1980_F and B-Pic-2432_R primer set and the A-Pic-19_F and A-Pic-607_R primer set. *Rhadinovirus* screening was performed using the Q-Rhad-6777_F and Q-Rhad-7093_R primers.

Initial library R adenovirus screening was performed using the O-AdD-18404_F and O-AdD-18777_R primer set, followed by subsequent screening using the I-AdD-19347_F and I-AdD-19925_R primer set. Astrovirus screening was performed using the M-Ast-4596_F and M-Ast-4977_R primers. MLV screening was performed using the C-MLV-4016_F and C-MLV-4324_R primers. Initial *Picobirnavirus* screening was performed using the M-PiBiS2-788_F and M-PiBiS2-892_R primers via qPCR, followed by subsequent screening using the O-PiBiS2-711_F and O-PiBiS2-1475_R primers via endpoint PCR. Picornavirus (*Cardiovirus*) screening was performed using the N-Pic-248_F and N-Pic-461_R primers.

f. Phylogenetic analysis

Phylogenetic analysis was performed when ≥ 60% of the identified viral genome was recovered, and the virus had been PCR confirmed in at least one animal. Some exceptions were made to this where either no full genes were sequenced or samples formed an outgroup that compromised tree clarity- these are highlighted in the results where appropriate. Geneious Prime was used to produce an alignment of a reference sequence database and the samples to be analysed using the Clustal Omega translation alignment function, to align by translated amino acid sequence. The aligned sequences were then manually checked to ensure that all sequences were in the same translational frame and adjustments were made as required by manually deleting 1 or 2 bases from the 5' end of sequences that were out of frame as necessary. The sequences were then realigned, and this process was repeated until all sequences aligned were in the same frame. Using NCBI genome annotations for the reference sequence provided for the specific tree, genome loci for specific genes were identified and confirmed within the reference sequence within the alignment. All sequences were then manually trimmed to encompass the stated region of the selected gene, including a full gene wherever possible.

Phylogenetic analysis was then performed using iqtree2.2.2.6 for Windows using the command line interface[313]. Tree generation was performed using the aligned .fasta files exported from Geneious with 1000 bootstrap replicates. The iqtree model finder function was used to select the best Bayesian phylogenetic model from 484 options for each tree, and the selected model for each tree is described in the figure legend. Each tree was then visualised using FigTree v1.4.4, where bootstrap values of < 70% were omitted and host species illustrations were added[314].

5. NGS screening (historic samples)

a. GAPDH and TapeStation quality control

All historic sample RNA was quantified and the fragment size was assessed using the TapeStation 4200 as previously described. Any sample with a successful RNA extraction result (i.e. concentration > 0 ng/µl) then had cDNA produced using EcoDry (as previously described for the modern samples) and underwent GAPDH qPCR using primers designed in-house by Dr. Joseph Chappell. For each sample, 10 µl of Sybr Green PCR Master Mix (Thermofisher, USA), 8 pM of GAPDH-70-F (5'-ATTGACCTCAACTACATGGTCTACA) primer, 8 pM of GAPDH-70-R (5'-TGACBGTGCCYTTGAACTTGC), 8.2 µl of Dep-C water and 1 µl of template cDNA was added. The PCR cycles used were 95^{o}C for 2 minutes, followed by 40 cycles of 95^{o}C for 5 seconds, 63^{o}C for 30 seconds and a plate read for every cycle. A positive result produced a 70bp product. Where GAPDH-70 qPCR was successful, GAPDH-149 PCR was performed using the GAPDH-149-F (5'- TTGACCTCAACTACATGGTCTAC) and GAPDH-149-R (5'- CCATTTGATGTTGGCGGGA) primers, using the same reagent mix and cycles as for GAPDH 70 PCR. A positive result produced a 149bp product. Where GAPDH 149 PCR was successful, GAPDH 295 PCR was performed, using the same reagents and cycles as the other GAPDH PCRs, with the GAPDH-295-F (5'-TGACCTCAACTACATGGTCT) and GAPDH-295-R (5'- GTTCACRCCCATCACAAACA) primers. A positive result produced a 295bp product. All statistical analysis was performed using GraphPad Prism 10.0.3, and all graphs were generated in GraphPad Prism (GraphPad Software, USA).

b. rRNA depletion

For the first 4 sample batches, rRNA depletion was performed using the NEBNext® Ultra™ II rRNA Depletion Kit v2 following the manufacturer's instructions as previously described. For the 5th sample batch and those following, the same kit and protocol was used with some modifications to enhance capture of small fragments as recommended by Beckmann Coulter technical support (private communication). For the first bead clean up step, 100 µl of RNA sample purification beads were added (instead of the 90 µl recommended in the protocol), along with 100 µl of 99.5% pure isopropanol (IPA, Thermofisher, USA). For the second bead clean up, 160 µl of SPRISelect beads were added (instead of the 144 µl recommended in the protocol), along with 160 µl of IPA. All other steps were performed as recommended by the manufacturer.

c. Library preparation and indexing

For the first 3 sample batches library preparation and indexing was performed using the NEBNext® Ultra™ II RNA Library Prep Kit for Illumina® and NEBNext® Multiplex Oligos for Illumina® (Index Primers set 4) kits following the manufacturer's instructions as previously described. For the 4th sample batch and those following, the same kit and protocol was used with some modifications. For the first bead clean up step, 103 µl of SPRISelect beads were added (instead of the 87 µl recommended by the manufacturer) along with 103 µl of IPA. For the PCR enrichment of adaptor

ligated DNA sample amplification was increased by adding 5-8 additional cycles to the amount recommended by the manufacturer according to the input concentration. For the second bead cleanup, 55 µl of SPRISelect beads were added to the sample (instead of the 50 µl recommended by the manufacturer) along with 50 µl of IPA. The bead cleanup modifications were suggested by Beckmann Coulter technical support (private communication). All other steps were followed as recommended by the manufacturer. For the 6th batch samples, the same steps were followed as for the 4th batch samples, except adaptors were diluted to 1:10000 rather than 1:1000 as recommended by the protocol.

d. Library quality control

All libraries underwent quantification and quality assessment using the TapeStation 4200 as previously described. Due to the variability in RNA fragmentation and fragment size in the historic samples these libraries were considered to be successful if a library peak ranging from 200bp- 300bp in size was observed on the TapeStation trace. The samples from the 5th batch also underwent quantification at the second strand synthesis stage to ensure that cDNA was being successfully produced by this protocol, by using the TapeStation 4200 and high sensitivity DNA screentape according to the manufacturer's instructions.

e. Historic sample PCR

3 sets of primers for historic CoV screening were designed *in silico* and tested- NHM-CoV-5360_F (5'- ACTAARTTTTATGGTGGBTGGVA) and NHM-CoV-5451_R (5'- CAYTTAGGATARTCCCANCCCA), NHM-CoV-8969_F (5'- TGATRTWCAACAGTGGGGHT) and NHM-CoV-9080_R (5'- ACAHCGWGTCATAATDGCRTC), and NHM-CoV-C_F (5'- ATGCGWGTTWTACATTTTGGYGC) and NHM-CoV-C_R (5'- ACAAGAAGTGTGYCADYWGG). Standard qPCR reagent quantities and cycles were used, 2x negative controls were tested, and NL63 and OC43 CoV cDNA from clinical isolates was used as positive controls (diluted at 1:100). Historic bat samples were tested by qPCR using the NHM-CoV-8969 primer set, as the NHM-CoV-5360 primers were not successful. Potentially positive isolates were sent for Sanger sequencing as previously described. The NHM-CoV-C primers were not successful during testing and follow up PCR investigations were not performed on the samples due to limited sample availability. 2 sets of primers for the L segment of historic LASV were tested- NHM-LASVL-3939_F (5'- CAAGTTGGGMTTTRATGTATGAYTTCATC) and NHM-LASVL-4046_R (5'- TCAACAYTAYTAACRTGGCATATGCA), and NHM-LASVL-4130_F (5'- AWGGACACATCATWGGRCCCC) and NHM-LASVL-4241_R (5'- ACAAYGAGAARGAATTTGAMAATGCY). All reagents and quantities were as described above for standard qPCR reactions, using 2x negative controls and a synthetic LASV construct designed by Dr. Patrick McClure as a positive control. The NHM-LASVL-3939_F set of primers failed to give a positive result, but the NHM-LASVL-4130_F primer set were successful when tested, and these primers were used to screen any historic *M. natalensis* samples that had a positive GAPDH result and/or a RIN value. Potentially positive isolates were sent for Sanger sequencing as previously described.

Chapter 3- Modern sample RNA extraction results, library preparation and filtering

1. Successful RNA extraction and cDNA synthesis

RNA was successfully extracted from all samples. Purity (as defined by the ratios of absorbance measured at 260 nm/280 nm and 260nm/230 nm) were determined using a Nanodrop spectrophotometer. Ratios between 2.00 and 2.20 were considered optimal, although samples were not disregarded if outside of this range. Samples were named BV (bank vole; *Myodes glareolus*), FV (field vole; *Microtus agrestis*), YM (yellow-necked mouse; *Apodemus flavicollis*), WM (wood mouse; *Apodemus sylvaticus*), V (vole) or W (least weasel; *Mustela nivalis*) according to reported morphology upon sample collection, although this was not always accurate upon species identification PCR. The liver samples for WM6 and WM7 were extracted twice due to low concentrations and sub-optimal purities in the first instance (260 nm/280 nm values of 2.39 and 2.51, respectively, and 260 nm/230 nm values of 0.56 and 1.42 respectively), resulting in increased but still sub-optimal purities upon the second extraction (260 nm/280 nm values of 1.8 and 2.07 respectively, and 260 nm/230 nm values of 1.25 and 1.89 respectively). RNA concentrations for gut samples were on average 916.21 ng/µl, ranging from 2138.8 ng/µl in sample YM5 to 131.8 ng/µl in sample WM62. RNA concentrations for liver samples were typically slightly lower than gut samples, with an average concentration of 838.4 ng/µl, ranging from 3294.1 ng/µl in sample YM6 to 16.5 ng/µl in sample WM6. Bank voles and field voles had higher average RNA concentrations across both tissues (1177.94 ng/µl and 1158.31 ng/µl respectively) than yellow-necked mice and wood mice (848.43 ng/µl and 816.84 ng/µl, respectively). Sample concentrations and Nanodrop 260 nm/280 nm and 260 nm/230 nm values are shown in Table 5 for all samples.

Following RNA extraction, cDNA synthesis was performed on all liver and gut samples for all animals. Approximately 1500 ng of each sample was diluted in dH_2O to give an optimal reaction concentration of 75-100 ng/µl. GAPDH screening was positive for both tissues in all samples, confirming that both RNA extraction and cDNA synthesis was successful for all samples (data not shown).

Species identification was performed using cytochrome B PCR as described in the methods section, and cytochrome B sequences were analysed by NTBLAST to allow for specific speciation. 129/140 animals were as expected according to field identification. Of the 13 reported bank voles, 11 were identified as bank voles by NTBLAST, whilst BV4 and BV7 were identified as field voles by NTBLAST. Of the 15 reported field voles, 6 were identified by NTBLAST as field voles, 7 were identified as wood mice (FV1, FV2, FV3, FV4, FV6, FV8 and FV9), and 2 were identified as bank voles (FV5 and FV7). Sample V1 was identified as a field vole by NTBLAST and all other samples were identified by NTBLAST as matching the species reported during field identification. The final species counts were 13 bank voles, 7 field voles, 17 yellow-necked mice, 100 wood mice, and 1 least weasel (Table 5). This was performed after library generation and submission- accordingly, libraries A and B are mixed species

libraries, whilst libraries C-R are unaffected. Samples were not renamed to prevent confusion and human error.

Table 5- Key information for all Welsh rodent samples.
RNA concentration and Nanodrop 260 nm/280 nm and 260 nm/230 nm absorbance ratios are shown for all samples, as is sampling location. Thick line indicates the split between first and second cohort of samples. For WM6 and WM7 liver, left number represents first extraction and right number represents second extraction. For species identification, samples shown in red are those for which the cytochrome B PCR based species identification does not match the reported field identification.

	Liver			Gut				
Sample	RNA concentration (ng/µl)	260 nm/280 nm ratio	260 nm/230 nm ratio	RNA concentration (ng/µl)	260 nm/280 nm ratio	260 nm/230 nm ratio	Cytochrome B PCR result	Sampling location
BV1	489.1	2.12	2.29				Bank vole	Frongoch
BV2	634.7	2.09	2.22	572.3	2.11	2.33	Bank vole	Frongoch
BV3	839.3	2.08	2.24	310	2.09	2.01	Bank vole	Frongoch
BV4	718.2	2.1	2.02	1407.4	2.13	2.33	Field vole	Nant Y Mwyn
BV5	1467.1	2.11	2.19	712.9	2.06	2.26	Bank vole	Nant Y Mwyn
BV6	2513.2	2.11	2.3	1682.2	2.11	2.36	Bank vole	Nant Y Mwyn
BV7	157.3	2.08	2.11	1384.8	2.07	2.28	Field vole	Nant Y Mwyn
BV8	280.9	2.09	2.2	879.2	2.07	2.33	Bank vole	Nant Y Mwyn
FV1	1412.5	2.11	2.32	222	2.09	2.13	Wood mouse	Frongoch
FV2	916.3	2.08	2.29	461.8	2.11	2.27	Wood mouse	Frongoch
FV3	3262.9	2.11	2.27	1078.4	2.08	2.34	Wood mouse	Frongoch
FV4	2417.3	2.14	2.33	978.2	2.11	2.33	Wood mouse	Frongoch
FV5	1921.7	2.11	2.25	494.8	2.11	2.03	Bank vole	Frongoch
FV6	1905.7	2.12	2.32	312.4	2.09	2.24	Wood mouse	Frongoch
FV7	2801.5	2.12	2.3	686.6	2.07	2.24	Bank vole	Frongoch
FV8	365.9	2.1	2.25	765.7	2.09	2.32	Wood mouse	Frongoch
FV9	612.8	2.1	2.31	1334.9	2.09	2.25	Wood mouse	Frongoch

FV10	1372.6	2.1	2.31	876	2.07	2.34	Field vole	Frongoch
YM1	855.8	2.08	2.33	490.8	2.1	2.33	Yellow-necked mouse	Nant Y Mwyn
YM2	319.1	2.09	1.91	1148.5	2.1	2.05	Yellow-necked mouse	Nant Y Mwyn
YM3	481.4	2.1	2.23	511.6	2.12	2.31	Yellow-necked mouse	Nant Y Mwyn
YM4	263.9	2.1	1.54	508.9	2.08	2.12	Yellow-necked mouse	Nant Y Mwyn
YM5	1870.5	2.08	2.22	2138.8	2.06	2.32	Yellow-necked mouse	Nant Y Mwyn
YM6	3294.1	2.12	2.32	553.3	2.14	2.38	Yellow-necked mouse	Nant Y Mwyn
YM7	166.4	2.07	1.98	518.8	2.02	2.31	Yellow-necked mouse	Nant Y Mwyn
YM8	1658.7	2.11	2.33	429.7	2.12	2.13	Yellow-necked mouse	Nant Y Mwyn
WM1	121.6	2.09	2.21	289.6	2.11	1.99	Wood mouse	Frongoch
WM2	284.7	2.11	2	1682	2.15	2.39	Wood mouse	Frongoch
WM3	910.7	2.14	2.26	520.7	2.14	2.4	Wood mouse	Frongoch
WM4	207.3	2.11	2.08	927.8	2.13	2.42	Wood mouse	Frongoch
WM5	461	2.13	2.2	554	2.07	2.3	Wood mouse	Frongoch
WM6	16.5/6.5	2.39/1.8	0.56/1.25	1498.2	2.14	2.42	Wood mouse	Frongoch

WM7	14.3/19.8	2.51/2.07	1.42/1.89	517.7	2.04	2.35	Wood mouse	Frongoch
WM8	2502.2	2.11	2.3	353.8	2.12	1.92	Wood mouse	Frongoch
WM9	160.2	2.19	0.17	960.4	2.1	2.3	Wood mouse	Frongoch
WM10	2560.2	2.1	2.3	515.3	2.06	2.01	Wood mouse	Frongoch
WM11	3045.8	2.12	2.31	539.3	2.15	2.3	Wood mouse	Frongoch
WM12	135.4	2.06	2.09	349.9	2.1	2.22	Wood mouse	Frongoch
WM13	994.8	2.11	2.24	427.7	2.1	1.86	Wood mouse	Frongoch
WM14	130.3	2.07	2.07	342.3	2.1	2.14	Wood mouse	Frongoch
WM15	1709.1	2.12	2.31	1651.1	2.13	2.33	Wood mouse	Frongoch
WM16	189.7	2.09	2.11	492.6	2.09	2.26	Wood mouse	Frongoch
WM17	187.9	2.07	2.18	1296.2	2.09	2.28	Wood mouse	Frongoch
WM18	149	2.06	2.18	526	2.15	1.9	Wood mouse	Frongoch
WM19	90.2	2.06	1.98	837.7	2.07	2.37	Wood mouse	Frongoch
WM20	211.2	2.1	2.21	499	2.14	2.28	Wood mouse	Frongoch
WM21	2840.8	2.14	2.36	980.9	2.15	2.32	Wood mouse	Frongoch
WM22	118.6	2.08	3.14	452.1	2.11	2.29	Wood mouse	Frongoch
WM23	283	2.08	2.6	235.4	2.08	2.05	Wood mouse	Frongoch
WM24	3036.7	2.13	2.24	682	2.12	2.32	Wood mouse	Frongoch

WM25	1418.2	2.1	2.36	490	2.11	2.29	Wood mouse	Frongoch
WM26	1121.2	2.09	2.37	1034.6	2.09	2.22	Wood mouse	Frongoch
WM27	469.8	2.09	2.37	499.3	2.11	2.26	Wood mouse	Frongoch
WM28	1730.8	2.08	2.2	387.1	2.11	2.04	Wood mouse	Frongoch
WM29	541.6	2.11	1.66	210.4	2.1	2.22	Wood mouse	Frongoch
WM30	337.5	2.08	2.48	485	2.11	2.27	Wood mouse	Frongoch
WM31	483.6	2.1	2.48	741.5	2.09	2.22	Wood mouse	Frongoch
WM33	122.3	2.07	2.22	921.2	2.11	2.34	Wood mouse	Frongoch
WM34	331.7	2.08	2.27	657.4	2.12	2.32	Wood mouse	Frongoch
WM35	1916.4	2.11	2.26	492	2.11	2.26	Wood mouse	Frongoch
WM36	1027.7	2.09	2.3	562.4	2.05	2.27	Wood mouse	Frongoch
WM37	2764.1	2.1	2.32	589.6	2.09	2.32	Wood mouse	Frongoch
WM38	91.9	2.06	2.24	655.8	2.09	2.28	Wood mouse	Frongoch
WM39	356.7	2.11	2.16	265	2.11	1.92	Wood mouse	Frongoch
WM40	642.4	2.07	2.13	661.4	2.08	2.28	Wood mouse	Frongoch
WM41	1216.3	2.1	2.25	844.8	2.1	2.03	Wood mouse	Frongoch
WM42	117.5	2.06	2	1190.8	2.11	2.27	Wood mouse	Nant Y Mwyn
WM43	794.9	2.11	2.32	1011.7	2.1	2.26	Wood mouse	Nant Y Mwyn

WM44	1017.6	2.11	2.23	455.9	2.13	2.3	Wood mouse	Nant Y Mwyn
WM45	150.8	2.07	1.91	672.4	2.1	2.34	Wood mouse	Nant Y Mwyn
WM46	569.7	2.05	2.3	431.2	2.13	2.22	Wood mouse	Nant Y Mwyn
WM47	421.2	2.11	2.26	343.9	2.11	2.25	Wood mouse	Nant Y Mwyn
WM48	380.6	2.1	2.27	1572.4	2.1	2.33	Wood mouse	Nant Y Mwyn
WM49	388.4	2.1	2.23	167.7	2.06	2.08	Wood mouse	Nant Y Mwyn
WM50	571.5	2.05	2.26	391.1	2.11	2.28	Wood mouse	Nant Y Mwyn
WM51	1096.3	2.11	2.27	669.1	2.11	2.28	Wood mouse	Nant Y Mwyn
WM52	1462.8	2.11	2.18	867.1	2.09	2.01	Wood mouse	Nant Y Mwyn
WM53	912.3	2.1	2.01	597.4	2.09	2.21	Wood mouse	Nant Y Mwyn
WM54	591.1	2.08	2.09	507.6	2.07	1.82	Wood mouse	Nant Y Mwyn
WM55	167.4	2.06	1.77	389.6	2.1	2.25	Wood mouse	Nant Y Mwyn
WM56	1301.8	2.1	2.3	141.5	2.08	2.26	Wood mouse	Nant Y Mwyn
WM57	811.6	2.11	2.24	862.8	2.08	2.28	Wood mouse	Nant Y Mwyn
WM58	1660.6	2.14	1.97	355.2	2.1	2.19	Wood mouse	Nant Y Mwyn
WM59	352.2	2.09	2.2	780.2	2.1	2.28	Wood mouse	Nant Y Mwyn
WM60	238.2	2.06	2.08	572.4	2.05	2.3	Wood mouse	Nant Y Mwyn
WM61	528.1	2.12	2.29	795.4	2.04	2.19	Wood mouse	Nant Y Mwyn

WM62	431.7	2.1	2.2	131.8	2.08	2.01	Wood mouse	Nant Y Mwyn
WM63	300.4	2.09	2.23	240.8	2.11	2.19	Wood mouse	Nant Y Mwyn
WM64	497.4	2.14	2.25	389.1	2.13	2.19	Wood mouse	Nant Y Mwyn
WM65	716.7	2.06	1.74	544.4	2.03	2.3	Wood mouse	Nant Y Mwyn
WM66	865.8	2.09	1.79	215.3	2.09	2.24	Wood mouse	Nant Y Mwyn
WM67	385.2	2.11	2.1	176.7	2.07	2.12	Wood mouse	Nant Y Mwyn
WM68	765.8	2.12	2.28	996.4	2.13	2.33	Wood mouse	Nant Y Mwyn
WM69	478.1	2.12	2.2	257.1	2.1	2.13	Wood mouse	Nant Y Mwyn
WM70	50.8	2.14	2.04	1494.6	2.08	2.2	Wood mouse	Nant Y Mwyn
WM71	1122.3	2.09	2.28	840.4	2.12	2.35	Wood mouse	Nant Y Mwyn
WM72	188.7	2.11	1.32	1088.1	2.1	2.32	Wood mouse	Nant Y Mwyn
WM73	2019.5	2.13	2.31	679.9	2.13	2.35	Wood mouse	Nant Y Mwyn
WM74	306.5	2.16	2.25	497.2	2.14	2.34	Wood mouse	Nant Y Mwyn
WM75	138	2.1	2.17	640.8	2.09	2.33	Wood mouse	Nant Y Mwyn
WM76	331.1	2.11	2.29	1547.1	2.04	2.25	Wood mouse	Nant Y Mwyn
WM77	1344	2.09	2.22	1003.1	2.08	2.35	Wood mouse	Nant Y Mwyn
WM78	464.4	2.12	2.28	447.8	2.1	2.26	Wood mouse	Nant Y Mwyn
WM79	269.9	2.12	2.24	1069.2	2.09	2.34	Wood mouse	Nant Y Mwyn

WM80	1032.8	2.12	2.3	1005.1	2.1	2.18	Wood mouse	Nant Y Mwyn
V1	164.4	2.07	2.15	1260.4	2.07	2.3	Field vole	Nant Y Mwyn
W1	1220.8	2.14	2.27	951.7	2.1	2.31	Least weasel	Nant Y Mwyn
BV9	661.7	2.09	2.06	1011.3	2.07	2.3	Bank vole	Nant Y Mwyn
BV10	304.2	2.11	2.03	1089	2.07	2.31	Bank vole	Nant Y Mwyn
BV11	1371.7	2.09	1.12	3566.8	2.13	2.29	Bank vole	Nant Y Mwyn
BV12	486.8	2.1	2.27	2174.3	2.14	2.33	Bank vole	Nant Y Mwyn
BV13	467.8	2.1	2.27	2029.5	2.14	2.33	Bank vole	Nant Y Mwyn
FV11	1463.2	2.14	1.93	642.5	2.05	2.02	Field vole	Nant Y Mwyn
FV12	501.6	2.1	2.29	2571.4	2.12	1.97	Field vole	Frongoch
FV13	332	2.08	2.25	1733.3	2.1	2.33	Field vole	Frongoch
FV14	362.5	2.09	2.26	2453.4	2.12	2.32	Field vole	Frongoch
FV15	1208.7	2.08	2.31	2239.8	2.12	2.32	Field vole	Frongoch
YM9	712.8	2.07	2.11	386.4	2.1	2.28	Yellow-necked mouse	Nant Y Mwyn
YM10	322.9	2.13	1.81	725.2	2.06	2.15	Yellow-necked mouse	Nant Y Mwyn
YM11	200.9	2.1	2.19	1143.5	2.07	2.31	Yellow-necked mouse	Nant Y Mwyn
YM12	174.7	2.1	2.18	318	2.12	2.29	Yellow-necked mouse	Nant Y Mwyn
YM13	244.5	2.08	2.19	931.1	2.11	2.34	Yellow-necked mouse	Nant Y Mwyn
YM14	2576.1	2.12	2.3	776.1	2.07	2.26	Yellow-necked mouse	Nant Y Mwyn
YM15	171	2.04	1.72	250	2.11	2.22	Yellow-necked mouse	Nant Y Mwyn

YM16	1462.2	2.1	2.3	504	2.12	2.33	Yellow-necked mouse	Nant Y Mwyn
YM17	279.3	2.1	1.93	2457.9	2.14	2.34	Yellow-necked mouse	Nant Y Mwyn
WM81	379.7	2.12	2.24	1532.5	2.11	2.34	Wood mouse	Nant Y Mwyn
WM82	408.5	2.12	1.81	911	2.06	2.26	Wood mouse	Nant Y Mwyn
WM83	663.6	2.04	2.23	1888.2	2.11	2.29	Wood mouse	Nant Y Mwyn
WM84	405.5	2.1	2.24	2221.2	2.14	2.34	Wood mouse	Nant Y Mwyn
WM85	613.1	2.04	2.25	1969.2	2.14	2.34	Wood mouse	Nant Y Mwyn
WM86	181.8	2.07	2.22	1476.7	2.1	2.34	Wood mouse	Nant Y Mwyn
WM87	742.2	2.06	2.09	3554.8	2.15	2.08	Wood mouse	Nant Y Mwyn
WM88	594.1	2.04	2.27	784.9	2.11	2.3	Wood mouse	Nant Y Mwyn
WM89	777	2.07	2.25	2439.1	2.14	2.33	Wood mouse	Nant Y Mwyn
WM90	170	2.08	2.28	2508.1	2.12	2.34	Wood mouse	Nant Y Mwyn
WM91	872	2.06	2.25	1940.5	2.14	2.33	Wood mouse	Frongoch
WM92	1293.8	2.1	2.21	1889.6	2.12	2.09	Wood mouse	Frongoch
WM93	198.9	2.09	1.99	754	2.08	2.32	Wood mouse	Frongoch
WM94	563.6	2.07	2.28	1137.6	2.1	2.34	Wood mouse	Frongoch

2. NGS library preparation

Prior to library preparation, samples were pooled. Up to 10 samples were pooled per library, with separate pools being produced for gut and liver tissue using the same constituent samples (Table 6). Pooling was performed prior to speciation and according to species field identification, although it was later found that pools A and B contained mixed species. All samples were diluted equally within each pool to 200-300 ng/µl according to number of samples in the pool, and the pool was then diluted to give a total concentration of approximately 67 ng/µl for library preparation. Pooled samples were analysed using the TapeStation electrophoresis platform prior to library preparation, and a RIN (RNA Integrity Number, a measure of RNA fragmentation that acts as a proxy indicator for RNA quality where 1 is highly fragmented and 10 is extremely intact) was obtained for each pool (Table 7). 30/36 pools were of high quality (RIN > 7.0), and 5 of these fell into the intermediate quality category (RIN of between 3 and 7). Only the Pool L gut library had a RIN below this threshold. Following technical advice from the library preparation kit provider (NEB, USA) it was decided to proceed to synthesise libraries from all 36 pools.

Table 6- Pooling scheme for all libraries.

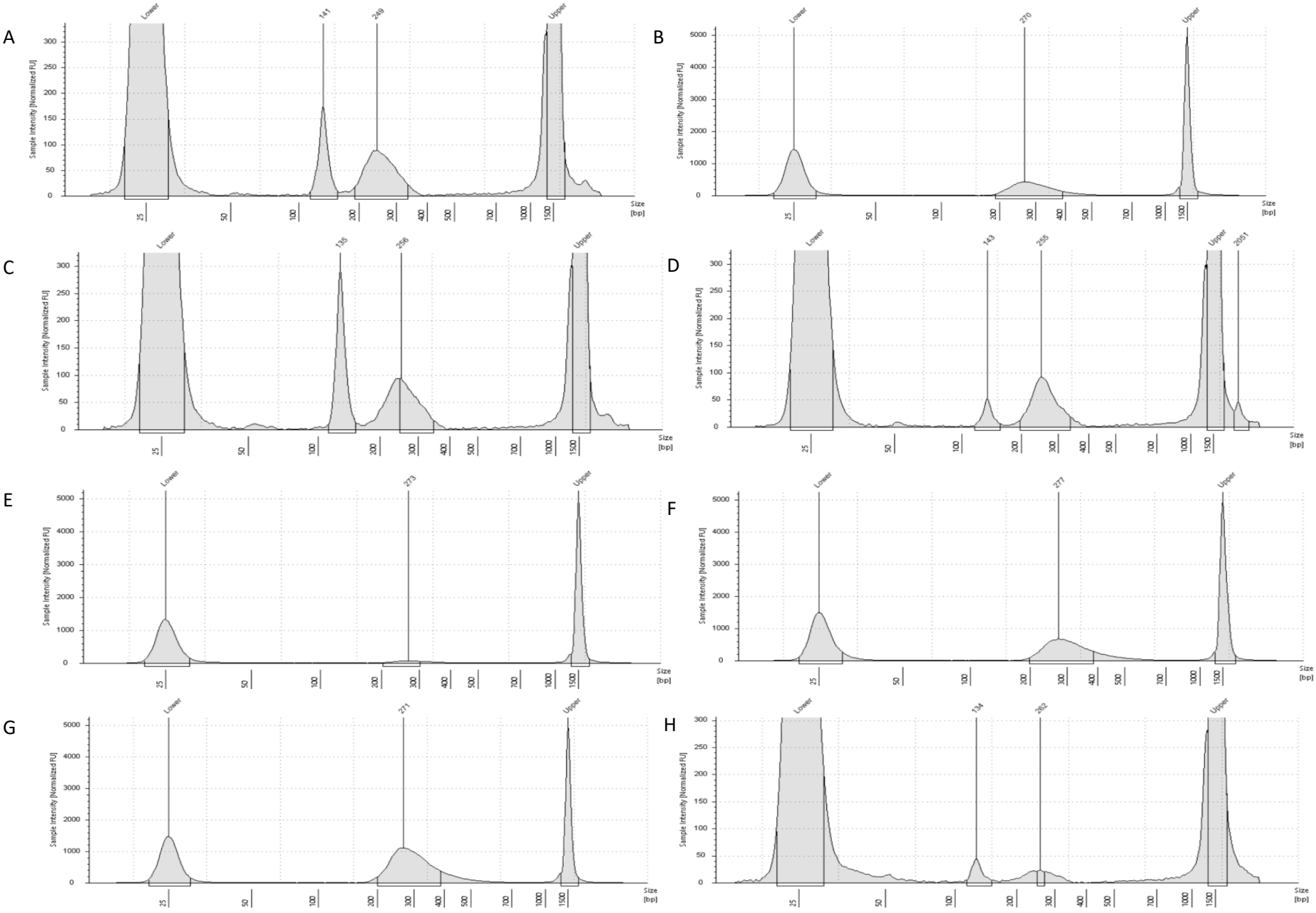
Pooling scheme for each library. Gut and liver pooling was performed separately, but with the same constituent samples comprising each library. All pools except for pool L were designed to contain only one species, according to the field identification of the animals. All samples were added at a concentration of 200-300 ng/µl according to the number of animals in the pools, and pools were then diluted to approximately 67 ng/µl for sequencing (resulting in a minimum concentration of 6.7 ng/µl per sample).

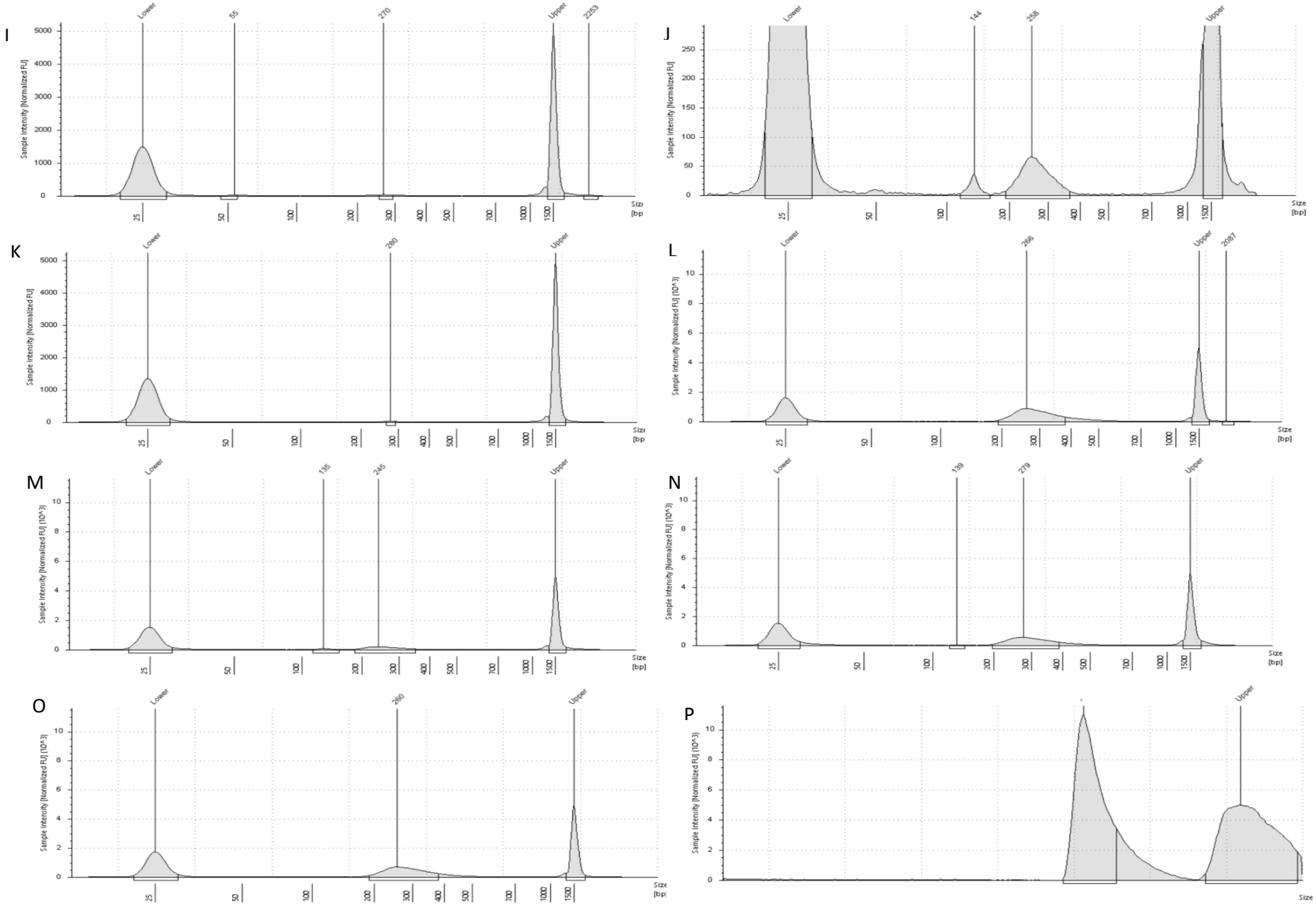
Pool	Samples added									
A	BV1	BV2	BV3	BV4	BV5	BV6	BV7	BV8		
B	FV1	FV2	FV3	FV4	FV5	FV6	FV7	FV8	FV9	FV10
C	YM1	YM2	YM3	YM4	YM5	YM6	YM7	YM8		
D	WM25	WM35	WM37	WM41	WM64	WM65	WM68	WM70	WM75	WM77
E	WM15	WM31	WM61	WM62	WM66	WM67	WM72	WM73	WM76	WM79
F	WM4	WM5	WM14	WM24	WM33	WM34	WM36	WM39	WM63	WM74
G	WM1	WM2	WM6	WM7	WM8	WM9	WM12	WM13	WM40	WM80
H	WM3	WM22	WM26	WM30	WM38	WM45	WM50	WM57	WM59	WM69
I	WM10	WM11	WM17	WM18	WM27	WM46	WM47	WM60	WM71	WM78
J	WM16	WM20	WM23	WM28	WM29	WM43	WM44	WM49	WM52	WM58
K	WM19	WM21	WM42	WM48	WM51	WM53	WM54	WM55	WM56	
L	V1	W1								
M	BV9	BV10	BV11	BV12	BV13					
N	FV11	FV12	FV13	FV14	FV15					
O	WM81	WM82	WM83	WM84	WM85	WM86	WM87			
P	WM88	WM89	WM90	WM91	WM92	WM93	WM94			
Q	YM9	YM10	YM11	YM12	YM13					
R	YM14	YM15	YM16	YM17						

Table 7- RIN values for each pool prior to library synthesis.
RIN values for each pool prior to library synthesis. RIN value is a measure of fragmentation which acts in this case as a proxy for indicator for RNA quality, where 10 is perfectly intact and high quality RNA and 1 is highly fragmented and low quality RNA. RIN of > 7.0 is considered to be high quality, RIN between 3 and 7 is considered to be of intermediate quality.

Pool	Gut RIN	Liver RIN	Pool	Gut RIN	Liver RIN
A	9.0	7.8	J	8.6	8.3
B	7.3	7.5	K	8.7	5.9
C	9.1	8.5	L	2.9	5.8
D	9.1	8.6	M	9.0	9.1
E	7.4	8.9	N	6.7	8.7
F	6.8	8.2	O	7.4	8.7
G	4.7	7.7	P	7.6	8.9
H	8.7	8.5	Q	8.1	8.3
I	7.5	7.9	R	7.3	9.0

NGS libraries were then produced for each library, and their quality was assessed by TapeStation analysis (Figure 4 for gut pool TapeStation traces and Figure 5 for liver pool TapeStation traces). Libraries were considered successful if a clear peak corresponding to between 240 and 300bp was identified on the trace. Gut library P and liver libraries G, I and K failed to synthesise an effective library in the first instance, and liver libraries I and K underwent fresh library preparations accordingly (I2 and K2 in Figure 5). Due to reagent limitations, gut library P and liver library G were not repeated. Some libraries showed a peak at approximately 135bp, indicating the presence of adapter dimer in addition to the peak between 240 and 300bp. Sequence from adapter dimers would be generated during sequencing, thereby reducing the overall yield of target reads. Accordingly, an extra bead clean-up step was performed to remove the adaptors for liver libraries A, B, C, D, E, F, H, I, J, K, N, O and R, and for gut libraries A, C, D, H, I, J, K, M, N, P, R. Due to sample volume limitations, these were not re-analysed on the TapeStation. Matching gut and liver libraries were then diluted and mixed in an equimolar manner to produce a final pool at 30 nM for each library containing both the gut and liver sequences. These were then submitted to Genewiz for Illumina NovaSeq sequencing. Due to all samples being processed on one NovaSeq S4 lane, individual index primers were assigned to the gut and liver pools for each sample set. This enabled identification of a target sequence and library demultiplexing.


A
Lower
141
249
Upper
Sample Intensity [Normalized FU]
Size [bp]
B
Lower
270
Upper
C
Lower
135
256
Upper
D
Lower
143
255
Upper
2051
E
Lower
273
Upper
F
Lower
277
Upper
G
Lower
271
Upper
H
Lower
134
262
Upper


I
Lower
65
270
Upper
2253
Sample Intensity [Normalized FU]
J
Lower
144
258
Upper
K
Lower
260
Upper
L
Lower
266
Upper
2087
Sample Intensity [Normalized FU] [10^3]
M
Lower
135
245
Upper
N
Lower
139
279
Upper
O
Lower
260
Upper
P
Upper
Size [bp]

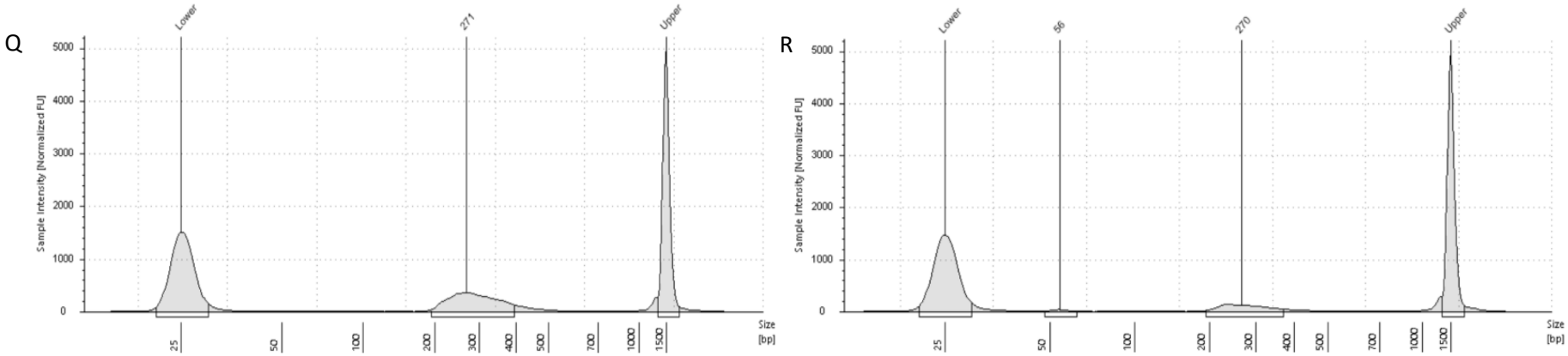
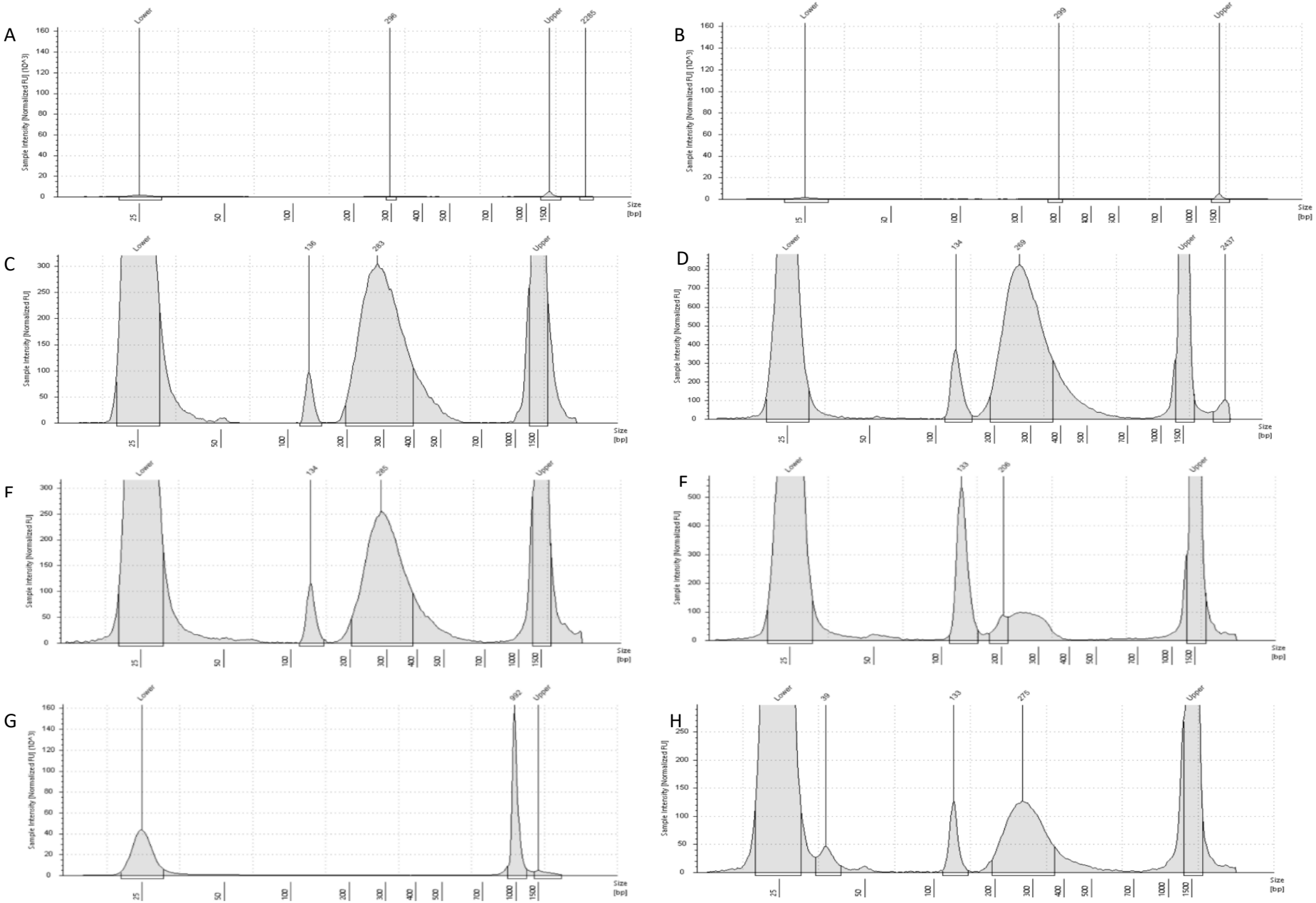
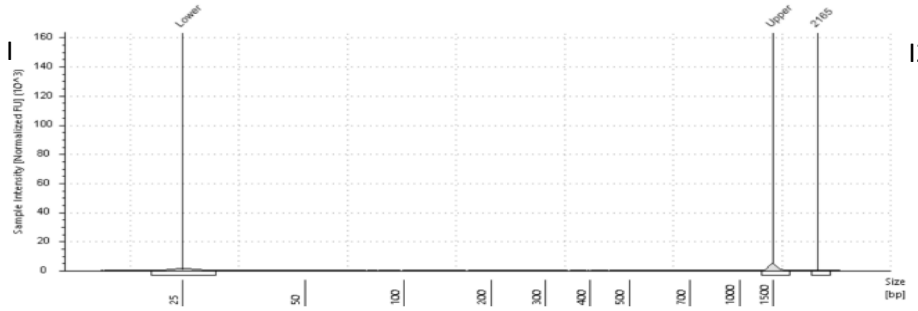


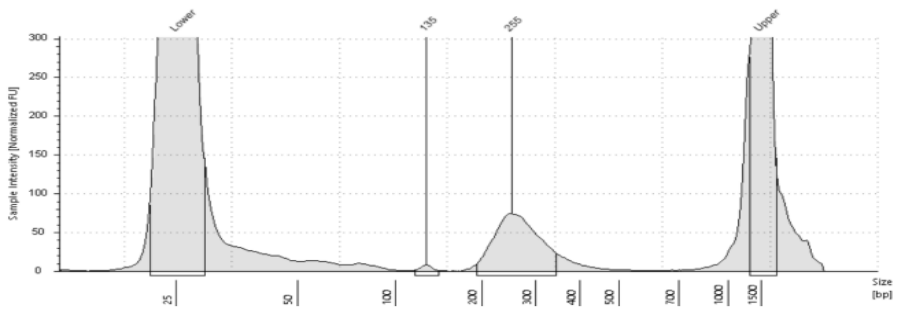
Figure 4- TapeStation trace of all Welsh rodent gut libraries.
Letter in top left corner of each trace denotes which pool generated the trace. X axis shows fragment size in base pairs, Y axis shows sample intensity in normalised fluorescent units. X axis scale is logarithmic, increasing from 0-1500 bp, whilst Y axis scale is variable and adjusted to each library. “Upper” and “Lower” peaks show upper and lower markers used to calibrate the software. Peaks at approximately 240-300bp show the presence of a successful target library, and peaks at approximately 135bp show the presence of adaptor dimer. All pools except for pool P showed a peak at 240-300bp, demonstrating successful library synthesis. Pools A, C, D, H, J, M and N showed the presence of adaptor dimer.


A
Lower
296
Upper
2285
Sample Intensity [Normalized FU] [10^3]
Size [bp]
B
Lower
299
Upper
Sample Intensity [Normalized FU] [10^3]
Size [bp]
C
Lower
136
283
Upper
Sample Intensity [Normalized FU]
Size [bp]
D
Lower
134
289
Upper
2437
Sample Intensity [Normalized FU]
Size [bp]
F
Lower
134
285
Upper
Sample Intensity [Normalized FU]
Size [bp]
F
Lower
133
296
Upper
Sample Intensity [Normalized FU]
Size [bp]
G
Lower
992
Upper
Sample Intensity [Normalized FU] [10^3]
Size [bp]
H
Lower
39
133
275
Upper
Sample Intensity [Normalized FU]
Size [bp]

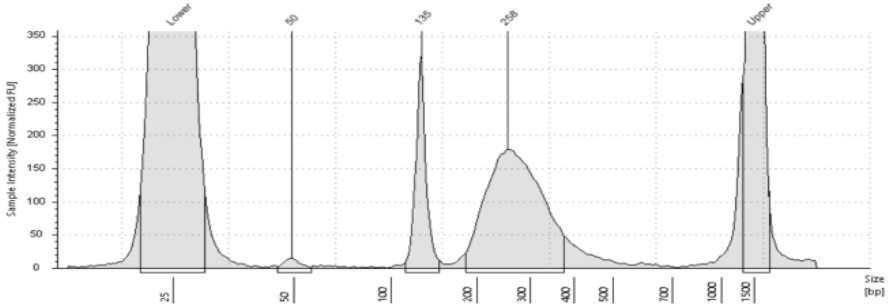
I


Lower
Upper
2165
Sample Intensity [Normalized FU] [10^3]
Size [bp]

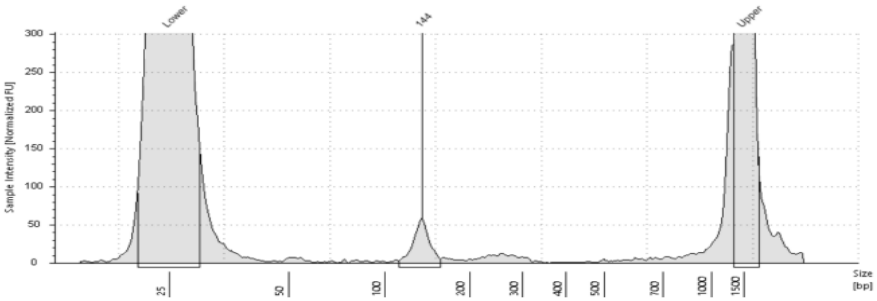
I2


Lower
135
255
Upper
Sample Intensity [Normalized FU]
Size [bp]

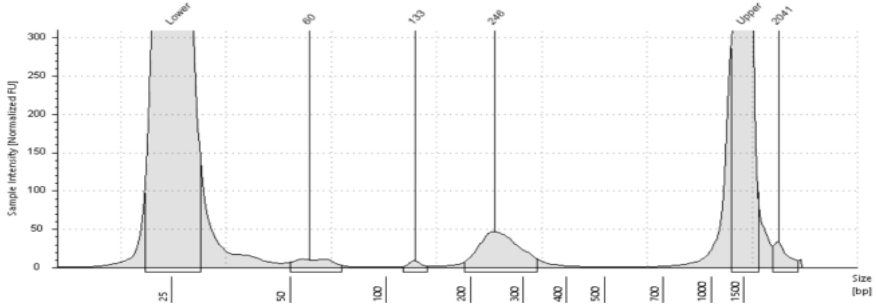
J


Lower
60
135
258
Upper
Sample Intensity [Normalized FU]
Size [bp]

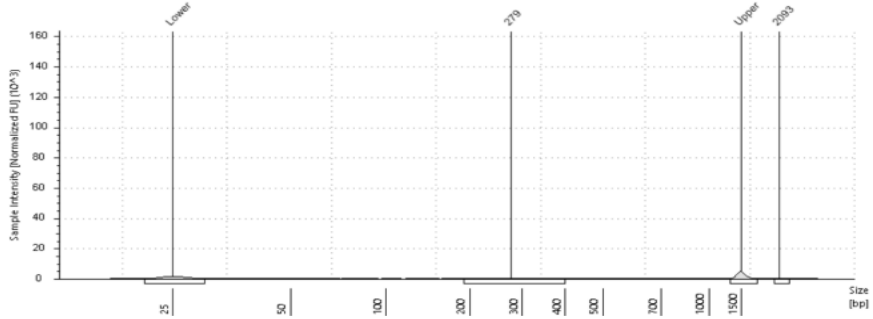
K


Lower
144
Upper
Sample Intensity [Normalized FU]
Size [bp]

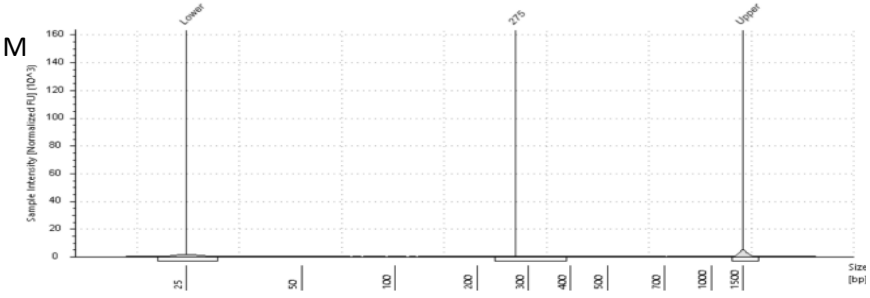
K2


Lower
60
133
246
Upper
2041
Sample Intensity [Normalized FU]
Size [bp]

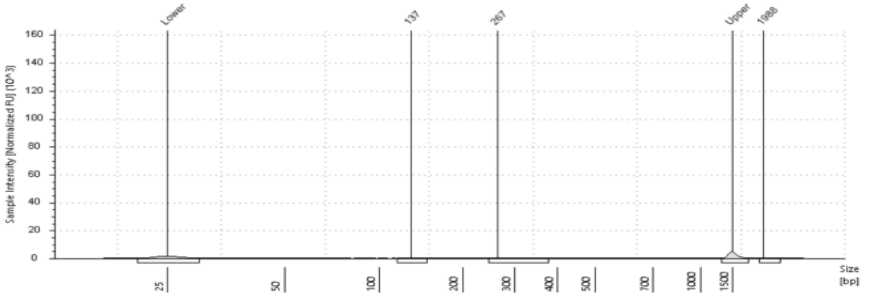
L


Lower
279
Upper
2093
Sample Intensity [Normalized FU] [10^3]
Size [bp]

M


Lower
275
Upper
Sample Intensity [Normalized FU] [10^3]
Size [bp]

N


Lower
137
267
Upper
1988
Sample Intensity [Normalized FU] [10^3]
Size [bp]

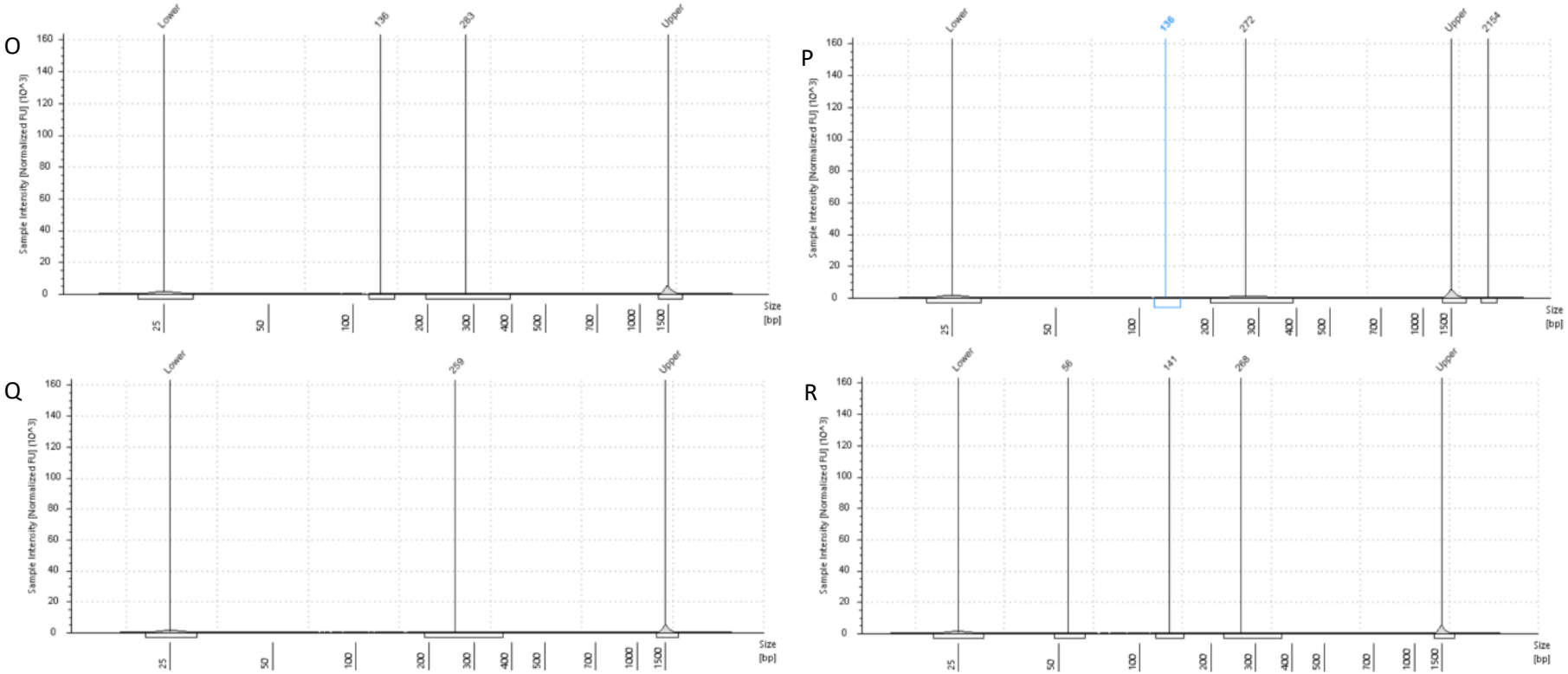


Figure 5- TapeStation trace of all Welsh rodent liver libraries.
Letter in top left corner of each trace denotes which pool generated the trace. Pools I2 and K2 represent show the second attempt at library preparation for these samples. X axis shows fragment size in base pairs, Y axis shows sample intensity in normalised fluorescent units. X axis scale is logarithmic, increasing from 0-1500 bp, whilst Y axis scale is variable and adjusted to each library. “Upper” and “Lower” peaks show upper and lower markers used to calibrate the software. Peaks at approximately 240-300bp show the presence of a successful target library, and peaks at approximately 135bp show the presence of adaptor dimer. All pools except for pools G, I and K showed a peak at 240-300bp, demonstrating successful library synthesis. Pools C, D, E, F, H, J and K showed the presence of adaptor dimer.

3. Metagenomic preparation and host filtering

Approximately 3.1 billion total reads were received from the NovaSeq S4 lane, which were demultiplexed by Genewiz as part of their service. Libraries ranged from approximately $1.35x10^8$ reads to $2.15x10^8$ reads, and were imported into Geneious Prime 2022.1.1, where reads were paired and merged as described in the methods. Filtration was applied to reduce each individual library to > $7.5x10^7$ paired reads or > $1.5x10^8$ unpaired reads as these are the caps for uploading samples to CZID. Where this was not possible filtering was performed to reduce read count to as close to cap as possible to minimise read loss, and libraries were then analysed by CZID despite exceeding the read cap. Initial filtering was performed by removing low quality reads, followed by host species reads (by removing reads that match to a reference host genome downloaded from NCBI Genbank), human reads, bacterial reads and fungal reads. Once sufficiently filtered, libraries were uploaded to CZID, where they were processed through the CZID pipeline 7.0 or 7.1[315]. Table 8 shows initial read numbers, post-filtration read numbers, and read numbers passing CZID controls. The library with the least initial reads was library K with 124,310,755 reads, and the library with the most initial reads was library N with 214,273,480 reads. Host and quality filtration removed between 99.42 % (library P) and 44.27% of reads (library A), with cohort 2 libraries consistently removing more reads during initial filtration than cohort 1 libraries. The CZID filtration then removed between 96.91% of reads (library C) and 16.12% of reads (library H), with substantial variability between libraries and no apparent link to the quantity of reads passing initial filtration. After both rounds of filtration, 14/18 libraries showed 100% of reads passing CZID quality control. 99.99% of library P reads passed quality controls, 86.73% of library A reads passed quality controls, 82.1% of library B reads passed quality controls, and library K had the lowest proportion of reads passing quality controls, with 75.95% of reads passing. Following CZID processing, $1.0x10^6$ reads were generated for each sample by CZID as a reportedly random snapshot of the processed reads and were analysed for plausible hits.

Table 8- Read counts and filtration efficiency for all libraries.

All values are in paired reads. Post-filtration read number represents the number of reads uploaded to CZID following internal host filtration and quality assessment. Reads numbers following CZID filtration represents reads that remained after passing through the CZID pipeline. % of reads passing CZID quality controls is relative to the number of reads that passed CZID filtration.

Library	Number of initial reads	Read numbers following internal host and quality filtration (% of initial reads)	Read numbers following CZID filtration (% of reads uploaded to CZID)	% of reads passing CZID quality controls	Pipeline version
A	134,568,804	75,000,000 (55.73%)	34,520,599 (46.03%)	86.73%	7.0
B	148,111,940	75,000,000 (50.64%)	11,230,837 (14.97%)	82.1%	7.0
C	162,293,069	75,000,000 (46.21%)	2,313,886 (3.09%)	100%	7.0
D	196,976,464	46,887,959 (23.80%)	3,747,439 (7.99%)	100%	7.1
E	175,750,467	43,911,729 (24.99%)	4,079,330 (9.29%)	100%	7.1
F	214,273,480	75,000,000 (35%)	47,069,833 (62.76%)	100%	7.1
G	162,237,188	75,000,000 (46.23%)	62,910,058 (83.88%)	100%	7.1
H	175,537,689	45,734,195 (26.05%)	5,726,284 (12.52%)	100%	7.1
I	150,244,780	64,816,262 (43.14%)	41,347,002 (63.79%)	100%	7.1
J	176,497,221	44,383,332 (25.15%)	6,123,341 (13.8%)	100%	7.1
K	124,310,755	49,102,980 (39.5%)	29,294,279 (59.66%)	75.95%	7.0
L	157,280,122	41,336,787 (26.28%)	18,185,977 (43.99%)	100%	7.1
M	167,766,623	11,903,979 (7.1%)	5,510,246 (46.29%)	100%	7.1
N	214,913,500	20,488,319 (9.53%)	12,167,327 (59.39%)	100%	7.1
O	188,279,603	7,048,368 (3.74%)	5,123,789 (72.69%)	100%	7.1

P	181,902,286	1,046,725 (0.58%)	414,910 (39.64%)	99.99%	7.1
Q	159,868,124	7,961,039 (4.98%)	5,691,771 (71.5%)	100%	7.1
R	172,658,639	18,913,376 (10.95%)	3,140,672 (16.61%)	100%	7.1

4. Discussion

As stated above, pooling was performed prior to species confirmation. As a result, library A is not entirely comprised of bank voles as intended but instead of 6 bank voles and 2 field voles, and library B is not entirely comprised of field voles as intended but instead of 7 yellow-necked mice, 2 bank voles and 1 field vole. Host filtration may have been inefficient in these pools, as host filtration was performed by matching reads to a host genome from NCBI Genbank and removing reads that match. Accordingly, by nature of not knowing that mixed host species genomes were present, the unexpected species genomes have not been filtered against. Whilst it would have been possible to then filter the library again against the correct host species genome, as species identification was performed after metagenomics analysis and PCR screening a second round of filtration at this point was deemed to have limited analytical value and was not performed. This would likely cause the largest effect in library B, which was filtered against a field vole genome despite unknowingly only being composed of 10% field vole samples, and a milder effect on library A as 75% of sequences were bank vole sequences. Fortunately, all hosts were rodents that are phylogenetically similar, so host genomes will have a reasonable overlap that still allows for some filtration. If this project was repeated, species identification would be performed prior to pooling.

The CZID pipeline is similar to and demonstrates similar performance to other established pipelines such as those reported by Brinkmann and colleagues[27,71,316]. Brinkmann and colleagues reported on the efficacy of 14 different in-house pipelines used in bioinformatics research groups, using a variety of software for each step including those used in the CIZD pipeline such as Diamond and those not used in the CZID pipeline, such as Kraken[27,315,316]. The CZID process is split into three sections- host filtering, alignment and post processing. All processes performed in CZID pipeline 7.0 and 7.1 are identical and performed in the same order with no significant differences. The differences between the 2 pipelines are only in the underlying code where modifications have been made to increase efficiency, and reviewing these changes is beyond the scope of this project. Whilst it is very unlikely that pipeline version will have had a major effect on sample processing, pipeline usage has been recorded for posterity regardless.

During host filtering, the pipeline validates that the input is suitable (present in a .fastq file format), prior to using the STAR aligner to remove host mapped reads according to the host species information provided in the metadata at the point of upload. Following this, the trimmomatic system is used to remove adapter sequences, using the known sequences of the Illumina adaptors to identify these. Reads then pass through the Price Seq filter to remove low quality reads (> 10% N reads or > 15% low quality nucleotides). These are reasonable thresholds that allow for the removal of low quality reads without being too stringent and allowing for some flexibility in read quality. Trimmomatic is a well-established and effective program that is commonly used for removing Illumina adaptors and is capable of

accurately trimming adaptor sequences from target reads. Trimmomatic is therefore a good choice of trimming software to use here[317]. Following this, CZID-dedup is used to identify duplicate reads by matching the first 70 bases of each read, followed by the LZW step which uses the Lempel-Ziv-Welch algorithm to remove low complexity reads and reduce sequence clutter. Bowtie2 is then used to remove remaining host reads, prior to the subsampling stage where CZID subsamples 1 million reads to continue through the pipeline. CZID purports that this is random, but in our experience of reprocessing the same library 3 times identical results were obtained across all 3 uploads, therefore it appears that this process is not random after all. This suggests that it is possible that only the first 1,000,000 reads are being analysed and the rest are not being examined by CZID, which could lead to a significant loss of data or low copy number viruses that are not represented in the subsampled reads, which is a major drawback to the CZID pipeline. After subsampling the 1 million reads are passed through the STAR aligner and Bowtie2 filter system once again, followed by a pass through the GSNAP filtration system, which is designed to remove contaminating human reads. After this, the samples proceed to the alignment stage[315].

During the alignment stage, the sample reads take multiple paths to reach an overall alignment (Figure 6). The "Alignment Minimap2" stage uses the minimap2 system to align against the NCBI nucleotide database. The "Call Hits Minimap2" stage then assigns matched accessions from the minimap2 alignment to specific taxon identities. The "Alignment Diamond" stage uses the Diamond system to align against the NCBI non-redundant protein database to align reads at the amino acid level. The "Call Hits Diamond" stage assigns the aligned reads from the Diamond alignment to specific taxons. Using both Diamond and minimap2 is wise as Diamond uses protein alignments to call hits whilst Minimap2 uses nucleotide alignments. By using both methods it allows for both high confidence nucleotide alignments whilst still allowing for detection of microorganisms that have undergone synonymous mutations that do not alter the amino acid sequence and may be missed by the nucleotide alignment[318,319]. After all alignments are performed, taxon counts from both the minimap2 alignment process and Diamond process are combined to give a single set of overall taxon counts, and .fasta files are generated for both identified and unidentified reads. These .fasta files then proceed to the post-processing stage[315].

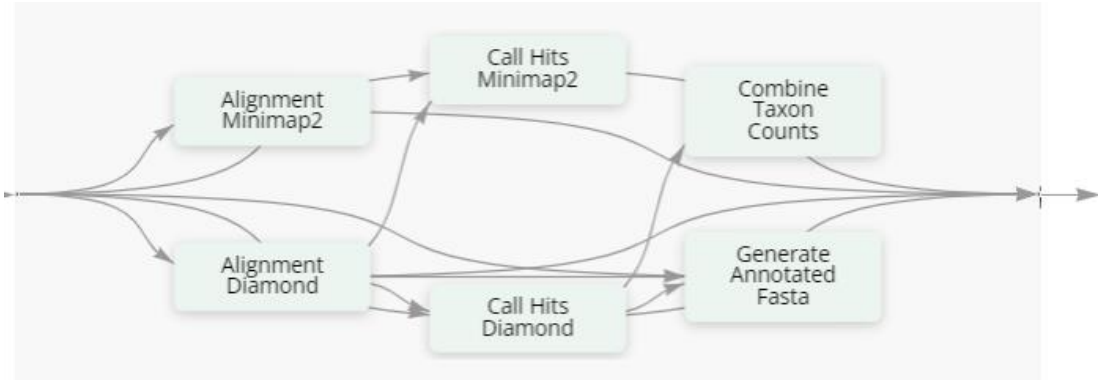


Figure 6- Progression of reads through the CZID alignment process.
Screenshot showing the pathing of the reads through the CZID alignment stage. Filtered reads are represented by the arrows on the far left and aligned reads being passed to the post processing stage are shown on the far right. The Diamond and Minimap2 processes use protein and nucleotide alignments respectively to allow for high confidence microorganism detection[318,319].

As with the alignment stage, samples take multiple paths through the post-processing stage (Figure 7). Rapsearch2 is the output file from the Diamond aligner-accordingly, the "Download Accessions Rapsearch2 Accessions" stage is matching the Diamond output with specific accession numbers. Similarly, Gsnap represents the output file from the minimap2 alignment, so the "Download Accessions Gsnap Accessions" stage is matching the minimap2 output with specific accession numbers. The "Assembly" stage uses the SPAdes to assemble individual reads into contigs and uses bowtie2 to link the original reads and their associated contigs[27,315]. SPAdes is often considered to be the gold standard approach for bioinformatics read assembly, where it has been shown to be reliable and accurate[320]. However, faster programs which can utilise longer k-mers for assembly such as SKESA may have provided better assembly coverage at approximately the same speed, and are specifically designed for microbial genomic assembly, therefore whilst the use of SPAdes in this pipeline is still valid and reliable it is possibly sub-optimal[320]. The "Generate Coverage Stats" step uses the assembled reads to estimate contig coverage of the identified genomes. The two "Blast Contigs" stages are performed together, first to identify the contigs produced by the minimpap2 alignment using the NCBI NTBLAST database and then to identify the contigs produced by the Diamond alignment using the NCBI non-redundant protein database (NCBI NRBLAST) in order to refine the overall output. Taxon counts are then combined from the two alignments and the "Compute Merged Taxon Counts" stage creates hit summary files, before the "Combine Json" stage produces Json output files, which are Javascript files to allow for easier usage. The "Generate Annotated Fasta" stage generates both annotated read fasta files and unidentified reads fasta files, which then passes to the "Generate Taxid Fasta" stage, which generates taxonomic ID summaries. Finally, the "Generate Taxid Locator" stage attaches the taxonomic ID summary to the non-host reads, finishing the process.

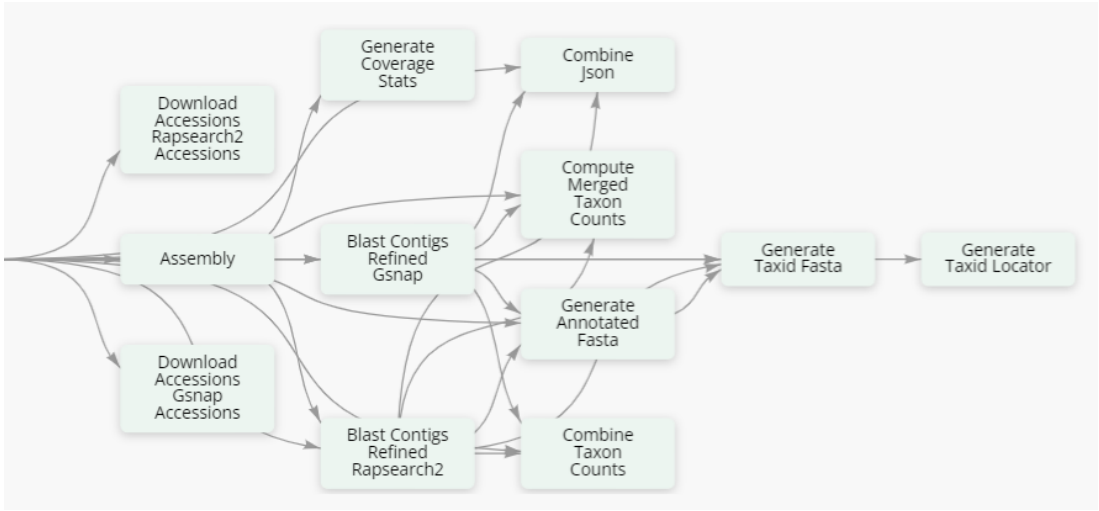


Figure 7- Progression of reads through the CZID post-processing stage.
Screenshot showing the pathing of the reads through the post-processing stage. Aligned reads enter the flowchart in the arrows on the far left. Through multiple pathing, the CZID post-processing steps allows for accurate linking of alignments generated in the alignment stage to specific hits, produces specific contigs, and converts the data into a more user-friendly UI.

For cohort 1 (libraries A-L), a relatively powerful (64GB RAM) local PC was used to process the libraries. However, due to the size of the libraries, this was inefficient- the entire process from importing the libraries through to filtering sufficiently for CZID upload took approximately 5-7 days per library. For the second cohort (libraries M-R), Dr. Stuart Astbury processed the samples through a multi core HPC at the University of Nottingham. As well as being significantly quicker (the entire process was completed overnight for 6 libraries, M-R), this appears to have increased the stringency of the filtration, removing more reads prior to upload to CZID, despite using the same programs to filter (STAR and bowtie2, repeated for two passes). It is unclear why this occurred, although it is possible that some of the stringency parameters which are not reported by CZID may be different to those used by Dr. Astbury. Regardless, this resulted in more manageable datasets and likely reduced false positive results, which is beneficial for downstream analysis.

This goes some way as to explaining the substantial difference in reads post-filtration between the two cohorts. However, this does not explain the differences in read number between libraries of the same cohort. Whilst efforts were made to minimise adaptor dimer contamination where possible, it is not possible to remove 100% of adaptor dimers prior to library sequencing, which will be filtered out in the read quality assessment stage, in turn reducing the total number of reads passing filtration. Operator experience and familiarity with the library preparation kit may also have an effect, as shown by the first 3 libraries prepared (libraries A, B and K) having the lowest % of reads passing quality filters. It is possible that increasing experience allowed for enhanced precision and accuracy when judging how dry beads are during the bead purification steps as the project progressed, which can in turn affect both yield and quality of sequences. For these libraries, there was no clear link between the RIN of the constituent pools of the libraries and either quality or read filtration efficacy. There is also no apparent link between initial library size and either raw number or % of reads passing filters. Regardless, some libraries were better quality and gave a better yield than others, perhaps in part due to the inherent randomness of working with animal samples.

Dr. Astbury also used the HPC to reanalyse some of the previously processed libraries using Kraken2, a well-known taxonomic identification software for metagenomics analysis[34]. The idea behind this approach was that using Kraken2 analysis on the HPC would not result in the non-random read cap imposed by CZID and prevent essentially the loss of all reads after the first 1,000,000 reads from the rest of the data. However, the Kraken2 results were near identical to the CZID results in the 3 libraries assessed in this manner, improving confidence in the results provided by CZID. As Kraken2 uses a different k-mer count and database to the NCBI databases used throughout the CZID pipeline, this is reassuring and suggests that the analysis of the libraries is robust and reproducible via two different bioinformatics approaches[315,321]. Whilst it is possible that any very low copy number viruses were still not detected (as is the case with all metagenomics projects), this validation from

an alternative method supports that the majority of viruses would be identified by the CZID approach[65]. Whilst the HPC method was significantly quicker, the output is in a less user friendly interface and is more difficult to interpret. It is unfortunate that our access to the HPC became available after the completion of most of heavy computational work for this project. However, for future work or other metagenomics projects within the laboratory, the HPC approach will be used.

Generating and processing the libraries for metagenomics analysis was highly successfuland produced large quantities of high quality data. This allowed for the later analysis of the metagenomics data with a high degree of confidence in the reads and allowed for specific virus detection PCRs to be designed and performed based on these libraries.

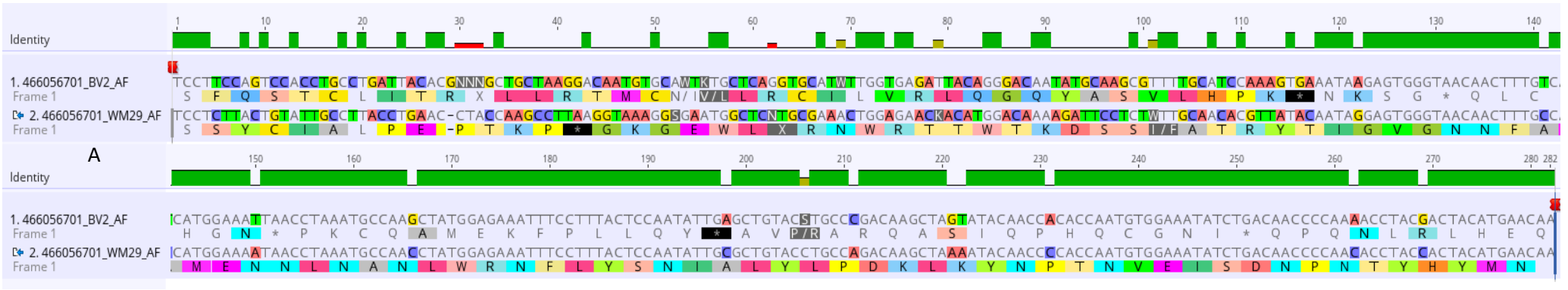
Chapter 4- Modern sample degenerate PCR screening results (Prong 1)

After the successful isolation of RNA and the generation of cDNA from liver and gut tissue of 140 rodents, the next step was to attempt to identify evidence of infection. Prior to receiving metagenomics data, this was performed by PCR screening of these samples using pan-family or pan-genus primer sets designed to detect key viral species.

1. Identification of 2 adenoviruses

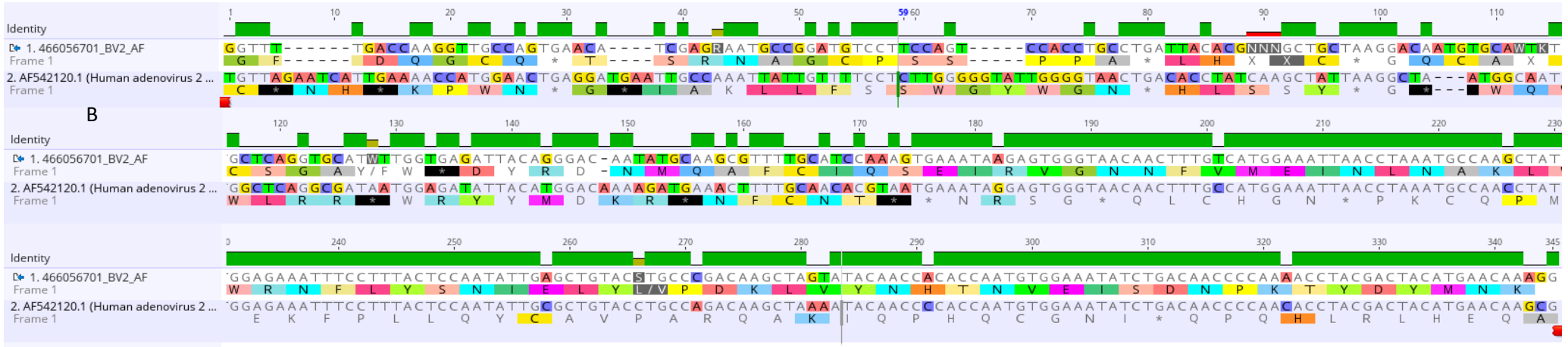
Initial adenovirus screening using the AdHex PCR primers yielded no positive results in any gut samples. BV2H and WM29H were the only positive samples (2/140), and all controls were as expected, i.e. the positive adenovirus control yielded a positive result and the negative control was negative and yielded no bands. As both positive samples were liver samples, it is unlikely that environmental food contamination has occurred here. Using NCBI NTBLAST both were identified as most similar to human *Mastadenovirus C*, and not identified as closely related to rodent *Mastadenovirus* isolates by NTBLAST. Following the trimming of low quality 5' and 3' end bases BV2H yielded a 329 base fragment, and WM29H yielded a 281 base fragment. Upon alignmen, these two sequences were not identical, sharing 67.02% identity, with reduced similarity at the 5' end of the alignment (the hexon CDS) and increased similarity at the 3' end of the alignment (the hexon gene) (Figure 8). BV2H was 67.32% similar to the most closely related adenovirus sequence (accession number AF542120) and WM29H was 84.47% similar to the most closely related adenovirus sequence (accession number LC720425), and in both alignments the 5' end was more similar than the 3' end (Figure 8).

Attempts were made to perform further PCR investigations to identify more of the adenovirus genome and allow for full gene phylogenetic analysis. The Alt_Adeno, Alt_Adeno_2 and Alt-Adeno_3 primers all failed to generate a product at any temperature upon gradient PCR investigations. This process exhausted the positive control material and further investigations were not performed. Despite this, phylogenetic analysis was performed on the available hexon sequence relative to bases 20,051-20,392 of a human adenovirus C reference sequence (NC_001405). Both BV2H and WM29H formed a clade amongst human adenovirus sequences, albeit with long branch lengths and weak bootstrap support.

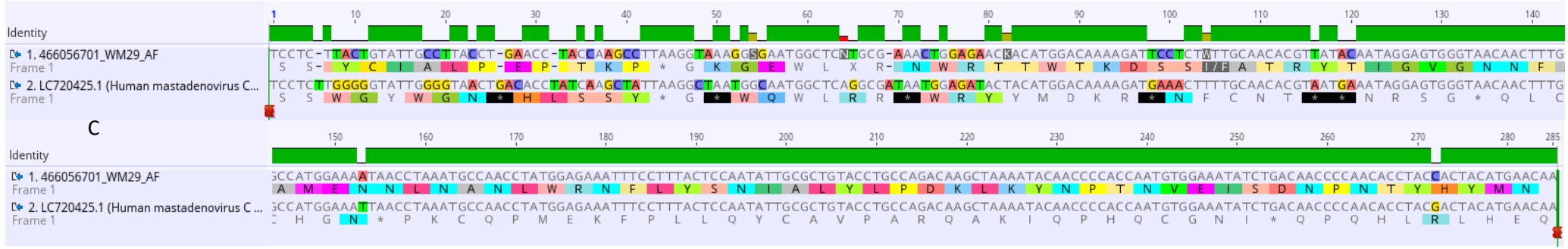

Identity
1. 466056701_BV2_AF
Frame 1
2. 466056701_WM29_AF
Frame 1

A

Identity
1. 466056701_BV2_AF
Frame 1
2. 466056701_WM29_AF
Frame 1


Identity
1. 466056701_BV2_AF
Frame 1
2. AF542120.1 (Human adenovirus 2 ...
Frame 1

B


Identity
1. 466056701_WM29_AF
Frame 1
2. LC720425.1 (Human mastadenovirus C...
Frame 1

C

Identity
1. 466056701_WM29_AF
Frame 1
2. LC720425.1 (Human mastadenovirus C ...
Frame 1

Figure 8- Alignments of the BV2H and WM29H adenovirus isolates, and their most similar viruses as identified by NTBLAST analysis.
In all alignments, red shapes indicate trimming of samples to match sequence length and remove low quality bases. In the identity bar, green bar represents 100% match, yellow bar represents a 50% chance of a match, a red bar represents 25% chance of a match and a blank space indicates a 0% chance of a match. Bar height also decreases as likelihood of match decreases. A) Alignment of BV2H isolate (top sequence) and WM29H isolate (bottom sequence) across the entire length of the sequenced region of the WM29H isolate. These isolates are 67.02% similar. B) Alignment of BV2H isolate (top sequence) and AF542120 (bottom sequence), across the entire length of the sequenced region of the BV2H isolate (bases 1227-1571 of AF542120 reference sequence). These sequences are 67.32% similar. C) Alignment of WM29H isolate (top sequence) and LC720425 (bottom sequence) across the entire length of the sequenced region of the WM29H isolate (bases 20,039-20,323 of LC720425 reference sequence). These sequences are 84.47% similar.

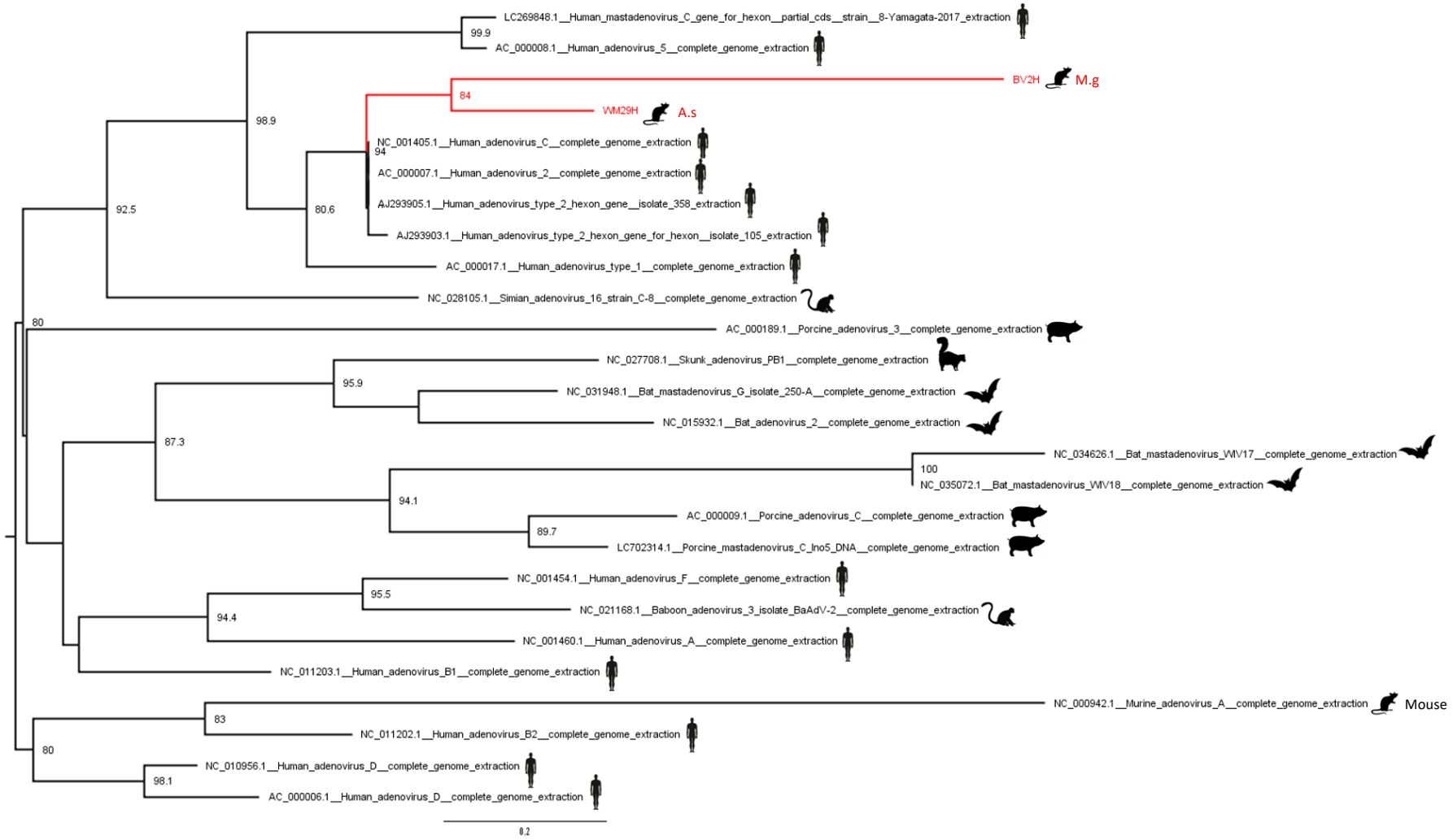

LC269848.1__Human_mastadenovirus_C_gene_for_hexon__partial_cds__strain__8-Yamagata-2017_extraction
99.9
AC_000008.1__Human_adenovirus_5__complete_genome_extraction
BV2H
M.g
84
VM29H
A.s
98.9
NC_001405.1__Human_adenovirus_C__complete_genome_extraction
94
AC_000007.1__Human_adenovirus_2__complete_genome_extraction
AJ293905.1__Human_adenovirus_type_2_hexon_gene__isolate_358_extraction
92.5
80.6
AJ293903.1__Human_adenovirus_type_2_hexon_gene_for_hexon__isolate_105_extraction
AC_000017.1__Human_adenovirus_type_1__complete_genome_extraction
NC_028105.1__Simian_adenovirus_16_strain_C-8__complete_genome_extraction
80
AC_000189.1__Porcine_adenovirus_3__complete_genome_extraction
NC_027708.1__Skunk_adenovirus_PB1__complete_genome_extraction
95.9
NC_031948.1__Bat_mastadenovirus_G_isolate_250-A__complete_genome_extraction
NC_015932.1__Bat_adenovirus_2__complete_genome_extraction
87.3
NC_034626.1__Bat_mastadenovirus_WIV17__complete_genome_extraction
100
NC_035072.1__Bat_mastadenovirus_WIV18__complete_genome_extraction
94.1
AC_000009.1__Porcine_adenovirus_C__complete_genome_extraction
89.7
LC702314.1__Porcine_mastadenovirus_C_Ino5_DNA__complete_genome_extraction
NC_001454.1__Human_adenovirus_F__complete_genome_extraction
95.5
NC_021168.1__Baboon_adenovirus_3_isolate_BaAdV-2__complete_genome_extraction
94.4
NC_001460.1__Human_adenovirus_A__complete_genome_extraction
NC_011203.1__Human_adenovirus_B1__complete_genome_extraction
NC_000942.1__Murine_adenovirus_A__complete_genome_extraction
Mouse
83
NC_011202.1__Human_adenovirus_B2__complete_genome_extraction
80
NC_010956.1__Human_adenovirus_D__complete_genome_extraction
98.1
AC_000006.1__Human_adenovirus_D__complete_genome_extraction
0.2

Figure 9- Midpoint rooted phylogenetic tree of degenerate PCR identified sequences. Relative to adenovirus hexon gene bases 20,051- 20,392 relative to reference sequence NC_001405 (human adenovirus C). Tree generated using TIM3+F+I+G4 model with 1000 bootstrap replicates. Node values represent % bootstrap support. Red branches and text highlight sample libraries, host species is shown next to each branch. Scale bar represents 0.2 substitutions per site. M.g= *Myodes glareolus*, A.s= *Apodemus sylvaticus*, "Mouse" represents a mouse host where the specific species was not recorded.

2. Identification of an old-world hantavirus sequence

All samples were screened using both the established Han_L primers[175] and the OW_HAN primers, designed in-house. The OW_HAN primers yielded no positive results, whilst the Han_L primers yielded 1/140 positive liver samples (in sample V1H, a field vole). In all reactions the positive control produced a clear positive band whilst the negative control produced no bands of any size, therefore all controls were as expected. The sequence chromatogram was of low quality, but the sequence was identified as most similar to a *Black Creek Canal Orthohantavirus* RdRp (accession number GU997097) with 87.61% similarity by NCBI NTBLAST based on the final 73 bases of the sequence. Only the final 73 bases were analysed by NTBLAST as the chromatogram trace of this region was of significantly higher quality than the rest of the trace and could be reliably basecalled. No further investigations were performed due to the return of the NGS data shortly after these results were obtained.

3. Other PCR screening panels

The initial CoV screening with the Woo_CoV primers yielded 1/140 positive liver samples (a wood mouse, WM18H). The positive control yielded a clear positive result, whilst the negative controls produced no PCR products of any size. Accordingly, WM18H was sent for sequencing, but failed to sequence on 3 separate occasions, exhausting the available material. Follow up PCR investigations were then performed on WM18H using both the Alt_CoV and Woo_CoV_2 primers. For both primer sets, the OC43 positive control was positive and the WM18H sample was negative. The initial PCR using the Woo_CoV primers was then repeated on WM18H and yielded a negative result, despite a clear positive result being observed in the positive control. Accordingly, this was deemed to be a negative sample and the initial screening was deemed to have yielded a false positive result.

Rubivirus screening yielded no positive results. Following a successful primer test gradient PCR which demonstrated that the Rub_1 primers effectively bound to the RubV vaccine extract at 59^{o}C, all liver and gut samples were screened using the Rub_1 primers. All animal samples were negative, the positive control produced a clear positive result and the negative control reactions produced no PCR product of any size.

Rotavirus PCR screening was unsuccessful. Both the R_VP1 primers and the R_VP7 primers failed to produce a positive result using established *Rotavirus* clinical isolate positive control material as a template when tested by gradient PCR. After attempting the gradient PCR 3 times with each primer set for both denatured and not denatured control material, the control material was exhausted and no further investigations were performed.

4. Discussion

The rationale behind performing degenerate PCR screening was primarily based on the increased sensitivity of a targeted PCR relative to an unbiased metagenomics screening[65]. As sufficient RNA and cDNA was available for all samples, degenerate screening was a worthwhile endeavour, especially whilst waiting for the return of the NGS data. Additionally, we considered that by targeting conserved regions of target genomes whilst also incorporating the necessary degeneracy to develop a pan-genus screening PCR where possible, we may identify more novel "dark matter" virus strains that may be too divergent to be identified by traditional metagenomics alignment and identification software such as kraken2[6]. Finally, an additional goal of this screening approach was to potentially develop an element of a multiplex screening panel for future virus discovery projects and clinical samples. Due to time limitations, a multiplex panel was not produced, although the successful primer sets that gave valid positive control PCR hits- i.e. the AdHex primers, the Woo_CoV primers and the Rub1 primers- may be incorporated into a multiplex screening panel at a later date. The viruses screened for were chosen based on a variety of factors which will be discussed below for individual viruses. The primary screening selection criteria included likelihood of identification in these samples and potential clinical significance if identified. Accordingly, screening primers were designed for both OW and NW arenaviruses. Samples were not screened for the presence of NW arenaviruses due to the extreme unlikelihood of finding any positive samples, as they are near exclusively found in the Americas[130]. OW arenavirus screening was not performed due to a lack of positive control material rendering results unreliable. The metagenomics data did not identify any arenaviruses of either type in any sample, therefore omitting this screening is unlikely to have resulted in any viruses being missed.

The adenovirus screening identified 2 positive samples, which were then found to be most similar to human *Mastadenovirus C* by NCBI NTBLAST. Although only small sections of the genomes were obtained, these sections were part of the Hexon gene which has previously been used for adenovirus typing by assessing hypervariable regions within this gene[113]. Whilst Hexon typing alone is not sufficient to accurately identify all adenovirus strains- particularly in recombinant adenoviruses- it still provides useful information that can be used to approximate viral strain identity and relatedness between them[113]. The viruses identified here were relatively divergent across this region, with the WM29H sequence being only 84.47% similar to the most similar virus by NTBLAST and BV2H being only 67.32% similar to the most similar virus by NTBLAST. Amongst other criteria, the ICTV requires a 5-15% difference within the polymerase gene of the adenovirus genome for an adenovirus to be considered as a novel species [322]. Whilst the viruses found here are quite divergent and genetically distinct, without the isolation of further genomic sequence- particularly of the polymerase gene- it is not possible to accurately assess species demarcation for these viruses. Ideally, further investigations would have been performed to identify more of the genome for these viruses- perhaps using the multiple target approach

described by Wu and colleagues- but due to a lack of control material, this was not possible[113]. Whilst it is unexpected that these viruses from rodents were identified as most similar to human viruses by NTBLAST this could be due to the limited length of available sequences. These viruses also clustered amongst human adenoviruses within the phylogenetic tree with strong bootstrap support (Figure 9) supporting that these viruses are indeed similar to human adenovirus species, although it should be noted that phylogenetic analysis of such a small section of genome may not be entirely reliable.

Interestingly, these *Mastadenovirus C* positive sequences were not identified by CZID during the metagenomic screening. This could be due to a low grade infection resulting in a low viral copy number below the sensitivity threshold of metagenomics identification, or potentially due to excessive subsampling of reads by CZID. It is also impossible to rule out Adenovirus contamination from the operator during nucleic acid extraction, PCR, or sequencing, although the standard molecular laboratory control measures and cleaning regimens render this unlikely. Regardless, this demonstrates that PCR and the associated sensitivity of this method still has a place within virus discovery workflows, even in the metagenomics age[59,60]. However, the PCR screening did not identify any of the *Mastadenovirus D* sequences identified by the metagenomics data, and it was found that the designed pan-adenovirus primers may not effectively bind to and identify *Mastadenovirus D* viruses when aligned against *Mastadenovirus D* viruses specifically due to a 3' mismatch in the *Mastadenovirus D* reference sequence. This would in turn limit their value for virus discovery and would limit their use in clinical diagnostics as some *Mastadenovirus D* strains are known human pathogens[113]. Future work regarding these viruses should focus on increasing the genomic coverage of the sequences identified here, in particular aiming to cover the polymerase gene and re-designing the screening primers to increase the range of detectable adenoviruses. Adenoviruses were selected for screening due to being relatively common viruses with the potential for rapid transmission which are known to circulate in rodents[113,120].

The hantavirus screening identified one positive field vole, specifically V1H. Due to poor sequencing only 73 bases were usable, and the V1H sequence was found to be most similar to *Black Creek Canal Orthohantavirus* by NTBLAST analysis. *Black Creek Canal Orthohantavirus* is a NW hantavirus that is found in Hipsid cotton rats in Florida and has been responsible for a singular recorded case of HPS[179]. Whilst this is extremely unlikely to be found in a Welsh rodent, the 73bp region found here was in the L segment of the hantavirus which is a highly conserved region, therefore whilst the NTBLAST result does not serve to allow for identification of the virus found here to the species level it does serve to illustrate that an *Orthohantavirus* is present in this animal[11]. A species level identification and phylogenetic analysis may have been possible if more of the hantavirus genome had been recovered here. Whilst further PCR investigations of the sample with alternative primer sets would allow for this, improving the quality of the sequence would also enable further analysis. One such

approach entails cloning the recovered insert into a bacterial vector and then sequencing multiple bacterial copies of the clone, thus yielding sufficient high quality sequence to allow for further analysis (although this was not performed due to time constraints). The metagenomics data also indicated the presence of a hantavirus (more specifically TATV) within library L which was later confirmed by specific PCR to be present in V1H, supporting the degenerate PCR result and confirming this finding as a true positive virus hit and allowing for species identification as TATV. However, the metagenomics data also indicated the presence of hantaviruses in libraries I, M and N, which were not detected by the degenerate PCR screening. This is likely due to an attempt to develop pan-hantavirus primers that covered both NW and OW hantaviruses, resulting in excessive degeneracy and a lack of specificity in the primer design, in turn leading to limited primer sensitivity. If these primers were to be used in any diagnostic capacity further refinement would be necessary. The rationale behind screening for hantavirus is that TATV has previously been identified in voles in Chester, UK, and the Kielder Forest, UK, which are approximately 45 and 225 miles away from the sampling sites for this study respectively, and Seoul orthohantavirus has also been identified within the UK in rats[11,12]. As such the detection of TATV within these samples acts to expand the known range for TATV within the UK, and it is now reasonable to assess that TATV infection is present within the northern half of England and Wales as a minimum[12].

The coronavirus screening did not yield any positive results, despite a false positive in the initial screening. This false positive was likely a contamination event for that specific PCR, as there were several CoV molecular biology projects being undertaken in the same molecular laboratory as this project at the same time. The metagenomics data indicated that a coronavirus may have been present in library J, but the potentially positive sample identified by degenerate screening (WM18H) is not present in library J. No CoVs were found during the degenerate screening of library J samples and specific confirmatory PCR investigations based on the metagenomics data were also unsuccessful, suggesting that this result may be a false positive. CoVs were investigated due to rodents being well known host species' for α and β coronaviruses within the United Kingdom, and due to current global interest in coronaviruses and their spillover and zoonotic transmission potential[29]. The Woo primers are well established screening primers, but they were initially designed and published in 2005 and due to the rapid evolution of CoVs it was decided that these primers should be updated to include more divergent and novel viruses that have been identified since this assay was published[153,156]. During gradient PCR investigations the Woo_CoV_2 primers performed as well as Woo_CoV primers, suggesting that whilst the modifications made did not reduce either the sensitivity or specificity of the primers, it may not necessarily have improved them either. Further investigations into the capacity of the Woo_CoV_2 primers to detect more divergent and varied CoVs would be beneficial to assess their value as potential diagnostic or evolutionary assessment screening primers.

All samples were negative when screened for *Rubivirus*es. As all controls were as expected and the primers provided a clear positive result when tested on the gradient PCR, this is believed to be a true negative result and is supported by the lack of *Rubivirus* sequence found in the metagenomics data. This is unsurprising as rodent-borne *Rubiviruses* have never yet been detected within the UK[301]). *Rubivirus* screening was performed due to a previous record of the presence of RusV in yellow-necked mice and wood mice, of which 17 and 100 respectively were investigated here[301]. RusV has also previously been found in Germanic Europe, but there is no published literature available regarding any investigations into rustrela presence within the UK therefore this investigation was designed to investigate the presence of RusV in the UK and to fill that knowledge gap[91]. Whilst our study is only confined to one region of the UK, our results suggest that RusV and other *Rubivirus*es are not circulating within UK rodent populations within this region of Wales.

In the metagenomics data *Rotavirus* reads were found in libraries B and I, albeit at very low read numbers, and confirmatory PCRs were unsuccessful for all samples within these libraries. One possibility to explain this is that the degenerate screening primers used were ineffective, despite targeting two different *Rotavirus* segments including the VP7 segment commonly used for *Rotavirus* typing. Designing more specific and less degenerate *Rotavirus* primers (i.e. designing *Rotavirus A* specific primers, *Rotavirus B* specific primers etc. as a multiplex panel) may have resolved this issue, although this was not performed. Another possibility is that the heat denaturation step may have been incorrect in some manner, either by overly denaturing to the extent that the primers didn't bind or by not denaturing enough so that the RNA was still double stranded, as this denaturation step was largely empirical[290]. Alternatively, it is possible that the metagenomics results simply represent low level contamination from the molecular laboratory where *Rotavirus* work has previously been conducted. *Rotavirus* investigations were performed due to the timing of the degenerate screening being undertaken in the winter, as prior to the COVID-19 pandemic *Rotavirus*es showed strong seasonality with an increase in cases in the winter and spring in the UK and USA, and *Rotaviruses* have also shown seasonal prevalence patterns within rodents[294,323]. Therefore, it was believed that there was a reasonable chance of isolating a *Rotavirus* during this timeframe.

Throughout the primer design process for this screening, great care was taken to ensure that primers are extremely unlikely to form any hairpins within 15°C of their annealing temperature, as it was rationalised that this would likely reduce binding and lead to false negative results. However, self-dimerisation was not taken into account at this stage of the project, which may have resulted in reduced primer efficiency and may explain some of the examples where a virus that was not found via degenerate screening was identified in the metagenomics data. Similarly, the rationale behind this screening was to make primers with broad target ranges, where possible targeting a conserved region of the genome to target the entire species, genus or even family. As a result, many of these primers are highly degenerate (up

to 96x in some cases), which may have proved too degenerate and impacted primer concentrations and efficacies. Accordingly, more positive results may have been obtained by designing primers that were more target specific and narrower in scope. Upon repeating this project, primers that are narrower in scope and free of self-dimers would be designed, ideally whilst retaining moderate degeneracy to still identify reasonably divergent viruses, and the sample screening would be repeated accordingly. Additionally, if this project was repeated, sensitivity testing would be performed, where specific primer sets would be tested by attempting to identify targets in a serial dilution of reducing target DNA amongst a background of irrelevant DNA. This would allow for the accurate assessment of sensitivity breakpoints prior to testing specific samples and using valuable cDNA.

Despite the primer design issues, the positive adenovirus results and the hantavirus identification demonstrated the value of PCR investigations in the field of virus discovery and will hopefully form the basis of further investigations into these samples[59,60]. This degenerate screening was performed in part to provide a frame of reference for the metagenomics data as a sense-check, and whilst it may not have been successful in that regards, it proved useful as a minimum to troubleshoot the primer design process to allow for more successful primer design for confirming metagenomics hits.

Chapter 5- Analysis of metagenomics data, PCR confirmation results and phylogenetic analysis of virus hits (prong 2)

After the successful generation of 18 metagenomics libraries and the filtration of the NGS data 1,000,000 reads were analysed for each library. This allowed for the identification of multiple viruses in each library, and for PCR investigations to confirm viral hits and to assess viral prevalence.

1. Hits by virus species

a. Alignments and reference sequences

Once reads were identified as potential virus hits in CZID, they were exported into Geneious Prime and aligned to a reference sequence relevant to the virus in question prior to specific primer design and PCR. All reference sequence accession numbers are shown in Table 9. Reference sequences were selected according to quality of sequence available, genome annotations and logical relatedness to sample (i.e. rodent viruses where available). Whenever possible, reference sequences used were those highlighted in the “refseq” section of NCBI Genbank. This is because reference sequences are required to meet a certain standard and to be validated by NCBI, therefore these were perceived to be more reliable and more likely to be accurate.

Table 9- Accession numbers for reference sequences used for NGS analysis.
Reference sequences selected to match virus identified by CZID hits, and where possible to select a high quality sequence ideally identified as a refseq by NCBI genbank (shown by the NC_ prefix in accession number).

Virus	Reference sequence accession number
Adenovirus	NC_010956
Arterivirus	NC_048210
Astrovirus	NC_036583
Beta papillomavirus	NC_005134
Bocaparvovirus	KY927869
Chapparvovirus	NC_055465
Coronavirus	NC_012936
CMV	NC_006273
Dependoparvovirus	NC_002077
Hepacivirus F	NC_038427
MLV	NC_001501
Orbivirus	Segment 1: NC_027533 Segment 2: NC_027539 Segment 3: NC_027540 Segment 4: NC_027541 Segment 5: NC_027542 Segment 6: NC_027543 Segment 7: NC_027544 Segment 8: NC_027545 Segment 9: NC_027546 Segment 10: NC_027547
Paramyxovirus	KY370098
Pegivirus	NC_021154
Picobirnavirus	Segment 1: NC_007026 Segment 2: NC_007027
Picornavirus- *Cardiovirus*	KY432930
Picornavirus- *Enterovirus C*	MN914205
Picornavirus- *Kunsagivirus*	NC_038317
Picornavirus- *Parechovirus*	NC_034453
Picornavirus- *Rosavirus*	NC_024070
Polyomavirus	NC_055556
Protoparvovirus	NC_038545
Rhadinovirus	NC_055233
Rodent *Hepacivirus*	NC_021153
Rotavirus	Segment 1: NC_011507 Segment 2: NC_011506 Segment 3: NC_011508 Segment 4: NC_011510 Segment 5: NC_011500 Segment 6: NC_011509 Segment 7: NC_011501

	Segment 8: NC_011502 Segment 9: NC_011503 Segment 10: NC_011504 Segment 11: NC_011505
TATV	S segment: NC_055635 M segment: NC_055637 L segment: NC_055636

A total of 114 viruses were identified by CZID during the metagenomics screening, with an average of 6.39 viruses per library. 76/114 viruses (66.67%) were confirmed by PCR in at least 1 animal within the target library. The remaining 38 viruses were often identified with low read numbers and low genomic coverage and were unable to be confirmed by PCR. Table 10 shows all viral reads detected in all libraries. Both CZID and Geneious prime attempt to assemble contigs when mapping to reference, therefore some samples have low read values but high genomic coverage where contig assembly has been highly successful. If a hit had ≥ 30% genomic coverage then this was deemed sufficient for an accurate NTBLAST analysis, and if it had ≥ 60% genomic coverage this was deemed sufficient for phylogenetic analysis, provided all or most of at least 1 informative gene was present within the available sequence. Viruses with genomic coverage of < 30% were still assessed by NTBLAST, but the results were found to be unreliable as they often best matched to extremely short regions of reference sequences- these are not reported here. Genome coverage was assessed by mapping available reads and contigs to the appropriate reference sequence, and then estimating the total % covered by mapped reads i.e. number of bases with reads or contigs mapped divided by the total number of bases in that reference sequence. Whilst this measurement does not account for contig length or gaps, it was deemed sufficient for approximate genomic coverage estimation. PCR confirmation was performed using specific primers designed based on the metagenomics data wherever possible, as described in chapter 2.

Table 10- Summary of viruses identified by CZID.

All viruses detected in all libraries by NGS, the primary host species of that library in which the virus was found (BV=bank vole, FV= field vole, WM= wood mouse, YM= yellow-necked mouse, W= weasel) the number of reads (or contigs when sufficiently assembled) per library, approximate % genome coverage, and prevalence within individual samples in the library as confirmed by PCR. Where ≥ 30% genome was recovered, NTBLAST analysis was performed, and best match and % identity is shown in those cases. "Host" represents a sample where the NTBLAST result identified the reads as most likely to be originating from host genomic carryover. Library L *Picobirnavirus* S2 was identified as either a *Picobirnavirus* or host chromosomal reads with equal probability by NTBLAST and is hence highlighted. Orbivirus S10 represents segment 10.

Virus	Library	Host species	Number of reads	% genome coverage	Proportion positive within library (by PCR)	Best match (NTBLAST)	Similarity to closest match
Adenovirus	A	BV	2	2	3/8		
	B	BV	1	1	1/8		
	C	YM	1	10	7/8		
	E	WM	9	8	3/10		
	G	WM	1635	80	3/10	DQ630759 (*Ovine adenovirus 6* strain WV419/75)	81.86%
	H	WM	10	8	2/10		
	I	WM	124	25	0/10		
	J	WM	52	15	1/10		
	K	WM	19	8	1/9		
	M	BV	13	10	1/5		
	N	FV	6	3	0/5		
	O	WM	1101	60	0/7	NC_014899 (*Murine adenovirus 2*)	85.39%
	P	WM	4	2	2/7		
	Q	YM	7	2	4/5		

	R	YM	5	2	0/4		
Arterivirus	L	FV/W	79	15	1/2		
	N	FV	85	40	1/5	KC862571.1 (*Porcine reproductive and respiratory syndrome virus* isolate DK-2003-6-5)	77.78%
Astrovirus	B	WM	5	25	0/10		
	C	YM	68	15	1/8		
	D	WM	13	10	1/10		
	E	WM	1	10	2/10		
	F	WM	90	65	1/10	LC460091.1 (Astrovirus MLB1 FT1601M3)	94.64%
	H	WM	92	25	4/10		
	I	WM	147	70	2/10	OR043647 (Raccoon dog astrovirus isolate KOR/18-026/intestine/2022)	84.3%
	J	WM	268	25	2/10		
	K	WM	5504	60	3/9	KT946735 (Rodent astrovirus isolate HN-014)	96%
	L	FV/W	1	1	0/2		
	M	BV	885	80	1/5	OR261080 (Bovine astrovirus strain BAstV/T996-2600/France/2020)	92.16%
	N	FV	216	75	1/5	MN626433 (Astrovirus sp. isolate AV/UKMa1_TT)	95%
	O	WM	4	10	1/7		
	P	WM	18	25	0/7		
	Q	YM	10	20	0/5		
	R	YM	16	20	0/4		
Beta papillomavirus	P	WM	4	7	0/7		
	Q	YM	1	1	0/5		

	R	YM	2	2	0/4		
Bocaparvovirus	D	WM	2	5	1/10		
	P	WM	15	25	1/7		
Chapparvovirus	P	WM	1	2	0/7		
Coronavirus	J	WM	5	2	0/10		
CMV	I	WM	10	1	0/10		
Dependoparvovirus	O	WM	228	50	1/7	NC_055486 (*Murine adeno-associated virus 2* isolate MAAV2/NYC/Manhattan/poolF1)	91.57%
Hepacivirus F	A	BV	13	30	1/8	Host	
	B	BV	1	3	1/8		
	E	WM	5	10	2/10		
	I	WM	2	5	0/10		
	L	FV/W	154	70	0/2	Host	
	M	BV	4163	95	1/5	MN242372.1 (*Hepacivirus myodae* isolate MgHV5, complete genome)	73.01%
	N	FV	37	15	2/5		
	P	WM	2	5	1/7		
MLV	A	BV	4	15	8/8		
	C	YM	88	25	7/8		
	D	WM	19	25	10/10		
	E	WM	104	25	10/10		
	F	WM	30	15	10/10		
	H	WM	169	35	10/10	Host	
	I	WM	23	15	10/10		
	J	WM	154	25	10/10		
	K	WM	13	15	2/10		

	L	FV/W	73	25	2/2		
	O	WM	1	2	7/7		
	Q	YM	69	20	5/5		
	R	YM	393	40	4/4	Host	
Orbivirus	E	WM	S10: 42	10	6/10		
Paramyxovirus	C	YM	18	15	1/8		
	D	WM	43	35	1/10	NC_005339 (*Mossman virus*)	94%
Pegivirus	L	FV/W	1	1	0/2		
	N	FV	86	30	1/5	MW897328 (Phaiomys leucurus *Pegivirus* isolate XZS)	75.56%
Picobirnavirus	E	WM	S1: 95 S2: 402	60	1/10	S2: MW930262 (*Picobirnavirus* sp. isolate YS23 RNA-dependent RNA-polymerase gene)	S2: 85.12%
	F	WM	S1: 310 S2: 445	95	2/10	S1: MW930266 (*Picobirnavirus* sp. isolate YS27 RNA-dependent RNA-polymerase gene) S2: MT150089 (*Rabbit Picobirnavirus* isolate rab049pbv01 RNA-dependent RNA polymerase gene)	S1: 89.37% S2: 81.9%
	G	WM	S1: 316 S2:404	95	1/10	S2: QXV86671 (RNA- dependent RNA-polymerase (*Picobirnavirus* sp.))	S2: 73.52 %
	H	WM	S1: 141 S2: 50	20	1/10		
	I	WM	S1: 46 S2: 90	60	0/10	S2: MZ556509 (MAG: *Picobirnavirus* sp. isolate R57-k141_224336)	S2: 93.62%
	J	WM	S1: 12 S2: 11	20	0/10		

	L	FV/W	S1: 51 S2: 48	60	1/2	S2: MW977276 (Porcine *Picobirnavirus* isolate 15213_ NODE_ 440_ len_ 1675_ cov_ 49.185185 RNA-dependent RNA polymerase (RdRp) gene), OR host	S2: 80.12%
	M	BV	S1: 3 S2: 11	20	0/5		
	N	FV	S1: 10 S2: 4	25	0/5		
	O	WM	S1: 99 S2: 135	80	1/7	S2: KY399057 (*Picobirnavirus* dog/ KNA/ 2015 strain PBV/ Dog/ KNA/ RVC7/ 2015 RNA-dependent RNA polymerase gene)	S2: 89.51%
	P	WM	S1: 2 S2: 5	20	0/7		
	R	YM	S1: 10 S2: 2	20	1/4		
Picornavirus-*Cardiovirus*	A	BV	9	10	5/8		
	B	BV	3	10	2/8		
	D	WM	34	60	1/10	JX257003 (*Encephalomyocarditis virus* type 2 isolate RD 1338 (D28/05) polyprotein gene)	91.38%
	E	WM	5	10	1/10		
	G	WM	559	90	2/10	OP381184 (*Encephalomyocarditis virus* isolate UK2016)	98%
	I	WM	2	95	1/10	ON136175 (*Cardiovirus* species)	82.22%
	J	WM	1	5	0/10		
	M	BV	174	60	1/5	NC_075977 (*Cardiovirus* F1 isolate RtMruf-PicoV/ JL2014-1 polyprotein (QKJ43_gp1) gene, complete cds)	84.8%

	N	FV	3	5	0/5		
	O	WM	3	8	0/7		
	P	WM	5	10	0/7		
	Q	YM	5	10	0/5		
	R	YM	5	10	0/4		
Picornavirus-*Enterovirus C*	M	BV	1	2	0/5		
Picornavirus-*Kunsagivirus*	P	WM	116	75	2/7	ON136180 (*Kunsagivirus* species)	91.24%
Picornavirus-*Parechovirus*	L	FV/W	4	2	0/2		
Picornavirus-*Rosavirus*	B	BV	10	50	2/8	NC_038880 (*Rosavirus* M-7 polyprotein)	84.06%
	M	BV	1	2	0/5		
Polyomavirus	F	WM	41	50	1/10	NC_055556 (*Apodemus flavicollis polyomavirus 1* isolate 3346)	96.24%
Protoparvovirus	F	WM	4	15	3/10		
	I	WM	4	25	7/10		
	O	WM	3	5	3/7		
Rhadinovirus	O	WM	722	45	5/7	EF495130 (Wood mouse herpesvirus strain Brest/An711)	99.39%
	Q	YM	258	5	3/5		
Rodent *Hepacivirus*	N	FV	37	15	1/5		
Rotavirus	B	WM	1-3 per segment	15	0/10		
	I	WM	1 per segment	2	0/10		

Tatenale virus	E	WM	S: 1 L: 1	2	0/10		
	F	WM	M: 1	1	0/10		
	I	WM	M: 1	1	0/10		
	L	FV/W	S: 1 M: 2 L: 2	5	1/2		
	M	BV	S: 1 M: 1 L: 1	1	0/5		
	N	FV	S: 264 M: 268 L: 358	95	1/5	S: NC_055635 (*Tatenale orthohantavirus* strain Upton_Heath segment S) M: MK883759 (*Tatenale orthohantavirus* strain Norton_Juxta segment M) L: MK883761 (*Tatenale orthohantavirus* strain Norton_Juxta segment L)	S: 90.85% M: 89.85% L: 86.89%

b. Adenoviruses were frequently identified

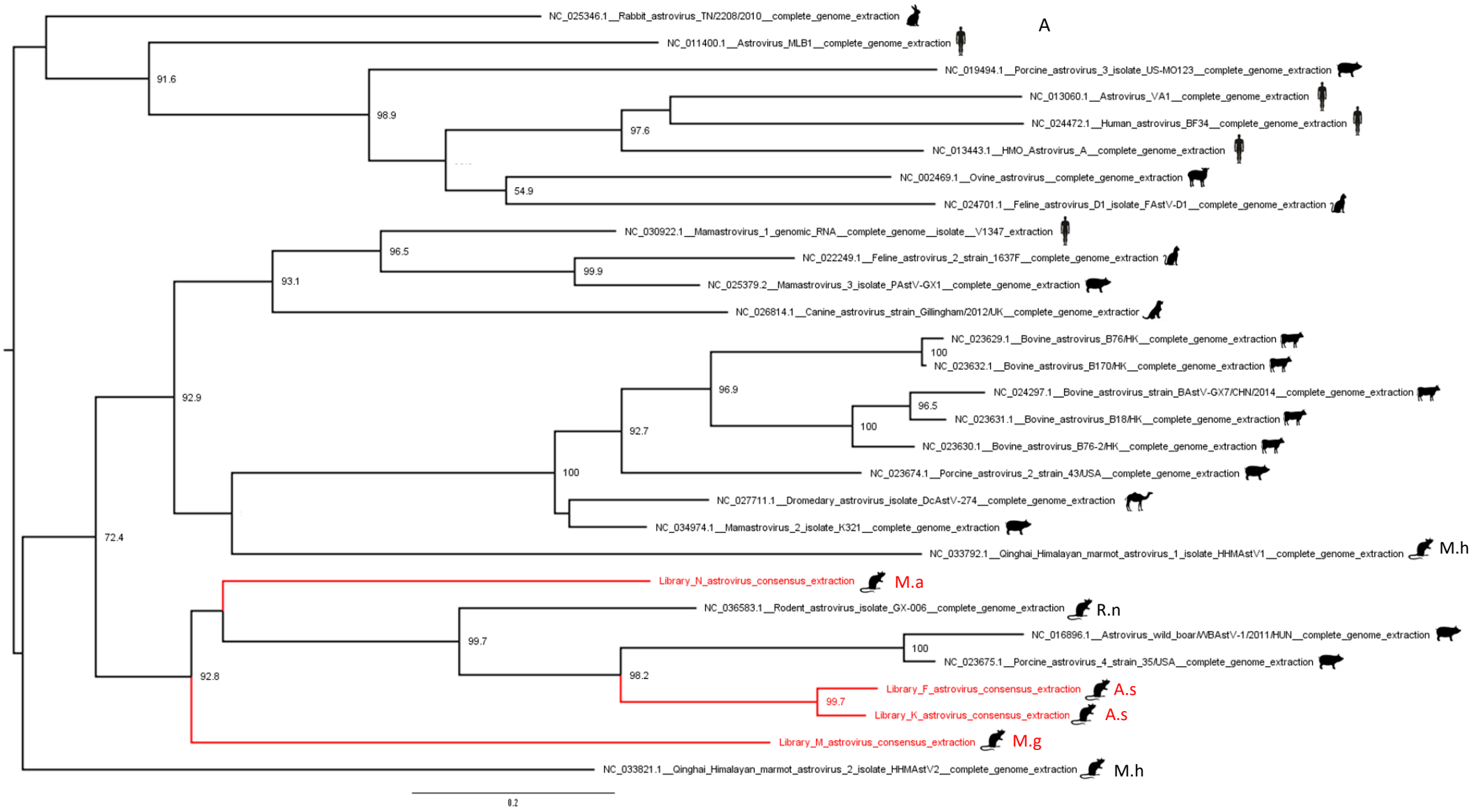
Adenoviruses were PCR confirmed in a total of 28 individual animals (20% of total animals) and adenovirus reads were identified in a total of 15 libraries. Hits were confirmed in a total of 12/15 libraries, with libraries I, N and O failing to provide PCR hits. Sufficient genomic coverage for NTBLAST analysis was only present in libraries G and O, with other libraries ranging from approximately 1-25% genomic coverage (Table 10). The library G sequence provided approximately 80% genomic coverage and by NTBLAST the best match was *ovine adenovirus 6* (DQ630759) with 81.86% similarity. The library O sequence provided approximately 60% coverage and by NTBLAST the best match was *murine adenovirus 2* (NC_014899) with approximately 85.39% similarity. Phylogenetic analysis was attempted on these libraries, but due to large sequence gaps throughout key genes this was not possible.

c. Astroviruses were frequently identified

Astroviruses were PCR confirmed within 24 individual animals (17.1% of total animals), and astrovirus reads were found in a total of 16 libraries. Hits were successfully PCR confirmed within 11/16 libraries, and of these 5 libraries provided sufficient genomic coverage for NTBLAST analysis (libraries F, I, K, M and N). The remaining libraries had genomic coverage ranging from approximately 1-25% (Table 10). The library M sequence provided approximately 80% genomic coverage and the best match by NTBLAST was a bovine astrovirus strain (OR261080) with approximately 92.16% similarity. The library N sequence provided approximately 75% genomic coverage and the best match was an astrovirus species sequence (MN626433) with approximately 95% similarity. The library I sequence provided approximately 70% genomic coverage with a best match of a racoon dog astrovirus (OR043647) with approximately 84.3% similarity. The library F sequence provided approximately 65% coverage and best matched an astrovirus species sequence (LC460091) with approximately 94.64% similarity, and the library K sequence provided approximately 60% genomic coverage with a best match of a rodent astrovirus sequence (KT946735), with 96% similarity.

Phylogenetic analysis of the entire highly conserved astrovirus RdRp gene (bases 3181-3885 relative to the rodent Astrovirus reference sequence NC_036583) and the less conserved capsid N gene (bases 4419-5341 relative to the rodent astrovirus reference sequence NC_036583) was performed on the library F, I, K, M and N sequences (Figures 10A and 10B, respectively). The library I sequence formed outgroups in both genes which substantially reduced the clarity of the trees and was therefore omitted. In the RdRp gene, all libraries clustered amongst other rodent astroviruses, and the libraries F and K sequences formed an individual clade that clustered amongst rodent and porcine astroviruses. The libraries M and N sequences were more divergent, forming individual branches, although still amongst rodent astroviruses. In the capsid gene the library M and N sequences formed individual branches that were most closely related to a rodent astrovirus, whilst the library F and K sequences formed a clade that clustered with the same porcine astroviruses as

for the RdRp gene. In both genes, library F and K sequences formed a clade, whilst libraries M and N sequences were quite divergent relative to the library F and K clade, suggesting that least two astrovirus species or strains may be present. Further ORF2 genomic information is required to assess species demarcation[324].


A
NC_025346.1__Rabbit_astrovirus_TN/2208/2010__complete_genome_extraction
NC_011400.1__Astrovirus_MLB1__complete_genome_extraction
NC_019494.1__Porcine_astrovirus_3_isolate_US-MO123__complete_genome_extraction
NC_013060.1__Astrovirus_VA1__complete_genome_extraction
NC_024472.1__Human_astrovirus_BF34__complete_genome_extraction
NC_013443.1__HMO_Astrovirus_A__complete_genome_extraction
NC_002469.1__Ovine_astrovirus__complete_genome_extraction
NC_024701.1__Feline_astrovirus_D1_isolate_FAstV-D1__complete_genome_extraction
NC_030922.1__Mamastrovirus_1_genomic_RNA__complete_genome__isolate__V1347_extraction
NC_022249.1__Feline_astrovirus_2_strain_1637F__complete_genome_extraction
NC_025379.2__Mamastrovirus_3_isolate_PAstV-GX1__complete_genome_extraction
NC_026814.1__Canine_astrovirus_strain_Gillingham/2012/UK__complete_genome_extraction
NC_023629.1__Bovine_astrovirus_B76/HK__complete_genome_extraction
NC_023632.1__Bovine_astrovirus_B170/HK__complete_genome_extraction
NC_024297.1__Bovine_astrovirus_strain_BAstV-GX7/CHN/2014__complete_genome_extraction
NC_023631.1__Bovine_astrovirus_B18/HK__complete_genome_extraction
NC_023630.1__Bovine_astrovirus_B76-2/HK__complete_genome_extraction
NC_023674.1__Porcine_astrovirus_2_strain_43/USA__complete_genome_extraction
NC_027711.1__Dromedary_astrovirus_isolate_DcAstV-274__complete_genome_extraction
NC_034974.1__Mamastrovirus_2_isolate_K321__complete_genome_extraction
NC_033792.1__Qinghai_Himalayan_marmot_astrovirus_1_isolate_HHMAstV1__complete_genome_extraction
M.h
Library_N_astrovirus_consensus_extraction
M.a
NC_036583.1__Rodent_astrovirus_isolate_GX-006__complete_genome_extraction
R.n
NC_016896.1__Astrovirus_wild_boar/WBAstV-1/2011/HUN__complete_genome_extraction
NC_023675.1__Porcine_astrovirus_4_strain_35/USA__complete_genome_extraction
Library_F_astrovirus_consensus_extraction
A.s
Library_K_astrovirus_consensus_extraction
A.s
Library_M_astrovirus_consensus_extraction
M.g
NC_033821.1__Qinghai_Himalayan_marmot_astrovirus_2_isolate_HHMAstV2__complete_genome_extraction
M.h
91.6
98.9
97.6
54.9
96.5
99.9
93.1
92.9
100
96.9
92.7
96.5
100
72.4
99.7
100
98.2
99.7
92.8
0.2

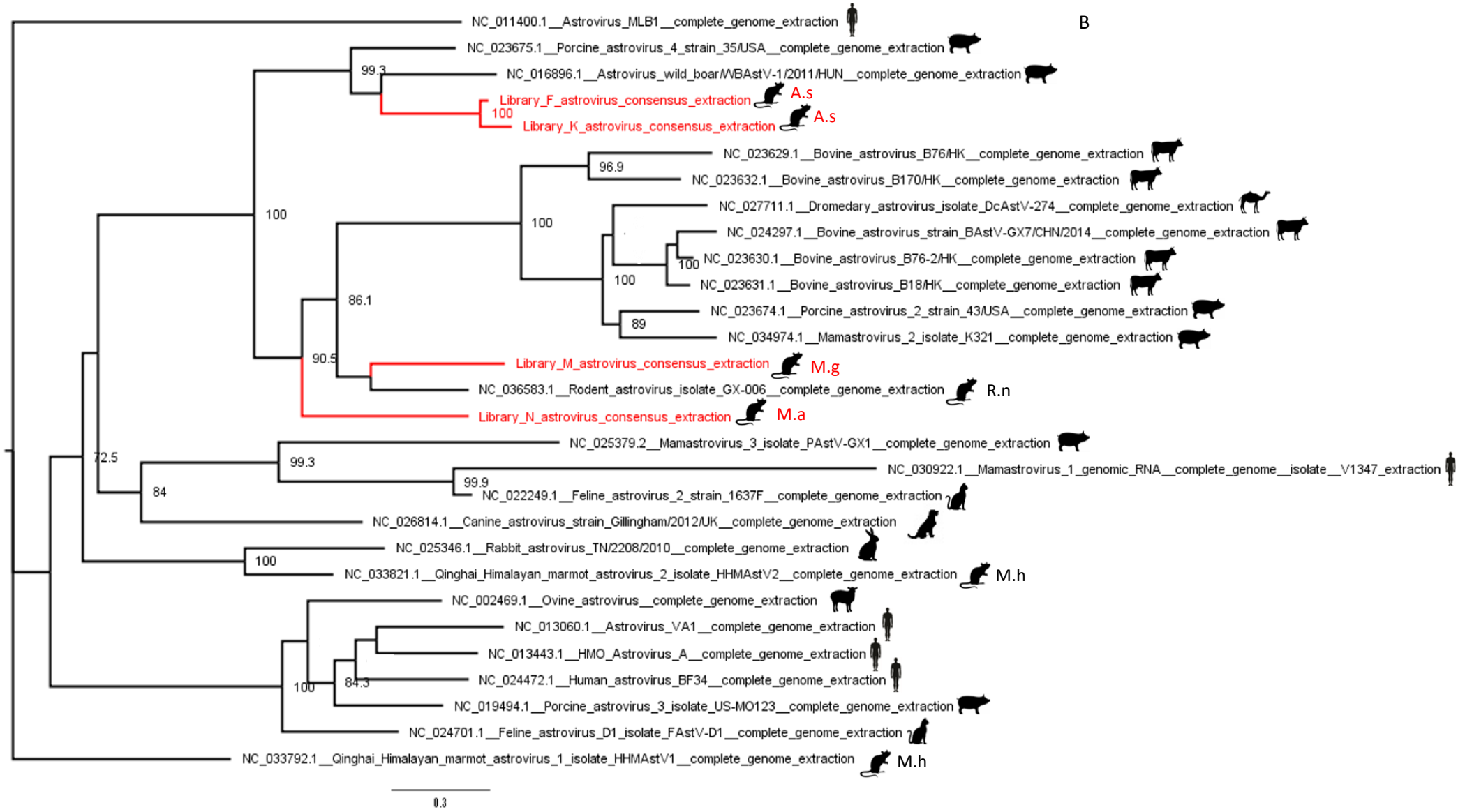
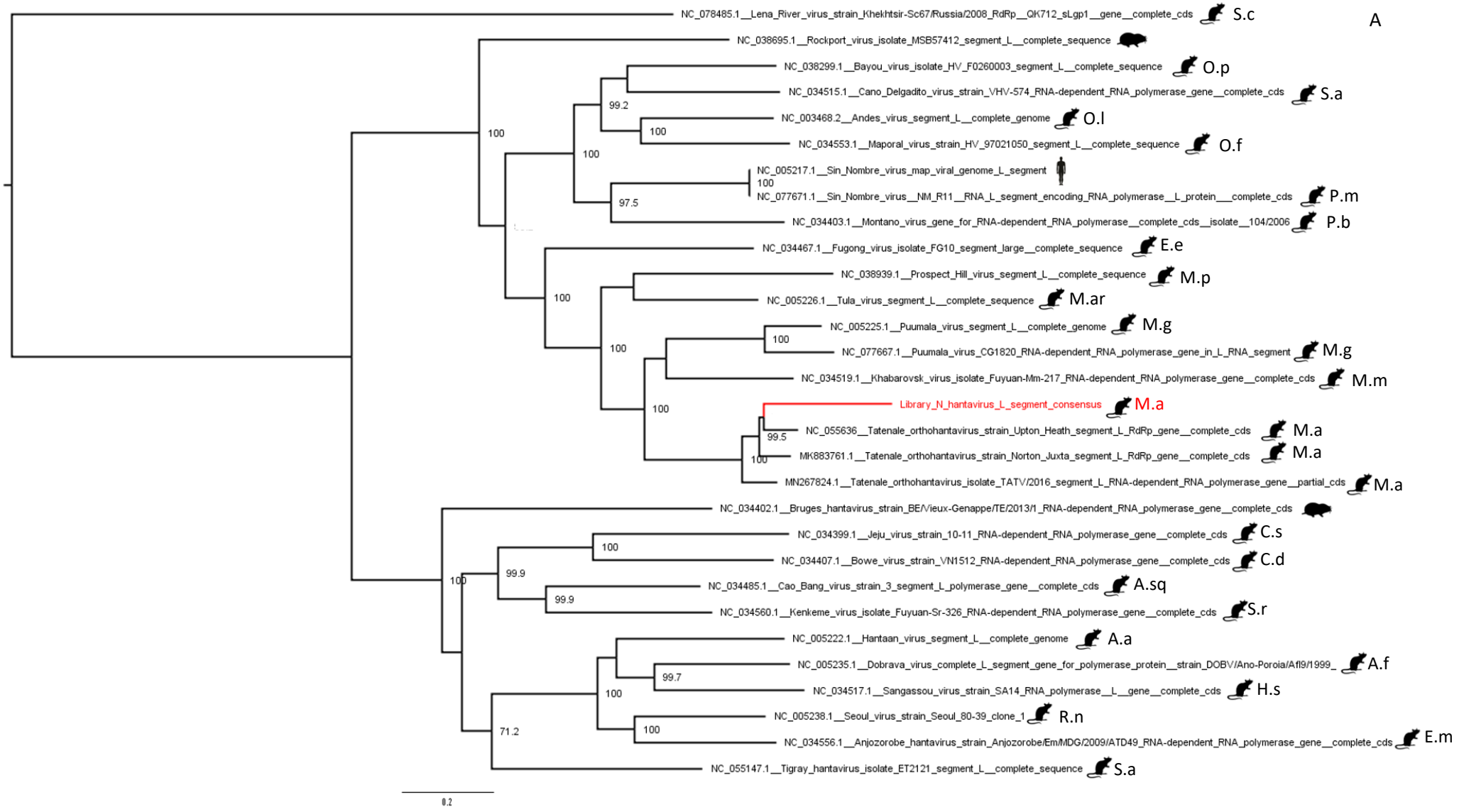

B
NC_011400.1__Astrovirus_MLB1__complete_genome_extraction
NC_023675.1__Porcine_astrovirus_4_strain_35/USA__complete_genome_extraction
NC_016896.1__Astrovirus_wild_boar/WBAstV-1/2011/HUN__complete_genome_extraction
Library_F_astrovirus_consensus_extraction
A.s
Library_K_astrovirus_consensus_extraction
A.s
NC_023629.1__Bovine_astrovirus_B76/HK__complete_genome_extraction
NC_023632.1__Bovine_astrovirus_B170/HK__complete_genome_extraction
NC_027711.1__Dromedary_astrovirus_isolate_DcAstV-274__complete_genome_extraction
NC_024297.1__Bovine_astrovirus_strain_BAstV-GX7/CHN/2014__complete_genome_extraction
NC_023630.1__Bovine_astrovirus_B76-2/HK__complete_genome_extraction
NC_023631.1__Bovine_astrovirus_B18/HK__complete_genome_extraction
NC_023674.1__Porcine_astrovirus_2_strain_43/USA__complete_genome_extraction
NC_034974.1__Mamastrovirus_2_isolate_K321__complete_genome_extraction
Library_M_astrovirus_consensus_extraction
M.g
NC_036583.1__Rodent_astrovirus_isolate_GX-006__complete_genome_extraction
R.n
Library_N_astrovirus_consensus_extraction
M.a
NC_025379.2__Mamastrovirus_3_isolate_PAstV-GX1__complete_genome_extraction
NC_030922.1__Mamastrovirus_1_genomic_RNA__complete_genome__isolate__V1347_extraction
NC_022249.1__Feline_astrovirus_2_strain_1637F__complete_genome_extraction
NC_026814.1__Canine_astrovirus_strain_Gillingham/2012/UK__complete_genome_extraction
NC_025346.1__Rabbit_astrovirus_TN/2208/2010__complete_genome_extraction
NC_033821.1__Qinghai_Himalayan_marmot_astrovirus_2_isolate_HHMAstV2__complete_genome_extraction
M.h
NC_002469.1__Ovine_astrovirus__complete_genome_extraction
NC_013060.1__Astrovirus_VA1__complete_genome_extraction
NC_013443.1__HMO_Astrovirus_A__complete_genome_extraction
NC_024472.1__Human_astrovirus_BF34__complete_genome_extraction
NC_019494.1__Porcine_astrovirus_3_isolate_US-MO123__complete_genome_extraction
NC_024701.1__Feline_astrovirus_D1_isolate_FAstV-D1__complete_genome_extraction
NC_033792.1__Qinghai_Himalayan_marmot_astrovirus_1_isolate_HHMAstV1__complete_genome_extraction
M.h
99.3
100
96.9
100
100
100
86.1
89
90.5
72.5
99.3
99.9
84
100
100
84.3
0.3

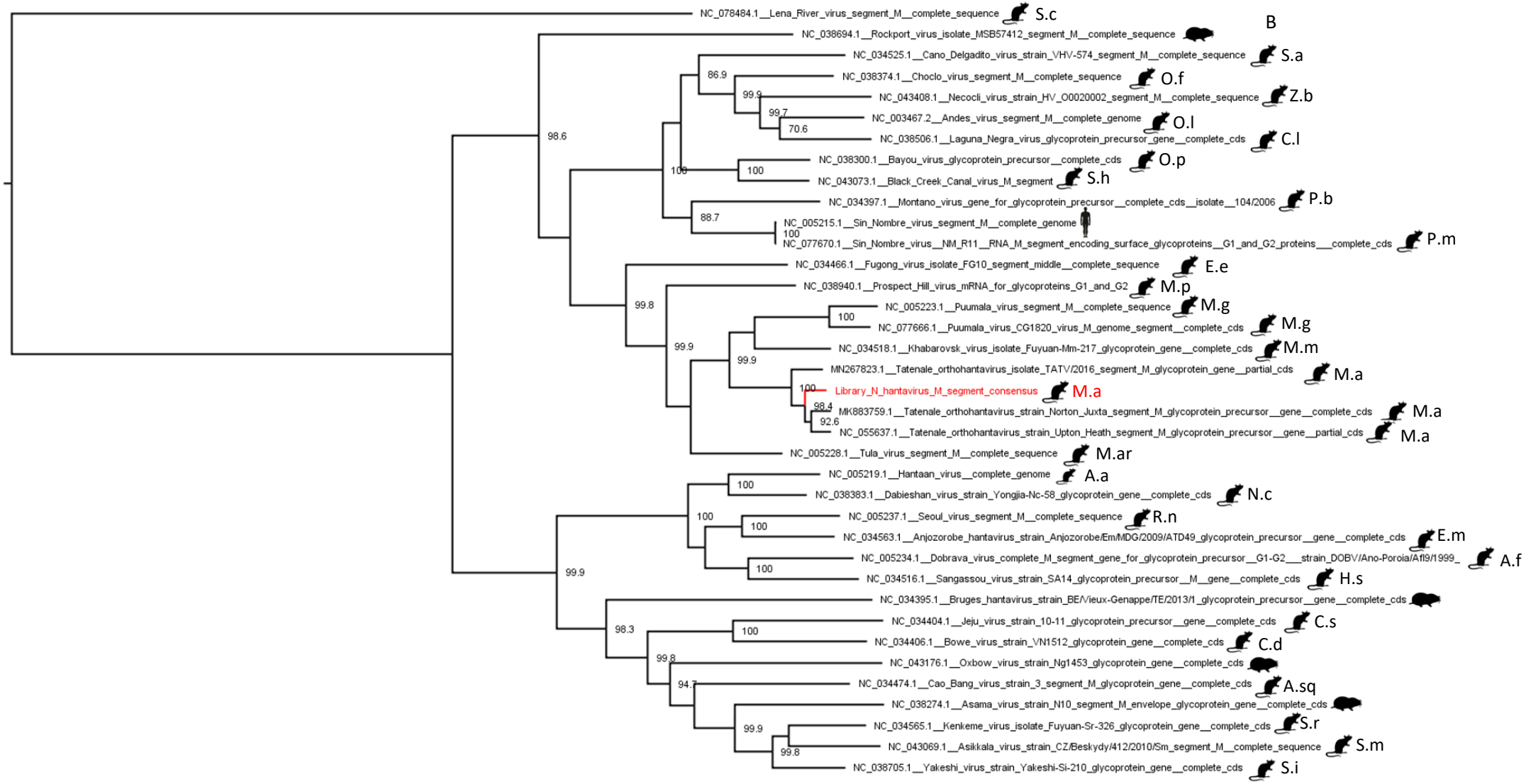
Figure 10- Phylogenetic analysis of the NGS identified astrovirus sequences.
A shows midpoint rooted phylogenetic tree of complete astrovirus RdRp gene, relative to bases 3181-3885 of reference sequence NC_036583 (rodent astrovirus). Tree generated using TVMe+I+G4 model with 1000 bootstrap replicates. Scale bar represents 0.2 substitutions per site. B shows midpoint rooted phylogenetic tree of astrovirus capsid N gene bases 4419-5339 relative to reference sequence NC_036583 (rodent astrovirus). Tree generated using TVMe+G4 model with 1000 bootstrap replicates. Scale bar represents 0.3 substitutions per site. For both trees, node values represent % bootstrap support, red branches and text highlight sample libraries, and host species is shown next to each branch. M.h= *Marmota himalayana*, M.a= *Microtus agrestis*, R.n= *Rattus norvegicus*, A.s= *Apodemus sylvaticus*, M.g= *Myodes glareolus*.

d. A near full hantavirus genome was recovered

Hantavirus reads were identified as TATV by CZID in 6/18 libraries (E, F, I, L, M and N) (Table 10). Libraries E, F, I and M gave 1-2% genomic coverage with between 2 and 5 reads per library and were unable to be confirmed by PCR. 2 field voles tested positive for the hantavirus L segment by PCR (1.4% of total animals), 1 of which was in library L and the other in library N. The library L sequence provided approximately 5% genomic coverage and therefore was insufficient for NTBLAST analysis. The library N sequences provided approximately 95% genomic information (full M and S segments and approximately 95% of the L segment). Upon NTBLAST analysis the S segment was identified as most similar to TATV strain Upton Heath (NC_055635) with 90.85% similarity, the M segment was identified as most similar to TATV strain Norton Juxta (MK883759) with 89.85% similarity, and the L segment was identified as most similar to TATV strain Norton Juxta (MK883761) with 86.89% similarity.

Phylogenetic analysis of all segments of the library N hantavirus was performed (Figures 11 A, 11 B and 11 C show the phylogenetic analysis of the L, M and S segments respectively). In all segments the library N segment formed a clade amongst TATVs. By BLASTX (translated protein BLAST) amino acid analysis, the M segment is 98.08% identical to a known TATV (QIA61110) and the S segment is 97.2% identical to another known TATV (QIA61108).


A
NC_078485.1__Lena_River_virus_strain_Khekhtsir-Sc67/Russia/2008_RdRp__QK712_sLgp1__gene__complete_cds
S.c
NC_038695.1__Rockport_virus_isolate_MSB57412_segment_L__complete_sequence
NC_038299.1__Bayou_virus_isolate_HV_F0260003_segment_L__complete_sequence
O.p
NC_034515.1__Cano_Delgadito_virus_strain_VHV-574_RNA-dependent_RNA_polymerase_gene__complete_cds
S.a
NC_003468.2__Andes_virus_segment_L__complete_genome
O.l
NC_034553.1__Maporal_virus_strain_HV_97021050_segment_L__complete_sequence
O.f
NC_005217.1__Sin_Nombre_virus_map_viral_genome_L_segment
NC_077671.1__Sin_Nombre_virus__NM_R11__RNA_L_segment_encoding_RNA_polymerase__L_protein__complete_cds
P.m
NC_034403.1__Montano_virus_gene_for_RNA-dependent_RNA_polymerase__complete_cds__isolate__104/2006
P.b
NC_034467.1__Fugong_virus_isolate_FG10_segment_large__complete_sequence
E.e
NC_038939.1__Prospect_Hill_virus_segment_L__complete_sequence
M.p
NC_005226.1__Tula_virus_segment_L__complete_sequence
M.ar
NC_005225.1__Puumala_virus_segment_L__complete_genome
M.g
NC_077667.1__Puumala_virus_CG1820_RNA-dependent_RNA_polymerase_gene_in_L_RNA_segment
M.g
NC_034519.1__Khabarovsk_virus_isolate_Fuyuan-Mm-217_RNA-dependent_RNA_polymerase_gene__complete_cds
M.m
Library_N_hantavirus_L_segment_consensus
M.a
NC_055636__Tatenale_orthohantavirus_strain_Upton_Heath_segment_L_RdRp_gene__complete_cds
M.a
MK883761.1__Tatenale_orthohantavirus_strain_Norton_Juxta_segment_L_RdRp_gene__complete_cds
M.a
MN267824.1__Tatenale_orthohantavirus_isolate_TATV/2016_segment_L_RNA-dependent_RNA_polymerase_gene__partial_cds
M.a
NC_034402.1__Bruges_hantavirus_strain_BE/Vieux-Genappe/TE/2013/1_RNA-dependent_RNA_polymerase_gene__complete_cds
NC_034399.1__Jeju_virus_strain_10-11_RNA-dependent_RNA_polymerase_gene__complete_cds
C.s
NC_034407.1__Bowe_virus_strain_VN1512_RNA-dependent_RNA_polymerase_gene__complete_cds
C.d
NC_034485.1__Cao_Bang_virus_strain_3_segment_L_polymerase_gene__complete_cds
A.sq
NC_034560.1__Kenkeme_virus_isolate_Fuyuan-Sr-326_RNA-dependent_RNA_polymerase_gene__complete_cds
S.r
NC_005222.1__Hantaan_virus_segment_L__complete_genome
A.a
NC_005235.1__Dobrava_virus_complete_L_segment_gene_for_polymerase_protein__strain_DOBV/Ano-Poroia/Afl9/1999_
A.f
NC_034517.1__Sangassou_virus_strain_SA14_RNA_polymerase__L__gene__complete_cds
H.s
NC_005238.1__Seoul_virus_strain_Seoul_80-39_clone_1
R.n
NC_034556.1__Anjozorobe_hantavirus_strain_Anjozorobe/Em/MDG/2009/ATD49_RNA-dependent_RNA_polymerase_gene__complete_cds
E.m
NC_055147.1__Tigray_hantavirus_isolate_ET2121_segment_L__complete_sequence
S.a
100
99.2
100
100
97.5
100
100
100
99.5
100
100
99.9
99.9
100
100
99.7
71.2
100
0.2


NC_078484.1__Lena_River_virus_segment_M__complete_sequence
S.c
B
NC_038694.1__Rockport_virus_isolate_MSB57412_segment_M__complete_sequence
NC_034525.1__Cano_Delgadito_virus_strain_VHV-574_segment_M__complete_sequence
S.a
86.9
NC_038374.1__Choclo_virus_segment_M__complete_sequence
O.f
99.9
NC_043408.1__Necocli_virus_strain_HV_00020002_segment_M__complete_sequence
Z.b
99.7
NC_003467.2__Andes_virus_segment_M__complete_genome
O.l
70.6
NC_038506.1__Laguna_Negra_virus_glycoprotein_precursor_gene__complete_cds
C.l
98.6
NC_038300.1__Bayou_virus_glycoprotein_precursor__complete_cds
O.p
100
100
NC_043073.1__Black_Creek_Canal_virus_M_segment
S.h
NC_034397.1__Montano_virus_gene_for_glycoprotein_precursor__complete_cds__isolate__104/2006
P.b
88.7
NC_005215.1__Sin_Nombre_virus_segment_M__complete_genome
100
NC_077670.1__Sin_Nombre_virus__NM_R11__RNA_M_segment_encoding_surface_glycoproteins__G1_and_G2_proteins__complete_cds
P.m
NC_034466.1__Fugong_virus_isolate_FG10_segment_middle__complete_sequence
E.e
NC_038940.1__Prospect_Hill_virus_mRNA_for_glycoproteins_G1_and_G2
M.p
99.8
NC_005223.1__Puumala_virus_segment_M__complete_sequence
M.g
100
NC_077666.1__Puumala_virus_CG1820_virus_M_genome_segment__complete_cds
M.g
99.9
NC_034518.1__Khabarovsk_virus_isolate_Fuyuan-Mm-217_glycoprotein_gene__complete_cds
M.m
99.9
MN267823.1__Tatenale_orthohantavirus_isolate_TATV/2016_segment_M_glycoprotein_gene__partial_cds
M.a
100
Library_N_hantavirus_M_segment_consensus
M.a
98.4
MK883759.1__Tatenale_orthohantavirus_strain_Norton_Juxta_segment_M_glycoprotein_precursor__gene__complete_cds
M.a
92.6
NC_055637.1__Tatenale_orthohantavirus_strain_Upton_Heath_segment_M_glycoprotein_precursor__gene__partial_cds
M.a
NC_005228.1__Tula_virus_segment_M__complete_sequence
M.ar
NC_005219.1__Hantaan_virus__complete_genome
A.a
100
NC_038383.1__Dabieshan_virus_strain_Yongjia-Nc-58_glycoprotein_gene__complete_cds
N.c
100
NC_005237.1__Seoul_virus_segment_M__complete_sequence
R.n
100
NC_034563.1__Anjozorobe_hantavirus_strain_Anjozorobe/Em/MDG/2009/ATD49_glycoprotein_precursor__gene__complete_cds
E.m
99.9
NC_005234.1__Dobrava_virus_complete_M_segment_gene_for_glycoprotein_precursor__G1-G2___strain_DOBV/Ano-Poroia/Afl9/1999_
A.f
100
NC_034516.1__Sangassou_virus_strain_SA14_glycoprotein_precursor__M__gene__complete_cds
H.s
NC_034395.1__Bruges_hantavirus_strain_BE/Vieux-Genappe/TE/2013/1_glycoprotein_precursor__gene__complete_cds
98.3
NC_034404.1__Jeju_virus_strain_10-11_glycoprotein_precursor__gene__complete_cds
C.s
100
NC_034406.1__Bowe_virus_strain_VN1512_glycoprotein_gene__complete_cds
C.d
99.8
NC_043176.1__Oxbow_virus_strain_Ng1453_glycoprotein_gene__complete_cds
94.7
NC_034474.1__Cao_Bang_virus_strain_3_segment_M_glycoprotein_gene__complete_cds
A.sq
NC_038274.1__Asama_virus_strain_N10_segment_M_envelope_glycoprotein_gene__complete_cds
99.9
NC_034565.1__Kenkeme_virus_isolate_Fuyuan-Sr-326_glycoprotein_gene__complete_cds
S.r
NC_043069.1__Asikkala_virus_strain_CZ/Beskydy/412/2010/Sm_segment_M__complete_sequence
S.m
99.8
NC_038705.1__Yakeshi_virus_strain_Yakeshi-Si-210_glycoprotein_gene__complete_cds
S.i
0.3

C

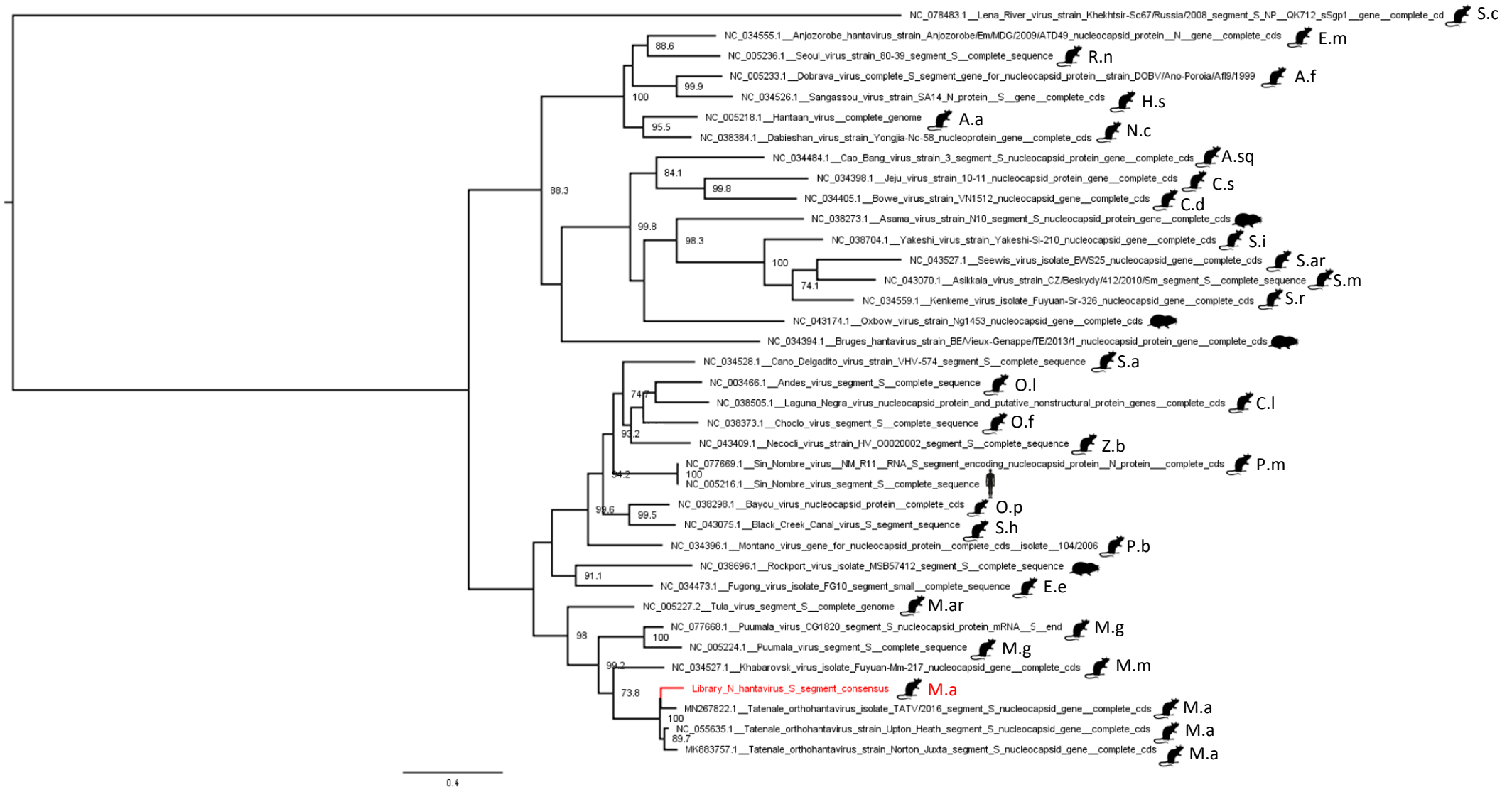

NC_078483.1__Lena_River_virus_strain_Khekhtsir-Sc67/Russia/2008_segment_S_NP__QK712_sSgp1__gene__complete_cd
S.c
NC_034555.1__Anjozorobe_hantavirus_strain_Anjozorobe/Em/MDG/2009/ATD49_nucleocapsid_protein__N__gene__complete_cds
E.m
88.6
NC_005236.1__Seoul_virus_strain_80-39_segment_S__complete_sequence
R.n
NC_005233.1__Dobrava_virus_complete_S_segment_gene_for_nucleocapsid_protein__strain_DOBV/Ano-Poroia/Afl9/1999
A.f
99.9
100
NC_034526.1__Sangassou_virus_strain_SA14_N_protein__S__gene__complete_cds
H.s
NC_005218.1__Hantaan_virus__complete_genome
A.a
95.5
NC_038384.1__Dabieshan_virus_strain_Yongjia-Nc-58_nucleoprotein_gene__complete_cds
N.c
NC_034484.1__Cao_Bang_virus_strain_3_segment_S_nucleocapsid_protein_gene__complete_cds
A.sq
84.1
NC_034398.1__Jeju_virus_strain_10-11_nucleocapsid_protein_gene__complete_cds
C.s
99.8
88.3
NC_034405.1__Bowe_virus_strain_VN1512_nucleocapsid_gene__complete_cds
C.d
NC_038273.1__Asama_virus_strain_N10_segment_S_nucleocapsid_protein_gene__complete_cds
99.8
NC_038704.1__Yakeshi_virus_strain_Yakeshi-Si-210_nucleocapsid_gene__complete_cds
S.i
98.3
NC_043527.1__Seewis_virus_isolate_EWS25_nucleocapsid_gene__complete_cds
S.ar
100
NC_043070.1__Asikkala_virus_strain_CZ/Beskydy/412/2010/Sm_segment_S__complete_sequence
S.m
74.1
NC_034559.1__Kenkeme_virus_isolate_Fuyuan-Sr-326_nucleocapsid_gene__complete_cds
S.r
NC_043174.1__Oxbow_virus_strain_Ng1453_nucleocapsid_gene__complete_cds
NC_034394.1__Bruges_hantavirus_strain_BE/Vieux-Genappe/TE/2013/1_nucleocapsid_protein_gene__complete_cds
NC_034528.1__Cano_Delgadito_virus_strain_VHV-574_segment_S__complete_sequence
S.a
NC_003466.1__Andes_virus_segment_S__complete_sequence
O.l
74.7
NC_038505.1__Laguna_Negra_virus_nucleocapsid_protein_and_putative_nonstructural_protein_genes__complete_cds
C.l
NC_038373.1__Choclo_virus_segment_S__complete_sequence
O.f
93.2
NC_043409.1__Necocli_virus_strain_HV_O0020002_segment_S__complete_sequence
Z.b
NC_077669.1__Sin_Nombre_virus__NM_R11__RNA_S_segment_encoding_nucleocapsid_protein__N_protein__complete_cds
P.m
94.2
100
NC_005216.1__Sin_Nombre_virus_segment_S__complete_sequence
NC_038298.1__Bayou_virus_nucleocapsid_protein__complete_cds
O.p
99.5
NC_043075.1__Black_Creek_Canal_virus_S_segment_sequence
S.h
NC_034396.1__Montano_virus_gene_for_nucleocapsid_protein__complete_cds__isolate__104/2006
P.b
NC_038696.1__Rockport_virus_isolate_MSB57412_segment_S__complete_sequence
91.1
NC_034473.1__Fugong_virus_isolate_FG10_segment_small__complete_sequence
E.e
NC_005227.2__Tula_virus_segment_S__complete_genome
M.ar
98
NC_077668.1__Puumala_virus_CG1820_segment_S_nucleocapsid_protein_mRNA__5__end
M.g
100
NC_005224.1__Puumala_virus_segment_S__complete_sequence
M.g
99.2
NC_034527.1__Khabarovsk_virus_isolate_Fuyuan-Mm-217_nucleocapsid_gene__complete_cds
M.m
73.8
Library_N_hantavirus_S_segment_consensus
M.a
MN267822.1__Tatenale_orthohantavirus_isolate_TATV/2016_segment_S_nucleocapsid_gene__complete_cds
M.a
100
NC_055635.1__Tatenale_orthohantavirus_strain_Upton_Heath_segment_S_nucleocapsid_gene__complete_cds
M.a
89.7
MK883757.1__Tatenale_orthohantavirus_strain_Norton_Juxta_segment_S_nucleocapsid_gene__complete_cds
M.a
0.4

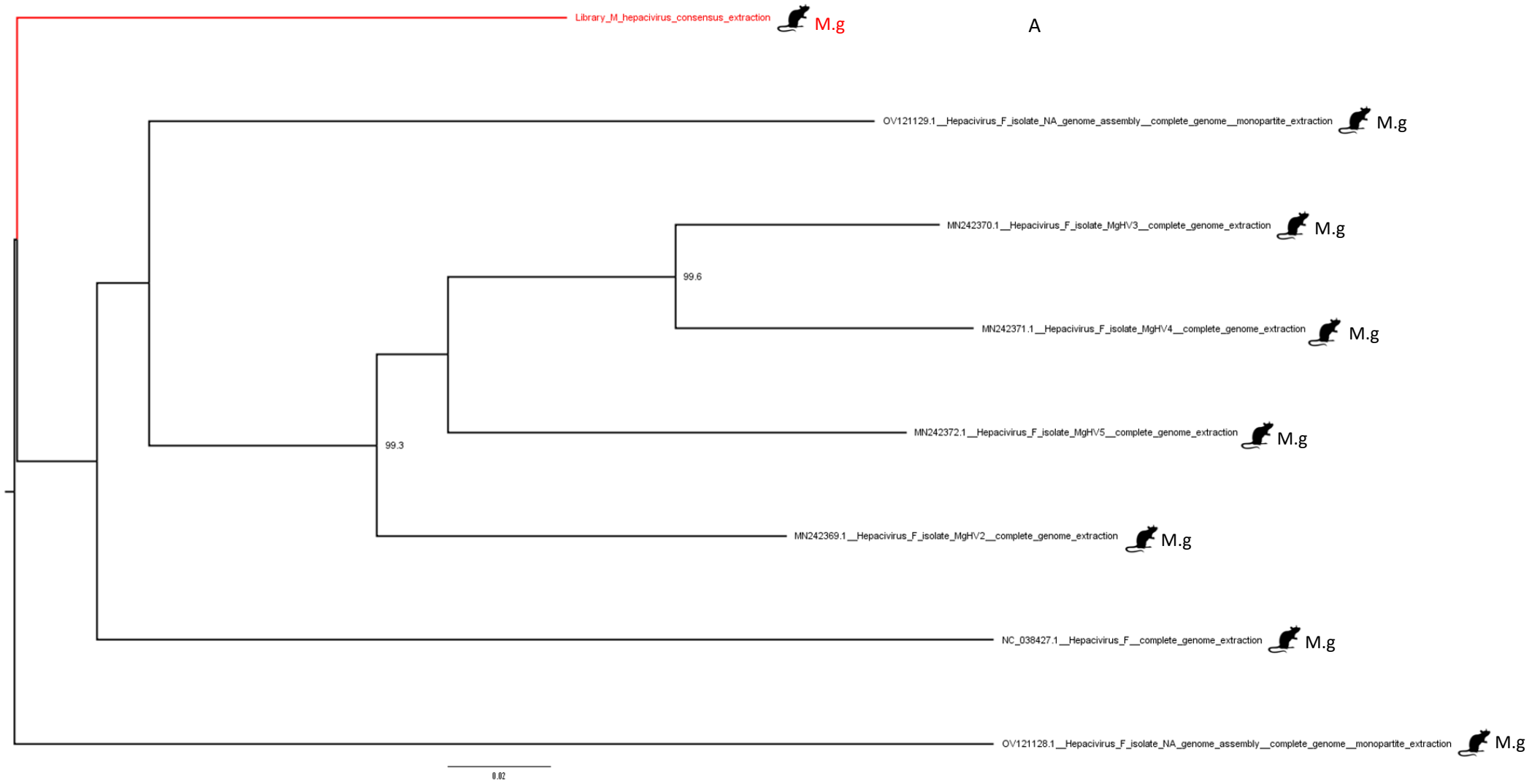
Figure 11- Phylogenetic analysis of NGS identified hantavirus.
A shows midpoint rooted phylogenetic tree of hantavirus L segment (RdRp gene) CDS, bases 1-6465 relative to TATV *orthohantavirus* strain Upton-Heath reference sequence (NC_055636). Tree generated using GTR+F+I+R5 model with 1000 bootstrap replicates. Scale bar represents 0.2 substitutions per site. B shows midpoint rooted phylogenetic tree of entire hantavirus M segment (glycoprotein precursor gene), bases 1-3355 relative to TATV *orthohantavirus* strain Upton-Heath reference sequence (NC_055637). Tree generated using GTR+F+I+R5 model with 1000 bootstrap replicates. Scale bar represents 0.3 substitutions per site. C shows midpoint rooted phylogenetic tree of entire hantavirus S segment (nucleocapsid gene), bases 1-1302 relative to TATV *orthohantavirus* strain Upton-Heath reference sequence (NC_055635). Tree generated using GTR+F+I+R5 model with 1000 bootstrap replicates. Scale bar represents 0.4 substitutions per site.
For all trees, node values represent % bootstrap support, red branches and text highlights sample libraries, and host species are shown next to each branch. S.c= *Sorex caecutiens*, O.p= *Oryzomys palustris*, S.a= *Sigmodon alstoni*, O.l= *Oligoryzomys longicaudatus*, O.f= *Oligoryzomys fulvescens*, P.m= *Peromyscus maniculatus*, P.b= *Peromyscus beatae*, E.e= *Eothenomys Eleusis*, M.p= *Microtus pennsylvanicus*, M.ar= *Microtus arvalis*, M.g= *Myodes glareolus*, M.m= *Microtus maximowiczii,* M.a= *Microtus agrestis*, C.s= *Crocidura shantungensis*, C.d= *Crocidura douceti*, A.sa= *Anourosorex squamipes*, S.r= *Sorex roboratus,* A.a= *Apodemus agrarius*, A.f= *Apodemus flavicollis*, H.s= *Hylomyscus simus*, R.n= *Rattus norvegicus*, E.m= *Eliurus majori*, S.a= *Stenocephalemys albipes*, Z.b= *Zygodontomys brevicauda*, C.l= *Calomys laucha*, S.h= *Sigmodon hispidus*, N.c= *Niviventer confucianus*, S.m= *Sorex minutus*, S.i= *Sorex isodon*, S.ar= *Sorex Araneus*.

e. *Hepacivirus* hits were PCR confirmed

Hepacivirus F reads were identified in 8 libraries and were PCR confirmed in 8 animals (5.7% of total animals)- 3 wood mice, 3 bank voles, and 2 field voles. The library I sequence provided only approximately 5% genomic information and could not be confirmed by PCR. The library L sequence provided approximately 70% genomic information, but could not be confirmed by PCR in either animal and upon NTBLAST analysis the library L reads were identified as probable host genomic reads. The libraries B, E, N and P sequences provided approximately 3, 10, 15 and 5% genomic coverage respectively, and were all PCR confirmed in at least 1 animal. The library A sequence provided approximately 30% genomic coverage and was PCR confirmed in 1 bank vole but upon NTBLAST analysis was found to be most closely related to host chromosomal reads. Library M provided approximately 95% genomic coverage and was most similar to *Hepacivirus myodae* isolate MgHV5 (MN242372). *Hepacivirus myodae* is an alternative name for *Hepacivirus F*[325].

Phylogenetic analysis of the library M sequence was performed using all *Hepacivirus F* sequences available on Genbank, assessing the *Hepacivirus* RdRp (bases 7344-9098 relative to the *Hepacivirus F* reference genome NC_038427) and E2 genes (bases 1494-2372 relative to the same reference sequence), shown in Figures 12 A and 12 B respectively. Upon phylogenetic analysis of the RdRp gene the library M sequence formed an outgroup, and upon analysis of the E2 genes the library M sequence clustered amongst the *Hepacivirus F* sequences and formed a clade with another *Hepacivirus F* virus (OV121128).

37 rodent *Hepacivirus* reads spanning approximately 15% of the genome were also identified in library N and were PCR confirmed in 1 field vole (0.7% of total animals) (Table 10). There was insufficient data to either accurately confirm this hit by NTBLAST or by phylogenetic analysis, therefore whilst it is reasonable to suggest some form of *Hepacivirus* was present it is difficult to provide any further information or to distinguish it from the library N *Hepacivirus F* virus with confidence.


Library_M_hepacivirus_consensus_extraction
M.g
A
OV121129.1__Hepacivirus_F_isolate_NA_genome_assembly__complete_genome__monopartite_extraction
M.g
MN242370.1__Hepacivirus_F_isolate_MgHV3__complete_genome_extraction
M.g
99.6
MN242371.1__Hepacivirus_F_isolate_MgHV4__complete_genome_extraction
M.g
MN242372.1__Hepacivirus_F_isolate_MgHV5__complete_genome_extraction
M.g
99.3
MN242369.1__Hepacivirus_F_isolate_MgHV2__complete_genome_extraction
M.g
NC_038427.1__Hepacivirus_F__complete_genome_extraction
M.g
OV121128.1__Hepacivirus_F_isolate_NA_genome_assembly__complete_genome__monopartite_extraction
M.g
0.02

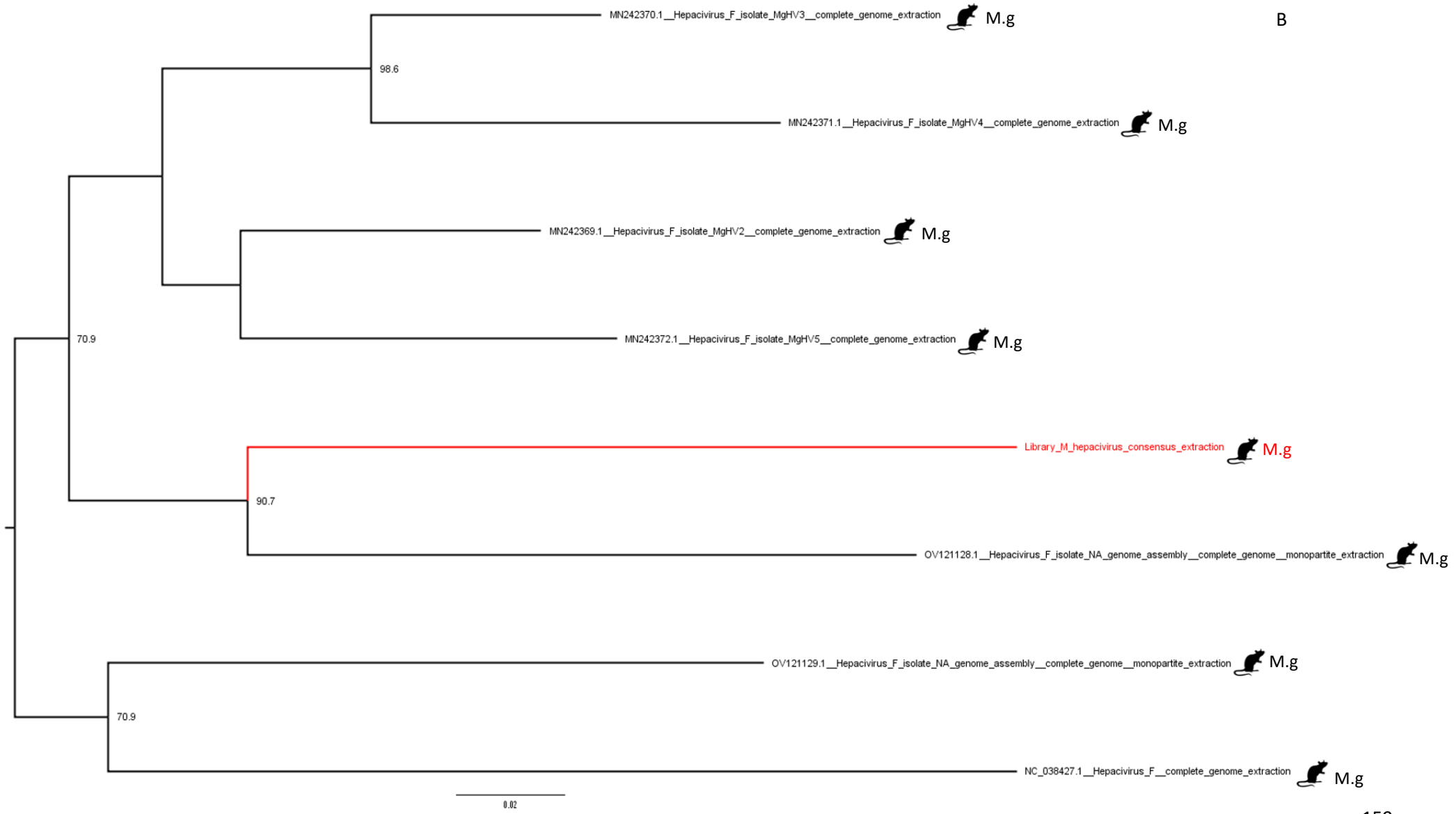

B
MN242370.1__Hepacivirus_F_isolate_MgHV3__complete_genome_extraction
M.g
98.6
MN242371.1__Hepacivirus_F_isolate_MgHV4__complete_genome_extraction
M.g
MN242369.1__Hepacivirus_F_isolate_MgHV2__complete_genome_extraction
M.g
70.9
MN242372.1__Hepacivirus_F_isolate_MgHV5__complete_genome_extraction
M.g
Library_M_hepacivirus_consensus_extraction
M.g
90.7
OV121128.1__Hepacivirus_F_isolate_NA_genome_assembly__complete_genome__monopartite_extraction
M.g
OV121129.1__Hepacivirus_F_isolate_NA_genome_assembly__complete_genome__monopartite_extraction
M.g
70.9
NC_038427.1__Hepacivirus_F__complete_genome_extraction
M.g
0.02

Figure 12- Phylogenetic analysis of NGS identified *Hepacivirus F* sequences.
A shows midpoint rooted phylogenetic tree of entire *Hepacivirus F* NS5B (RdRP) gene, bases 7344-9098 relative to *Hepacivirus F* reference genome (NC_038427). Tree generated using TPM2+I+G4 model with 1000 bootstrap replicates. Scale bar represents 0.02 substitutions per site. B shows midpoint rooted phylogenetic tree of entire *Hepacivirus F* E2 gene, bases 1494-2372 relative to *Hepacivirus F* reference genome (NC_038427). Scale bar represents 0.02 substitutions per site. Tree generated using TPM2u+F+G4 model with 1000 bootstrap replicates. For both trees, node values represent % bootstrap support, red branches and text highlights sample libraries, and host species is shown next to each branch. M.g= *Myodes glareolus*.

f. MLV was the most common virus identified

MLV reads were identified in a total of 13 libraries, and was PCR confirmed in at least 1 animal in 13/13 libraries (Table 10). A total of 94 animals (67.1% of total animals) were MLV positive by PCR including 69 wood mice, 16 yellow-necked mice, 6 bank voles, 3 field voles, and the least weasel. All sequences except for those of libraries H and R provided < 30% genomic coverage so did not undergo NTBLAST analysis. The library H sequence provided approximately 35% genomic coverage and was identified as probable host chromosomal sequence upon NTBLAST, and the library R sequence provided approximately 40% genomic sequence and was also identified as probable host chromosomal sequence by NTBLAST. Investigations to distinguish between MLV EVEs and exogenous MLV viruses were not performed due to time constraints.

g. 1 section of *Orbivirus* was found

Orbivirus reads were found in library E, where the entire segment 10 was recovered and no reads were recovered for any other segments. This was PCR confirmed in 6 wood-mice (4.3% of all animals tested) (Table 10). No match was provided by NTBLAST, but by BLASTX this sequence was identified as most similar to *Kemorovo virus* NS1 protein (AYM94264) with 31.87% similarity. Phylogenetic analysis of the full segment 10 was then performed (Figure 13), where the library E *Orbivirus* was found to form an outgroup with significant branch length and therefore divergence relative to other *Orbiviruses*. Due to time constraints, no further genomic segments were identified.

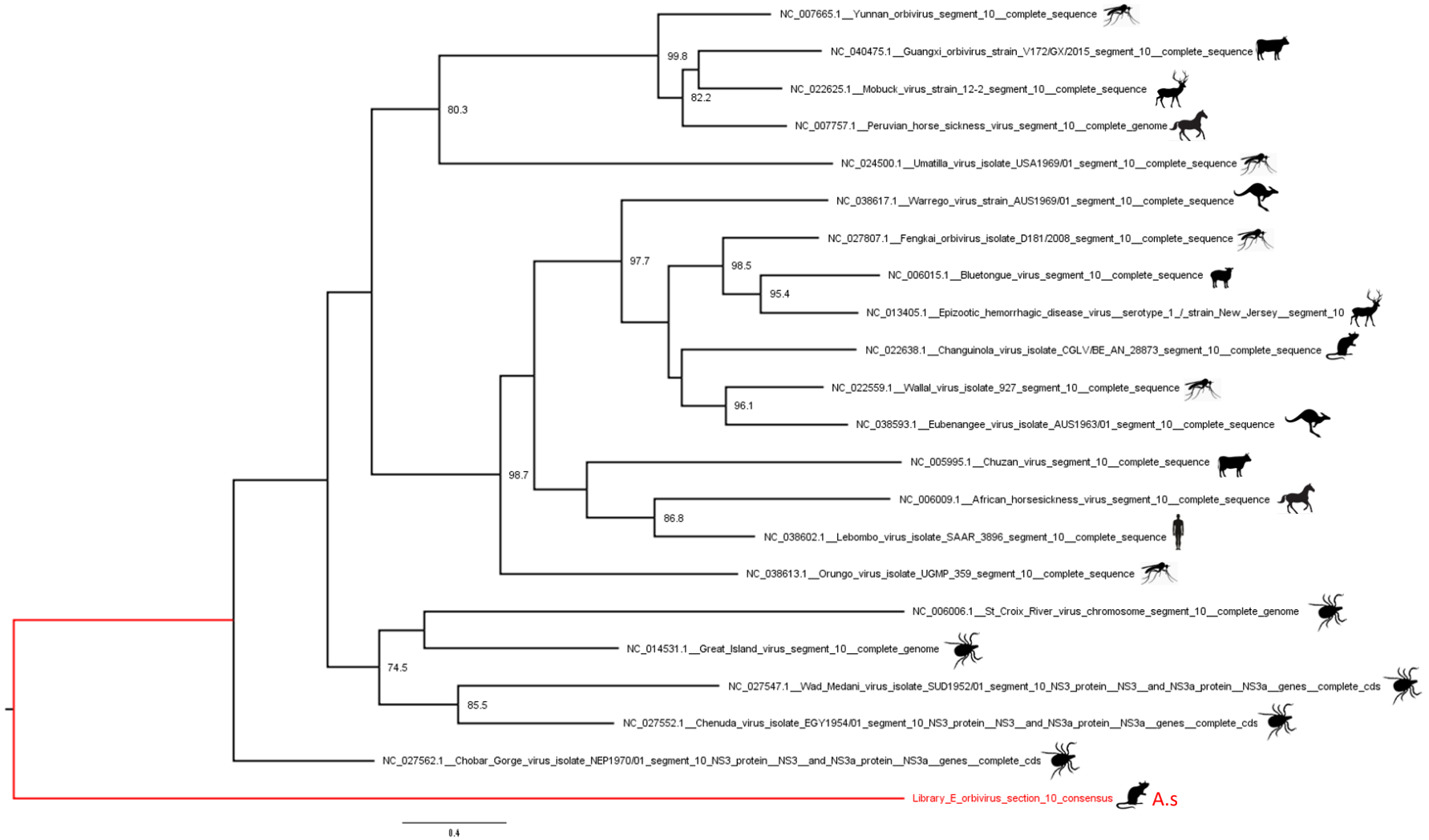

NC_007665.1__Yunnan_orbivirus_segment_10__complete_sequence
NC_040475.1__Guangxi_orbivirus_strain_V172/GX/2015_segment_10__complete_sequence
NC_022625.1__Mobuck_virus_strain_12-2_segment_10__complete_sequence
NC_007757.1__Peruvian_horse_sickness_virus_segment_10__complete_genome
NC_024500.1__Umatilla_virus_isolate_USA1969/01_segment_10__complete_sequence
NC_038617.1__Warrego_virus_strain_AUS1969/01_segment_10__complete_sequence
NC_027807.1__Fengkai_orbivirus_isolate_D181/2008_segment_10__complete_sequence
NC_006015.1__Bluetongue_virus_segment_10__complete_sequence
NC_013405.1__Epizootic_hemorrhagic_disease_virus__serotype_1_/_strain_New_Jersey__segment_10
NC_022638.1__Changuinola_virus_isolate_CGLV/BE_AN_28873_segment_10__complete_sequence
NC_022559.1__Wallal_virus_isolate_927_segment_10__complete_sequence
NC_038593.1__Eubenangee_virus_isolate_AUS1963/01_segment_10__complete_sequence
NC_005995.1__Chuzan_virus_segment_10__complete_sequence
NC_006009.1__African_horsesickness_virus_segment_10__complete_sequence
NC_038602.1__Lebombo_virus_isolate_SAAR_3896_segment_10__complete_sequence
NC_038613.1__Orungo_virus_isolate_UGMP_359_segment_10__complete_sequence
NC_006006.1__St_Croix_River_virus_chromosome_segment_10__complete_genome
NC_014531.1__Great_Island_virus_segment_10__complete_genome
NC_027547.1__Wad_Medani_virus_isolate_SUD1952/01_segment_10_NS3_protein__NS3__and_NS3a_protein__NS3a__genes__complete_cds
NC_027552.1__Chenuda_virus_isolate_EGY1954/01_segment_10_NS3_protein__NS3__and_NS3a_protein__NS3a__genes__complete_cds
NC_027562.1__Chobar_Gorge_virus_isolate_NEP1970/01_segment_10_NS3_protein__NS3__and_NS3a_protein__NS3a__genes__complete_cds
Library_E_orbivirus_section_10_consensus
A.s
99.8
82.2
80.3
97.7
98.5
95.4
96.1
98.7
86.8
74.5
85.5
0.4

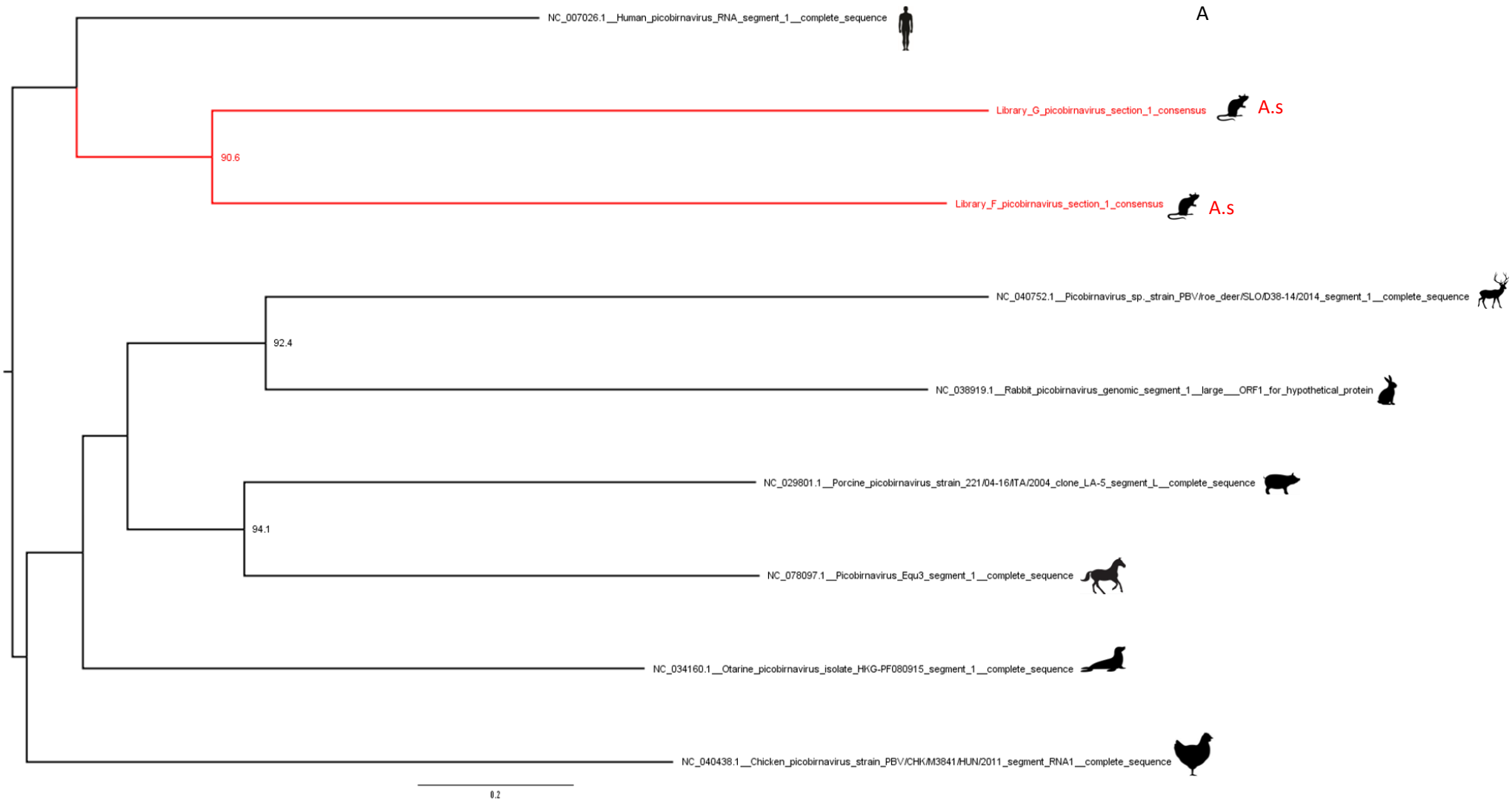
Figure 13- Phylogenetic analysis of NGS identified *Orbivirus* sequence.
Midpoint rooted phylogenetic tree of full *Orbivirus* segment 10, relative to bases 1-809 of *changuinola virus* segment 10 reference genome (NC_022638). Tree generated using TPM3u+F+I+R3 model. Node values represent % bootstrap support. Red branch and text highlights sample library. Scale bar represents 0.4 substitutions per site. A.s= *Apodemus sylvaticus*.

h. *Picobirnaviruses* found in multiple libraries

Picobirnavirus reads were identified in 12 libraries and were PCR confirmed in 5 wood mice, 1 yellow-necked mouse and 1 field vole (5% of total animals) across 7 libraries (E, F, G, H, L, O, and R) (Table 10). For the library E virus approximately 60% of the total genome was identified, and whilst insufficient segment 1 sequence was recovered for NTBLAST analysis, segment 2 was found to be most similar to *Picobirnavirus* sp. isolate YS23 (MW930262) with approximately 85.12% similarity. For the library F virus approximately 95% of the total genome was recovered and by NTBLAST analysis segment 1 was most similar to *Picobirnavirus* spp. isolate YS27 RdRp gene (MW930266) with 89.37% similarity. The library F segment 2 was identified as most similar to Rabbit *Picobirnavirus* isolate rab049pbv01 RdRp gene (MT150089) by NTBLAST with 82.51% similarity. The library G PBV segment 1 was not identified by NTBLAST despite approximately 95% genome coverage, yet was identified as most similar to a marmot *Picobirnavirus* putative capsid (AVX53463) by BLASTX with approximately 44.31% similarity. The library G PBV segment 2 was identified as most similar to a *Picobirnavirus* species RdRp (QXV86671) with approximately 73.52% similarity by NTBLAST. Approximately 55% of the segment 1 genome and 80% of the segment 2 genome for the library O *Picobirnavirus* was provided by the NGS data. Segment 1 provided no results following NTBLAST analysis but by BLASTX analysis was found to best match a marmot *Picobirnavirus* capsid gene (AVX53810) with approximately 47.32% similarity, whilst segment 2 was identified as most similar to the *Picobirnavirus* dog/ KNA/ 2015 strain PBV/ Dog/ KNA/ RVC7/ 2015 RdRp gene isolate (KY399057) with approximately 89.51% similarity. The library H virus only provided approximately 20% genomic coverage for each segment and could not be analysed by NTBLAST. Library I *Picobirnavirus* sequences could not be confirmed by PCR, but approximately 60% of the total genome and nearly all of segment 2 was recovered which was identified as best matching *Picobirnavirus* sp. isolate R57-k141_224336 (MX556509) by NTBLAST. Approximately 60% of the library L *Picobirnavirus* genome was recovered, and whilst not enough of segment 1 was recovered to undergo NTBLAST analysis segment 2 was identified as most similar to Porcine *Picobirnavirus* isolate 15213_ NODE_ 440_ len_ 1675_ cov_ 49.185185 RdRp (MW977276) or host chromosomal sequence with identical similarity of 80.12%. By BLASTX analysis this segment was identified as most similar to a chicken *Picobirinavirus* RdRp (AXL64612.1) with 72.97% similarity. Finally, library R provided approximately 20% genomic coverage, but neither segment provided sufficient genome for accurate NTBLAST analysis.

Phylogenetic analysis of segment 1 was performed on the library F and G viruses, as these were the only libraries where sufficient genomic information available (Figure 14 A). The entire S1 segment was analysed, and libraries F and G formed their own clade with strong bootstrap support amongst other *Picobirnavirus*es. Phylogenetic analysis of the entire segment 2 was performed on the library E, F, G, I, L and O viruses, as these all provided sufficient genomic coverage (Figure 14 B). The library E and O sequences formed a highly supported clade with a chicken PBV isolate

(NC_040439). The library L, I and G sequences formed their own clade, most closely clustering with a dog PBV isolate (NC_030526). The library F sequence formed a clade of its own, but clustered amongst other PBV sequences.


A
NC_007026.1__Human_picobirnavirus_RNA_segment_1__complete_sequence
Library_G_picobirnavirus_section_1_consensus
A.s
90.6
Library_F_picobirnavirus_section_1_consensus
A.s
NC_040752.1__Picobirnavirus_sp._strain_PBV/roe_deer/SLO/D38-14/2014_segment_1__complete_sequence
92.4
NC_038919.1__Rabbit_picobirnavirus_genomic_segment_1__large___ORF1_for_hypothetical_protein
NC_029801.1__Porcine_picobirnavirus_strain_221/04-16/ITA/2004_clone_LA-5_segment_L__complete_sequence
94.1
NC_078097.1__Picobirnavirus_Equ3_segment_1__complete_sequence
NC_034160.1__Otarine_picobirnavirus_isolate_HKG-PF080915_segment_1__complete_sequence
NC_040438.1__Chicken_picobirnavirus_strain_PBV/CHK/M3841/HUN/2011_segment_RNA1__complete_sequence
0.2

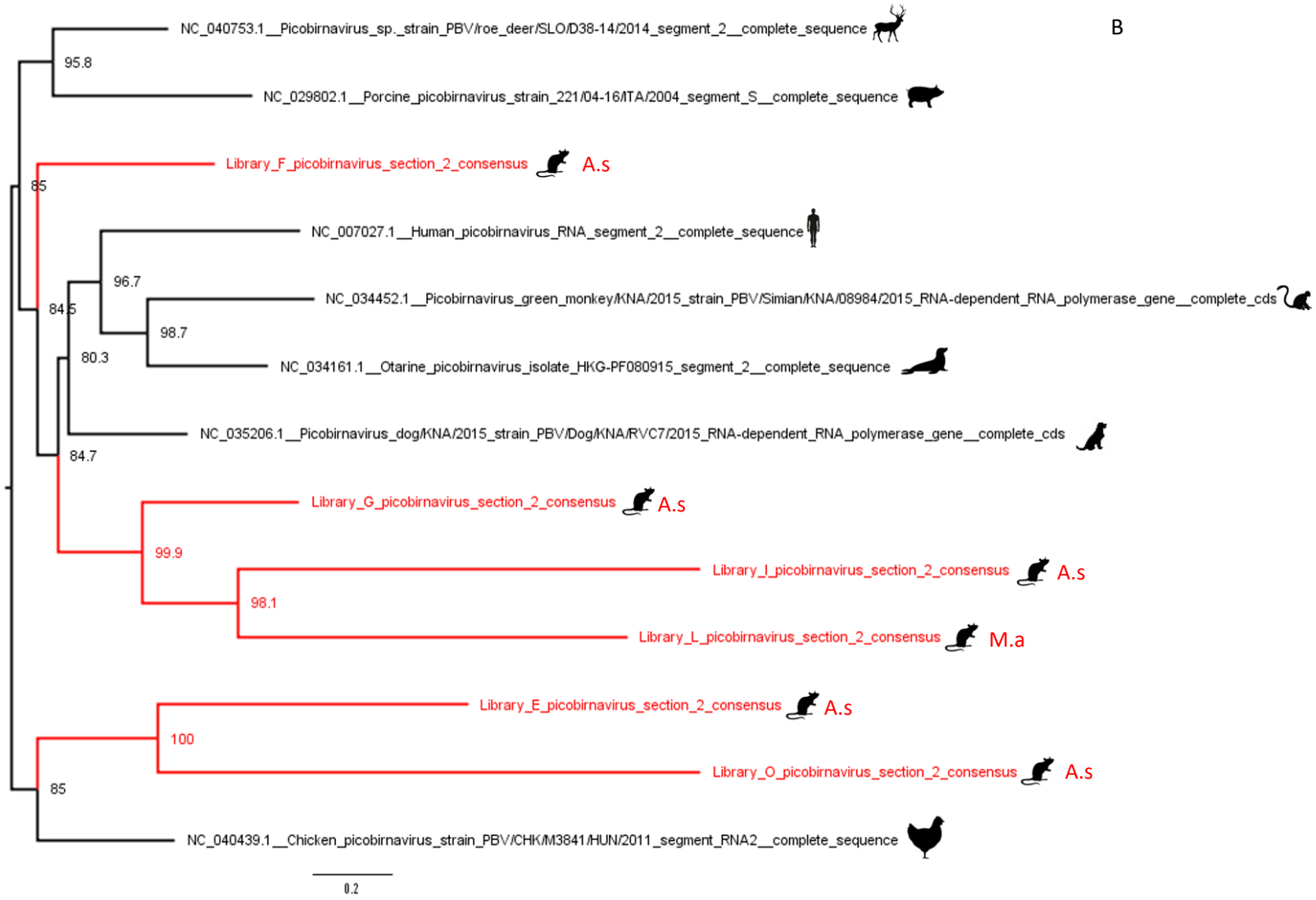
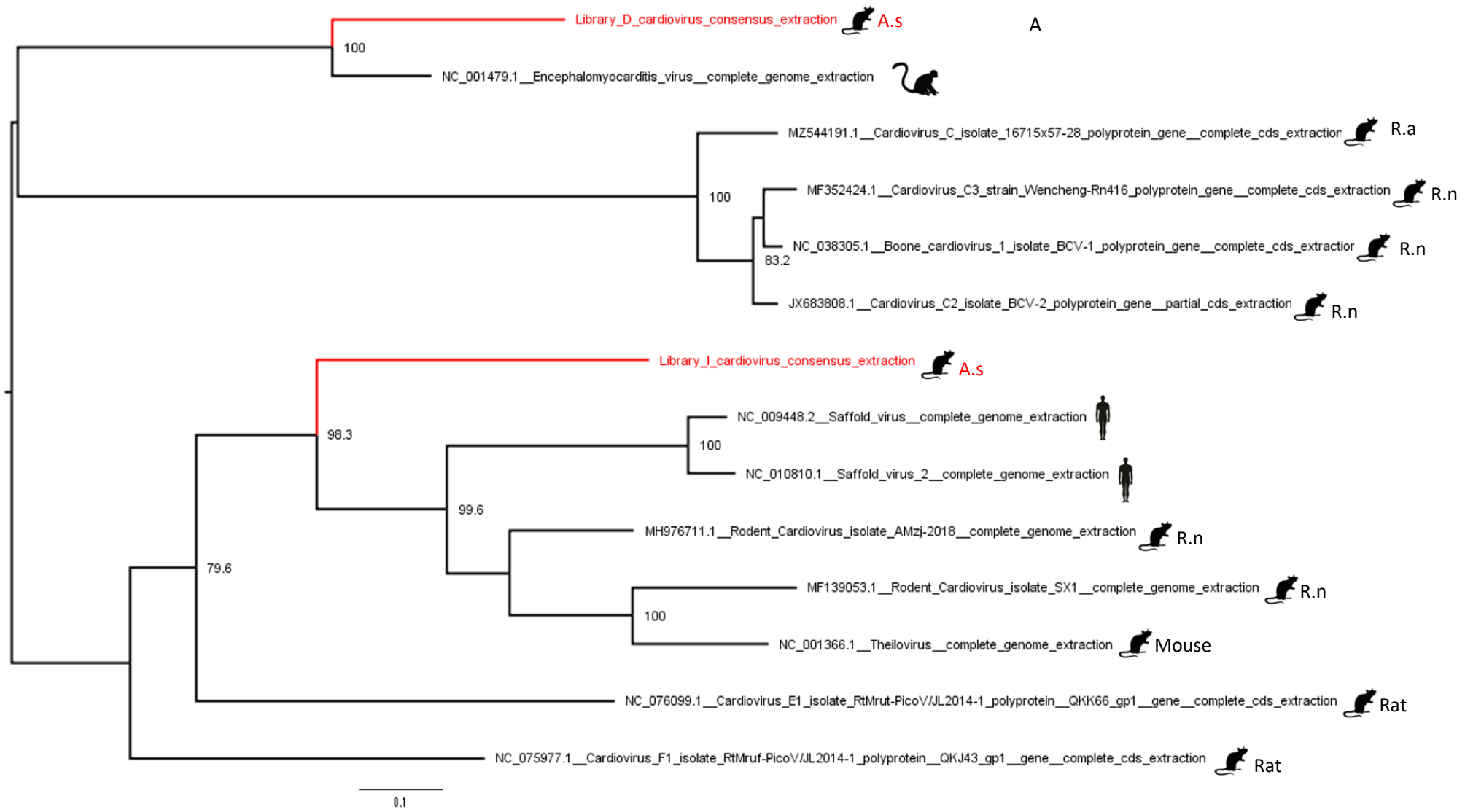

B
NC_040753.1__Picobirnavirus_sp._strain_PBV/roe_deer/SLO/D38-14/2014_segment_2__complete_sequence
95.8
NC_029802.1__Porcine_picobirnavirus_strain_221/04-16/ITA/2004_segment_S__complete_sequence
Library_F_picobirnavirus_section_2_consensus
A.s
85
NC_007027.1__Human_picobirnavirus_RNA_segment_2__complete_sequence
96.7
NC_034452.1__Picobirnavirus_green_monkey/KNA/2015_strain_PBV/Simian/KNA/08984/2015_RNA-dependent_RNA_polymerase_gene__complete_cds
84.5
98.7
80.3
NC_034161.1__Otarine_picobirnavirus_isolate_HKG-PF080915_segment_2__complete_sequence
NC_035206.1__Picobirnavirus_dog/KNA/2015_strain_PBV/Dog/KNA/RVC7/2015_RNA-dependent_RNA_polymerase_gene__complete_cds
84.7
Library_G_picobirnavirus_section_2_consensus
A.s
99.9
Library_I_picobirnavirus_section_2_consensus
A.s
98.1
Library_L_picobirnavirus_section_2_consensus
M.a
Library_E_picobirnavirus_section_2_consensus
A.s
100
Library_O_picobirnavirus_section_2_consensus
A.s
85
NC_040439.1__Chicken_picobirnavirus_strain_PBV/CHK/M3841/HUN/2011_segment_RNA2__complete_sequence
0.2

Figure 14- Phylogenetic analysis of NGS identified *Picobirnavirus* sequences.
A shows midpoint rooted phylogenetic tree of *Picobirnavirus* segment 1, relative to full S1 segment (bases 1-2666) of the porcine *Picobirnavirus* strain reference sequence (NC029801). Generated using TPM3u+F+R2 model with 1000 bootstrap replicates. Scale bar represents 0.2 substitutions per site. B shows midpoint rooted phylogenetic tree of PiBiS2, relative to full S2 segment (bases 1-1730) of the porcine *Picobirnavirus* strain reference sequence (NC029802). Generated using TVM+F+G4 model with 1000 bootstrap replicates. Scale bar represents 0.2 substitutions per site. For both trees, node values represent % bootstrap support, red branches and text highlight sample libraries and host species is shown next to each branch. A.s= *Apodemus sylvaticus*, M.a= *Microtus agrestis*.

i. Picornaviruses of three genera were identified

Cardiovirus reads were identified in a total of 13 libraries and were PCR confirmed in 7/13 (Table 10). A total of 13 animals (9.3% of total animals) were *Cardovirus* positive by PCR, including 7 bank voles, 4 wood mice and 2 field voles. The sequences from libraries J, N, O, P, Q and R all provided low read numbers and low genomic coverage and failed to be confirmed by PCR. The library A, B and E sequences all also provided limited genomic coverage but were PCR confirmed in at least 1 animal per library. Library D and M viruses each provided approximately 60% genomic coverage, and their sequencess were found to be most similar to EMCV type 2 isolate RD 1338 (D28/05) (JX257003) with 91.38% similarity and *Cardiovirus* F1 isolate RtMruf-PicoV/ JL2014-1 (NC_075977) with 84.8% similarity by NTBLAST, respectively. The library G sequence yielded approximately 90% of a *Cardiovirus* genome and was identified as most similar to EMCV isolate UK2016 (OP381184) with 98% similarity by NTBLAST, and the library I sequence yielded approximately 95% of a *Cardiovirus* genome and was best matched to a *Cardiovirus* species (ON136175) by NTBLAST with 82.22% similarity.

Phylogenetic analysis was performed on libraries D, G, I and M. Figure 15 A shows a phylogenetic tree of the highly conserved RdRp gene relative to bases 6599-7976 of the rodent *Cardiovirus* isolate SX1 (MF139053), where the library I sequence forms a clade of its own amongst other *Cardioviruses* and the library D sequence forms a clade with an EMCV isolate (NC_001479). For this tree the library G and M sequences were omitted as they formed outgroups that significantly reduced the clarity of the rest of the tree. Figure 15 B shows a phylogenetic tree of the less conserved capsid gene (bases 1579-2222 relative to the same reference sequence), where the library G and I sequences formed a clade which then clustered with another clade consisting of the library D sequence and the same EMCV isolate, whilst the library M sequence formed another clade with a different *Cardiovirus* isolate (NC_075977). In this tree all library sequences clustered amongst rodent *Cardioviruses*.


A
Library_D_cardiovirus_consensus_extraction
A.s
100
NC_001479.1__Encephalomyocarditis_virus__complete_genome_extraction
MZ544191.1__Cardiovirus_C_isolate_16715x57-28_polyprotein_gene__complete_cds_extraction
R.a
100
MF352424.1__Cardiovirus_C3_strain_Wencheng-Rn416_polyprotein_gene__complete_cds_extraction
R.n
NC_038305.1__Boone_cardiovirus_1_isolate_BCV-1_polyprotein_gene__complete_cds_extraction
R.n
83.2
JX683808.1__Cardiovirus_C2_isolate_BCV-2_polyprotein_gene__partial_cds_extraction
R.n
Library_I_cardiovirus_consensus_extraction
A.s
NC_009448.2__Saffold_virus__complete_genome_extraction
98.3
100
NC_010810.1__Saffold_virus_2__complete_genome_extraction
99.6
MH976711.1__Rodent_Cardiovirus_isolate_AMzj-2018__complete_genome_extraction
R.n
79.6
MF139053.1__Rodent_Cardiovirus_isolate_SX1__complete_genome_extraction
R.n
100
NC_001366.1__Theilovirus__complete_genome_extraction
Mouse
NC_076099.1__Cardiovirus_E1_isolate_RtMrut-PicoV/JL2014-1_polyprotein__QKK66_gp1__gene__complete_cds_extraction
Rat
NC_075977.1__Cardiovirus_F1_isolate_RtMruf-PicoV/JL2014-1_polyprotein__QKJ43_gp1__gene__complete_cds_extraction
Rat
0.1

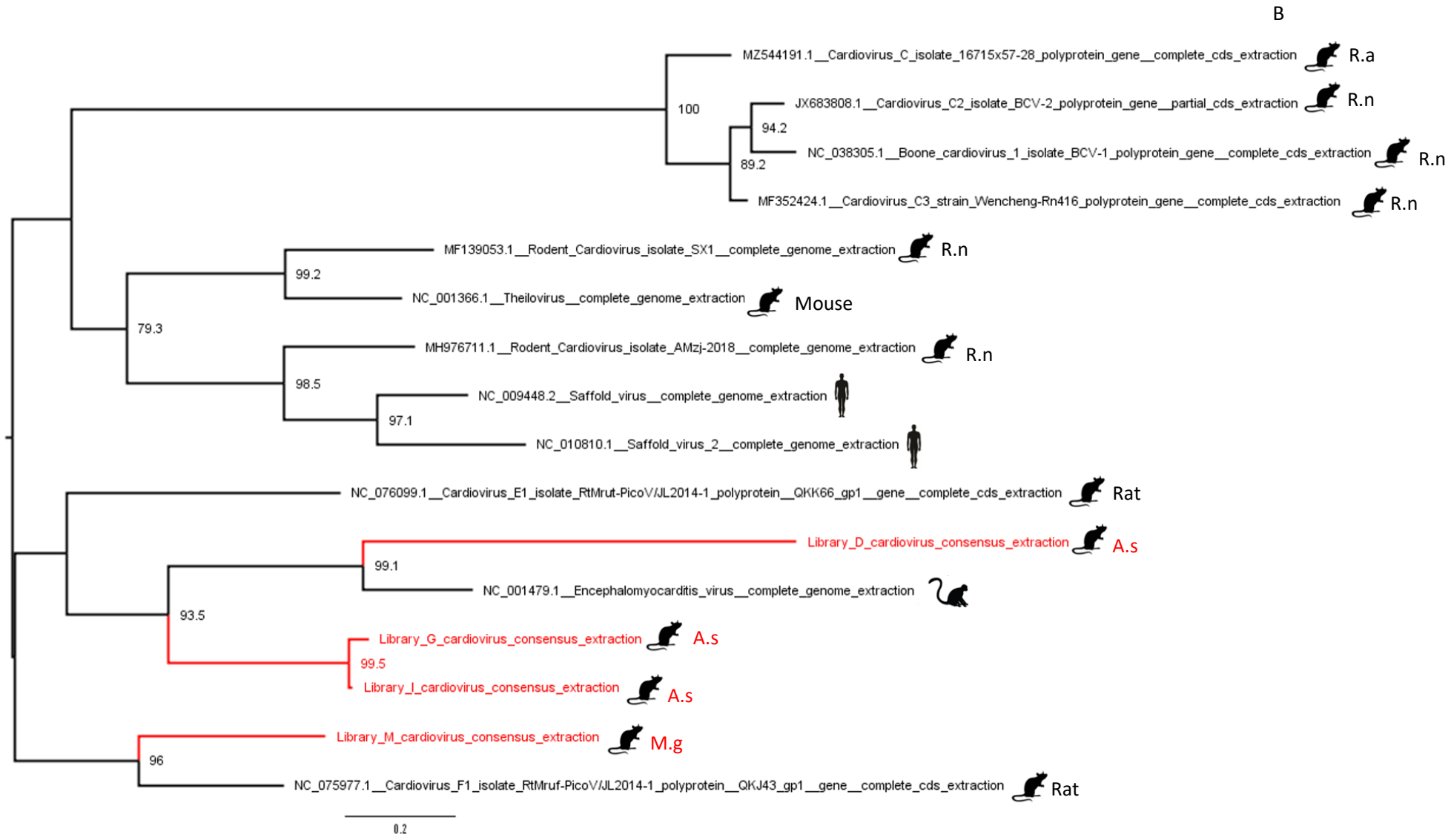
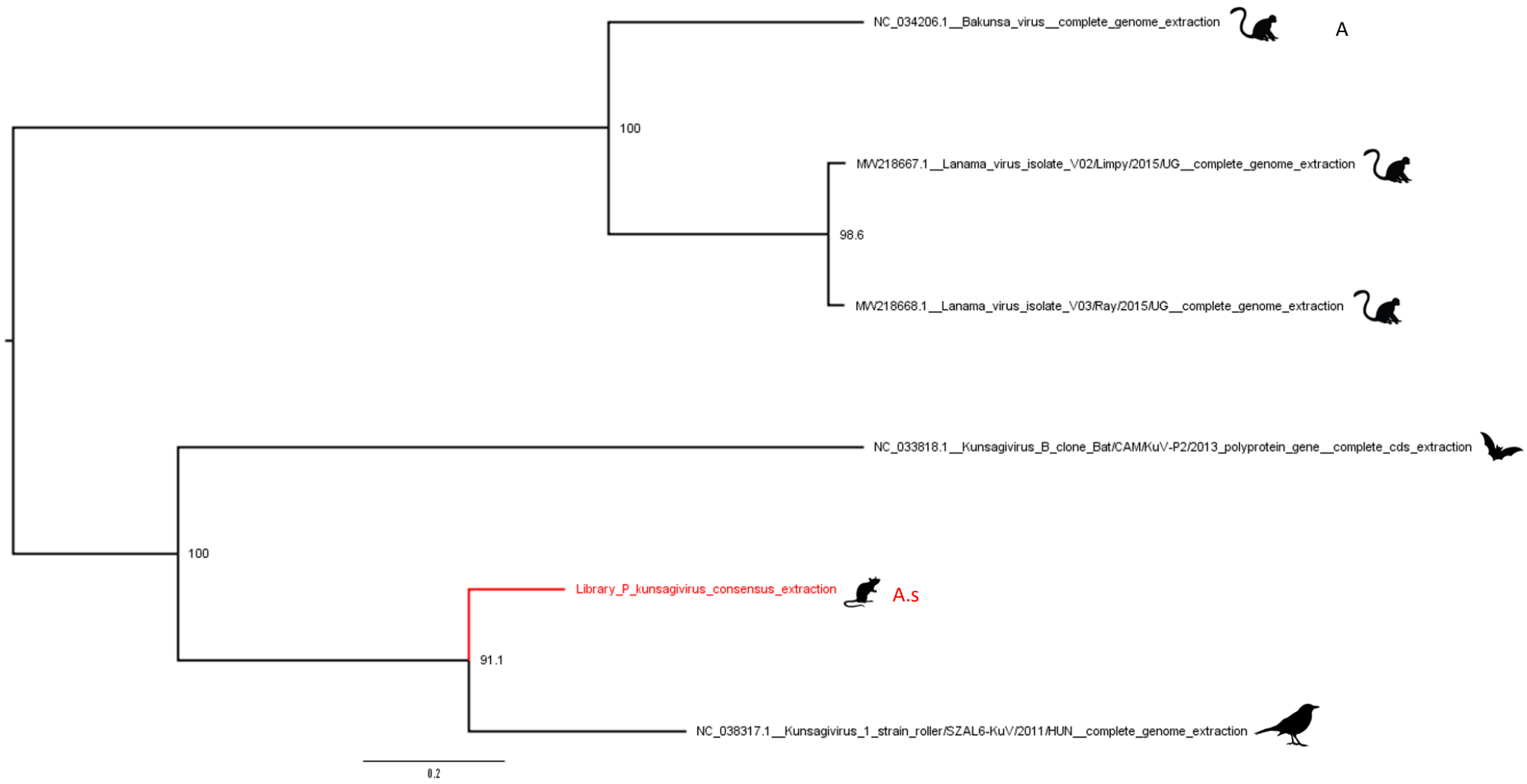

B
MZ544191.1__Cardiovirus_C_isolate_16715x57-28_polyprotein_gene__complete_cds_extraction
R.a
JX683808.1__Cardiovirus_C2_isolate_BCV-2_polyprotein_gene__partial_cds_extraction
R.n
NC_038305.1__Boone_cardiovirus_1_isolate_BCV-1_polyprotein_gene__complete_cds_extraction
R.n
MF352424.1__Cardiovirus_C3_strain_Wencheng-Rn416_polyprotein_gene__complete_cds_extraction
R.n
MF139053.1__Rodent_Cardiovirus_isolate_SX1__complete_genome_extraction
R.n
NC_001366.1__Theilovirus__complete_genome_extraction
Mouse
MH976711.1__Rodent_Cardiovirus_isolate_AMzj-2018__complete_genome_extraction
R.n
NC_009448.2__Saffold_virus__complete_genome_extraction
NC_010810.1__Saffold_virus_2__complete_genome_extraction
NC_076099.1__Cardiovirus_E1_isolate_RtMrut-PicoV/JL2014-1_polyprotein__QKK66_gp1__gene__complete_cds_extraction
Rat
Library_D_cardiovirus_consensus_extraction
A.s
NC_001479.1__Encephalomyocarditis_virus__complete_genome_extraction
Library_G_cardiovirus_consensus_extraction
A.s
Library_I_cardiovirus_consensus_extraction
A.s
Library_M_cardiovirus_consensus_extraction
M.g
NC_075977.1__Cardiovirus_F1_isolate_RtMruf-PicoV/JL2014-1_polyprotein__QKJ43_gp1__gene__complete_cds_extraction
Rat
100
94.2
89.2
99.2
79.3
98.5
97.1
99.1
93.5
99.5
96
0.2

Figure 15- Phylogenetic analysis of NGS identified *Cardiovirus* sequences.
A shows midpoint rooted phylogenetic tree of *Cardiovirus* RdRp-like protein gene (bases 6599-7976) relative to rodent Cardiovirus isolate SX1 (MF139053). Generated using GTR+F+I+G4 model with 1000 bootstrap replicates. Scale bar represents 0.1 substitutions per site. B shows midpoint rooted phylogenetic tree of Cardiovirus capsid gene (bases 1579-2222) relative to rodent *Cardiovirus* isolate SX1 (MF139053). Generated using TIM2+F+I+G4 with 1000 bootstrap replicates. Scale bar represents 0.2 substitutions per site. For both trees, node values represent % bootstrap support, red branches and text highlight sample libraries and host species are shown at branch ends. A.s= *Apodemus sylvaticus*, R.a= *Rattus Argentiventer*, R.n= *Rattus norvegicus*, M.g= *Myodes glareolus*. "Mouse" represents a mouse host where the specific species was not recorded, "rat" represents a mouse host where the specific species was not recorded.

Kunsagivirus reads were found in library P and *Kunsagivirus* was PCR confirmed in 2 wood mice (1.4% of total animals, Table 10). Approximately 75% of the *Kunsagivirus* genome was recovered and when analysed by NTBLAST was identified as most similar to a *Kunsagivirus* species isolate (ON136180) with 91.24% similarity. Phylogenetic analysis of this isolate was then performed using the highly conserved *Kunsagivirus* RdRp gene (bases 6117-7155 relative to the *Kunsagivirus 1* strain roller/SZAL6-KuV/2011/HUN reference sequence NC_038317) and the less conserved *Kunsagivirus* capsid-like gene (bases 1695-2042 also relative to NC_038317), shown in Figures 16 A and 16 B respectively. All *Kunsagivirus* genomes available on Genbank were used for this analysis. Within the RdRp gene the library P sequence formed a clade with the *Kunsagivirus* 1 strain (NC_038317), whereas within the capsid-like gene this virus formed a clade with a rodent *Kunsagivirus* strain (ON136180).

Rosavirus reads were identified in libraries B and M and were PCR confirmed in 2 bank voles (1.4% of total animals), both of which were in library B (Table 10). The library M sequence only provided 1 *Rosavirus* read covering approximately 2% of the genome and could not be confirmed by PCR. The library B sequence provided approximately 50% *Rosavirus* genome coverage and upon NTBLAST analysis was found to most closely match a *Rosavirus* M-7 (NC_038880) isolate with 84.06% similarity. Insufficient genomic information was available to perform phylogenetic analysis on this isolate.


NC_034206.1__Bakunsa_virus__complete_genome_extraction
A
100
MW218667.1__Lanama_virus_isolate_V02/Limpy/2015/UG__complete_genome_extraction
98.6
MW218668.1__Lanama_virus_isolate_V03/Ray/2015/UG__complete_genome_extraction
NC_033818.1__Kunsagivirus_B_clone_Bat/CAM/KuV-P2/2013_polyprotein_gene__complete_cds_extraction
100
Library_P_kunsagivirus_consensus_extraction
A.s
91.1
NC_038317.1__Kunsagivirus_1_strain_roller/SZAL6-KuV/2011/HUN__complete_genome_extraction
0.2

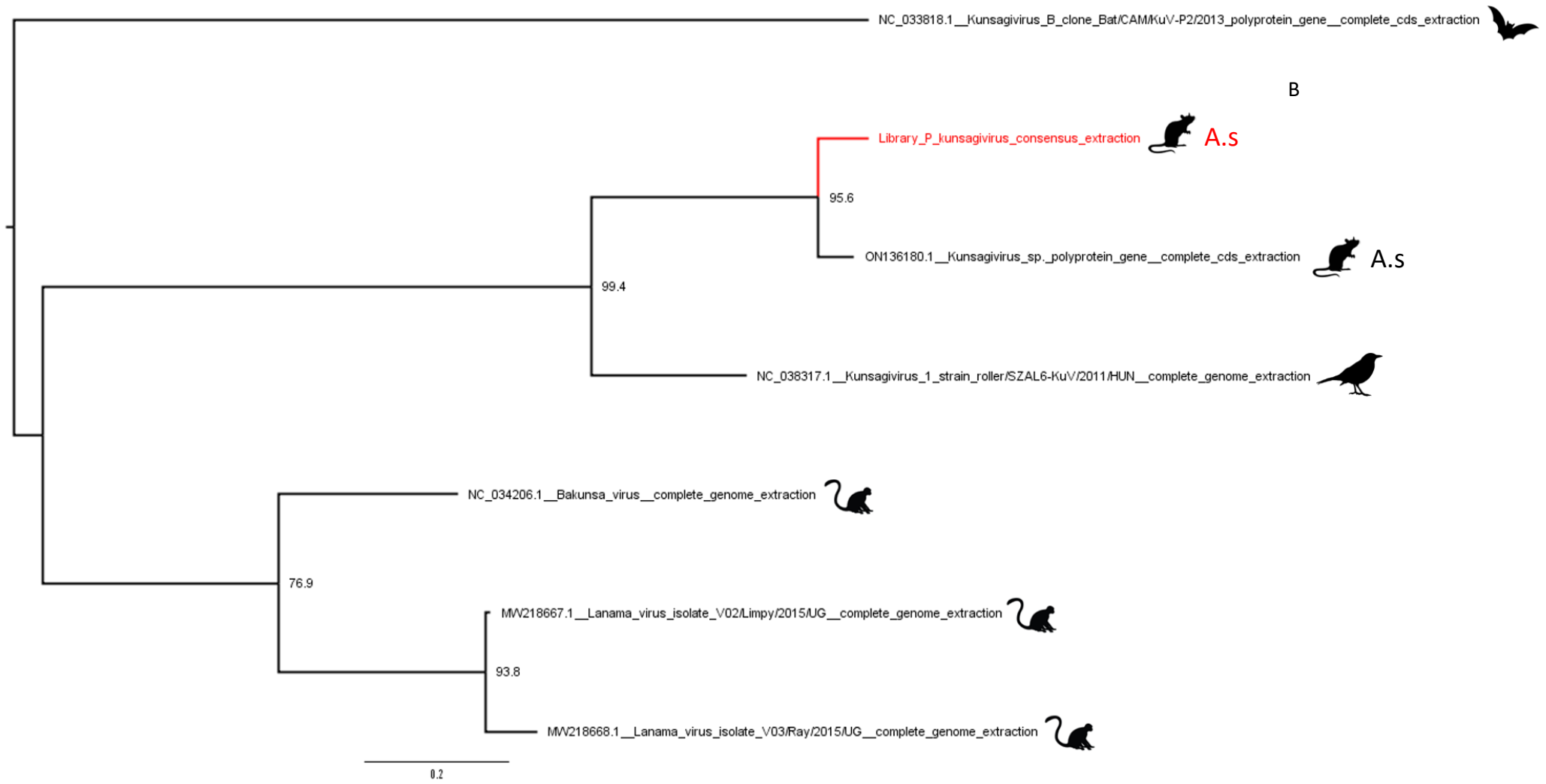

NC_033818.1__Kunsagivirus_B_clone_Bat/CAM/KuV-P2/2013_polyprotein_gene__complete_cds_extraction
B
Library_P_kunsagivirus_consensus_extraction
A.s
95.6
ON136180.1__Kunsagivirus_sp._polyprotein_gene__complete_cds_extraction
A.s
99.4
NC_038317.1__Kunsagivirus_1_strain_roller/SZAL6-KuV/2011/HUN__complete_genome_extraction
NC_034206.1__Bakunsa_virus__complete_genome_extraction
76.9
MW218667.1__Lanama_virus_isolate_V02/Limpy/2015/UG__complete_genome_extraction
93.8
MW218668.1__Lanama_virus_isolate_V03/Ray/2015/UG__complete_genome_extraction
0.2

Figure 16- Phylogenetic analysis of a NGS identified *Kunsagivirus* sequence.
A shows midpoint rooted phylogenetic tree of *Kunsagivirus* RdRp gene (bases 6117-7155) relative to the *Kunsagivirus 1* strain roller/SZAL6-KuV/2011/HUN reference sequence NC_038317. Generated using TIM3+F+G4 model with 1000 bootstrap replicates. Scale bar represents 0.2 substitutions per site. In both trees node values represent % bootstrap support, red branches and text highlights sample libraries and host species are shown at branch ends. B shows midpoint rooted phylogenetic tree of *Kunsagivirus* capsid-like gene (bases 1695- 2042) relative to the *Kunsagivirus 1* strain roller/SZAL6-KuV/2011/HUN reference sequence NC_038317. Generated using HKY+F+G4 model with 1000 bootstrap replicates. Scale bar represents 0.2 substitutions per site. A.s= *Apodemus sylvaticus*.

j. Other PCR confirmed hits

Arteriviruses were found within 2 libraries and a total of 2 field voles (1.4% of total animals), with 1 positive animal per library (Table 10). The library L genomic coverage was approximately 15% and was therefore not analysed by NTBLAST. The library N sequence genomic coverage was approximately 40% and the best match by NTBLAST was a PRRSV isolate (KC862571) with 77.78% similarity.

Bocaparvovirus reads were identified in 2 libraries (D and P) and were PCR confirmed in 2 wood mice (1.4% of total animals, Table 10). 1 PCR positive animal was found in each library and genome coverage ranged from approximately 2-25% and was therefore insufficient for reliable NTBLAST analysis or phylogenetic analysis. A *Dependoparvovirus* was PCR confirmed within 1 wood mouse (0.7% of total animals) and *Dependoparvovirus* reads were only found in library O, where 228 reads covering approximately 50% of the genome were found (Table 10). NTBLAST identified this virus as most similar to a murine AAV isolate (NC_055486) with 91.57% similarity. Due to limited genomic information phylogenetic analysis was not performed. *Protoparvovirus* reads were identified in libraries F, I and O, and were successfully PCR confirmed in all 3 libraries, giving a total of 13 positive wood mice (9.3% of total animals tested). The library O sequence provided approximately 5% genomic coverage, the library F sequence provided approximately 15% genomic coverage and the library I sequence provided approximately 25% genomic coverage, therefore accurate assessment by NTBLAST would not have been possible for these viruses (Table 10).

Paramyxovirus reads were found in libraries C and D, and were PCR confirmed in 1 yellow-necked mouse and 1 wood mouse (1.4% of total animals), with 1 successful PCR confirmation per library (Table 10). The library C paramyxovirus sequence provided only approximately 15% of the total genome which was insufficient for NTBLAST analysis. The library D paramyxovirus sequence provided approximately 35% of the total genome which whilst insufficient for phylogenetic analysis did allow for NTBLAST analysis, where the best match was identified as most similar to *Mossman virus* (NC_005339) with 94% similarity. Insufficient genome was recovered to perform reliable phylogenetic analysis.

Pegivirus reads were found in libraries L and N and were PCR confirmed in 1 field vole (0.7% of total animals) (Table 10). Library L provided only 1 read and approximately 1% genomic coverage and could not be PCR confirmed. The library N *Pegivirus* provided approximately 30% genomic coverage and was identified by NTBLAST as most similar to *Phaiomys leucurus Pegivirus* isolate XZS (MW897328) with approximately 75.56% similarity. Due to limited genome coverage phylogenetic analysis was not performed.

Polyomavirus reads covering approximately 50% of the genome were identified in library F and the presence of a polyomavirus was PCR confirmed in 1 wood mouse (0.7% of total animals) (Table 10). Whilst there was insufficient genomic coverage to

perform phylogenetic analysis, coverage was sufficient for NTBLAST analysis which identified this virus as most similar to *Apodemus flavicollis* polyomavirus 1 isolate 3346 (NC_055556) with 96.24% similarity.

Rhadinovirus reads were identified in libraries O and Q and were PCR confirmed in a total of 5 wood mice and 3 yellow-necked mice (5.7% of total animals), confirming hits in both libraries (Table 10). The library Q sequence provided 258 reads, although due to the size of the *Rhadinovirus* genome this only provided approximately 5% genome coverage and therefore could not undergo NTBLAST analysis. The library O sequence provided 722 reads covering approximately 45% of the genome and when analysed by NTBLAST was identified as most closely related to a wood mouse herpesvirus (EF495130) with 99.39% similarity.

k. Other unconfirmed hits

A single *Chapparvovirus* read was identified in library P with approximately 2% genomic coverage and was unable to be confirmed by PCR in any animals. 5 CoV reads were identified in a single library (library J) providing approximately 2% of the genome with 1x read coverage at each point. Neither degenerate CoV primers nor specifically designed library J CoV primers provided any hits by PCR. 10 CMV reads approximating ≤ 1% of the genome were identified in library I. Efforts to confirm this hit by PCR were not successful and no animals tested positive for CMV. A single *Enterovirus C* read was found in library M with approximately 2% genomic coverage, although this could not be PCR confirmed. 4 *Parechovirus* reads were identified in library L providing approximately 2% genomic coverage, but these could also not be confirmed by PCR. *Rotavirus* reads were found in library B, with 1-3 reads per segment being identified and resulting in approximately 15% of the overall genome being identified. However, all samples were negative by PCR. A single *Rotavirus* read was also identified in library I and was unable to be confirmed by PCR (Table 10). Finally, cucumber green mottle mosaic virus was identified at relatively high read numbers in all libraries, likely a contaminant from the cucumber used to bait the traps. As this is a plant virus these reads were not investigated further.

2. Hits by host species

Positivity proportion values were assessed by PCR screening all animals that were represented in the pool in which a viral hit was found. An animal was deemed positive if the virus was PCR positive by one set of primers in either or both target tissues and confirmed by sequencing and NTBLAST analysis of the PCR product. PCR positive animals within the species were then quantified, and the poroprotion positive value was calculated as a proportion of PCR positive animals relative to the total number of animals of the host species.

a. 24 bank vole viruses

A total of 13 bank voles were tested throughout this project, and 24 viral hits were found in total in bank voles representing an average of 1.85 viruses per animal. Within these samples, *Cardioviruses* were PCR confirmed identified in 7 animals, MLV in 6, adenoviruses in 5, *Hepacivirus F* in 3, *Rosaviruses* in 2 and an astrovirus in 1. *Cardiovirus* and *Rosavirus* were identified using *Cardiovirus* or *Rosavirus* specific primers rather than generic picornavirus primers. A total of 10 viruses were found in liver samples and 23 in gut samples (Table 11).

Table 11- Viruses found within bank voles.

All virus hits confirmed within bank vole samples, the tissue in which they were found, and proportion positive values for each virus within the libraries. Only PCR confirmed hits are shown. Total number of viral hits shown in red.

Virus	Liver hits	Gut hits	Positive animals (% proportion positive)
Adenovirus	0	5	5/13 (38.4%)
Astrovirus	0	1	1/13 (7.7%)
Hepacivirus F	3	2	3/13 (23.1%)
MLV	4	6	6/13 (46.2%)
Picornavirus-*Cardiovirus*	2	7	7/13 (53.8%)
Picornavirus-*Rosavirus*	1	2	2/13 (15.4%)
Total hits	10	23	24

Other potential hits include *Rotavirus*, an unidentified picornavirus, *Enterovirus C*, and/or TATV as at least 1 read for each of these were identified within the NGS data. These were unable to be confirmed by PCR and therefore it is not possible to verify these hits or to provide proportion positivity values (Table 10).

b. 15 field vole viruses

A total of 9 field voles were tested throughout this project, and 15 viral hits were identified by PCR within field voles representing an average of 1.67 hits per animal. Within these samples, MLV was PCR confirmed within 3 animals, arteriviruses, *Cardioviruses*, TATV and *Hepacivirus F* in 2 animals, and an astrovirus, *Pegivirus*, *Picobirnavirus* and rodent *Hepacivirus* in 1 animal per virus. *Hepacivirus F* and rodent *Hepacivirus* were identified as two different viruses by CZID and were confirmed using *Hepacivirus F* and rodent *Hepacivirus* primers respectively. A total of 11 viruses were found in liver samples and 14 in gut samples (Table 12).

Table 12- Viruses found within field voles.
All virus hits confirmed within field vole samples, the tissue in which they were found, and proportion positive values for each virus within the libraries. Only PCR confirmed hits are shown here. Total number of viral hits shown in red.

Virus	Positive liver samples	Positive gut samples	Positive animals (% proportion positive)
Arterivirus	2	2	2 (22.2%)
Astrovirus	0	1	1 (11.1%)
Hepacivirus F	2	1	2 (22.2%)
MLV	2	3	3 (33.3%)
Pegivirus	1	1	1 (11.1%)
Picobirnavirus	1	1	1 (11.1%)
Picornavirus-*Cardiovirus*	0	2	2 (22.2%)
Rodent *Hepacivirus*	1	1	1 (11.1%)
TATV	2	2	2 (22.2%)
Total hits	11	14	15

Other potential hits include an adenovirus, *Rotavirus*, *Parechovirus*, paramyxovirus, and/or a *Cardiovirus* as at least 1 read for each of these were identified within the NGS data. These were unable to be confirmed by PCR and therefore it is not possible to verify these hits or to provide proportion positivity values (Table 10).

c. 17 yellow-necked mouse viruses

A total of 17 yellow-necked mice were tested throughout this project, and a total of 33 viral hits were PCR confirmed within these animals representing an average of 1.94 hits per animal. Within these samples, MLV was PCR confirmed within 16 animals, adenoviruses were identified within 11, *Rhadinovirues* in 3 and an astrovirus, paramyxovirus and *Picobirnavirus* within 1 each. A total of 20 viruses were found in liver samples and 23 in gut samples (Table 13).

Table 13- Viruses found within yellow-necked mice.

All virus hits confirmed within yellow-necked mouse samples, the tissue in which they were found, and proportion positive values for each virus within the libraries. Only PCR confirmed hits are shown here. Total number of viral hits shown in red.

Virus	Positive liver samples	Positive gut samples	Positive animals (%proportion positive)
Adenovirus	2	9	11/17 (64.7%)
Astrovirus	0	1	1/17 (5.9%)
MLV	16	9	16/17 (94.1%)
Paramyxovirus	1	1	1/17 (5.9%)
Picobirnavirus	0	1	1/17 (5.9%)
Rhadinovirus	1	2	3/17 (17.6%)
Total hits	20	23	33

Other potential hits include a *Beta papillomavirus* and a *Cardiovirus* as at least 1 read for each of these were identified within the NGS data. These were unable to be confirmed by PCR and therefore it is not possible to verify these hits or to provide proportion positivity values (Table 10).

d. 144 wood mouse viruses

A total of 100 wood mice were tested throughout this project, and a total of 144 viral hits were confirmed by PCR representing an average of 1.44 hits per animal. Within these samples, MLV was PCR confirmed within 69 animals, astroviruses within 20, *Protoparvoviruses* in 13, adenoviruses in 12, *Orbiviruses* within 6, and *Picobirnaviruses* and *Rhadinoviruses* within 5 animals each. Additionally, *Cardiovirus*es were PCR confirmed within 4 animals, *Hepacivirus F* within 3 animals, *Bocaparvoviruses* and *Kunsagiviruses* within 2 animals each, and a paramyxovirus, AAV and polyomavirus were PCR confirmed within 1 animal each. A total of 102 viruses were found in liver samples and 135 in gut samples (Table 14).

Table 14- Viruses found within wood mice.

All virus hits confirmed within wood mouse samples, the tissue in which they were found, and proportion positive values for each virus within the libraries. Only PCR confirmed hits are shown here. Total number of viral hits shown in red.

Virus	Positive liver samples	Positive gut samples	Positive animals (% proportion positive)
AAV	1	1	1 (1%)
Adenovirus	3	10	12 (12%)
Astrovirus	4	18	20 (20%)
Bocaparvovirus	2	1	2 (2%)
Hepacivirus F	2	1	3 (3%)
MLV	68	69	69 (69%)
Orbivirus	6	4	6 (6%)
Paramyxovirus	1	1	1 (1%)
Picobirnavirus	0	5	5 (5%)
Picornavirus-*Cardiovirus*	2	4	4 (%)
Picornavirus-*Kunsagivirus*	1	2	2 (2%)
Polyomavirus	1	1	1 (1%)
Protoparvovirus	8	11	13 (13%)
Rhadinovirus	3	5	5 (5%)
Total hits	102	135	144

Other potential hits include a coronavirus, *Chapparvovirus*, CMV, and a hantavirus as at least 1 read for each of these were identified within the NGS data. These were unable to be confirmed by PCR and therefore it is not possible to verify these hits or to provide proportion positivity values (Table 10).

e. 2 least weasel viruses

Only 1 least weasel was tested during this project, as it was accidentally caught whilst trying to catch the target animals. MLV was identified from both the liver and the gut of the weasel and an astrovirus was identified within the liver, representing a total of 2 viral hits confirmed. No further viruses were identified (Table 15).

Table 15- Viruses found within a least weasel.
All virus hits confirmed within the least weasel sample, the tissue in which they were found, and proportion positive values for each virus within the libraries. Only PCR confirmed hits are shown here. Total number of viral hits shown in red.

Virus	Positive liver samples	Positive gut samples	Positive animals (% proportion positive)
Astrovirus	1	0	1 (100%)
MLV	1	1	1 (100%)
Total hits	2	1	2

Other potential hits include *Hepacivirus F* and a *Parechovirus* as at least 1 read for each of these were identified within the NGS data. These were unable to be confirmed by PCR and therefore it is not possible to verify these hits or to provide proportion positivity values (Table 10).

f. Summary of viral hits

In total, 144 viral hits were identified from liver tissue and 193 from gut tissue across 140 animals giving a total of 216 viral hits and an average of 1.54 viral hits per animal (Table 16). The most common virus by far was MLV, which was PCR confirmed in 67.1% of animals (94/140), followed by adenoviruses and astroviruses, which were PCR confirmed in 20% (28/140) and 17.1% (24/140) of animals respectively. A total of 17 picornaviruses were PCR confirmed, with 9.3% of animals testing positive for *Cardiovirus*es (13/140), and 1.4% of animals testing positive for *Kunsagiviruses* and *Rosaviruses* (2/140 per virus species). *Protoparvoviruses* were PCR confirmed in a total of 9.3% of animals (13/140), whilst *Rhadinoviruess* and *Hepacivirus F* were each PCR confirmed in 5.7% of animals (8/140 per virus). *Picobirnavirus* hits were PCR confirmed in 5% of animals (7/140), *Orbiviruses* within 4.3% of animals (6/140), and TATV, arteriviruses and *Bocaparvoviruses* were each PCR confirmed in 1.4% of animals (2/140 per virus). Finally, a paramyxovirus, *Pegivirus*, polyomavirus, AAV and rodent *Hepacivirus* were each PCR confirmed in 0.7% of animals (1/140 animal each). Other hits were found within the NGS data but were unable to be confirmed by specific PCR. A total of 21 animals were negative for all tested viruses and gave no hits- these were 4 field voles (44.4% of field voles), 2 bank voles (15.4% of bank voles) and 15 wood mice (15% of wood mice).

Table 16- Virus found across all species.
All virus hits confirmed throughout all samples, the tissue in which they were found, *and proportion positive values* for each virus within the libraries. Only PCR confirmed hits are shown here. Total number of viral hits shown in red.

Virus	Positive liver samples	Positive gut samples	Positive animals (% proportion positive)
AAV	1	1	1 (0.7%)
Adenovirus	5	24	28 (20%)
Astrovirus	5	21	24 (17.1%)
Arterivirus	2	2	2 (1.4%)
Bocaparvovirus	2	1	2 (1.4%)
Hepacivirus F	7	4	8 (5.7%)
MLV	91	88	94 (67.1%)
Orbivirus	6	4	6 (4.3%)
Paramyxovirus	1	1	1 (0.7%)
Pegivirus	1	1	1 (0.7%)
Picobirnavirus	1	7	7 (5%)
Picornavirus-*Cardiovirus*	4	13	13 (9.3%)
Picornavirus-*Kunsagivirus*	1	2	2 (1.4%)
Picornavirus-*Rosavirus*	1	2	2 (1.4%)
Polyomavirus	1	1	1 (0.7%)
Protoparvovirus	8	11	13 (9.3%)
Rhadinovirus	4	7	8 (5.7%)
Rodent *Hepacivirus*	1	1	1 (0.7%)
TATV	2	2	2 (1.4%)
Total hits	144	193	216

3. Discussion

MLV was by far the most common virus found within the samples and was present in 67.1% of animals. However, MLV is known to incorporate into host species genomes as an endogenous provirus, which may have been misidentified by the CZID "mapping to reference" process as an exogenous retrovirus[192,193]. Similarly, as PCR confirmation was performed on animal cDNA that included the host genome it is possible that the PCR primers would bind to both endogenous MLV proviruses and exogenous MLV viruses, rendering it difficult to distinguish between viral and genomic PCR positives. Difficulty distinguishing between endogenous and exogenous viral sequences (potentially due to co-evolution of virus and host) may also explain why all MLV sequences were suggested to be probable host genomic contamination by NTBLAST, further complicating the assessment of MLV sequences[59]. Specialist software such as RetroTector is availsble to distinguish between endogenous and exogenous retroviruses, however many of these programmes are specific to human genomes and would not be suitable for this project [326]. Accordingly, due to time constraints and the relative value of these investigations vs. further characterising other genomes investigations into the endogenous or exogenous nature of the MLV hits found was not performed, therefore it is difficult to state that the proportion positivity estimates given here are definitive and accurate representations of the MLV burden within these species. However, not all animals within the same species were positive by MLV PCR and MLV was not found in all libraries. Therefore it stands to reason that MLV proviruses were not found in all animals and that it is probable that some or all of MLV PCR hits were true positives. Similarly, the identification of MLV in all 5 host species suggests that at least some of the MLV hits are likely viral in origin, as it is unlikely that MLV has integrated and persisted within only some members of 5 distinct host species that were trapped in the same location, although in the weasel in particular it is possible that the MLV reads came from food contamination as weasels are known to eat small rodents as a major component of their diet[327].

There are similar issues in interpreting any identified *Dependoparvovirus*, *Chapparvovirus* and *Protoparvovirus* hits, as all can integrate into the host genome and form either proviruses or EVEs[217,218]. Whilst an in-depth analysis of endogenous vs exogenous parvoviruses is beyond the scope of this work, it is possible to draw some relatively simple conclusions here. For example, the *Chapparvovirus* found here was only identified as a single read in 1 library and could not be confirmed by PCR, which suggests that it is a genomic read that mapped to the *Chapparvovirus* reference sequence and is likely not a valid hit. The *Protoparvovirus* hits are more likely to be true positives, as these were PCR confirmed in multiple (but not all) animals in the libraries in which the reads were found. Unlike MLV which was found in all 5 host species, *Protoparvovirus* reads were only identified in wood mice, therefore it is more difficult to confidently assert that these reads are not of endogenous viruses. The *Dependoparvovirus* reads were only PCR confirmed in 1 animal, but were identified by NTBLAST as a murine AAV, somewhat supporting that

this virus is likely to be an exogenous infection rather than an endogenous provirus or an EVE. The *Dependoparvovirus* is perhaps the most important of these, as with further investigations and genomic recovery this could be developed into an effective gene therapy vector[212,217,226].

Adenoviruses were the second most common virus found here. The sequences with sufficient genomic coverage for NTBLAST analysis were identified as most similar to adenovirus species, which when considered with the successful PCR confirmation in these libraries suggests that these are true positive adenovirus hits. Despite having significant genomic coverage for the library G and O viruses there were large gaps in the polymerase and hexon genes that are often used for species demarcation in both viruses, in turn preventing accurate phylogenetic analysis[322]. Neither of the samples that were positive for adenoviruses by degenerate PCR (BV1H and WM29H) were positive by confirmatory PCR following NGS analysis despite both being found in libraries where other adenovirus hits have been confirmed in other animals. This could be due to the high diversity of adenoviruses limiting the efficacy of the PCR primers, although as adenovirus primers were designed based on the NGS data specifically this should have mitigated the impact of this to some degree[109]. It is also possible that the two hits in BV1H and WM29H were false positives as both were identified as most similar to human *Mastadenovirus C* despite originating from rodent tissue, although this could be a result of attempting NTBLAST analysis on a small segment of approximately 300 bases.

Astrovirus reads were PCR confirmed in 11/16 libraries in which reads were found and were the third most common virus identified here. Phylogenetic analysis of both the RdRp gene and the capsid N gene suggest that these viruses are indeed astroviruses, as they all cluster comfortably amongst the astrovirus tree with reasonable bootstrap support and with similar topography across both genes. All hits were identified as highly similar to known astrovirus species by NTBLAST, supporting this identification. It is also reasonable to infer that the PCR confirmed astrovirus hits from other libraries are true positive hits as the same PCR primers were used to identify phylogenetically confirmed hits. The library M and N viruses appear to be somewhat divergent, but due to extremely unclear ICTV guidelines for astrovirus species demarcation that are currently being reworked it is very difficult to assess whether or not these are potentially novel species[322]. It does appear that there are likely to be at least two astrovirus species or strains (separated by host species) that were detected here as two clades were formed in the capsid phylogenetic tree (Figure 10 B), although this may simply illustrate the relative diversity of this gene instead[19]. The recovery and analysis of further genomic information would help to confirm this. It is also interesting that the library F and library K viruses formed a clade in both trees, as library F was primarily composed of animals from Frongoch, and library K was primarily composed of animals from Nant Y Mwyn. This suggests that the Astrovirus genes analysed were similar across this

distance, in turn suggesting that the same species and strain of Astrovirus is likely to be present across these regions of Wales.

Hantavirus reads were found in 6 libraries but could only be PCR confirmed in 2 libraries. For the 4 libraries where the hits could not be confirmed it is likely that low grade contamination from our laboratory has occurred due to a history of TATV work on previous projects[12]. For the library N TATV nearly a full genome was recovered, and all 3 segments were found to be closest to TATVs by NTBLAST, and formed clades with the TATV reference sequences during hantavirus genus phylogenetic analysis. Therefore, it is likely that this virus is a TATV (Figures 11 A, 11 B and 11 C). According to ICTV criteria, new hantavirus species must be > 7% different at the amino acid level across the M and S segments[328]. BLASTX analysis was performed to assess this, as this translates the submitted nucleotide sequence into protein prior to alignment with other sequences, in turn allowing for the matching of sequences with synonymous mutations resulting in the same protein sequence. By BLASTX analysis neither the M nor S segment were greater than 3% divergent from the most similar viruses, suggesting that this virus is not a novel species. Due to previous evidence of TATV in the UK within field voles and the identification of the library N virus as TATV it is reasonable to assume that the library L virus is also TATV, although more genomic information would be recovered to definitively confirm this[12]. The tentatively confirmed TATV hit in the V1H sample within library L also supports the TATV hit in sample V1H degenerate screening, further validating this hit.

As is common in metagenomics projects, many picornaviruses have been identified in this project[251]. Based on the phylogenetic analysis shown in Figures 15 A and 15 B, the *Cardiovirus*es identified in libraries, D, G, I and M are indeed *Cardiovirus*es, as they cluster well within the *Cardiovirus* tree for all both genes. Amongst both trees the library D virus always clusters with an EMCV isolate, and therefore is likely to be more specifically an EMCV. Across the RdRp gene the library I virus appears to be the most divergent and forms its own individual clade amongst the tree. By BLASTX analysis of the entire polyprotein, the library I virus is 93.39% similar to the most similar virus (*Cardiovirus* polyprotein, UZQ18722) and the library G virus is 93.09% similar to the same reference sequence. ICTV species demarcation criteria requires greater than 30% divergence across the polyprotein amino acid sequence, therefore these viruses are *Cardioviruses*[322].

Reads belonging to 4 picornavirus species other than *Cardiovirus* were detected throughout this project. The most informative of these were the *Kunsagivirus* reads of library P which covered approximately 75% of the genome. Whilst it is safe to assess the library P virus as definitely belonging to the genus *Kunsagivirus*, not enough of P1 gene was recovered to assess whether this is a new species within this genus according to the ICTV criteria[322]. In the RdRp gene the library P virus shared a clade with the roller strain reference sequence (NC_038317, Figure 16 A), although this may possibly be due to the rodent reference sequence (ON136180) not providing RdRp coverage, as this virus formed a clade with the rodent reference sequence upon

phylogenetic analysis of the capsid-like gene (Figure 16 B). If confirmed, this would lend significant support to the theory that the first *Kunsagivirus* discovered was not infecting the roller bird that produced the faeces in which it was found but was instead infecting a rodent that was consumed by the roller bird, and support that rodents are the true hosts of *Kunsagiviruses*[254]. The library B *Rosavirus* sequence recovered was also insufficient for assessment of species demarcation by the ICTV standards, although the NTBLAST result and the positive PCR confirmation suggests that the virus is indeed a member of the *Rosavirus* genus[322]. More genomic information would allow for phylogenetic analysis to the species level and assessment of novelty. To our knowledge, this is the first time that a *Rosavirus* has been identified within the UK, and the first example of *Rosavirus* found in a bank vole, although more definitive confirmation and genomic characterisation would be required to confirm this[95,252,254,278].

The library M *Rosavirus* reads, the library M *Enterovirus C* reads and the library L *Parechovirus* reads were all unable to be confirmed by PCR and were all identified by 4 reads or fewer. It is possible that these are legitimate hits and were truly present in the samples, particularly considering the ubiquitous and diverse nature of picornaviruses[111,250]. However, even amongst these highly diverse viruses there are more conserved regions such as the polymerase gene, and reads from 1 picornavirus species such as a *Cardiovirus* may map to a conserved region of a reference sequence for an alternative picornavirus reference sequence, such as a *Parechovirus*[111]. Accordingly, due to low read numbers and the presence of *Cardiovirus*es within library M this is likely to be the case for the library M *Rosavirus* reads and the library M *Enterovirus C* reads, and these hits are likely to be artefacts of alignment. Library L did not contain any other picornaviruses, therefore this logic does not apply for ruling out the *Parechovirus* hit, however due to a lack of PCR confirmation and minimal genomic coverage it is unlikely to be a true positive hit. Alternatively, it is possible that a highly divergent picornavirus is present in library L which did not map to a reference sequence due to significant genomic differences, although more genome recovery would be necessary to investigate this theory[6,250]. It is possible that the reads associated with the unconfirmed *Cardiovirus* in library P could be an erroneous mapping of a *Kunsagivirus* read to a conserved region as a *Kunsagivirus* was PCR confirmed within library P.

Hepacivirus F reads were found in 8 libraries, but it is difficult to definitively state that they are true positive results in any library other than library M. The library M *Hepacivirus F* was identified as most similar to a known *Hepacivirus* F virus by NTBLAST, and upon phylogenetic analysis clustered amongst *Hepacivirus F* viruses in the E2 gene, supporting its identification as such (Figure 12 B). Whilst the library M virus formed an outgroup in the RdRp gene tree (Figure 12 A) it should be noted that the rate of substitutions per site is extremely low, so a small amount of divergence may lead to major changes in tree topography in this instance. Whilst *Hepacivirus F* has previously been identified in mainland Europe, to our knowledge this is the first

time that *Hepacivirus F* has been identified within the UK, suggesting a potential expansion to the range of this virus[100,105]. The ICTV guidelines for a new *Hepacivirus* species state that it must be ≥ 25% divergent across a conserved region of the NS3B gene (amino acids 1123-1566), amongst other criteria[322]. By BLASTX analysis the library M *Hepacivirus* virus is 95.01% similar to the most closely related *Hepacivirus F* virus (CAH0532115, a bank vole from Germany), suggesting that this is not a novel *Hepacivirus* and further supporting the identification of this virus as a member of the *Hepacivirus F* species. Whilst the exact cross-species transmission route of *Hepaciviruses* is currently unknown, one hypothesis involves the action of a biting insect as a vector. If this is the case, then it is possible that in infected insect crossed over from Europe (either carried on high winds or via commercial travel methods) and in turn infected the animals found here, supporting the validity of these findings[329].

For other libraries it is difficult to confirm that *Hepacivirus F* has been found, as the library A and L sequences were identified as most likely to be host genomic contamination by NTBLAST. For the library A sequence this may be an artefact of attempting to analyse a somewhat fragmented sequence with relatively low genomic coverage by NTBLAST, although it is also possible that this is indeed a false positive due to genomic contamination. The positive PCR result in 1 animal does not clarify this, as specific *Hepacivirus F* PCR primers were designed based on the library A sequence and would therefore bind to the sequence regardless of whether it was viral or genomic in origin, and as the cDNA used for PCR screening included the host genome it is impossible to determine whether this PCR result was viral or genomic in origin. The library L sequence yielded approximately 70% *Hepacivirus F* genomic coverage and could not be confirmed by PCR. It is possible that due to the lack of specific host filtration that genomic least weasel reads mapped to the *Hepacivirus F* reference sequence and no viral *Hepacivirus* reads were present, although this is somewhat unlikely as most of the *Hepacivirus F* genome was covered by detected reads. It is unclear why this sequence could not be PCR confirmed as the significant genomic coverage (that includes the primer sequences) suggests a sufficient viral load for PCR detection, as PCR is more sensitive than NGS[65]. Further genomic recovery and phylogenetic analysis may clarify whether these viruses are indeed *Hepaciviruses* or whether host genome contamination was detected here.

The overall *Hepacivirus F* proportion positivity stated here may not be accurate and may in fact be anywhere from 1/140 animals (if only the library M hit is a true positive) to 8/140 animals (if all hits are true positives). Either of these values are lower than the previously published prevalence value of 10.99%, although this value was based on samples from mainland Europe so may not be directly comparable to this study[105]. The recovery of more *Hepacivirus F* genome for these animals would allow for phylogenetic analysis and more accurate confirmation of these hits. The rodent *Hepacivirus* hit of library N was confirmed by PCR using specific rodent *Hepacivirus* primers, but insufficient genomic information is available to provide a

more in depth analysis. Accordingly, it is possible that this virus could be any *Hepacivirus* from species *E-J* as these are all rodent *Hepacivirus*es, including *Hepacivirus F*, although further genomic information would be required to identify which species this virus belongs to[105].

PBV reads were found in 12 libraries and upon phylogenetic analysis the viruses in libraries E, F, G, L and O clustered amongst other PBVs (Figures 14 A and 14 B respectively). NTBLAST identified the viruses from these libraries as most similar to PBVs in all cases except for when assessing the library L virus, further supporting the identification of these viruses as PBVs. The library L PBV segment 2 was reported by NTBLAST analysis to be either a PBV segment 2 or host genomic contamination, although BLASTX analysis identified this as most similar to a chicken *Picobirnavirus* suggesting that this may indeed be a true positive PBV hit. Interestingly, the segment 1 NTBLAST result for the library F virus identified this as a probable PBV RdRp, which is unexpected as the PBV RdRP gene is found on segment 2[245]. It is unclear why this has occurred when the same reference sequence was used for this alignment and all other *Picobirnavirus* alignments and accurately differentiated between segment 1 and segment 2 reads in all other libraries. Additionally, the library F PBV segment 1 had near full genomic coverage, therefore minimising the interference associated with missing genomic segments. By BLASTX analysis the library F segment 1 is similar to a marmot *Picobirnavirus* capsid gene (AVX53485) with 75.96% similarity and 48% query coverage, and is 76.71% similar to a *Picobirnavirus* RdRp gene (QXV86692) with only 17% query coverage. This suggests that perhaps the NTBLAST misidentification of this segment as a RdRp gene is due to high similarity across a small segment rather than a slightly less precise match across a larger query region. The library J, N and P hits could not be confirmed by PCR or analysed by NTBLAST or phylogenetics suggesting these viruses were not true positive hits, perhaps due to host reads misaligning as PBV reads. The library I PBV reads could not be confirmed by PCR despite having significant genomic coverage and undergoing PCR using 4 separate primer pairs, all of which have yielded positive results elsewhere, although it should be noted that the PBV genome is extremely variable which may lead to reduced primer efficiency in some cases[242,243]. This is surprising as sufficient genomic coverage was available to perform phylogenetic analysis on the library I segment 2, which clustered amongst the other PBV sequences from this project within the tree. With PCR confirmation this would be a clear true positive result, but due to a lack of PCR confirmation this cannot be stated definitively despite the support from the phylogenetic and NTBLAST analyses. Accordingly, the prevalence of PBVs may be greater than the proportion positive reported here for these species.

There are no specific ICTV guidelines for *Picobirnavirus* species demarcation, therefore it is difficult to accurately assess whether any of the viruses found here represent novel species or genogroups[322]. Whilst PBVs have been reported within rodents previously, to our knowledge this is the first time that PBVs have been reported in wood mice, yellow-necked mice or field voles, potentially representing

an expansion of the known host species for PBVs, and potentially representing novel PBV species which may be specific to the UK[8,41,241].

Orbivirus reads were found in library E, and whilst the entirety of segment 10 was recovered and not a single read mapping to any other segment was identified. By BLASTX analysis this sequence was found to be most similar to a Kemerovo virus, and although this was with very low similarity of 31.71% no other hits were identified by BLASTX, suggesting that this is a highly divergent *Orbivirus* hit. However, this is not sufficient to reasonably assess species demarcation as ICTV species demarcation assessment requires analysis of the *Orbivirus* segments 1 and 3, as well as recombination studies[322]. Whilst *Orbivirus*es have been reported in most of the world native *Orbivirus*es have never been reported within the UK or UK wildlife to our knowledge, therefore if confirmed through further genome recovery and analysis this could represent a major expansion of the known range of *Orbivirus*es with potentially significant implications for the UK agricultural industry and livestock management[198,199]. If this was the case, the geographical isolation of the UK *Orbivirus* would also explain the substantial divergence seen here, and would suggest that the virus found here would likely represent a new species, although the recovery of at least all of segment 1 and 3 and ideally the whole genome would be required to confirm this.

Arterivirus reads were recovered from 2 libraries and were PCR confirmed in 1 animal in each library. There was insufficient genomic coverage in either library to allow for phylogenetic analysis, but the NTBLAST result for library N of most similarity to PRRSV suggests that for library N at least this was a true arterivirus positive. Considering the library L PCR positive result was confirmed using the same primer set, it is reasonable to suggest that this is also a true positive result. Insufficient genomic sequence was recovered to accurately assess potential novelty of these arteriviruses. The main value of these findings is the proportion positive of 22.2% within field voles (2/9 animals), which despite the small sample size suggests that there could be a large arterivirus burden within field voles. This could have agricultural implications due to the potential cross-species transmission of arterivruses from small mammals to livestock animals, where arteriviruses such as PRRSV can cause severe disease[141].

Bocaparvovirus reads were found in 2 libraries and were PCR confirmed in 2 animals. Due to the absence of any NTBLAST data or phylogenetic analysis due to insufficient genome recovery it is difficult to definitively state that these were true positive hits, although PCR confirmation with *Bocaparvovirus* specific primers suggests that this is the case. Both hits were found in wood mice, and whilst *Bocaparvovirus*es have previously been found in other members of the *Apodemus* genus this is to our knowledge the first time that they have been reported within *A. flavicollis* specifically[8,224]. Whilst more genomic sequence is required to validate this hit and to perform phylogenetic analysis, if confirmed this suggests that *Bocaparvovirus*es may infect a broader range of rodent host species than currently believed.

Paramyxovirus reads were detected in 2 libraries, and insufficient genome was recovered from either library for phylogenetic analysis, although the library D (wood-mouse) sequence was identified by NTBLAST as being most similar to *Mossman virus*. *Mossman virus* is a paramyxovirus of mice suggesting that this paramyxovirus hit is likely to be a true positive result, and as the same PCR primers were used for the library C hit this implies that this is also likely to be a true positive result[330,331]. Many other paramyxoviruses have also been detected in a variety of rodent species, supporting the possibility of finding paramyxoviruses within the rodents tested here[331].

Pegivirus reads were identified in library L and library N. For library L, only a single read was found and could not be PCR confirmed. Accordingly, this is likely an erroneous read, or perhaps a *Hepacivirus* read to a conserved section of genome that has been misidentified as a *Pegivirus* read due to the genetic similarity of *Hepaciviruses* and *Pegiviruses*[185,237]. The library N *Pegivirus* was PCR confirmed, and whilst insufficient genomic coverage was available to perform phylogenetic analysis this virus was identified as most similar to a *Pegivirus* found in Blyth's vole (*Phaiomys leucurus*) by NTBLAST, suggesting that this is a true positive *Pegivirus* identification. To our knowledge, this is the first time that a *Pegivirus* has been reported in bank voles, although *Pegivirus*es have been reported in a variety of other rodent species[234,237]. Whilst this is unlikely to have major implications for either human or animal welfare due to the believed inability of *Pegivirus*es to undergo cross-species transmission, if confirmed by further genomic analysis this would nonetheless represent a new host species for *Pegivirus* infection[235,237].

Polyomavirus reads were found and PCR confirmed in 1 wood mouse within library F, providing approximately 50% genomic coverage. This was found to be most closely related to an *A. flavicollis* polyomavirus, which when considered with the positive PCR result suggests that this is a true positive polyomavirus hit. To our knowledge this is the first time that polyomaviruses have been reported from *A. sylvaticus* specifically, although polyomaviruses have previously been reported from other rodent species including *A. flavicollis*, therefore it is perhaps unsurprising that polyomaviruses can also infect the closely related wood mouse[103,281]. Further genomic recovery would be necessary to definitively confirm the presence of a wood mouse polyomavirus and to allow for species demarcation and novelty assessment.

Rhadinovirus reads were found in two libraries and the library O sequence was identified as most similar to a wood mouse herpesvirus by NTBLAST with > 99% similarity, which is reasonable considering *Rhadinovirus*es are within the *Herpesviridae* family[169]. Considering this and the PCR confirmation it is reasonable to assume that these are true positive *Rhadinovirus* hits. For the library Q virus insufficient genome was recovered for accurate NTBLAST analysis, but PCR hits were confirmed using the same primers as for the library O hits, providing reasonable evidence that these are true positive hits. Whilst many human *Rhadinoviruses* can

have clinical consequences there is little evidence that rodent *Rhadinoviruses* are significant, or that they can easily undergo cross-species transmission[169,287].

Many other virus hits were identified within the NGS data which could not be confirmed by PCR. For example, the CoV reads found could not be PCR confirmed, and are likely contamination from previous CoV work within our laboratory. This is likely also the case for the *Rotavirus* reads identified here. The CMV reads may represent a false positive, or potentially a few reads of an endogenous or latent CMV infection in 1 animal- regardless, this could not be PCR confirmed and insufficient genome was recovered to investigate any further and this was discounted from any further analysis[332]. Cucumber green mottle mosaic virus was also identified in all pools, in many cases with significant genomic coverage, although this was not investigated further due to the fact that these are exclusively plant viruses that cannot infect mammals[333]. As with all metagenomics studies, a variety of bacterial, fungal and archaeal reads were detected, but the analysis of these was beyond the scope of the project. Bacteriophage reads were also detected in all libraries, but as stated in the introduction these were not analysed as this was beyond the scope of this project[5,59].

Proportion positivity values provided here must not be interpreted as broadly applicable prevalence values. To make prevalence estimates such as these, many further factors require consideration, including sample size vs. overall host species population, geographic distribution, seasonal distribution, and the health of the animal host (where diseased animals are more likely to be caught than healthy animals potentially leading to a prevalence over-estimation)[334]. Here, there is a reasonable chance of an overestimation and an underestimation of proportion positivity for the *Hepacivirus F* and *Picobirnaviruses* respectively. For all species except for wood mice the sample size was somewhat small, potentially reducing the accuracy of any prevalence estimates given, hence the alternative use of proportion positivity. Additionally, the vast majority of these samples were collected in September or October, and as rodent viral burdens have previously been shown to vary by season any prevalence data presented may have been inaccurate during summer and winter[41]. The animals sampled represented a range of adult and juvenile rodents, therefore the age of the animal is unlikely to be a significant source of error in this work.

UK wildlife has been under sampled to date, and to our knowledge this is the first time that an in-depth analysis of the virome of Welsh rodents has been performed[23]. Accordingly, these proportion positivity values do not have reasonable points of comparison within the UK at the time of writing[23]. As Great Britain is an island and is therefore limited in cross-species interaction, prevalence estimates for other parts of the world- including mainland Europe- may not be effective comparators for this study. Therefore whilst the proportion positivity estimates provided here are subject to the usual limitations of wildlife studies such as inevitable sampling bias, these are still valid and reasonably accurate proportion positivity estimates, although further

research and data regarding host population sizes would be required to convert these values into accurate prevalence estimates[23,28]. More viruses were identified in gut tissue than liver tissue, perhaps due to viruses present within food consumed by the animal, or possibly due to simply extracting larger gut tissue sections than liver sections allowing for the recovery of increased viral copy numbers and increased sensitivity[23,65]. Interestingly, the average number of hits per animal ranged from 1.44 (in wood mice) to 1.94 (in yellow-necked mice; discounting the single least weasel sample which had 3 hits), suggesting that no individual species of those tested here had a greater overall viral richness or burden than any other. Whilst this is somewhat unexpected, relatively high viral burdens and substantial viral richness is often found in rodents, therefore it is not too surprising that all species tested provided hits for many viruses[7].

It is unlikely that there are incidental viruses present in liver tissue that are not derived from the host. However, this is likely to occur when considering viruses found in the gut tissue, due to the potential detection of viruses present in food consumed by the host animal rather than directly infecting the animal host itself[23]. For example, the presence of cucumber green mottle mosaic virus that was most likely derived from the cucumber used to bait the traps that were used to capture the animals illustrates this clearly, as this is a plant virus that cannot infect mammalian hosts[333]. Therefore in instances where the only PCR confirmation for a specific virus is in the gut tissue without phylogenetic data or NTBLAST data to support this- such as in the yellow-necked mouse astrovirus in library C- it is impossible be entirely certain that the virus in question is indeed present in the rodent host tissue rather than the food of the host. Whilst this could inflate proportion positivity values, it is debatable as to whether this is a valid hit as it is reasonable to argue that the consumption of infected food is probable within the lifecycle of the host animal. This leads to a potential source of infection, and therefore suggests that the virus is still a valid hit and will be present in this manner in some wild hosts[5]. Furthermore, despite the exact source of the virus being unclear within this specific situation, it is still reasonable that these viruses be investigated as they have been discovered regardless of source and are still valid hits from a virus discovery perspective.

Some of the animals tested here were negative for all viruses and did not appear to give any specific viral hits either by PCR or NGS. That they are truly infected by no viruses is very unlikely due to the ubiquitous nature of viruses infecting all mammalian life[5]. It is possible that the viruses present in these samples were present at very low copy numbers below the sensitivity threshold of NGS, or as bacterial and fungal background reads were not considered in this project it is possible that a significant bacterial presence was preferentially sequenced, reducing the reads available for amplification of viral sequences[65]. Alternatively, the diluting effect of pooling the samples may have reduced the number of reads per animal to the point where low copy number viruses were not detected, although an average of approximately 22.14 million reads were sequenced per animal rendering this

unlikely[41]. This diluting effect may also explain to some extent why no truly novel viruses have been definitively identified here, as any low copy number divergent reads may have been diluted out and not mapped to reference sequences, in turn being lost as "dark matter"[32]. Another potential issue with the pooling approach taken here is that often multiple animals within the same pool tested positive for the same virus. Accordingly, when analysed by metagenomics, there is a risk that multiple strains/species of the same genus were analysed and compiled into one consensus genome, which would either lead to artificial diversity and variability relative to a reference or a low copy number divergent virus being masked by a higher copy number and less divergent virus[59]. This could be further elucidated by primer walking and full genome characterisaiton of each virus found, although this was neither financially nor practically viable.

The pooling dilution effect and the random subsampling by CZID may also explain the absence of commonly found viruses such as γ-herpesviruses and LCMV[99.335]. Whilst no LCMV reads were found, sporadic individual γ-herpesvirus reads were occasionally found, albeit never with more than one read per library. Accordingly, these were not investigated and were discounted from further analysis. However, with greater read depth and/or effective degenerate screening primers it may have been possible to screen all samples for these viruses, and it is likely that a proportion positivity estimate could have been gained for these viruses. This was not performed due to time and financial constraints.

There were both advantages and disadvantages to the primer design approach taken here. By designing specific primers for individual viruses based on the NGS data, in theory primers should be perfectly accurate for the specific virus detected within the library and be able to detect very low copy numbers of this virus. Indeed, if enough copies are present for NGS sequencing, PCR should be able to amplify these viruses[65]. This was largely seen here, as over 66% of viruses detected by the NGS analysis were PCR confirmed. However, if there were multiple species of the same genus or strains of the same species within the library some viruses may have been preferentially amplified over others, essentially masking closely related hits within the same library. Another issue is that whilst this approach was taken at the start, only a few primer sets were designed for each virus based upon up to 4 libraries (excluding cardiovirus primers, where more sets were developed). In situations where viruses were found in most libraries- for example, when considering adenovirus or astrovirus hits- designing specific primers for each library may have increased the chances of confirming a hit. However, this was neither cost effective nor efficient, therefore designing up to 4 primer sets for an individual virus was deemed to be a reasonable middle ground between confirming as many hits as possible whilst being practically feasible. It is also true that by taking this approach highly divergent viruses may not have been detected, and viruses with highly variable genomes such as PBVs show variability across the binding sites between libraries, reducing primer sensitivity and leading to false negative results[242,243].

With more genomic coverage for many of these viruses more in-depth phylogenetic analysis could be performed, and increased genomic coverage may allow for the analysis of more samples, as this would likely reduce the amount of sequences that formed outgroups reducing the clarity of the tree that resulted in their omission. Ideally primer walking PCR would have been used to increase genomic coverage of key genes and recover full genomes. This would have allowed for improved phylogenetic analysis and species demarcation for many of the viruses found here, although this was not performed due to time and budget constraints[47]. Other approaches for enhancing phylogenetic analysis were considered, such as concatenating reads within the library O adenovirus to produce pseudo-genomes that were more suitable for phylogenetic analysis, although in this specific case all of the important genes for phylogenetic analysis such as the hexon gene were still missing significant amounts of sequence rendering phylogenetic assessment unreliable.

Future work for these samples would ideally involve recovering more genomic sequence in positive animals by overlapping PCRs and primer walking, as this would then allow for more in depth phylogenetic analysis and increased confidence when confirming hits[47]. Priority should be given to the *Kunsagivirus* and *Rosavirus* genomes, as these are relatively rare picornaviruses which may lead to the identification of novel species[255,278]. Enhanced confirmation and phylogenetic analysis should also be performed on those viruses which have been found in a new host for the first time, including the *Hepacivirus F* viruses, PBVs and *Orbiviruses* found here. This would also allow the elucidation of which of these viruses have zoonotic potential, as improved phylogenetic analysis would allow for the identification of how similar the viruses are to known zoonotic viruses. It is not reasonable to infer that any of the viruses found here are zoonotic without futher investigations. However, many of the genera found here including adenoviruses, *Cardioviruses* and PBVs include species that have previously been shown to undergo zoonotic transmission, and therefore it is possible that some of the viruses of these species found here may also have zoonotic potential[95,119,246]. Although rare, some species of polyomavirus also have zoonotic potential and therefore the polyomavirus found here is worth investigating futher, and some species of paramyxovirus have also been shown to undergo zoonotic transmission suggesting that the paramyxovirus found here also warrants further investigation[207,280].

Whilst this work only considers the virology of the metagenomics, many reads for bacteriophages, bacteria, fungi and archea were also identified in the specimens. It is beyond the scope of this thesis to analyse this data, but this data will be shared with other research groups and the University of Nottingham to allow for in-depth analysis and investigation into these reads, before eventually being made publically available.

Chapter 6- Historic virus investigation results (Prong 3)

1. Sample collection

In total, 21 *Epomops franqueti* (Franquet's epauletted fruit bat), 9 *Epomops buettikoferi* (Buettikofer's epauletted fruit bat), 6 *Hypsignathus monstrosus* (Hammerheaded fruit bat), 4 *Miniopterus inflatus* (Greater long-fingered bat) and 2 *Rhinolophus affinis* (Intermediate horseshoe bat) were sampled, giving a total of 42 unique bat specimens sampled. 18 *Mastomys natalensis* (Natal multimammate mouse), 2 *Mastomys erythroleucus* (Guinea multimammate mouse), 1 *Mastomys spp.* that is either *Mastomys coucha* (Southern multimammate mouse) or *M. erythroleucus*, 3 *Praomys tullbergi* (Tullberg's soft-furred mouse) and 1 *Mastomys spp.* were sampled, giving a total of 25 unique rodent specimens, and a total of 67 unique specimens overall (Table 2). *Praomys* and *Mastomys* are closely related genera both morphologically and phylogenetically and have previously been considered to be the same genus prior to a relatively recent reclassification, hence the double labelling on some of the specimen jars, however it is safe to assume that any specimen containing the species name *natalensis* is a member of the *Mastomys* genus[336]. *M. coucha* and *M. natalensis* are also phylogenetically and morphologically very similar, therefore it is not too surprising that initially determining the species of these specimens was unsuccessful[336]. Cytochrome B speciation was not performed due to the lack of sample material available- accordingly, some specimens were not identified to the species level and the morphologically stated species name was not confirmed. The minimum age of sample is based on the date of entry into the NHM collection rather than the actual collection date of the sample itself by the expedition, as this information is not always available. Accordingly, this information is the minimum time since collection and samples are often likely to be older than the value provided here. The specific specimens that batch 1 came from were not recorded- accordingly, information about age and specimen number is not available.

2. RNA extraction, cDNA synthesis and GAPDH analysis

a. RNA extraction

RNA extraction was performed for all samples. Following a successful extraction, each sample RNA underwent quantity and quality assessment using the Nanodrop spectrophotometer, and quantification and fragmentation assessment using the TapeStation electrophoresis platform. As described in chapter 3, optimal quality values for Nanodrop 260 nm/280 nm and 260 nm/230 nm ratios are between 2.00 and 2.2, but for these archival and degraded samples a reduction in these purity ratios was expected and purity ratios outside of this range did not preclude samples from further processing. Similarly, historic RNA was expected to be highly degraded and fragmented due to its age and a low RIN value was expected, although where the TapeStation could not provide a RIN and gave a result of "N/A" the sample was considered to be too degraded to generate a metagenomics library. Table 17 shows the RNA extraction information for each sample.

A total of 79 RNA extractions were attempted throughout this project. The batch 1 and 2 samples are not counted amongst these, as these were extracted prior to the author joining the project and the data was not available to the author- if these were counted, the total would be 89 extractions. Of these 79 extractions, 68 were successful. 2 extractions (1867.4.12.324G and 1867.4.12.324H) were unsuccessful due to physical degradation of the sample, where the sample was so friable that during transport it broke down into tiny tissue chunks that could not be used. 9 extractions (shown as "extraction unsuccessful" in Table 17) appeared to be successful (i.e they were able to be processed and proceed to the quantification step of the extraction process) but failed to show any RNA concentration data during either Nanodrop or TapeStation assessment. It is unclear why these samples failed to extract, although it is interesting to note that sample 1966.3503 was successfully extracted in batches 3 and 4 but failed to be extracted in batch 5. cDNA synthesis was then attempted on all samples that gave either a positive RIN or a positive quantity of RNA upon TapeStation analysis. cDNA synthesis success was assessed via GAPDH PCR.

Table 17- Historic sample RNA extraction.

Shows a successful or failed RNA extraction for each sample, and if successful the Nanodrop concentration and 260/280 and 260/230 purity ratio, the TapeStation concentration and RIN value and GAPDH 70 PCR result. “Sample too degraded to extract” represents a physically degraded sample, “extraction unsuccessful” illustrates a sample which successfully completed the RNA extraction process but gave no RNA concentration upon either Nanodrop or TapeStation analysis. For GAPDH analysis, + represents a positive sample assessed via endpoint PCR rather than qPCR, therefore no Ct value is available. - represents a negative sample. Samples with blacked out boxes were not analysed GAPDH PCR.

Sample	Nanodrop concentration (ng/µl)	260/280	260/230	TapeStation concentration (ng/µl)	RIN	GAPDH 70 Ct
1984.1654G	4.5	1.49	0.13	0.355	N/A	-
1984.1655G	8.2	1.53	0.54	0.292	N/A	-
1966.3502G	15.1	1.47	0.35	0.0763	N/A	+
1966.3503G	16.1	1.41	0.25	0.461	2.5	+
1966.3504G	Extraction unsuccessful					
1966.3505G	Extraction unsuccessful					
1966.3506G	185.9	1.87	1.50	64.8	2.1	+
1880.7.21.3G	Extraction unsuccessful					
1880.7.21.1G	15.2	1.49	0.56	0.656	2.6	+
1880.7.21.4G	5.0	1.59	0.24	0.647	2.4	-
1948.598G	29.7	1.45	0.67	5.52	N/A	-
1867.4.12.324G	54.9	1.20	0.26	6.29	N/A	-
1932.1.17.14G	Extraction unsuccessful					
1932.1.17.15G	28.6	1.63	0.51	7.26	N/A	+
1979.1229G	19.5	1.41	0.63	3.920	2.3	-
1979.1230G	Extraction unsuccessful					
1979.1240G	11.9	1.41	0.31	1.640	1.7	-
1966.3503G2	33.7	1.10	0.26	0.296	N/A	-
1966.3503H	39.5	1.14	0.27	0.590	3.6	36
1966.3506G2	516.1	2.01	1.84	389.6	2.3	-
1966.3506H	954.4	2.07	2.02	624	2.2	27
1880.7.21.1G2	23.5	1.11	0.28	0.117	N/A	-
1880.7.21.1H	48.3	1.15	0.29	0.171	N/A	-
1880.7.21.4G2	49.8	1.18	0.32	1.700	3.3	38
1880.7.21.4H	101.0	1.11	0.28	0.165	N/A	> 40
1979.1229G2	33.6	1.45	0.68	1.806	1.4	-
1979.1229H	23.8	1.34	0.48	0.644	1.8	38
1979.1240G2	34.0	1.43	0.97	7.30	N/A	-
1979.1240H	26.7	1.43	0.79	1.260	1.9	> 40
1867.4.12.324G2	Sample too degraded to extract					
1867.4.12.324H	Sample too degraded to extract					
1966.3503H2	Extraction unsuccessful					

1966.3506H2	1040.9	2.03	1.62	114	1.8	
1979.1229H2	21.2	1.46	0.69	1.68	1.6	
1979.1240H2	25.7	1.35	0.44	2.07	1.7	
1880.7.21.4H2	62.5	1.11	0.29	4.98	N/A	
1984.1655H	2.2	1.12	0.08	0.661	1.7	
66.3498	358.8	1.98	1.79	104	2.4	32
69.963	29.1	1.24	0.31	10.2	3.9	36
59.205	122.2	1.8	1.47	56.3	1.9	-
47.588/B62	26.6	1.27	0.37	4.97	N/A	-
79.428	119.5	1.65	1.47	31.4	1.3	-
62.1819	239.6	1.82	1.56	105	1.5	-
68.962	24.9	1.22	0.29	0.337	N/A	36
68.966	29.6	1.11	0.25	5.15	N/A	-
79.434	22	1.57	1	3.22	1.7	-
62.1818	5.8	1	0.11	0.241	N/A	-
68.357	9.5	1.35	0.53	0.277	N/A	37
66.3509	135.3	1.84	1.19	29.7	2.7	36
68.356	1.1	1.06	0.16	0.831	N/A	-
68.970	41.5	1.42	0.46	10.6	3.6	31
64.59	209.4	1.92	1.66	28.6	2.7	29
68.971	208.2	1.91	1.48	36.2	3.2	29
84.800	31.7	1.57	0.78	10.4	2.4	35
71.607	Extraction unsuccessful					
79.424	239.4	1.74	1.63	84.7	1.5	34
72.285	38.3	1.61	0.77	7.18	N/A	-
61.1642	40.8	1.6	0.7	9.20	N/A	36
61.16.43	15.7	1.49	0.38	0.062	N/A	-
76.1540	15.1	1.58	0.42	0.675	1	39
79.1241	5.4	1.63	0.46	0.378	N/A	-
76.1547	38	1.52	1.04	7.01	N/A	-
48.1163	98.9	1.55	1.1	13.0	2.3	-
79.1181	8.2	1.56	0.65	0.432	2	-
79.1198	7.9	1.5	0.63	0.172	N/A	-
1988.167	16.8	1.45	0.61	0.502	2.1	39
79.1295	4.3	1.57	0.39	0.115	N/A	-
79.1297	3.1	1.75	0.39	0.768	N/A	-
79.1294	7.5	0.97	0.13	N/A	N/A	
79.1537	14.7	1.56	0.66	3.08	1.7	38
1988.169	7.9	1.47	0.98	0.748	N/A	-
77.3425	15.9	1.49	0.73	0.728	1.7	38
77.3427	27.6	1.6	0.89	7.01	N/A	38
68.652	12.7	1.29	0.28	N/A	N/A	
71.393	13	1.47	0.98	0.142	N/A	-
71.402	Extraction unsuccessful					
68.654	Extraction unsuccessful					

12.11.25.16	6.5	1.17	0.25	N/A	N/A	
9.1.9.25	16.4	1.48	0.48	0.858	N/A	-

b. GAPDH analysis

GAPDH PCR was used to assess cDNA synthesis and as a proxy measurement for RNA fragmentation. GAPDH analysis was not performed on batch 2 or 5 samples due limited RNA and sample RNA was preserved to use for library preparation instead. Ct data is unavailable for batch 1 and 3 samples as these experiments were conducted using endpoint PCR rather than qPCR. All samples that gave a positive RNA concentration upon TapeStation analysis underwent GAPDH PCR (Table 18). GAPDH 70 PCR was performed first and GAPDH 70 positive samples were then tested for GAPDH 149, and GAPDH 149 positive samples were tested for GAPDH 295 as another proxy measurement of RNA fragmentation. Fragment sizes were estimated to be somewhere between the largest product size GAPDH positive result and smallest product size GAPDH negative result for each sample.

Table 18- Historic sample GAPDH PCR results.

Shows samples that underwent GAPDH investigations, which GAPDH investigations were performed and Ct values for positive samples. + indicates a positive result without a Ct value as these samples were analysed using endpoint PCR rather than qPCR. - represents a negative sample. Blacked out boxes represent no experiment performed for that sample using those GAPDH primers.

Sample	GAPDH 70 Ct	GAPDH 149 Ct	GAPDH 295 Ct	Sample	GAPDH 70 Ct	GAPDH 149 Ct	GAPDH 295 Ct
Epo1	+	+	+	47.588/B62	-		
Epo2	+	+	+	79.428	-		
Min1	-			62.1819	-		
Min2	-			68.962	36	-	
Hyp1	-			68.966	-		
Hyp2	-			79.434	-		
1984.1654G	-			62.1818	-		
1984.1655G	-			68.357	37	-	
1966.3502G	+			66.3509	36	-	
1966.3503G	+			68.356	-		
1966.3506G	+			68.970	31	39	-
1880.7.21.1G	+			64.59	29	39	> 40
1880.7.21.4G	-			68.971	29	40	-
1948.598G	-			84.800	35	-	
1867.4.12.324G	-			79.424	34	-	
1932.1.17.15G	+			72.285	-		
1979.1229G	-			61.1642	36	-	
1979.1240G	-			61.16.43	-		
1966.3503G2	-			76.1540	39	38	39
1966.3503H	36			79.1241	-		
1966.3506G2	-			76.1547	-		

1966.3506H	27	[redacted]	48.1163	-	[redacted]
1880.7.21.1G2	-		79.1181	-	
1880.7.21.1H	-		79.1198	-	
1880.7.21.4G2	38		1988.167	39	-
1880.7.21.4H	> 40		79.1295	-	
1979.1229G2	-		79.1297	-	
1979.1229H	38		79.1537	38	-
1979.1240G2	-		1988.169	-	
1979.1240H	> 40		77.3425	38	-
66.3498	32	-	77.3427	38	-
69.963	36	-	71.393	-	
59.205	-		9.1.9.25	-	

Whilst conventionally qPCR would use a certain Ct cutoff value (often 35) to identify positive samples, due to the age and degradation of these samples any positive Ct value was considered as a positive result, including for samples that had not reached the Ct threshold but were beginning to show a peak at a Ct > 40. In essence, Ct value requirements were relaxed relative to most reactions to ensure sufficient sensitivity. A total of 50 RNA extractions were attempted on bat samples, resulting in 44 eligible for GAPDH screening. Of these, 21 were GAPDH 70 positive, 5 were GAPDH 149 positive, and 3 were GAPDH 295 positive. 29 RNA extractions were attempted on rodents, resulting in 22 samples eligible for GAPDH screening. Of these, 8 were GAPDH 70 positive, 1 was GAPDH 149 positive, and 1 was GAPDH 295 positive. In total, 66 samples were eligible for GAPDH screening, and of these 29 were GAPDH 70 positive, 6 were GAPDH 149 positive and 4 were GAPDH 295 positive. Results breakdowns by species are shown in Table 19.

Table 19- GAPDH PCR success by species.
All %s are of the total number of attempted extractions for that species, excluding those from batches 2 and 5. *E. franqueti, E. franqueti Tomes* and *E. franqueti strepitans* are grouped for this analysis, as are *M. inflatus* and *M. inflatus africanus*, and *M. natalensis* and M. *(Praomys) natalensis*. For this analysis, sample 48.1163 is considered as *M. erythroleucus*.

Species	Attempted extractions	Eligible for GAPDH PCR (%)	GAPDH 70 positive (%)	GAPDH 149 positive (%)	GAPDH 295 positive (%)
E. franqueti	29	24 (82.76%)	13 (44.83%)	2 (6.90%)	2 (6.90%)
E. buettikoferi	9	9 (100%)	3 (33.33%)	0 (0%)	
H. monstrosus	6	6 (100%)	4 (66.66%)	3 (50%)	1 (16.16%)
M. inflatus	4	3 (75%)	0 (0%)		
R. affinis	2	2 (100%)	1 (50%)	0 (0%)	
Bat samples	50	44 (88%)	21 (42%)	5 (10%)	3 (6%)
M. natalensis	22	18 (81.82%)	8 (36.36%)	1 (4.55%)	1 (4.55%)
M. eyrthroleucus	3	1 (33.33%)	0 (0%)		
P. tullbergi	3	2 (66.66%)	0 (0%)		
M. spp.	1	1 (100%)	0 (0%)		
Rodent samples	29	22 (75.86%)	8 (27.59%)	1 (3.45%)	1 (3.45%)
All samples	79	66 (83.54%)	29 (36.71%)	6 (7.59%)	4 (5.06%)

All data was tested for normality via Shapiro-Wilk normality test, and all data was found to follow a non-normal distribution pattern. Overall, there was no link between RNA concentration and GAPDH 70 success rate, suggesting that the quality of the sample is not directly linked to RNA extraction quantity. The average concentration of GAPDH positive samples was 101.02 ng/µl by Nanodrop and 37.09 ng/µl by TapeStation, whereas for the GAPDH negative samples the average Nanodrop concentration was 51.29 ng/µl and the average TapeStation concentration was 20.47 ng/µl. However, these differences were not statistically significant ($p=0.076$ for Nanodrop concentration and $p=0.208$ for TapeStation concentration by Mann-Whitney test). Nanodrop and TapeStation concentrations were also proportional for most samples (Figure 17 A). The average minimum age of GAPDH 70 positive samples was 65.79 years vs. 62.53 years for GAPDH negative samples, showing no significant differences ($p=0.538$, Mann-Whitney test). There was no correlation between age and GAPDH success ($p=0.70$, linear regression), nor between age and GAPDH 70 Ct value ($p=0.80$) (Figures 17 B and 17 C, respectively). For calculation purposes, N/A RIN results were considered to be 0. GAPDH 70 positive samples had an average RIN value of 1.83 whereas GAPDH 70 negative samples had an average RIN of 0.58, and this value was statistically significantly different ($p \leq 0.0001$) (Figure 17 D). There was no correlation between RIN and age ($p=0.79$, linear regression).

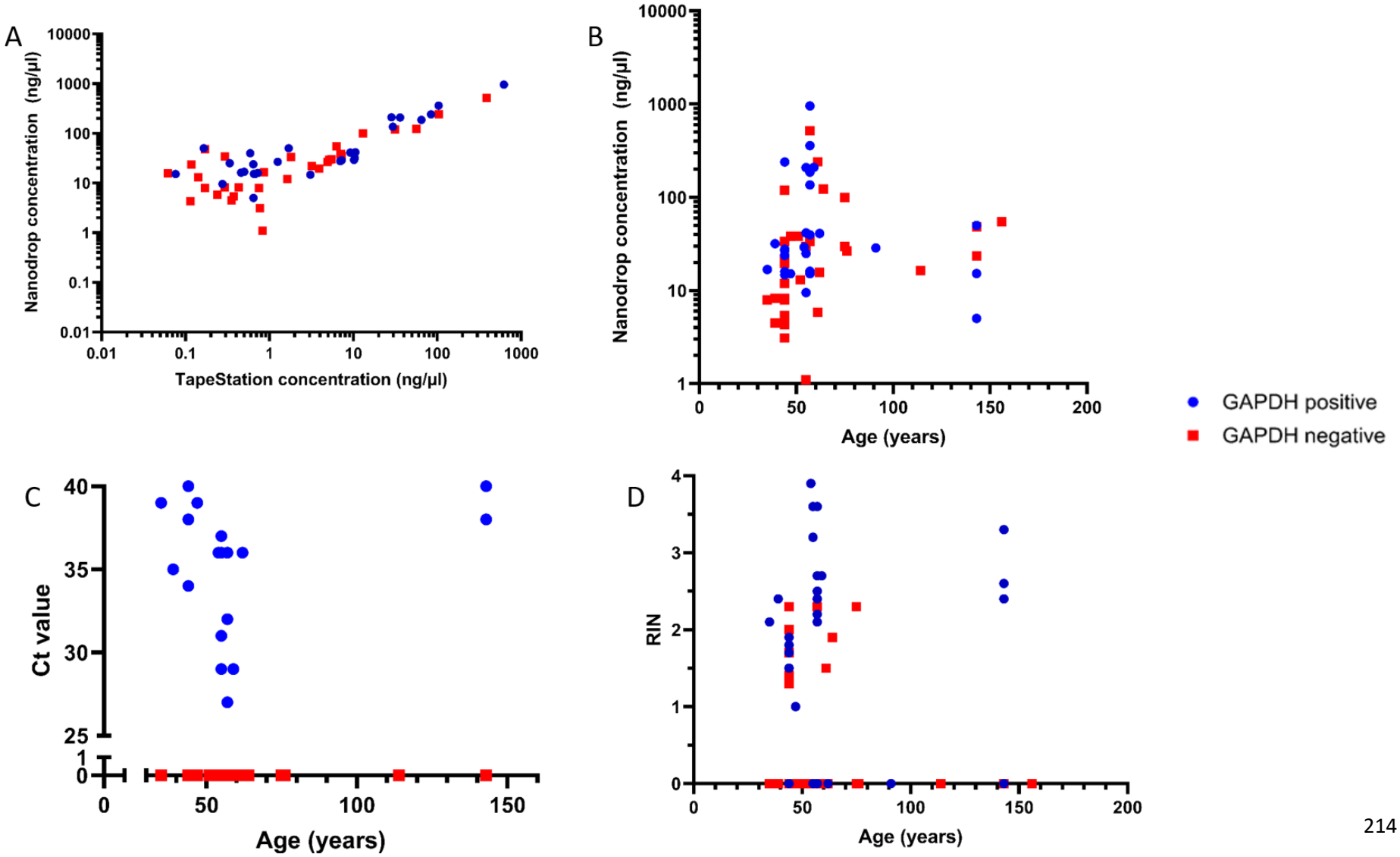

A
Nanodrop concentration (ng/µl)
TapeStation concentration (ng/µl)
B
Nanodrop concentration (ng/µl)
Age (years)
GAPDH positive
GAPDH negative
C
Ct value
Age (years)
D
RIN
Age (years)

Figure 17- Analysis of historic sample concentration, age, RIN and GAPDH 70 status. For all figures, blue circles are GAPDH positive samples, and red squares are GAPDH negative samples. For RIN calculations, N/A RIN results were counted as 0 for calculation purposes. A) Nanodrop vs TapeStation concentrations for each sample. B) Nanodrop concentration vs minimum age of each sample, and whether each sample was GAPDH positive. C) Age vs GAPDH 70 Ct value. D) RIN vs minimum age of each sample. No statistically significant relationships were found except for a positive correlation between RIN value and GAPDH PCR success ($p \leq 0.0001$, t-test), and an increased average RIN value was associated with GAPDH 70 success ($p \leq 0.001$, Mann-Whitney test).

3. Library preparation

1966.3506G and 1966.3506H initially underwent the Illumina library preparation process as described in the manufacturer's protocol with no modifications. These samples were chosen as they had the highest concentrations of all samples according to both the Nanodrop and the TapeStation, whilst having acceptable RINs by archival RNA standards (2.3 and 2.2, respectively). The library preparations were unsuccessful, and no library was visible on the TapeStation traces. If a library had been successfully synthesised a peak would be visible between 240 and 300bp, which was not the case in either library. The peak at 138bp in library 3506H suggests a small but quantifiable amount of primer dimer. Figure 18 shows the TapeStation traces for these attempted libraries.

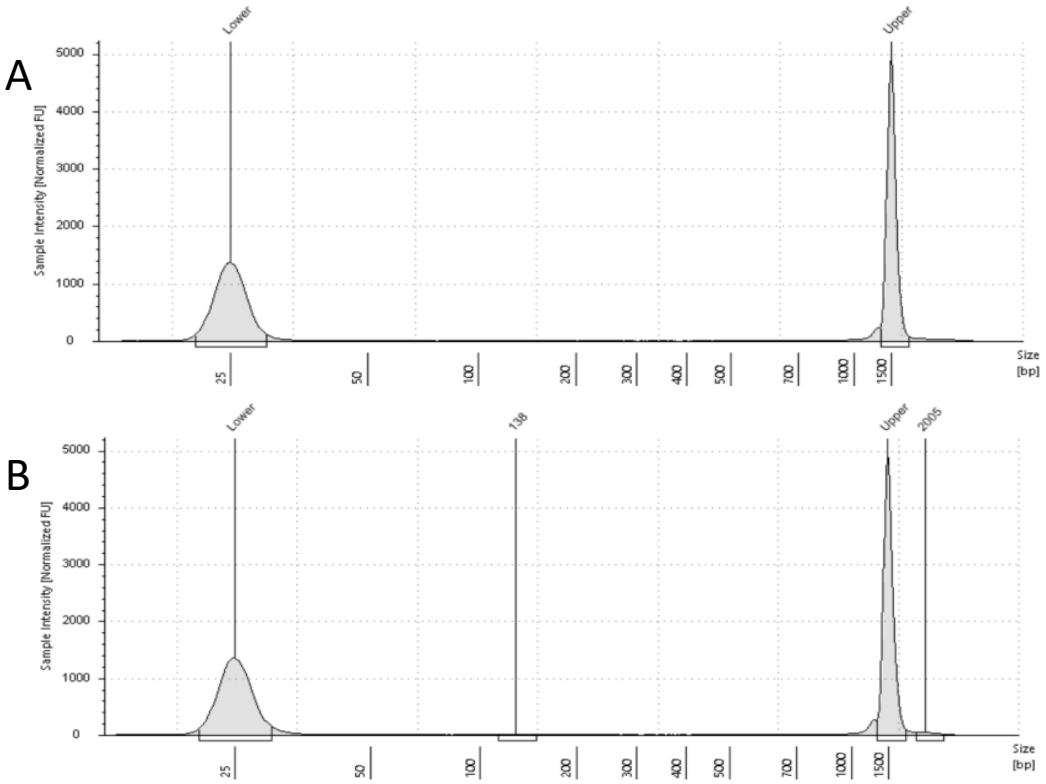
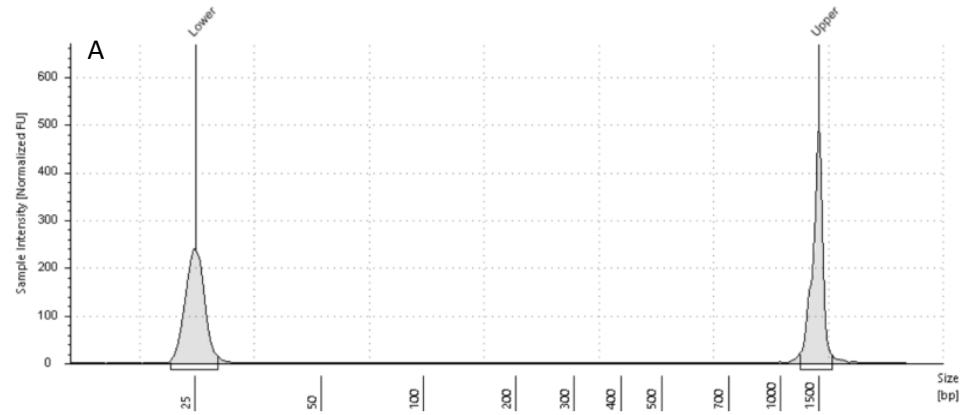


Figure 18- TapeStation traces for sample 1966.3506G and 1966.3506H attempted libraries.

A shows 1966.3506G and B shows 1966.3506H. “Upper” and “Lower” peaks show upper and lower markers used to calibrate the software. A successful library would show a peak between 200 and 300bp, which is absent in both traces. The indication of a peak at 138bp for 3506H would suggest a small amount of adaptor dimer.

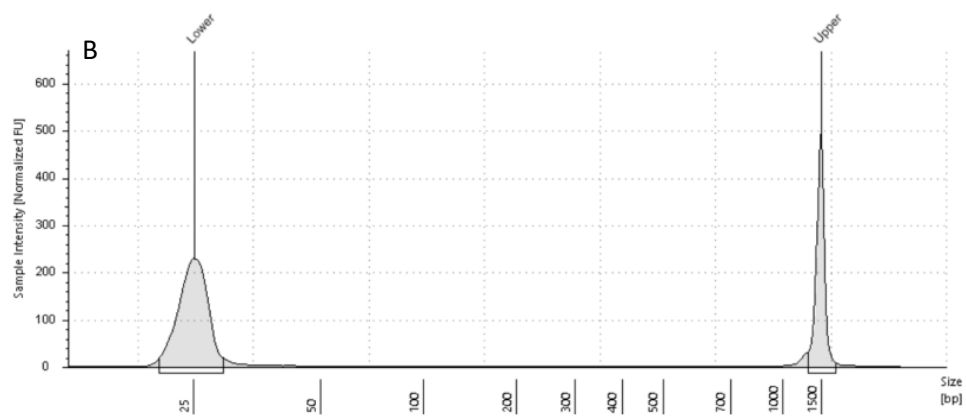
Based on these results and further GAPDH 70 investigations, it was hypothesised that due to the extensive RNA fragmentation within the samples that the fragments were too small to be undergo library preparation via conventional Illumina methods and were being filtered out by the bead clean up steps, as these filter out small fragments. Accordingly, modifications were made to the protocol by adding IPA to the bead clean up steps as recommended by Beckmann Coulter technical support. Beads and IPA were added in a 1:1 ratio, and the quantity of beads added was also increased from a ratio of 1.9x sample volume to 2.1x sample volume. Additional amplification cycles were also added to the protocol for future samples beyond this point.

Following the modified protocol described above libraries were attempted of 4 additional samples- 1979.1240H2, 1984.1655H, 1979.1229H2 and 1966.3506H2. These were the 5th collection samples that yielded both a positive RNA concentration and a RIN value. To ensure that second strand synthesis was producing the complementary DNA strand as expected, these 4 samples were analysed on the TapeStation after the second strand synthesis was complete. As this step is performed prior to PCR amplification and concentrations were expected to be low but quantifiable, these were considered to be successful enough to continue to the end of the library preparation if either a visible trace was present, or there was a measurable sample concentration as determined by the software. Accoridngly, second strand synthesis was successful for the libraries 1979.1229H2 and 1966.3506H2, as a concentration of 0.547 pg/µl was called by the software for library 1979.1229H2 and a visible trace was observable for library 1996.3506H2 (Figure 19). These libraries proceeded to completion. The second strand synthesis reactions for 1979.1240H2 and 1984.1655H were unsuccessful as no trace or concentration was available for either library (Figure 19), so these libraries were not completed to conserve reagents.



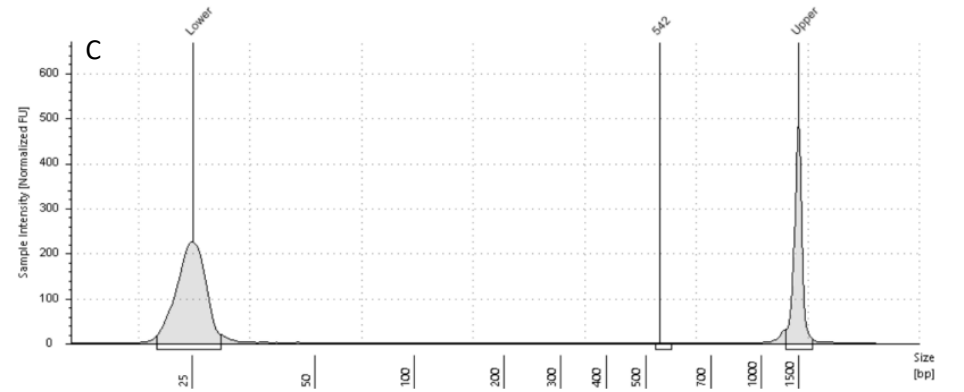
Sample Table

Well	Conc. [pg/µl]	Sample Description	Observations
A1		1240	



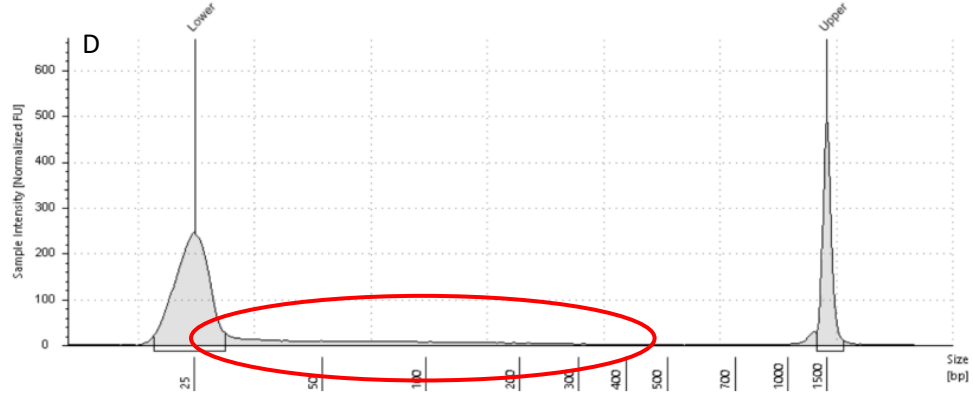
Sample Table

Well	Conc. [pg/µl]	Sample Description	Observations
B1		1655	



Sample Table

Well	Conc. [pg/µl]	Sample Description	Observations
C1	0.547	1229	



Sample Table

Well	Conc. [pg/µl]	Sample Description	Observations
D1		3506	

Figure 19- Second strand synthesis TapeStation traces.

A= 1979.1240H2, B= 1984.1655H, C= 1979.1229H2, D=1966.3506H2. Upper" and "Lower" peaks show upper and lower markers used to calibrate the software. Red circles highlight regions where the criteria for successful second strand synthesis are met i.e either a visible trace of any concentration is present or a concentration as called by the software. Concentration for C= 0.547 pg/µl. Libraries A and B were unsuccessful, whereas libraries C and D successfully underwent second strand synthesis.

The libraries for 1979.1229H2 and 1966.3506H2 were completed as described above and underwent quality control analysis on the TapeStation (Figure 20). Library synthesis for 1966.3506H2 was unsuccessful and the TapeStation trace gave no indication of a library, whereas 1979.1229H2 gave a small peak at 204bp, suggesting successful generation of a library containing approximately 70-75bp inserts. However, a significantly larger peak was observed at 133bp in the TapeStation trace for 1979.1229H2, suggesting a large amount of adaptor dimer contamination which was approximately 12-15x greater than the target library. This rendered this library unsuitable for sequencing as only 6-8% of sequenced reads would represent target sequence rather than adaptor dimer.

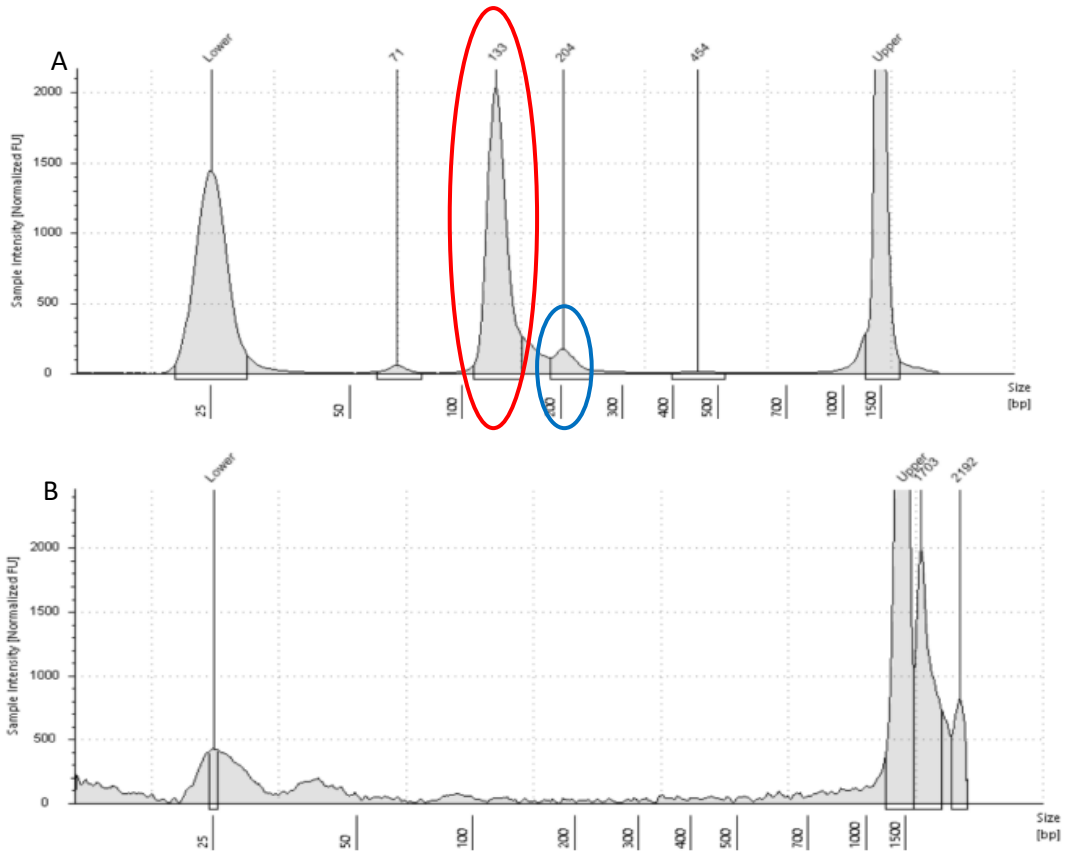


Figure 20- Post-library preparation TapeStation traces for 1979.1229H2 (A) and 1966.3506H2 (B).
"Upper" and "Lower" peaks show upper and lower markers used to calibrate the software. Red circle in A highlights significant adaptor dimer at 133bp, blue circle represents ~75bp insert library shown by a peak at 204bp. Peak at 71bp in A represents single adaptor contamination. Absence of peaks at approximately 200-300bp in B suggests a lack of successful library generation. Low level jagged trace represents background noise from the library preparation process.

Following the partial success of library 1979.1229H2 attempts were made to reduce adaptor contamination by increasing adaptor dilution. 4 more libraries were attempted using batch 6 samples- 66.3498, 68.963, 59.205 and 76.1540. These were chosen as they all gave acceptable RIN values and extraction concentrations, all except for 59.205 were GAPDH 70 positive, and these samples represented 3 different target species across a minimum of 20 years of age. Of these samples 66.3498 and 59.205 failed to generate any form of library, as shown by the absence of any peak at 200-300bp. Sample 68.963 generated an approximately 75bp library with a peak at 200bp but with approximately 6.5x greater adaptor dimer at 128bp, whilst sample 76.1540 also generated an approximately 75bp library peak at 200bp with 7.5x greater adaptor dimer at 125bp. These are still unsuitable for sequencing as only approximately 13.3% and 11.8% respectively of reads represent the target library, but this is still more successful than previous attempts and closer to an analysable library. TapeStation traces for these samples are shown in Figure 21. No further attempts at library preparations were performed due to the limited quantity of available sample and the significant cost of reagents.

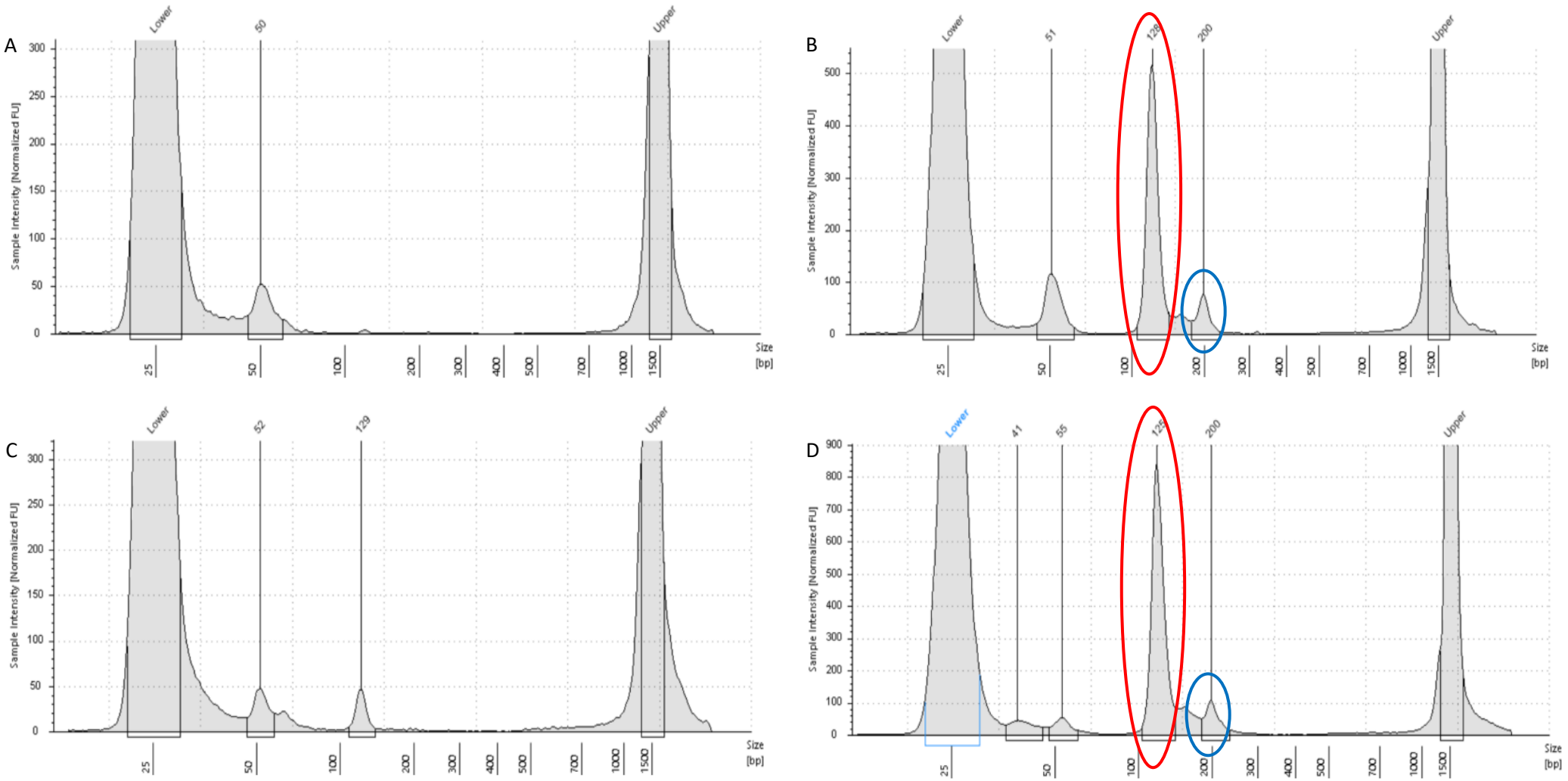


Figure 21- Post-library preparation TapeStation traces for 66.3498 (A), 68.963 (B), 59.205 (C) and 76.1540 (D).
"Upper" and "Lower" peaks show upper and lower markers used to calibrate the software. Red circle represents adaptor dimer, blue circle represents ~75bp insert library. Library A showed no peak at 200-300bp, therefore no library was synthesised. Library B showed a peak at 200bp representing successful synthesis of a library, albeit at low concentration. The large peak at 128bp in library B represents substantial adaptor dimer contamination. Library C showed no peak at 200-300bp, therefore no library was synthesised. A small peak was visible at 128bp, representing a small amount of adaptor dimer contamination. Library D showed a peak at 200bp representing successful synthesis of a library, albeit at low concentration. The large peak at 128bp in library B represents substantial adaptor dimer contamination.

4. Coronavirus presence in historic samples shown by PCR

qPCR screening for CoVs was performed using the NHM-CoV-8969_F and NHM-CoV-9080_R primers. A total of 22 samples were tested including all of the extraction batch 5 samples and all of the batch 6 samples that were GAPDH 70 positive. These primers target an 111bp fragment, therefore a GAPDH 70 positive result suggests sufficiently large fragments for the PCR to be successful may be present. Due to the historic nature of these samples any positive Ct value was considered to be a potential positive result- in essence, the Ct threshold was ommitted to allow for maximum sensitivity, albeit at an increased risk of false-negative results. Any samples with a positive Ct value were sent for sequencing. 6 samples were potentially positive, which were 1966.3503H2, 1966.3506H2, 68.970, 84.800, 76.1540 and 68.962, with Ct values of 34, 39, 39, 36, 38 and > 40, respectively (Table 20). 2x negative controls, an OC43 CoV positive control and an NL63 CoV positive control were also tested and gave results as expected (Ct values of 0, 0, 33 and 38, respectively).

Table 20- Historic sample CoV screening.

Screened using the NHM-CoV-8969_F and 9080_R qPCR results. - represents a negative result. Results shown in red produced a PCR product but were not identified as CoVs by sequencing. Results shown in black produced a PCR product which were identified as CoVs by sequencing.

Sample	CoV PCR Ct	Sample	CoV PCR Ct
1966.3503H2	34	68.971	-
1966.3506H2	39	84.800	36
1880.7.21.1H	-	79.424	-
1880.7.21.4H2	-	76.1540	38
1979.1229H2	-	1988.167	-
1979.1240H2	-	79.1537	-
66.3498	-	77.3425	-
68.963	-	68.962	> 40
66.3509	-	68.357	-
68.970	39	61.1642	-
64.59	-	77.3427	-

All positive samples except for 68.970 were successfully sequenced, and of these 1966.3506H2, 68.962 and 84.800 were identified as most similar to CoVs by NTBLAST. Due to the small product size produced by the primers, after the trimming of the primer sequences and low quality bases the remaining sequence was quite short (81bp, 64bp and 80bp for 1966.3506H2, 68.962 and 84.800 respectively). For all positives, NTBLAST identified 2 human coronavirus OC43 matches (accession numbers OR266949 and OQ828657) and 18 canine respiratory coronavirus matches (top two accession numbers OQ621727 and OQ621726) as the best matches, with equal quality E values of 4.54^{-20} for all 20 matches due to all matches having identical sequences across the target region. This is likely because of the highly conserved nature of the polymerase region of ORF1ab gene that the primers target rather than the presence of either a human or canine coronavirus in bat samples[337]. 1966.3506H2 was 94.44% similar to these hits, 68.962 was 93.10% similar, and 84.800 was 94.60% similar. Across the 3 samples the majority of the sequence is similar, excluding at position 65 as shown in Figure 22 A-C where the isolate from sample 68.962 had a cytosine base and the other two samples had an adenine base. There may be more differences towards the 3' end of the read, but due to poor read quality it is difficult to be certain of these. All three samples were then aligned by translation relative to bases 18500-18700 of one of the canine CoV matches identified by NTBLAST (OQ621727) and bases 18600-18800 of one of the human CoV matches identified by NTBLAST (OR266949) (Figure 22 D). Across this region the two reference sequences were 100% identical to each other. Excluding the extreme 3' and 5' ends where the sequence quality was lower and was therefore less reliable, all 3 historic CoV samples were identical to the reference sequences and each other except for at positions 80, 83 and 86 relative to the alignment shown in Figure 22 D. A distance matrix for the similarity between these samples is shown in Table 21, although these values may be unreliable due to the low read quality at the 3' end of all samples. Phylogenetic analysis was not performed on these samples due to the short and highly conserved segments observed. Attempts to confirm the hits with other genomic regions were not made due to a lack of effective primers and time constraints.

Table 21- Similarity analysis of the historic CoVs detected.
Similarity matrix between the 3 historic CoV sequences, and 2 of their closest matches (OR266949, a human coronavirus OC43, and OQ621727, a canine respiratory coronavirus). The two reference sequences were 100% identical across this region, and all historic CoVs are 85-96.15% similar to each other and reference sequences across this region. For this historic CoVs, similarities are potentially underestimated due to some low quality chromatogram at the 3' and 5' ends of the sequence.

	1966.3506	68.962	84.800	OQ621727	OR266949
1966.3506		88.33%	96.15%	85.9%	85.9%
68.962	88.33%		88.33%	85%	85%
84.800	96.154%	88.33%		85.9%	85.9%
OQ621727	85.897%	85%	85.9%		100%
OR266949	85.897%	85%	85.9%	100%	

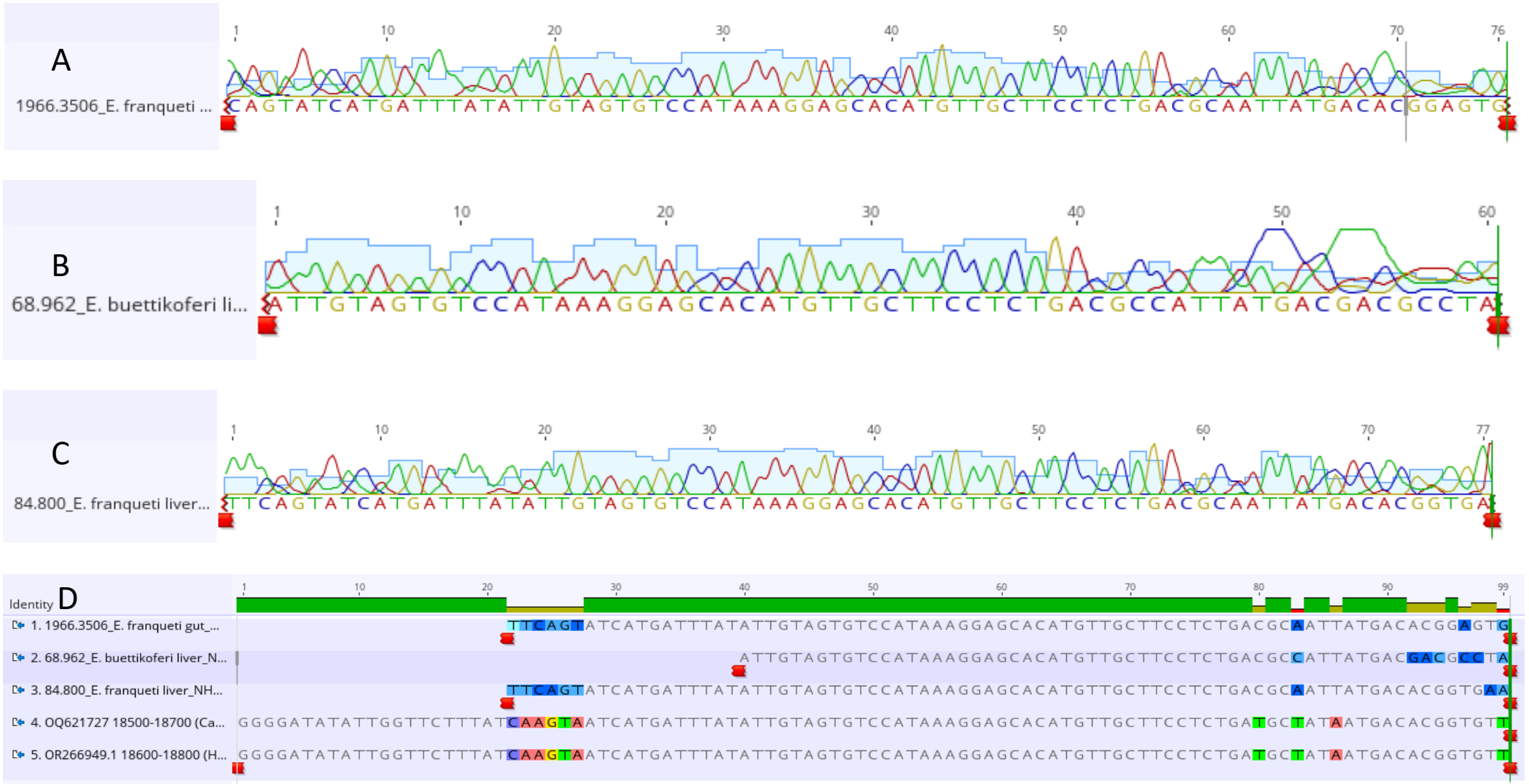


Figure 22- Analysis of the historic CoV hits and canine and human reference sequences.
In all sequences and the alignment, a red shape represents where bases have been trimmed, either to remove low quality sequence or to show the alignment of the same region. A shows the chromatogram and sequence for the historic CoV isolated from 1966.3506, B shows the chromatogram and sequence for the historic CoV isolated from 68.962, C shows the chromatogram and sequence for the historic CoV isolated from 84.800. D shows the alignment of the 3 historic CoV isolates to a human CoV isolate (OR266949) and a canine CoV isolate (OQ621727), two of the closest matches reported by NTBLAST. Identity bar is dark green for regions of ≥ 70% similarity, light green/yellow for regions of 70-30% similarity and red for regions of ≤ 30% similarity. Reference sequences were 100% identical to one another across the aligned region, and excluding the 3' and 5' ends where the chromatogram were of lower quality only differed from historic CoV sequences at bases 80, 83 and 86.

LASV screening qPCR was also performed on all historic rodent samples which had a RIN greater than 0 and/or a positive GAPDH 70 PCR result. 2x negative controls and 2 positive control reactions at different concentrations using a synthetic positive control were also included in all reactions. The positive controls worked as expected and the negative controls showed low level non-specific binding (Ct values of 40 and 40). Of the 12 samples tested, 5 produced Ct values suggesting a potential positive, but all 5 were unable to be validated by sequencing (Table 22). Whilst it is possible that some of these results may be true positives, the non-specific binding in the negative controls and their similar Ct values to the potentially positive sample results renders this unlikely. Due to time constraints further LASV PCR investigations were not performed using alternative primer sets.

Table 22- Historic samples LASV qPCR.

- represents a negative result by qPCR. Ct values are shown for PCR positive samples, although no positives found here were able to be validated by sequencing.

Sample	LASV PCR Ct	Sample	LASV PCR Ct
1932.1.17.15	38	77.3425	37
1979.1229G	-	77.3427	39
1979.1229H	-	48.1163	-
1979.1240G	-	79.1181	-
1979.1240H	-	NC1	40
76.1540	39	NC2	40
1988.167	-	Positive control 10^{-10}	34
79.1537	39	Positive control 10^{-12}	40

5. Discussion

One of the main concerns of all ancient and historic molecular work is contamination with modern nucleic acids[81]. All reasonable efforts were taken to avoid this including working in a separate laboratory to all other projects, deep cleaning of all areas and equipment both before and after experiments, using separate equipment wherever possible and separate lab coats, and using new and sterile reagents for historic molecular work and then ensuring that those reagents are only used for this work. Despite these efforts it is not 100% possible to rule out molecular contamination throughout this process, as is the case in all historic and ancient genomics work[82]. Future work on this project should be performed in an historic sample paleo-molecular lab such as the recently opened aeDNA (ancient and environmental DNA) facility at the University of Nottingham.

Specimen species were chosen according to two main criteria. The most important was that the specimens or their phylogenetically similar relatives were members of species that are known reservoirs of important viruses. For example, *E. franqueti* has been shown to be an important reservoir for CoVs, and *M. natalensis* is the major reservoir species for LASV, two viruses of importance to both wildlife and human health[45,159]. The second criterion was that there was a reasonable number of the species available in the NHM archives, so that sampling these species will not cause significant damage to the overall collection. Additionally, pregnant specimens or infant specimens were excluded, both due to difficulty with sampling and due to their potentially great value for other projects.

Despite the care taken with specimen selection, the very nature of the samples caused some issues regarding tissue collection- this is a common issue with historic genomics projects[77]. The first and most important issue was that it was impossible to accurately predict the quality of a specimen and its tissue prior to sampling it. Some specimens were relatively easy to sample, whilst others had highly friable tissue that was difficult or impossible to extract which led to failed RNA extractions for some samples. This is likely due to a variety of factors, including the actual age of the sample, the time between the death of the animal and preservation, the method of initial preservation of the animal, the transport conditions to the NHM archives, and whether the sample had ever dried out within the jar[77,83,87]. Similarly, some animals will have been formalin fixed and others will not have due to their age predating formalin usage, and as it is impossible to know which are which, all samples were treated as if formalin fixation had occured, which is likely not optimal for some of these specimens[84,338]. Indeed, formalin fixation status is not recorded for any of these samples, and it is therefore unknown whether many of the specimens were or were not fixed. It is believed that the mass manufacturing of formalin was first performed in 1891, and therefore this work assumes that mass adopton of formalin fixation followed shortly after this[309]. Accordingly, samples collected prior to 1891 are considered as likely unfixed, and those after 1891 are considered likely to have been formalin fixed. Without further metadata it is impossible to draw definitive

conclusions between any of these factors and sample quality. Another factor was the ease of sampling- for example, in larger animals such as *M. natalensis* and *H. monstrosus* organ identification and sampling was significantly easier than for smaller animals such as *M. inflatus*, where their small organs made precise sampling difficult. Larger animals tended to have larger organs, which allowed for larger tissue chunks to be taken for RNA extraction, which made the RNA extraction process more straightforward, even if this did not necessarily relate to quality of RNA. Due to the unpredictability of the specimens, the best approach for obtaining high quality samples was to simply sample as many animals as possible, as only ~37% of samples gave reasonable quality samples without excessive fragmentation (indicated by GAPDH 70 PCR). This approach is common amongst aDNA work, as it is well established that sample quality assessment prior to extraction is largely impossible with current technology and methodology[77].

Sample section also affected RNA extraction success. For example, for specimen 1966.3503 the first liver tissue RNA extraction was successful and the second was unsuccessful, despite being from the same organ from the same animal. This suggests that even within the same organ there must be some variability regarding quality of RNA preservation and fixation, as has been observed within other aRNA studies[83]. The cause of this is unclear, although it is possibly due to variable penetration of the fixative or the speed of fixation. Sampling the entire organ would likely have solved this problem, but due to the historical importance of these specimens and their irreplaceability this was not an option. The somewhat empirical nature of the tissue sampling may also have affected RNA extraction success, as larger and more intact tissue is easier to process. Similarly, the tissue type will have affected the success of the extraction for any given sample. Liver tissue RNA extraction was frequently more successful and easier than gut RNA extraction due to the elastic nature of gut tissue rendering it difficult to break down adequately. Focussing on liver tissue in the later collection batches also removed the risk of incidental virus discovery from food contaminants within the gut, and thus liver tissue was the predominant tissue type investigated[23].

There was no major difference in the success rate of extractions between species, although bat extractions had an overall slightly better success rate than rodent extractions (88% vs 75.86%). However, 42% of bat samples were GAPDH 70 positive vs 27.59% of rodent samples, showing a small but appreciable difference in quality and success rate by this metric. The reason for this is not entirely clear, although as the rodents were bigger than the bats (excluding the *H. monstrosus* specimens) it is possible that the fixation time was longer due to slower fixative penetration, possibly resulting in increased genomic fragmentation. There was no clear link between sample age and extraction success or GAPDH 70 success. This was surprising, as it stands to reason that older samples would have undergone more degradation prior to extraction, yet this was not observed. The minimum age of the specimens tested here is 35 years, suggesting that RNA degradation may occur over time and may have

plateaued by this point. To test this more samples with ages between 6 months and 35 years old could undergo RNA extraction to see if there is a link between age and RNA degradation across this timeframe. However, there was a statistically significant link between RIN and GAPDH 70 success, suggesting that GAPDH 70 PCR can act as an effective proxy indicator of genomic fragmentation and quality. Whilst this is unlikely to have a significant impact on the downstream analysis of these samples, this suggests that only one of these quality control analyses are necessary, and therefore sample RNA could be conserved and cost per sample could be reduced by omitting RIN analysis or GAPDH 70 assessment.

Whilst the Nanodrop analysis was useful for a baseline indication of the success of extraction, the purity values are unreliable when the concentration of the nucleic acid is < 20ng/µl as is the case for approximately half of the samples here[339]. Using a Qubit system instead of a Nanodrop may have increased quantification accuracy, although it is often considered sensible to assess purity using a Nanodrop and concentration using a qubit which may not have been an option here due to sample scarcity[339,340]. Despite these drawbacks, for a given sample the Nanodrop concentration and TapeStation concentration tended to be in broad agreement (Figure 16 A). The fact that neither Nanodrop nor TapeStation concentration was statistically significantly linked to GAPDH 70 success but that a higher RIN was significantly associated with GAPDH 70 success suggests that the RNA degradation and fragmentation is likely the key factor dictating extraction success. The GAPDH results also showed that the fragment size is typically above 70bp but below 149bp for most samples. This is in broad agreement with other studies, such as that performed by Dux and colleagues who found their average RNA fragment size to be 95-136bp in a sample from 1912[78].

Library preparation was not successful for any of these samples. For some samples such as 66.3498, no indication of a library was seen on the TapeStation trace even when the GAPDH 70 PCR was positive and the input concentration was high (Figure 20). The reasons for this are unclear, although one possibility is that whilst the RNA is not too fragmented to produce a library (as indicated by the GAPDH 70 result) it is damaged in some other way. Historic DNA and RNA is known to undergo significant damage in a variety of ways, including the formation of abasic sites and atypical nucleotides, as well as misincorporation of bases such as uracil bases as a consequence of cytosine deamination[85,89]. This damage can lead to reduced interaction with the enzymatic steps of library preparation process, in turn preventing or reducing the effectiveness of library preparation, and whilst aDNA can be repaired via a variety of methods there are currently no published methodologies or available protocols for aRNA repair[77,89]. Whilst some damage of this nature is inevitable and is indeed useful for assessing that the sampled RNA is in fact historic in origin rather than a modern contaminant, to assess this software such as mapDamage 2.0 is used, which requires NGS data available to analyse, which in turn requires a successful library generation[89]. It should be noted that it is difficult to

distinguish genuine damage from random chance nucleotide misincorporations, even with this or equivalent software[83]. Another option is that despite efforts to enhance the capture of small fragments, they were simply filtered out as part of the library preparation process- a known issue with the NEB Ultra II library DNA preparation kit which may also occur in the RNA kit[77]. Other library preparation approaches such as the Santa Cruz reaction were considered but were not used as these tended to increase the amount of adaptor dimer present, which was already posing a significant problem[77].

Adaptor dimerisation is common during library preparation and occurs to some extent in all libraries. Adaptor dimers reduce the read count when the library is sequenced as the adaptor dimer is recognised as a target read by the software, diluting the quantity of target reads accordingly[341]. For samples which synthesised a library, there was always a significant excess of adaptor dimers ranging from 6.5 fold greater than the library to 15 fold greater, in turn diluting the reads available for the actual library to down approximately 13-6%. Accordingly, in a 100 million read library, only between 6 and 13 million reads would represent the sample rather than the adaptor dimer. By the time low quality reads, host reads, genomic reads, bacterial and fungal reads and other contaminants have been removed, this will result in very few reads remaining in which to detect viruses. As historic viruses are likely present at low copy numbers due to preservation process and nucleic acid degradation over time, the limited sensitivity of NGS may then result in failing to identify viruses that are present in the sample[65]. Substantial adaptor dimer contamination is unsurprising, as adaptor dimers are more likely to form when the initial RNA input is low which was the case for many of these libraries[341]. Similarly, whilst it was necessary to increase the amplification cycles within the library preparation beyond the manufacturer recommendations due to the low input concentrations this is also likely to have increased the amount of adaptor dimer, as the adaptor is the primer binding target within these steps regardless of whether attached to the target DNA or having formed a dimer[84]. It was not economically feasible to try and sequence libraries with such a significant adaptor dimer presence. Whilst it is theoretically possible to simply "overpower" the dimers with enough reads per sample, this would have required a full NovaSeq S4 lane and $3x10^9$ reads per library (or possibly a pool of two libraries), and at a cost of approximately £5000 per lane this was prohibitively expensive. This would also then require significant filtering to remove the adaptor dimer contamination, which in turn requires significant computing power, likely requiring the use of the HPC which also comes at a cost and complicates later analysis. Accordingly, even the most successful libraries synthesised here were not sequenced.

Various approaches were considered to overcome the adaptor dimer contamination which were not pursued due to time constraints. The first was to design custom adaptors that included a rare restriction digest site. The protocol would have been adapted to include a restriction digest step at the very end, before undertaking the

entire protocol from the adaptor ligation onwards a second time as the library fragments would also have lost their adaptors to degradation. This would in theory result in significantly more target sequences being present due to the previous round of amplification, resulting in a better ratio of target to adaptor. This approach was not pursued as custom adaptors are expensive and repeating the entire second half of the library preparation process for each sample is both expensive and inefficient. There were also concerns about the restriction digestion causing the loss of target DNA, although using a restriction enzyme with a 6 base recognition site would result in the cutting of approximately 1 in every 4096 fragments, or 0.024% of fragments, resulting in relatively negligible loss of target sequences, or of approximately 1 in every 65,536 fragments (0.0015%) using an enzyme with an 8 base recognition site. Finally, there were concerns regarding the potential bias after an extra amplification step, where more well represented fragments would be disproportionately amplified and may lead to a loss of less well represented fragments. However, it was accepted that in principle this would have been an acceptable risk in order to at least identify well represented viruses in the sample.

Another approach considered to minimise adaptor-dimer contamination was using custom adaptors containing uracil bases and then using uracil-N-glycosylase to remove the adaptors after amplification[342]. However, the standard adaptors already contain uracil which is cleaved as part of the standard process for strand separation, therefore adding uracils would lead to premature cleavage loss of adaptors. Gel extraction approaches were also considered, but due to the difficulty of extracting such small fragments from a gel and the risk of contamination from nucleic acids remaining in the gel tank this was not pursued further. Alternative adaptor schemes are available but tend to be targeted towards miRNA libraries rather than aRNA libraries, limiting their effectiveness for this project[341].

Whilst the metagenomics approach was being optimised and developed, qPCR screening of the samples was considered to be a useful approach for rapid historic virus investigation. The CoV PCR produced 3 positive results from 22 samples (13.6%). This is broadly in line with values reported in other papers, where positivity rates in modern African bats (including *E. franqueti*) were approximately 8.8%, supporting that these are likely to be true positive results[159]. Similarly, all negative controls were negative, suggesting that there was no contamination within the PCR, and all extractions were carried out under sterile conditions, suggesting that contamination did not enter the reaction at this point. However, as it was not possible to perform a confirmatory PCR at another point in the genome and it is impossible to rule out contamination with 100% certainty, there is still a small possibility that these results are false positives. Regardless, finding CoVs in these samples is reasonable considering the geographical profile and virological profile of the animals sampled, as CoVs have previously been found in African bats including *E. franqueti*[159].

Because of the highly conserved nature of the region targeted by the PCR and the small product size, the fact that the NTBLAST report identified the samples as human

and canine CoVs does not necessarily indicate contamination. This is supported by the alignment shown in Figure 22 D, where the reference sequences are identical, illustrating how highly conserved this region is. The fact that the two reference sequences were 100% identical across this region and the 3 historic CoV sequences were not suggests that evolution has occurred from the sequence seen in the historic CoVs to the sequence seen in the modern reference sequences across this highly conserved region. This evidence supports that the CoV sequences found here are indeed historic in origin. More genomic information is required to distinguish this further and to allow for species examination and phylogenetic analysis, as well as for definitively confirming the historic origin of these viruses. Whilst it would have been possible to clone the viruses into bacterial strains and then to sequence the clones to increase the read quality across the 3' and 5' regions, this would only have provided a small amount of further sequence which would have been similar to the designed primers and therefore unreliable, therefore this was not performed. Ideally, successful libraries would have been produced and sequenced for these samples, as these either would have provided more genomic information and thus confirmed the hits or had no CoV reads and refuted the hits. This would also have allowed mapDamage 2.0 to have been used on these samples to confirm that the hits were historic in origin[89]. CoVs have also recently been found in *Miniopterus* bats, suggesting that further screening of these could be worthwhile in the future[40].

None of the animals tested for LASV produced a positive PCR result, despite 5 animals producing a product with identifiable Ct values. This is somewhat surprising, as reported LASV prevalence in *M. natalensis* ranges from 5.9-14.5%, and 9/10 animals tested were *M. natalensis* giving a predicted 1/10 positivity rate[343]. However, this prevalence value fluctuates both by geographical location within Africa and whether the sample was collected in the wet or dry season, which is not information that is available for these samples[343]. Also, only 10 animals were tested here, which is a small enough sample size that all samples could be negative through chance alone. Testing alternative genomic locations with alternative primer sets may have increased the likelihood of identifying a positive sample, but this was not pursued due to time constraints. Additionally, as primers are designed based on modern virus genomes, there is a possibility that primers may not bind to highly divergent historic genomes even if primers are successfully tested on modern controls, limiting the effectiveness of this PCR approach[81]. This also serves to illustrate some of the limitations of analysis based on Ct values alone without confirmatory sequencing. For example, it is possible that a non-specific PCR product was produced and detected in the LASV "positive" samples, which would then give rise to a Ct value despite no actual virus being detected. Melt curve analysis would have improved the confidence of these results but was not performed in this case, and still would not have been a sufficient alternative to PCR product sequencing and analysis.

Overall, this project has made significant progress regarding developing the process of RNA extraction and library preparation for historic RNA samples, which is known

to be difficult and unreliable at best[77,84]. It should be noted that the protocol used here and the protocol used by Speer and colleagues are very similar, with the only significant difference being that Speer *et al.* used a bead beater to enhance tissue breakdown, whereas a bead beater was not used in this project[84]. It is possible that enhancing sample breakdown via bead beating may have allowed for the release of more RNA and RNA of better quality. Unfortunately, the RNA extractions for this project were performed before the publication of the study by Speer and colleagues, and therefore their methods could not be adopted here[84]. There are still issues to be addressed, primarily regarding adaptor dimer contamination within library preparations, although this is a well-known issue in ancient and historic library preparations by a variety of techniques, including more sensitive preparation methods such as the Santa Cruz reaction where adaptor dimers can form a large proportion of libraries[77]. Regardless, this project has generated a large collection of RNA samples to be tested once these issues are resolved. This project has also found evidence of CoVs in three historic bat samples illustrating both that virus discovery is plausible within archival samples and that key virus species can be detected, for which there is very little published evidence of to date[81]. This may later allow for more thorough genomic investigations and potentially enhanced phylogenetic and evolutionary biology studies. Work on this project is still ongoing within the laboratory group, with a 7th collection of approximately 125 samples having been performed in August 2023.

Chapter 7- Final conclusions and further work

The overall aim of this project was to take a three-pronged approach to virus discovery, by utilising 3 different approaches to expand our collective knowledge of the virome[5]. The first prong entailed using degenerate PCR screening of UK rodents to identify potentially novel viruses. The second prong involved using metagenomics to perform an unbiased investigation into the viruses within UK wildlife and their abundance. The third prong involved using a combination of PCR based and metagenomics investigations into historic viruses in an effort to provide evolutionary information on key virus families.

The first two prongs involved testing 140 UK rodents of 5 species, caught in a previously uninvestigated region of Wales. Before any investigations into the viral profile animals could be performed the RNA extraction and cDNA synthesis for these samples was required. This project used an established methodology for this, yielding high quality nucleic acid samples from all animals. This provided sufficient samples to perform reliable virus investigations in the first and second prongs.

The degenerate PCR screening approach was met with limited success in this project. Only 5 viruses were considered for due to time constraints, and of these, both sets of *Rotavirus* primers were unable to be validated, therefore only 4 viruses were successfully screened for- adenoviruses, hantaviruses, coronaviruses and *Rubiviruses*. The adenovirus screening was the most successful and adenoviruses were found in 1 bank vole and 1 wood mouse, although only a small section of the genome was recovered. Whilst these hits were validated by Sanger sequencing, they were not detected in the metagenomics data for these animals, although that may be due to issues regarding the sensitivity of NGS and is therefore unsurprising[65]. What is more surprising is that adenoviruses were detected in the NGS data and confirmatory screening PCR in 28 samples that were not detected by degenerate PCR. Whilst the exact reasons for this are unclear, it is likely due to a primer design issue, potentially that the primers were too degenerate and therefore would not effectively bind to many adenovirus species. This was likely an issue throughout the degenerate primer design process for all viruses.

The degenerate PCR did yield a positive hantavirus result in 1 field vole, which was then confirmed by the NGS data. However, the degenerate hantavirus primers resulted in a false negative result for another animal in which a near full TATV genome was recovered by the NGS, likely due to the same primer design issues as for the adenovirus primers above. As this appears to be consistent across both primer sets, this then calls into question the validity of the entirely negative CoV and *Rubivirus* screening results, as if there are systemic issues with primer design throughout this process then it is possible that these primers are also unreliable potentially leading to false negative results. However, the CoV primers and the *Rubivirus* primers tested here were effective in validation tests, and the *Rubivirus* negative result in particular is unsurprising, suggesting that these may be accurate results[301].

Whilst the aims of this prong were not met- i.e. no novel viruses were discovered, likely due to excessive primer degeneracy- valuable lessons were learned from this approach. Significantly more care was taken when designing primers for the other two facets of this project, resulting in better primers and more success in those approaches. The degenerate screening approach was also reasonable in principle, and by reworking the primers used and improving their design by removing some degeneracy whilst still maintaining enough degeneracy to detect novel primers, an effective virus discovery screening panel could be developed. Such a panel could also be expanded in future to include other viruses of interest.

The second approach taken here involved using NGS libraries to perform unbiased, high throughput screening of the entire liver and gut virome for the same samples. Whilst it is true that using other tissue such as spleen, lung or kidney may have improved the success of viruses with strong tissue tropism for other tissues, these tissues were unavailable for this project and this screening approach was still highly successful, allowing for the identification of many viruses. One virus identified via this approach was a near full genome of a TATV virus. This has a two-fold impact, in that it expands the known range of TATV within Great Britain to include Wales as a minimum and adds to the limited amount of TATV sequences available by identifying the sequence of an entire TATV M and S segment and most of a TATV L segment[12]. Sufficient genomic coverage for phylogenetic analysis was also recovered from a variety of other viruses found within these samples, including *Hepacivirus F*, astrovirus, *Orbivirus*, two genera of picornavirus and *Picobirnavirus*. Whilst none of these viruses were found to be novel species by phylogenetic analysis, many of these represent an expansion to the knowledge base for that virus in other ways. For example, this is the first time that *Hepacivirus F, Rosavirus* or *Orbivirus* viral hits have ever been reported within Great Britain, representing an expansion to the range of these viruses[105,199,252]. Similarly, this is the first time (to our knowledge) that a *Picobirnavirus* has been identified in a field vole, that a *Pegivirus* or *Rosavirus* has been identified in a bank vole, that a *Picobirnavirus* or a polyomavirus has been identified in a wood-mouse, or that a *Picobirnavirus* or a *Bocaparvovirus* has been identified in a yellow-necked mouse[41,224,234,252,254,281]. Each of these findings represents an expansion of the host range for the virus in question, and whilst the impact of any individual finding of these is somewhat minor, collectively they represent a major increase in the knowledge of the virome and potential host species within the UK. Further work is required to provide more information on these viruses and to characterise their genome fully, as due to the geographically isolated nature of the UK it is likely that at least one of these viruses will prove to be a novel species or genus upon full genome phylogenetic analysis.

The second prong approach also allowed for the estimation of the abundance of 19 species or genera of virus within the poorly sampled UK. Whilst some of these proportion positivity values may not be entirely accurate- for example, issues regarding endogenous vs exogenous viruses for MLV (or potentially MLV-like viruses-

insufficient sequence was recovered to distinguish between these) and difficulty validating some *Hepacivius F* hits may artificially inflate these values, although for the most part these are accurate proportion positivity estimates and often the first for the virus in question within the UK[192]. Whilst it is difficult to extrapolate these results across the entire UK or to apply them to other rodents within the UK, these estimates allow for an increase in the knowledge of the virome of UK rodents. Some of the findings also could have potential health or agricultural implications, such as the relatively high abundance of adenoviruses and astroviruses that could theoretically evolve to infect humans, and the high proportion positivity of arteriviruses within field voles. Further work for this approach has many applications, including sampling more animals to increase the sample size and therefore the accuracy of the proportion positivity estimates and to work towards gaining prevalence estimates, repeating this approach on another area of the UK to compare the proportion positivity and spatial virome dynamics between the two sites, and simply as a baseline for deciding which viruses are likely to be present within UK rodents for future wildlife screening projects. For example, performing similar studies at different times of the year may allow for a temporal comparison of the virome with this study, and this may also act to provide indications of which virus species may be found and in which host animals, whilst also providing a reasonable sample size to compare against.

The final prong involved investigating viruses within historic bat and rodent samples in an effort discover and characterise ancestral viruses. Whilst this goal was not successful, significant advances in the field of historic RNA virus discovery and methodology were made throughout this project. For example, by altering existing protocols and refining the methodology, a method for historic RNA extraction from mammalian samples was developed which is now over 90% reproducible in the latest batch of specimens and is highly successful considering the potentially degraded state of ancient and historic specimens[77,83]. Whilst this project was unable to successfully optimise a protocol for the production of a NGS library from ancient or historic samples, it was able to advance this field and make significant progress towards doing so. By modifying and adapting existing library preparation protocols to be compatible with highly fragmented and degraded genomes it was possible to demonstrate NGS library synthesis from 3 different historic specimens, albeit with significant adaptor dimer contamination. Whilst this was not an option for this project, similar libraries produced in research groups with more funding would be able to “overpower” the adaptor dimer contamination with sufficient sequencing power, and in turn be able to perform phylogenetic analysis on libraries of this quality on historic samples. Whilst future work should and will focus on refining the historic library preparation process, even at the current level of optimisation the methods developed here could lead to effective historic or ancient RNA virus discovery if widely adopted.

The work on historic samples also was able to identify the presence of historic CoVs in 3 different specimens by qPCR. Whilst the genome sections identified were extremely small, they were similar enough to be identified as CoVs by NTBLAST, but different enough across a highly conserved region to support that they are truly historic viruses that were present in the preserved animal at the time of preservation and not contamination from the laboratory process. To our knowledge, this is the first time that historic CoVs have ever been identified in archival samples, let alone by qPCR. Whilst this not only serves as effective proof of concept for historic virus discovery via conventional methods, this also allows for the expansion of the screening programme using this approach to include other important viruses of animals[38]. Whilst considerations must be made regarding sample scarcity and irreplaceability vs diversity and importance of viruses screened for, these results indicate that accurate virus discovery in historic samples is plausible and may eventually lead to significant evolutionary discoveries. It must also be noted that this project primarily screened for RNA viruses due to many viruses of potential human and livestock importance being RNA viruses, their tendency to evolve rapidly, and as screening for both DNA and RNA viruses would have doubled both input requirement from irreplacable samples and cost[17].

Overall, the 3 pronged approach to virus discovery was successful. As well as characterising a small part of the UK virome, this study identified viruses that are novel within their hosts and the UK, provided accurate proportion positivity values for many viruses within the UK and made significant advancements in the field of historic virus discovery. Whilst all of these results findings are potentially significant in their own right, a key lesson from this project is that virus discovery is most successful when a diverse variety of approaches and methods are utilised, and that metagenomics, PCR and historic virus work all have key roles to play in characterising the virome going forwards.

Professional internship reflective statement

For my placement, I worked with the research and specialist molecular development team within the Microbiology department at Queens Medical Centre, Nottingham, as a research assistant. My project was to develop an in-house 16S rRNA sequencing assay for use within the department. This assay was designed to be used on both fluid and tissue samples, but not bone samples.

Although I was not able to fully complete the project, my work has advanced the development of the 16S rRNA assay significantly, and it is now in late-stage development. Hopefully, another member of the department will be able to use the data and documentation that I generated to complete the project in the near future, as this project should reduce patient test result turn-around times and help them to secure a positive outcome once completed.

The first step of this project was to extract nucleic acids from samples. My role in this was to compare two extraction process (automated mechanical extraction using a Promega Maxwell RSC instrument or manual extraction using a QIAAmp DNA Mini Kit), and to refine and optimise the extraction process for high-throughput routine use in the laboratory. I found that the QIAAmp method was substantially more reliable than the automated extraction process, and that the workflow was more convenient for a busy lab.

The second step was to develop a PCR assay that would allow for diagnostic sequencing of the 16S rRNA gene. This involved identifying and testing a target range within the gene, then identifying and testing appropriate primers and cycling conditions to allow for accurate amplification of the target range. After significant development at this stage, the V1-3 region of the 16S rRNA gene was identified as the most reliable target, and a standard PCR assay was developed. Following PCR, all samples would be viewed on an agarose gel to identify positive results and any potential off-target amplification, and negative control tests would be assessed for any contamination.

The third step was to purify the sample to minimise any potential contamination, whilst also maintaining sufficient concentration for downstream sequencing. My role was to compare two purification methods- one automated (using the EXO-SAP IT enzyme system) and one manual (QIAquick Gel Extraction kit)- to identify which worked best. I found that the EXO-SAP IT system caused a significant reduction in sample concentration, which frequently made downstream sequencing impossible, and was less efficient at removing any off-target effects. Following purification, all samples were quantified using the Qubit system.

The final major step was to perform Sanger Sequencing using a SeqStudio analyser. My role was to actually perform the sequencing, including diluting the samples to the correct concentrations (where applicable), amplifying the sample, performing a bead purification and running the sequencing. I was then responsible for analysing the sequences, and running them through three databases (EZBIOCloud, SepsiTest and

RDP) that I was responsible for identifying at the start of the project to identify the bacteria present in the sample.

Whilst the basis for this pipeline was clear in the project brief, the majority of the specific process were developed throughout my placement. This pipeline was first tested on known bacterial samples stored in a pathogen bank, and was then tested on 50 patient diagnostic fluid and tissue samples. Although the pipeline took some time to develop, by the end of my placement the pipeline was broadly functioning as desired, and producing the results expected. Unfortunately, due to time constraints, insufficient samples were tested to validate and implement the pipeline- something that will hopefully be performed within the department in the near future.

My other responsibilities included general lab maintenance and stock checking, and writing change management documents and standard operating procedures for this pipeline for the rest of the team to use.

Throughout this project, I developed many skills that will be beneficial to me in the future. I learnt how to manually perform Sanger Sequencing, and re-enforced my understanding of the sequencing process. I also improved my knowledge of PCR reaction development, including exploring the concept of touchdown and nested PCRs. Additionally, this project was the first time that I've performed a gel extraction, which is a common laboratory technique that could be useful to me in the future. I also learnt how the NHS operates within a research framework, including how to handle confidentiality, quality etc. Finally, I significantly enhanced my document writing and management skills, and how to transcribe experimental results in a manner that someone else can then come along and follow up on them.

This placement was an extremely positive experience for me, and has shown me that in an ideal world my career will involve both research and diagnostics. It also re-enforced that I am happier in a lab than behind a desk, and I will take this information forwards with me into my future career.

References

1. Smith, S. E., Huang, W. Q., Tiamani, K., Unterer, M., Mirzaei, M. K., & Deng, L. (2022). Emerging technologies in the study of the virome. *Current Opinion in Virology, 54*.

2. Mushegian, A. R. (2020). Are there 10^{31} virus particles on Earth, or more, or fewer? *Journal of Bacteriology, 202*(9), 10.1128/jb.00052-00020.

3. Hendrix, R. W., Smith, M. C., Burns, R. N., Ford, M. E., & Hatfull, G. F. (1999). Evolutionary relationships among diverse bacteriophages and prophages: all the world's a phage. *Proc Natl Acad Sci U S A, 96*(5), 2192-2197.

4. Mokili, J. L., Rohwer, F., & Dutilh, B. E. (2012). Metagenomics and future perspectives in virus discovery. Curr Opin Virol, 2(1), 63-77.

5. Virgin, H. W. (2014). The virome in mammalian physiology and disease. *Cell, 157*(1), 142-150.

6. Bai, G. H., Lin, S. C., Hsu, Y. H., & Chen, S. Y. (2022). The human virome: viral metagenomics, relations with human diseases, and therapeutic applications. *Viruses-Basel, 14*(2).

7. Luis, A. D., Hayman, D. T. S., O'Shea, T. J., Cryan, P. M., Gilbert, A. T., Pulliam, J. R. C., . . . Webb, C. T. (2013). A comparison of bats and rodents as reservoirs of zoonotic viruses: are bats special? *Proceedings of the Royal Society B-Biological Sciences, 280*(1756).

8. Wu, Z. Q., Lu, L., Du, J., Yang, L., Ren, X. W., Liu, B., . . . Jin, Q. (2018). Comparative analysis of rodent and small mammal viromes to better understand the wildlife origin of emerging infectious diseases. *Microbiome, 6*.

9. Carlson, C. J., Gibb, R. J., Albery, G. F., Brierley, L., Connor, R. P., Dallas, T. A., . . . Seifert, S. N. (2022). The Global Virome in One Network (VIRION): an Atlas of Vertebrate-Virus Associations. *Mbio, 13*(2).

10. Carroll, D., Daszak, P., Wolfe, N. D., Gao, G. F., Morel, C. M., Morzaria, S., . . . Mazet, J. A. K. (2018). The Global Virome Project. *Science, 359*(6378), 872-874.

11. McElhinney, L. M., Marston, D. A., Pounder, K. C., Goharriz, H., Wise, E. L., Verner-Carlsson, J., . . . Fooks, A. R. (2017). High prevalence of Seoul hantavirus in a breeding colony of pet rats. *Epidemiology and Infection, 145*(15), 3115-3124.

12. Chappell, J. G., Tsoleridis, T., Onianwa, O., Drake, G., Ashpole, I., Dobbs, P., . . . McClure, C. P. (2020). Retrieval of the complete coding sequence of the UK-endemic Tatenale Orthohantavirus reveals extensive strain variation and supports its classification as a novel species. *Viruses-Basel, 12*(4).

13. Gravinatti, M. L., Barbosa, C. M., Soares, R. M., & Gregori, F. (2020). Synanthropic rodents as virus reservoirs and transmitters. *Revista Da Sociedade Brasileira De Medicina Tropical, 53*.

14. Bozkurt, F., Yousef, A., Baleanu, D., & Alzabut, J. (2020). A mathematical model of the evolution and spread of pathogenic coronaviruses from natural host to human host. *Chaos Solitons & Fractals, 138*.

15. Hemida, M. G., Alhammadi, M., Almathen, F., & Alnaeem, A. (2021). Exploring the potential roles of some rodents in the transmission of the Middle East respiratory syndrome coronavirus. *Journal of Medical Virology*.

16. Wolf, Y. I., Kazlauskas, D., Iranzo, J., Lucia-Sanz, A., Kuhn, J. H., Krupovic, M., . . . Koonin, E. V. (2018). Origins and evolution of the global RNA virome. *Mbio, 9*(6).

17. Duffy, S. (2018). Why are RNA virus mutation rates so damn high? *PLOS Biology, 16*(8).

18. Wang, H. W., Cui, X. Y., Cai, X. H., & An, T. Q. (2022). Recombination in positive-strand RNA viruses. *Frontiers in Microbiology, 13*.

19. Wohlgemuth, N., Honce, R., & Schultz-Cherry, S. (2019). Astrovirus evolution and emergence. *Infection, Genetics and Evolution, 69*, 30-37.

20. Simon-Loriere, E., & Holmes, E. C. (2011). Why do RNA viruses recombine? *Nature Reviews Microbiology, 9*(8), 617-626.

21. Nelson, M. I., Perofsky, A., McBride, D. S., Rambo-Martin, B. L., Wilson, M. M., Barnes, J. R., . . . Bowman, A. S. (2020). A heterogeneous swine show circuit drives zoonotic transmission of Influenza A viruses in the United States. *Journal of Virology, 94*(24).

22. Cholleti, H., de Jong, J., Blomstrom, A. L., & Berg, M. (2022). Characterization of Pipistrellus pygmaeus bat virome from Sweden. *Viruses-Basel, 14*(8).

23. Wallau, G. L., Barbier, E., Tomazatos, A., Schmidt-Chanasit, J., & Bernard, E. (2023). The virome of bats inhabiting Brazilian biomes: knowledge gaps and biases towards zoonotic viruses. *Microbiology Spectrum*.

24. Kwok, K. T. T., Nieuwenhuijse, D. F., Phan, M. V. T., & Koopmans, M. P. G. (2020). Virus metagenomics in farm animals: a systematic review. *Viruses-Basel, 12*(1).

25. Harvey, E., & Holmes, E. C. (2022). Diversity and evolution of the animal virome. *Nature Reviews Microbiology, 20*(6), 321-334.

26. Perez-Cataluna, A., Cuevas-Ferrando, E., Randazzo, W., & Sanchez, G. (2021). Bias of library preparation for virome characterization in untreated and treated wastewaters. *Science of the Total Environment, 767*.

27. Brinkmann, A., Andrusch, A., Belka, A., Wylezich, C., Hoper, D., Pohlmann, A., . . . Nitsche, A. (2019). Proficiency testing of virus diagnostics based on bioinformatics analysis of simulated In Silico high-throughput sequencing data sets. *Journal of Clinical Microbiology, 57*(8), 12.

28. Nusser, S. M., Clark, W. R., Otis, D. L., & Huang, L. (2008). Sampling considerations for disease surveillance in wildlife populations. *Journal of Wildlife Management, 72*(1), 52-60.

29. Tsoleridis, T., Onianwa, O., Horncastle, E., Dayman, E., Zhu, M. R., Danjittrong, T., . . . McClure, C. P. (2016). Discovery of novel Alphacoronaviruses in European rodents and shrews. *Viruses-Basel, 8*(3).

30. Firth, C., Bhat, M., Firth, M. A., Williams, S. H., Frye, M. J., Simmonds, P., . . . Lipkin, W. I. (2014). Detection of zoonotic pathogens and characterization of novel viruses carried by commensal Rattus norvegicus in New York City. *Mbio, 5*(5).

31. Yin, H. C., Wan, D. C., & Chen, H. Y. (2022). Metagenomic analysis of viral diversity and a novel astrovirus of forest rodent. *Virology Journal, 19*(1).

32. Olival, K. J., Hosseini, P. R., Zambrana-Torrelio, C., Ross, N., Bogich, T. L., & Daszak, P. (2017). Host and viral traits predict zoonotic spillover from mammals. *Nature, 546*(7660), 646-+.

33. Murphy, E. G., Williams, N. J., Bennett, M., Jennings, D., Chantrey, J., & McElhinney, L. M. (2019). Detection of Seoul virus in wild brown rats (Rattus norvegicus) from pig farms in Northern England. *Veterinary Record, 184*(17).

34. Rose, R., Constantinides, B., Tapinos, A., Robertson, D. L., & Prosperi, M. (2016). Challenges in the analysis of viral metagenomes. *Virus Evolution, 2*(2).

35. Mollentze, N., & Streicker, D. G. (2020). Viral zoonotic risk is homogenous among taxonomic orders of mammalian and avian reservoir hosts. *Proceedings of the National Academy of Sciences of the United States of America, 117*(17), 9423-9430.

36. Singh, B. B., Ward, M. P., Kostoulas, P., & Dhand, N. K. (2023). Zoonosis-Why we should reconsider "What's in a name? ". *Frontiers in Public Health, 11*.

37. Schulz, C., Fast, C., Wernery, U., Kinne, J., Joseph, S., Schlottau, K., . . . Beer, M. (2019). Camelids and cattle are dead-end hosts for Peste-des-Petits-Ruminants Virus. *Viruses-Basel, 11*(12).

38. Caron, A., Cappelle, J., Cumming, G. S., de Garine-Wichatitsky, M., & Gaidet, N. (2015). Bridge hosts, a missing link for disease ecology in multi-host systems. *Veterinary Research, 46*.

39. Haydon, D. T., Cleaveland, S., Taylor, L. H., & Laurenson, M. K. (2002). Identifying reservoirs of infection: A conceptual and practical challenge. Emerging Infectious Diseases, 8(12), 1468-1473.

40. Chen, Y.-M., Hu, S.-J., Lin, X.-D., Tian, J.-H., Lv, J.-X., Wang, M.-R., . . . Zhang, Y.-Z. (2023). Host traits shape virome composition and virus transmission in wild small mammals. *Cell*.

41. Raghwani, J., Faust, C. L., François, S., Nguyen, D., Marsh, K., Raulo, A., . . . Pybus, O. G. (2022). Seasonal dynamics of the wild rodent faecal virome. *bioRxiv*, 2022.2002.2009.479684.

42. Rupasinghe, R., Chomel, B. B., & Martinez-Lopez, B. (2022). Climate change and zoonoses: a review of the current status, knowledge gaps, and future trends. *Acta Tropica, 226*.

43. Beermann, S., Dobler, G., Faber, M., Frank, C., Habedank, B., Hagedorn, P., . . . Wilking, H. (2023). Impact of climate change on vector- and rodent-borne infectious diseases. J Health Monit, 8(Suppl 3), 33-61.

44. Musa, S. S., Zhao, S., Gao, D. Z., Lin, Q. Y., Chowell, G., & He, D. H. (2020). Mechanistic modelling of the large-scale Lassa fever epidemics in Nigeria from 2016 to 2019. *Journal of Theoretical Biology, 493*.

45. Mazzola, L. T., & Kelly-Cirino, C. (2019). Diagnostics for Lassa fever virus: a genetically diverse pathogen found in low-resource settings. *Bmj Global Health, 4*.

46. Jonas, O., & Seifman, R. (2019). Do we need a Global Virome Project? *Lancet Global Health, 7*(10), E1314-E1316.

47. Tsoleridis, T., Chappell, J. G., Onianwa, O., Marston, D. A., Fooks, A. R., Monchatre-Leroy, E., . . . Ball, J. K. (2019). Shared common ancestry of rodent Alphacoronaviruses sampled globally. *Viruses-Basel, 11*(2).

48. Ibrahim, Y. M., Werid, G. M., Zhang, H., Fu, L. Z., Wang, W., Chen, H. Y., & Wang, Y. (2022). Potential zoonotic swine enteric viruses: the risk ignored for public health. *Virus Research, 315*.

49. Narrod, C., Zinsstag, J., & Tiongco, M. (2012). A one health framework for estimating the economic costs of zoonotic diseases on society. *Ecohealth, 9*(2), 150-162.

50. McKibbin, W., & Fernando, R. (2023). The global economic impacts of the COVID-19 pandemic. Economic Modelling, 129, 106551.

51. Neves, E. S., Mendenhall, I. H., Borthwick, S. A., Su, Y. C. F., & Smith, G. J. D. (2021). Genetic diversity and expanded host range of astroviruses detected in small mammals in Singapore. *One Health, 12*.

52. Saad, S. M., Sanderson, R., Robertson, P., & Lambert, M. (2021). Effects of supplementary feed for game birds on activity of brown rats Rattus norvegicus on arable farms. Mammal Research, 66(1), 163-171.

53. Cito, F., Rijks, J., Rantsios, A. T., Cunningham, A. A., Baneth, G., Guardabassi, L., . . . Giovannini, A. (2016). Prioritization of companion animal transmissible diseases for policy intervention in Europe. *Journal of Comparative Pathology, 155*(1, Supplement 1), S18-S26.

54. Varanda, C., Felix, M. D., Campos, M. D., & Materatski, P. (2021). An Overview of the application of viruses to biotechnology. *Viruses-Basel, 13*(10).

55. Lauber, C., & Seitz, S. (2022). Opportunities and challenges of data-driven virus discovery. *Biomolecules, 12*(8).

56. Labadie, T., Batejat, C., Leclercq, I., & Manuguerra, J. C. (2020). Historical discoveries on viruses in the environment and their impact on public health. *Intervirology, 63*(1-6), 17-32.

57. Saleh, A., Qamar, S., Tekin, A., Singh, R., & Kashyap, R. (2021). Vaccine Development Throughout History. Cureus, 13(7), e16635.

58. Hematian, A., Sadeghifard, N., Mohebi, R., Taherikalani, M., Nasrolahi, A., Amraei, M., & Ghafourian, S. (2016). Traditional and Modern Cell Culture in Virus Diagnosis. Osong Public Health Res Perspect, 7(2), 77-82.

59. Greninger, A. L. (2018). A decade of RNA virus metagenomics is (not) enough. *Virus Research, 244*, 218-229.

60. Su, Z. Q., Labaj, P. P., Li, S., Thierry-Mieg, J., Thierry-Mieg, D., Shi, W., . . . Shi, L. M. (2014). A comprehensive assessment of RNA-seq accuracy, reproducibility and information content by the Sequencing Quality Control Consortium. *Nature Biotechnology, 32*(9), 903-914.

61. Chuang, L. Y., Cheng, Y. H., & Yang, C. H. (2013). Specific primer design for the polymerase chain reaction. *Biotechnology Letters, 35*(10), 1541-1549.

62. Tsoleridis, T., Chappell, J. G., Monchatre-Leroy, E., Umhang, G., Shi, M., Bennett, M., . . . Ball, J. K. (2020). Discovery and prevalence of divergent RNA viruses in European field voles and rabbits. *Viruses-Basel, 12*(1).

63. Schmidt, S., Saxenhofer, M., Drewes, S., Schlegel, M., Wanka, K., Frank, R., . . . Ulrich, R. (2016). High genetic structuring of Tula hantavirus. *Archives of Virology, 161*(5), 1135-1149.

64. Pounder, K. C., Begon, M., Sironen, T., Henttonen, H., Watts, P. C., Voutilainen, L., . . . McElhinney, L. M. (2013). Novel hantavirus in field vole, United Kingdom. *Emerging Infectious Diseases, 19*(4), 673-675.

65. Moser, L. A., Ramirez-Carvajal, L., Puri, V., Pauszek, S. J., Matthews, K., Dilley, K. A., . . . Shabman, R. S. (2016). A universal next-generation sequencing protocol to generate noninfectious barcoded cDNA libraries from high-containment RNA viruses. *Msystems, 1*(3).

66. Ramesh, A., Nakielny, S., Hsu, J., Kyohere, M., Byaruhanga, O., de Bourcy, C., . . . DeRisi, J. L. (2019). Metagenomic next-generation sequencing of samples from pediatric febrile illness in Tororo, Uganda. *PLOS ONE, 14*(6), 17.

67. Bergner, L. M., Mollentze, N., Orton, R. J., Tello, C., Broos, A., Biek, R., & Streicker, D. G. (2021). Characterizing and evaluating the zoonotic potential of novel viruses discovered in vampire bats. *Viruses-Basel, 13*(2).

68. Lin, B., Hui, J. A., & Mao, H. J. (2021). Nanopore technology and its applications in gene gequencing. *Biosensors-Basel, 11*(7).

69. Modi, A., Vai, S., Caramelli, D., & Lari, M. (2021). The Illumina sequencing protocol and the NovaSeq 6000 system. In A. Mengoni, G. Bacci, & M. Fondi (Eds.), *BACTERIAL PANGENOMICS, 2 EDITION: Methods and Protocols* (Vol. 2242, pp. 15-42).

70. Bowman, S. K., Simon, M. D., Deaton, A. M., Tolstorukov, M., Borowsky, M. L., & Kingston, R. E. (2013). Multiplexed Illumina sequencing libraries from picogram quantities of DNA. *Bmc Genomics, 14*.

71. Chappell, J. G., Byaruhanga, T., Tsoleridis, T., Ball, J. K., & McClure, C. P. (2019). Identification of infectious agents in high-throughput sequencing data sets is easily achievable using free, cloud-based bioinformatics platforms. *Journal of Clinical Microbiology, 57*(12).

72. Vilsker, M., Moosa, Y., Nooij, S., Fonseca, V., Ghysens, Y., Dumon, K., . . . de Oliveira, T. (2019). Genome Detective: an automated system for virus identification from high-throughput sequencing data. *Bioinformatics, 35*(5), 871-873.

73. Jeon, S. A., Park, J. L., Park, S. J., Kim, J. H., Goh, S. H., Han, J. Y., & Kim, S. Y. (2021). Comparison between MGI and Illumina sequencing platforms for whole genome sequencing. *Genes & Genomics, 43*(7), 713-724.

74. Chen, Z., Erickson, D. L., & Meng, J. H. (2021). Polishing the Oxford Nanopore long-read assemblies of bacterial pathogens with Illumina short reads to improve genomic analyses. *Genomics, 113*(3), 1366-1377.

75. Canuti, M., & van der Hoek, L. (2014). Virus discovery: are we scientists or genome collectors? *Trends in Microbiology, 22*(5), 229-231.

76. Perot, P., Lecuit, M., & Eloit, M. (2017). Astrovirus diagnostics. *Viruses-Basel, 9*(1).

77. Kapp, J. D., Green, R. E., & Shapiro, B. (2021). A fast and efficient single-stranded genomic library preparation method optimized for ancient DNA. *Journal of Heredity, 112*(3), 241-249.

78. Dux, A., Lequime, S., Patrono, L. V., Vrancken, B., Boral, S., Gogarten, J. F., . . . Calvignac-Spencer, S. (2020). Measles virus and rinderpest virus divergence dated to the sixth century BCE. *Science, 368*(6497).

79. Raxworthy, C. J., & Smith, B. T. (2021). Mining museums for historical DNA: advances and challenges in museomics. Trends in Ecology & Evolution, 36(11), 1049-1060.

80. Smith, O., Dunshea, G., Sinding, M.-H. S., Fedorov, S., Germonpre, M., Bocherens, H., & Gilbert, M. T. P. (2019). Ancient RNA from Late Pleistocene permafrost and historical canids shows tissue-specific transcriptome survival. *PLOS Biology, 17*(7), e3000166.

81. Ng, T. F. F., Chen, L.-F., Zhou, Y., Shapiro, B., Stiller, M., Heintzman, P. D., . . . Delwart, E. (2014). Preservation of viral genomes in 700-y-old caribou feces from a subarctic ice patch. *Proceedings of the National Academy of Sciences, 111*(47), 16842-16847.

82. Keller, A., Kreis, S., Leidinger, P., Maixner, F., Ludwig, N., Backes, C., . . . Meese, E. (2016). miRNAs in ancient tissue specimens of the Tyrolean Iceman. *Molecular Biology and Evolution, 34*(4), 793-801.

83. Marmol-Sanchez, E., Fromm, B., Oskolkov, N., Pochon, Z., Kalogeropoulos, P., Eriksson, E., . . . Friedlander, M. (2023). Historical RNA expression profiles from the extinct Tasmanian tiger. *Genome Research*.

84. Speer, K. A., Hawkins, M. T. R., Flores, M. F. C., McGowen, M. R., Fleischer, R. C., Maldonado, J. E., . . . Muletz-Wolz, C. R. (2022). A comparative study of RNA yields

from museum specimens, including an optimized protocol for extracting RNA from formalin-fixed specimens. *Frontiers in Ecology and Evolution, 10*.

85. Der Sarkissian, C., Allentoft, M. E., Ávila-Arcos, M. C., Barnett, R., Campos, P. F., Cappellini, E., . . . Orlando, L. (2015). Ancient genomics. *Philosophical Transactions of the Royal Society B: Biological Sciences, 370*(1660).

86. Guy, P. L. (2013). Ancient RNA? RT-PCR of 50-year-old RNA identifies peach latent mosaic viroid. *Archives of Virology, 158*(3), 691-694.

87. Camacho-Sanchez, M., Burraco, P., Gomez-Mestre, I., & Leonard, J. A. (2013). Preservation of RNA and DNA from mammal samples under field conditions. *Molecular Ecology Resources, 13*(4), 663-673.

88. Torres, M. G., Weakley, A. M., Hibbert, J. D., Kirstein, O. D., Lanzaro, G. C., & Lee, Y. (2019). Ethanol as a potential mosquito sample storage medium for RNA preservation. *F1000Research, 8*, 1431-1431.

89. Jónsson, H., Ginolhac, A., Schubert, M., Johnson, P. L. F., & Orlando, L. (2013). mapDamage2.0: fast approximate Bayesian estimates of ancient DNA damage parameters. *Bioinformatics, 29*(13), 1682-1684.

90. Muhlemann, B., Margaryan, A., Damgaard, P. D., Allentoft, M. E., Vinner, L., Hansen, A. J., . . . Jones, T. C. (2018). Ancient human parvovirus B19 in Eurasia reveals its long-term association with humans. *Proceedings of the National Academy of Sciences of the United States of America, 115*(29), 7557-7562.

91. de le Roi, M., Puff, C., Wohlsein, P., Pfaff, F., Beer, M., Baumgaertner, W., & Rubbenstroth, D. (2023). Rustrela Virus as putative cause of nonsuppurative meningoencephalitis in lions. *Emerging Infectious Diseases, 29*(5), 1042-1045.

92. Patrono, L. V., Vrancken, B., Budt, M., Dux, A., Lequime, S., Boral, S., . . . Calvignac-Spencer, S. (2022). Archival influenza virus genomes from Europe reveal genomic variability during the 1918 pandemic. *Nature Communications, 13*(1).

93. Thibault, P. A., Watkinson, R. E., Moreira-Soto, A., Drexler, J. F., & Lee, B. (2017). Zoonotic potential of emerging Paramyxoviruses: knowns and unknowns. In M. Kielian, T. C. Mettenleiter, & M. J. Roossinck (Eds.), *Advances in Virus Research, Vol 98* (Vol. 98, pp. 1-55).

94. He, X. Z., Wang, X., Fan, G. H., Li, F., Wu, W. P., Wang, Z. H., . . . Ma, X. J. (2022). Metagenomic analysis of viromes in tissues of wild Qinghai vole from the eastern Tibetan Plateau. Scientific Reports, 12(1).

95. Lau, S. K. P., Woo, P. C. Y., Li, K. S. M., Zhang, H. J., Fan, R. Y. Y., Zhang, A. J. X., . . . Yuen, K. Y. (2016). Identification of novel Rosavirus species that infects diverse

rodent species and causes multisystemic dissemination in mouse model. *PLOS Pathogens, 12*(10).

96. Phan, T. G., Kapusinszky, B., Wang, C., Rose, R. K., Lipton, H. L., & Delwart, E. L. (2011). The fecal viral flora of wild rodents. *PLOS Pathogens, 7*(9), e1002218.

97. Salisbury, A. M., Begon, M., Dove, W., Niklasson, B., & Stewart, J. P. (2014). Ljungan virus is endemic in rodents in the UK. *Archives of Virology, 159*(3), 547-551.

98. Nishiyama, S., Dutia, B. M., Stewart, J. P., Meredith, A. L., Shaw, D. J., Simmonds, P., & Sharp, C. P. (2014). Identification of novel anelloviruses with broad diversity in UK rodents. *Journal of General Virology, 95*, 1544-1553.

99. Fevola, C., Kuivanen, S., Smura, T., Vaheri, A., Kallio-Kokko, H., Hauffe, H. C., . . . Jääskeläinen, A. J. (2018). Seroprevalence of lymphocytic choriomeningitis virus and Ljungan virus in Finnish patients with suspected neurological infections. *Journal of Medical Virology, 90*(3), 429-435.

100. Schneider, J., Hoffmann, B., Fevola, C., Schmidt, M. L., Imholt, C., Fischer, S., . . . Ulrich, R. G. (2021). Geographical distribution and genetic diversity of bank vole Hepaciviruses in Europe. *Viruses-Basel, 13*(7).

101. Jeske, K., Hiltbrunner, M., Drewes, S., Ryll, R., Wenk, M., Spakova, A., . . . Ulrich, R. G. (2019). Field vole-associated Traemmersee hantavirus from Germany represents a novel hantavirus species. *Virus Genes, 55*(6), 848-853.

102. Fevola, C., Rossi, C., Rosso, F., Girardi, M., Rosa, R., Manica, M., . . . Hauffe, H. C. (2020). Geographical distribution of Ljungan Virus in small mammals in Europe. *Vector-Borne and Zoonotic Diseases, 20*(9), 692-702.

103. Tan, Z. Z., Gonzalez, G., Sheng, J. L., Wu, J. M., Zhang, F. Q., Xu, L., . . . He, B. (2020). Extensive genetic diversity of Polyomaviruses in sympatric bat communities: host switching versus coevolution. *Journal of Virology, 94*(9).

104. Van Brussel, K., & Holmes, E. C. (2022). Zoonotic disease and virome diversity in bats. Current Opinion in Virology, 52, 192-202.

105. Drexler, J. F., Corman, V. M., Müller, M. A., Lukashev, A. N., Gmyl, A., Coutard, B., . . . Drosten, C. (2013). Evidence for Novel Hepaciviruses in rodents. *PLOS Pathogens, 9*(6), e1003438.

106. DeBiasi, R. L., & Tyler, K. L. (2015). Orthoreoviruses and Orbiviruses. *Mandell, Douglas, and Bennett's Principles and Practice of Infectious Diseases*, 1848-1850 e1841.

107. Ancillotto, L., Santini, L., Ranc, N., Maiorano, L., & Russo, D. (2016). Extraordinary range expansion in a common bat: the potential roles of climate change and urbanisation. The Science of Nature, 103(3), 15.

108. Lelli, D., Prosperi, A., Moreno, A., Chiapponi, C., Gibellini, A. M., De Benedictis, P., . . . Lavazza, A. (2018). Isolation of a novel Rhabdovirus from an insectivorous bat (Pipistrellus kuhlii) in Italy. Virology Journal, 15(1), 37.

109. Benkő, M., Aoki, K., Arnberg, N., Davison, A. J., Echavarría, M., Hess, M., . . . Consortium, I. R. (2022). ICTV virus taxonomy profile: Adenoviridae 2022. *Journal of General Virology, 103*(3).

110. Diffo, J., Ndze, V. N., Ntumvi, N. F., Takuo, J. M., Mouiche, M. M. M., Tamoufe, U., . . . Lange, C. E. (2019). DNA of diverse adenoviruses detected in Cameroonian rodent and shrew species. *Archives of Virology, 164*(9), 2359-2366.

111. Lukashev, A. N. (2010). Recombination among picornaviruses. *Reviews in Medical Virology, 20*(5), 327-337.

112. Ismail, A. M., Lee, J. S., Lee, J. Y., Singh, G., Dyer, D. W., Seto, D., . . . Rajaiya, J. (2018). Adenoviromics: mining the human Adenovirus Species D genome. *Frontiers in Microbiology, 9*.

113. Wu, X., Zhang, J., Lan, W., Quan, L., Ou, J., Zhao, W., . . . Zhang, Q. (2022). Molecular typing and rapid identification of human Adenoviruses associated with respiratory diseases using universal PCR and sequencing primers for the three major capsid genes: penton base, hexon, and fiber. *Front Microbiol, 13*, 911694.

114. Chang, J. (2021). Adenovirus vectors: excellent tools for vaccine development. *Immune Network, 21*(1).

115. Kumakamba, C., N'Kawa, F., Kingebeni, P. M., Losoma, J. A., Lukusa, I. N., Muyembe, F., . . . Lange, C. E. (2020). Analysis of adenovirus DNA detected in rodent species from the Democratic Republic of the Congo indicates potentially novel adenovirus types. *New Microbes and New Infections, 34*.

116. Greenwood, A. G., & Sanchez, S. (2002). Serological evidence of murine pathogens in wild grey squirrels (Sciurus carolinensis) in North Wales. *Veterinary Record, 150*(17), 543-546.

117. Becker, S. D., Bennett, M., Stewart, J. P., & Hurst, J. L. (2007). Serological survey of virus infection among wild house mice (Mus domesticus) in the UK. *Laboratory Animals, 41*(2), 229-238.

118. Wellehan, J. F. X., Johnson, A. J., Harrach, B., Benko, M., Pessier, A. P., Johnson, C. M., . . . Jacobson, E. R. (2004). Detection and analysis of six lizard

adenoviruses by consensus primer PCR provides further evidence of a reptilian origin for the atadenoviruses. *Journal of Virology, 78*(23), 13366-13369.

119. Borkenhagen, L. K., Fieldhouse, J. K., Seto, D., & Gray, G. C. (2019). Are adenoviruses zoonotic? A systematic review of the evidence. *Emerg Microbes Infect, 8*(1), 1679-1687.

120. Garcia-Zalisnak, D., Rapuano, C., Sheppard, J. D., & Davis, A. R. (2018). Adenovirus ocular infections: prevalence, pathology, pitfalls, and practical pointers. *Eye & Contact Lens-Science and Clinical Practice, 44*, S1-S7.

121. Zheng, X. Y., Qiu, M., Ke, X. M., Guan, W. J., Li, J. M., Huo, S. T., . . . Chen, Q. (2016). Detection of novel adenoviruses in fecal specimens from rodents and shrews in southern China. *Virus Genes, 52*(3), 417-421.

122. Arvind, N., Sushma, M., & Krishnappa, J. (2019). Prevalence of Rotavirus and Adenovirus in the childhood gastroenteritis in a tertiary care teaching hospital. *Journal of Pure and Applied Microbiology, 13*(2), 1011-1015.

123. Ljubin-Sternak, S., Mestrovic, T., Luksic, I., Mijac, M., & Vranes, J. (2021). Seasonal Coronaviruses and other neglected respiratory viruses: a global perspective and a local snapshot. *Frontiers in Public Health, 9*.

124. Mousavi Nasab, S. D., Zali, F., Kaghazian, H., Aghasadeghi, M. R., Mardani, R., Gachkar, L., . . . Ghasemzadeh, A. (2020). Prevalence of astrovirus, adenovirus, and sapovirus infections among Iranian children with acute gastroenteritis. *Gastroenterol Hepatol Bed Bench, 13*(Suppl1), S122-s127.

125. Zhang, R. M., Wang, H. M., Tian, S. F., & Deng, J. K. (2021). Adenovirus viremia may predict adenovirus pneumonia severity in immunocompetent children. *Bmc Infectious Diseases, 21*(1).

126. Sato-Dahlman, M., LaRocca, C. J., Yanagiba, C., & Yamamoto, M. (2020). Adenovirus and immunotherapy: advancing cancer treatment by combination. *Cancers, 12*(5).

127. Radoshitzky, S. R., Buchmeier, M. J., Charrel, R. N., Clegg, J. C. S., Gonzalez, J. J., Günther, S., . . . Ictv Report, C. (2019). ICTV virus taxonomy profile: Arenaviridae. *J Gen Virol, 100*(8), 1200-1201.

128. Lan, X. H., Zhang, Y. L., Jia, X. Y., Dong, S. Q., Liu, Y., Zhang, M. M., . . . Wang, W. (2022). Screening and identification of Lassa virus endonuclease-targeting inhibitors from a fragment-based drug discovery library. *Antiviral Research, 197*.

129. Simulundu, E., Mweene, A. S., Changula, K., Monze, M., Chizema, E., Mwaba, P., . . . Bates, M. (2016). Lujo viral hemorrhagic fever: considering diagnostic capacity

and preparedness in the wake of recent Ebola and Zika virus outbreaks. *Reviews in Medical Virology, 26*(6), 446-454.

130. Sarute, N., & Ross, S. R. (2017). New world Arenavirus biology. In L. Enquist (Ed.), *Annual Review of Virology, Vol 4* (Vol. 4, pp. 141-158).

131. Kumar, S., Yadav, D., Singh, D., Shakya, K., & Rathi, B. (2023). Recent developments on Junin virus, a causative agent for Argentine haemorrhagic fever. *Reviews in Medical Virology, 33*(2).

132. Mateer, E. J., Huang, C., Shehu, N. Y., & Paessler, S. (2018). Lassa fever-induced sensorineural hearing loss: a neglected public health and social burden. *Plos Neglected Tropical Diseases, 12*(2).

133. Ali, Z., Cardoza, J. V., Basak, S., Narsaria, U., Bhattacharjee, S., Meher, G. U., . . . George, S. S. (2023). A multi-epitope vaccine candidate against Bolivian hemorrhagic fever caused by Machupo Virus. *Applied Biochemistry and Biotechnology*.

134. Mafayle, R. L., Morales-Betoulle, M. E., Romero, C., Cossaboom, C. M., Whitmer, S., Aguilera, C. E. A., . . . Montgomery, J. M. (2022). Chapare hemorrhagic fever and virus detection in rodents in Bolivia in 2019. *New England Journal of Medicine, 386*(24), 2283-2294.

135. Bonthius, D. J. (2012). Lymphocytic choriomeningitis virus: an underrecognized cause of neurologic disease in the fetus, child, and adult. *Semin Pediatr Neurol, 19*(3), 89-95.

136. Silva-Ramos, C. R., Montoya-Ruiz, C., Faccini-Martinez, A. A., & Rodas, J. D. (2022). An updated review and current challenges of Guanarito virus infection, Venezuelan hemorrhagic fever. *Archives of Virology, 167*(9), 1727-1738.

137. da Paz, T. Y. B., Hernandez, L. H. A., da Silva, S. P., da Silva, F. S., de Barros, B. C. V., Casseb, L. M. N., . . . Cruz, A. C. R. (2023). Novel rodent Arterivirus detected in the Brazilian Amazon. *Viruses-Basel, 15*(5).

138. Snijder, E. J., Kikkert, M., & Fang, Y. (2013). Arterivirus molecular biology and pathogenesis. *Journal of General Virology, 94*, 2141-2163.

139. Brinton, M. A., Gulyaeva, A. A., Balasuriya, U. B. R., Dunowska, M., Faaberg, K. S., Goldberg, T., . . . Gorbalenya, A. E. (2021). ICTV virus taxonomy profile: Arteriviridae 2021. *Journal of General Virology, 102*(8).

140. Vanmechelen, B., Vergote, V., Laenen, L., Koundouno, F. R., Bore, J. A., Wada, J., . . . Maes, P. (2018). Expanding the Arterivirus host spectrum: Olivier's Shrew Virus 1, a novel Arterivirus discovered in African Giant Shrews. *Scientific Reports, 8*.

141. Dastjerdi, A., Inglese, N., Partridge, T., Karuna, S., Everest, D. J., Frossard, J. P., . . . Stidworthy, M. F. (2021). Novel Arterivirus associated with outbreak of fatal encephalitis in European Hedgehogs, England, 2019. *Emerging Infectious Diseases, 27*(2), 578-581.

142. Prajapati, M., Acharya, M. P., Yadav, P., & Frossard, J. P. (2023). Farm characteristics and sero-prevalence of porcine reproductive and respiratory syndrome virus (PRRSV) antibodies in pigs of Nepal. *Veterinary Medicine and Science, 9*(1), 174-180.

143. Bailey, A. L., Lauck, M., Sibley, S. D., Friedrich, T. C., Kuhn, J. H., Freimer, N. B., . . . O'Connor, D. H. (2016). Zoonotic potential of simian Arteriviruses. *Journal of Virology, 90*(2), 630-635.

144. Lazic, S., Lupulovic, D., Gaudaire, D., Petrovic, T., Lazic, G., & Hans, A. (2017). Serological evidence of equine arteritis virus infection and phylogenetic analysis of viral isolates in semen of stallions from Serbia. *Bmc Veterinary Research, 13*.

145. Brown, J. R., Morfopoulou, S., Hubb, J., Emmett, W. A., Ip, W., Shah, D., . . . Breuer, J. (2015). Astrovirus VA1/HMO-C: an increasingly recognized neurotropic pathogen in immunocompromised patients. *Clinical Infectious Diseases, 60*(6), 881-888.

146. Arowolo, K. O., Ayolabi, C. I., Adeleye, I. A., Lapinski, B., Santos, J. S., & Raboni, S. M. (2020). Molecular epidemiology of astrovirus in children with gastroenteritis in southwestern Nigeria. *Archives of Virology, 165*(11), 2461-2469.

147. Wu, L. M., Teng, Z., Lin, Q. N., Liu, J., Wu, H. Y., Kuang, X. Z., . . . Xie, Y. H. (2020). Epidemiology and genetic characterization of classical human Astrovirus infection in Shanghai, 2015-2016. *Frontiers in Microbiology, 11*.

148. Hargest, V., Sharp, B., Livingston, B., Cortez, V., & Schultz-Cherry, S. (2020). Astrovirus replication is inhibited by Nitazoxanide In Vitro and In Vivo. *Journal of Virology, 94*(5).

149. Woo, P. C. Y., de Groot, R. J., Haagmans, B., Lau, S. K. P., Neuman, B. W., Perlman, S., . . . Yeh, S. H. (2023). ICTV virus taxonomy profile: Coronaviridae 2023. *Journal of General Virology, 104*(4).

150. Zhou, Q., Li, Y., Huang, J., Fu, N. S., Song, X., Sha, X., & Zhang, B. (2021). Prevalence and molecular characteristics of feline coronavirus in southwest China from 2017 to 2020. *Journal of General Virology, 102*(9).

151. Woo, P. C. Y., Lau, S. K. P., Lam, C. S. F., Lau, C. C. Y., Tsang, A. K. L., Lau, J. H. N., . . . Yuen, K. Y. (2012). Discovery of seven novel mammalian and avian Coronaviruses in the genus Deltacoronavirus supports bat Coronaviruses as the gene

source of Alphacoronavirus and Betacoronavirus and avian Coronaviruses as the gene source of Gammacoronavirus and Deltacoronavirus. *Journal of Virology, 86*(7), 3995-4008.

152. Monastiri, A., Martin-Carrillo, N., Foronda, P., Izquierdo-Rodriguez, E., Feliu, C., Lopez-Roig, M., . . . Serra-Cobo, J. (2021). First Coronavirus active survey in rodents from the Canary Islands. *Front Vet Sci, 8*.

153. Woo, P. C. Y., Lau, S. K. P., Chu, C. M., Chan, K. H., Tsoi, H. W., Huang, Y., . . . Yuen, K. Y. (2005). Characterization and complete genome sequence of a novel coronavirus, coronavirus HKU1, from patients with pneumonia. *Journal of Virology, 79*(2), 884-895.

154. Shahrajabian, M. H., Sun, W. L., & Cheng, Q. (2021). Product of natural evolution (SARS, MERS, and SARS-CoV-2); deadly diseases, from SARS to SARS-CoV-2. *Human Vaccines & Immunotherapeutics, 17*(1), 62-83.

155. Wong, L. H., Edgar, J. R., Martello, A., Ferguson, B. J., & Eden, E. R. (2021). Exploiting connections for viral replication. *Frontiers in Cell and Developmental Biology, 9*.

156. Korner, R. W., Majjouti, M., Alcazar, M. A. A., & Mahabir, E. (2020). Of mice and men: the Coronavirus MHV and mouse models as a translational approach to understand SARS-CoV-2. *Viruses-Basel, 12*(8).

157. Wasberg, A., Raghwani, J., Li, J., Pettersson, J. H.-O., Lindahl, J. F., Lundkvist, Å., & Ling, J. (2022). Discovery of a novel Coronavirus in Swedish bank voles (Myodes glareolus). *Viruses, 14*(6), 1205.

158. Dyrdak, R., Hodcroft, E. B., Wahlund, M., Neher, R. A., & Albert, J. (2021). Interactions between seasonal human coronaviruses and implications for the SARS-CoV-2 pandemic: a retrospective study in Stockholm, Sweden, 2009-2020. *Journal of Clinical Virology, 136*.

159. Kumakamba, C., Niama, F. R., Muyembe, F., Mombouli, J.-V., Kingebeni, P. M., Nina, R. A., . . . Lange, C. E. (2021). Coronavirus surveillance in wildlife from two Congo basin countries detects RNA of multiple species circulating in bats and rodents. *PLOS ONE, 16*(6), e0236971.

160. Eckerle, I., Lenk, M., & Ulrich, R. G. (2014). More novel Hantaviruses and diversifying reservoir hosts time for development of reservoir-derived cell culture models? *Viruses-Basel, 6*(3), 951-967.

161. Park, S., Lee, Y., Michelow, I. C., & Choe, Y. J. (2020). Global seasonality of human Coronaviruses: a systematic review. *Open Forum Infectious Diseases, 7*(11).

162. Li, Y., Wang, X., & Nair, H. (2020). Global Seasonality of Human Seasonal Coronaviruses: A Clue for Postpandemic Circulating Season of Severe Acute Respiratory Syndrome Coronavirus 2? J Infect Dis, 222(7), 1090-1097.

163. Chow, E. J., Uyeki, T. M., & Chu, H. Y. (2023). The effects of the COVID-19 pandemic on community respiratory virus activity. *Nature Reviews Microbiology, 21*(3), 195-210.

164. Schindell, B. G., Allardice, M., McBride, J. A. M., Dennehy, B., & Kindrachuk, J. (2022). SARS-CoV-2 and the Missing Link of Intermediate Hosts in Viral Emergence - What We Can Learn From Other Betacoronaviruses. Frontiers in Virology, 2.

165. Chen, C., Haupert, S. R., Zimmermann, L., Shi, X., Fritsche, L. G., & Mukherjee, B. (2022). Global prevalence of post-Coronavirus disease 2019 (COVID-19) condition or long COVID: a meta-analysis and systematic review. *Journal of Infectious Diseases, 226*(9), 1593-1607.

166. Temmam, S., Vongphayloth, K., Baquero, E., Munier, S., Bonomi, M., Regnault, B., . . . Eloit, M. (2022). Bat coronaviruses related to SARS-CoV-2 and infectious for human cells. Nature, 604(7905), 330-336.

167. Tarantini, F. S., Wu, S. Y., Jenkins, H., Lopez, A. T., Tomlin, H., Hyde, R., . . . Benest, A. V. (2022). Direct RT-qPCR assay for the detection of SARS-CoV-2 in saliva samples. *Methods and Protocols, 5*(2).

168. Shi, T., Robertson, C., & Sheikh, A. (2023). Effectiveness and safety of coronavirus disease 2019 vaccines. *Current Opinion in Pulmonary Medicine, 29*(3), 138-142.

169. Gatherer, D., Depledge, D. P., Hartley, C. A., Szpara, M. L., Vaz, P. K., Benkő, M., . . . Davison, A. J. (2021). ICTV virus taxonomy profile: Herpesviridae 2021. *J Gen Virol, 102*(10).

170. Roark, H. K., Jenks, J. A., Permar, S. R., & Schleiss, M. R. (2020). Animal models of congenital Cytomegalovirus transmission: implications for vaccine development. *J Infect Dis, 221*(Suppl 1), S60-s73.

171. Struble, E. B., Murata, H., Komatsu, T., & Scott, D. (2021). Immune prophylaxis and therapy for human Cytomegalovirus infection. *International Journal of Molecular Sciences, 22*(16).

172. Dietrich, M. L., & Schieffelin, J. S. (2019). Congenital Cytomegalovirus infection. *Ochsner Journal, 19*(2), 123-130.

173. Llah, S. T., Mir, S., Sharif, S., Khan, S., & Mir, M. A. (2018). Hantavirus induced cardiopulmonary syndrome: a public health concern. *Journal of Medical Virology, 90*(6), 1003-1009.

174. Thomason, A. G., Begon, M., Bradley, J. E., Paterson, S., & Jackson, J. A. (2017). Endemic hantavirus in field voles, Northern England. *Emerging Infectious Diseases, 23*(6), 1033-1035.

175. Klempa, B., Fichet-Calvet, E., Lecompte, E., Auste, B., Aniskin, V., Meisel, H., . . . Kruger, D. H. (2006). Hantavirus in African wood mouse, Guinea. *Emerging Infectious Diseases, 12*(5), 838-840.

176. Kuenzli, A. B., Marschall, J., Schefold, J. C., Schafer, M., Engler, O. B., Ackermann-Gaumann, R., . . . Staehelin, C. (2018). Hantavirus cardiopulmonary syndrome due to imported Andes Hantavirus infection in Switzerland: a multidisciplinary challenge, two cases and a literature review. *Clinical Infectious Diseases, 67*(11), 1788-1795.

177. Prist, P. R., Dandrea, P. S., & Metzger, J. P. (2017). Landscape, climate and Hantavirus cardiopulmonary syndrome outbreaks. *Ecohealth, 14*(3), 614-629.

178. Zhang, R., Mao, Z., Yang, J., Liu, S., Liu, Y., Qin, S., . . . Wang, Z. (2021). The changing epidemiology of hemorrhagic fever with renal syndrome in Southeastern China during 1963-2020: a retrospective analysis of surveillance data. *PLoS Negl Trop Dis, 15*(8), e0009673.

179. Billings, A. N., Rollin, P. E., Milazzo, M. L., Molina, C. P., Eyzaguirre, E. J., Livingstone, W., . . . Fulhorst, C. F. (2010). Pathology of Black Creek Canal Virus infection in juvenile hispid cotton rats (Sigmodon hispidus). *Vector-Borne and Zoonotic Diseases, 10*(6), 621-628.

180. Liu, R., Ma, H., Shu, J., Zhang, Q., Han, M., Liu, Z., . . . Wu, X. (2019). Vaccines and therapeutics against Hantaviruses. *Front Microbiol, 10*, 2989.

181. Zelena, H., Mrazek, J., & Kuhn, T. (2013). Tula Hantavirus infection in immunocompromised host, Czech Republic. *Emerging Infectious Diseases, 19*(11), 1873-1876.

182. Jiang, J. F., Hao, Y. Q., He, B. A., Su, L. H., Li, X. Z., Liu, X. X., . . . Tu, C. C. (2022). Severe acute hepatitis outbreaks associated with a novel Hepacivirus in Rhizomys pruinosus in Hainan, China. *Journal of Virology, 96*(17).

183. Rohrs, S., Begeman, L., Straub, B. K., Boadella, M., Hanke, D., Wernike, K., . . . Beer, M. (2021). The bank vole (Clethrionomys glareolus)-small animal model for Hepacivirus infection. *Viruses-Basel, 13*(12).

184. Smith, D. B., Becher, P., Bukh, J., Gould, E. A., Meyers, G., Monath, T., . . . Simmonds, P. (2016). Proposed update to the taxonomy of the genera Hepacivirus and Pegivirus within the Flaviviridae family. *Journal of General Virology, 97*, 2894-2907.

185. Simmonds, P., Becher, P., Bukh, J., Gould, E. A., Meyers, G., Monath, T., . . . Consortium, I. R. (2017). ICTV virus taxonomy profile: Flaviviridae. *Journal of General Virology, 98*(1), 2-3.

186. Porter, A. F., Pettersson, J. H., Chang, W. S., Harvey, E., Rose, K., Shi, M., . . . Holmes, E. C. (2020). Novel hepaci- and pegi-like viruses in native Australian wildlife and non-human primates. *Virus Evol, 6*(2), veaa064.

187. Hartlage, A. S., Cullen, J. M., & Kapoor, A. (2016). The strange, expanding world of animal Hepaciviruses. In L. W. Enquist (Ed.), *Annual Review of Virology, Vol 3* (Vol. 3, pp. 53-75).

188. Spearman, C. W., Dusheiko, G. M., Hellard, M., & Sonderup, M. (2019). Hepatitis C. *Lancet, 394*(10207), 1451-1466.

189. Rein, A. (2011). Murine leukemia viruses: objects and organisms. *Adv Virol, 2011*, 403419.

190. Coffin, J., Blomberg, J., Fan, H., Gifford, R., Hatziioannou, T., Lindemann, D., . . . Ictv Report, C. (2021). ICTV virus taxonomy profile: Retroviridae 2021. *J Gen Virol, 102*(12).

191. Khanna, M., Manocha, N., Himanshi, Joshi, G., Saxena, L., & Saini, S. (2019). Role of retroviral vector-based interventions in combating virus infections. *Future Virology, 14*(7), 473-485.

192. Fontes, F., Rocha, S., Sánchez, R., Pessina, P., Sebastian, M., Benavides, F., & Breijo, M. (2022). Detection of high antibodies titers against rat leukemia virus in an outbreak of reproductive disorders and lymphomas in Wistar rats. *Lab Anim, 56*(5), 437-445.

193. Hayward, A., Cornwallis, C. K., & Jern, P. (2015). Pan-vertebrate comparative genomics unmasks retrovirus macroevolution. *Proc Natl Acad Sci U S A, 112*(2), 464-469.

194. Arias, M., & Fan, H. (2014). The saga of XMRV: a virus that infects human cells but is not a human virus. *Emerg Microbes Infect, 3*(4), e.

195. Skorski, M., Bamunusinghe, D., Liu, Q., Shaffer, E., & Kozak, C. A. (2019). Distribution of endogenous gammaretroviruses and variants of the Fv1 restriction gene in individual mouse strains and strain subgroups. *PLOS ONE, 14*(7), e0219576.

196. Howard, T. M., Sheng, Z., Wang, M., Wu, Y., & Rasheed, S. (2006). Molecular and phylogenetic analyses of a new amphotropic murine leukemia virus (MuLV-1313). *Virol J, 3*, 101.

197. Ahasan, M. S., Subramaniam, K., Krauer, J. M. C., Sayler, K. A., Loeb, J. C., Goodfriend, O. F., . . . Lednicky, J. A. (2020). Three new Orbivirus species isolated from farmed white-tailed deer (Odocoileus virginianus) in the United States. *Viruses-Basel, 12*(1).

198. Drolet, B. S., van Rijn, P., Howerth, E. W., Beer, M., & Mertens, P. P. (2015). A review of knowledge gaps and tools for Orbivirus research. *Vector-Borne and Zoonotic Diseases, 15*(6), 339-347.

199. Wu, D., Tan, Q. Q., Zhang, H., Huang, P., Zhou, H. Q., Zhang, X., . . . Liang, G. D. (2020). Genomic and biological features of a novel orbivirus isolated from mosquitoes, in China. *Virus Research, 285*.

200. Noronha, L. E., Cohnstaedt, L. W., Richt, J. A., & Wilson, W. C. (2021). Perspectives on the changing landscape of epizootic hemorrhagic disease virus control. *Viruses-Basel, 13*(11).

201. Madani, H., Casal, J., Alba, A., Allepuz, A., Cetre-Sossah, C., Hafsi, L., . . . Napp, S. (2011). Animal diseases caused by Orbiviruses, Algeria. *Emerging Infectious Diseases, 17*(12), 2325-2327.

202. Rojas, J. M., Martin, V., & Sevilla, N. (2021). Vaccination as a strategy to prevent Bluetongue Virus vertical transmission. *Pathogens, 10*(11).

203. Vanmechelen, B., Meurs, S., Horemans, M., Loosen, A., Maes, T. J., Laenen, L., . . . Maes, P. (2022). The characterization of multiple novel paramyxovirus species highlights the diverse nature of the subfamily Orthoparamyxovirinae. *bioRxiv*, 2022.2003.2010.483612.

204. Zhang, Y. F., Zhang, J. T., Wang, Y. N., Tian, F., Zhang, X. L., Wang, G., . . . Zhang, X. A. (2023). Genetic diversity and expanded host range of J Paramyxovirus detected in wild small mammals in China. *Viruses-Basel, 15*(1).

205. Amarasinghe, G. K., Ayllon, M. A., Bao, Y., Basler, C. F., Bavari, S., Blasdell, K. R., . . . Kuhn, J. H. (2019). Taxonomy of the order Mononegavirales: update 2019. *Archives of Virology, 164*(7), 1967-1980.

206. Matsumoto, Y., Ohta, K., Kolakofsky, D., & Nishio, M. (2018). The control of paramyxovirus genome hexamer length and mRNA editing. *Rna, 24*(4), 461-467.

207. Thakur, N., & Bailey, D. (2019). Advances in diagnostics, vaccines and therapeutics for Nipah virus. *Microbes and Infection, 21*(7), 278-286.

208. Napp, S., Alba, A., Rocha, A. I., Sanchez, A., Rivas, R., Majo, N., . . . Busquets, N. (2017). Six-year surveillance of Newcastle disease virus in wild birds in north-eastern Spain (Catalonia). *Avian Pathology, 46*(1), 59-67.

209. Balkema-Buschmann, A., Fischer, K., McNabb, L., Diederich, S., Singanallur, N. B., Ziegler, U., . . . Colling, A. (2022). Serological Hendra Virus diagnostics using an indirect ELISA-based DIVA approach with recombinant Hendra G and N proteins. *Microorganisms, 10*(6).

210. Gastanaduy, P. A., Goodson, J. L., Panagiotakopoulos, L., Rota, P. A., Orenstein, W. A., & Patel, M. (2021). Measles in the 21st century: progress toward achieving and sustaining elimination. *Journal of Infectious Diseases, 224*, S420-S428.

211. Beleni, A. I., & Borgmann, S. (2018). Mumps in the vaccination age: global epidemiology and the situation in Germany. *International Journal of Environmental Research and Public Health, 15*(8).

212. Jager, M. C., Tomlinson, J. E., Lopez-Astacio, R. A., Parrish, C. R., & Van de Walle, G. R. (2021). Small but mighty: old and new parvoviruses of veterinary significance. *Virology Journal, 18*(1).

213. Väisänen, E., Fu, Y., Hedman, K., & Söderlund-Venermo, M. (2017). Human Protoparvoviruses. *Viruses, 9*(11).

214. Pénzes, J. J., Söderlund-Venermo, M., Canuti, M., Eis-Hübinger, A. M., Hughes, J., Cotmore, S. F., & Harrach, B. (2020). Reorganizing the family Parvoviridae: a revised taxonomy independent of the canonical approach based on host association. *Archives of Virology, 165*(9), 2133-2146.

215. Capozza, P., Buonavoglia, A., Pratelli, A., Martella, V., & Decaro, N. (2023). Old and novel enteric Parvoviruses of dogs. *Pathogens, 12*(5).

216. Chong, Y. L., & Ng, K. H. (2017). Genomic recombination in primate bocavirus: inconsistency and alternative interpretations. *Virus Genes, 53*(6), 774-777.

217. Hildebrandt, E., Penzes, J. J., Gifford, R. J., Agbandje-Mckenna, M., & Kotin, R. M. (2020). Evolution of dependoparvoviruses across geological timescales-implications for design of AAV-based gene therapy vectors. *Virus Evolution, 6*(2).

218. de Souza, W. M., Romeiro, M. F., Fumagalli, M. J., Modha, S., de Araujo, J., Queiroz, L. H., . . . Gifford, R. J. (2017). Chapparvoviruses occur in at least three vertebrate classes and have a broad biogeographic distribution. *J Gen Virol, 98*(2), 225-229.

219. Phan, T., & Nagaro, K. (2020). Cutavirus: A newly discovered parvovirus on the rise. *Infection Genetics and Evolution, 80*.

220. Vilibic-Cavlek, T., Tabain, I., Kolaric, B., Mihulja, K., Blazevic, L., Bogdanic, M., . . . Mrzljak, A. (2021). Parvovirus B19 in Croatia: a large-scale seroprevalence study. *Medicina-Lithuania, 57*(11).

221. Vilmane, A., Terentjeva, A., Tamosiunas, P. L., Suna, N., Suna, I., Petraityte-Burneikiene, R., . . . Nora-Krukle, Z. (2020). Human Parvoviruses may affect the development and clinical course of meningitis and meningoencephalitis. *Brain Sciences, 10*(6).

222. Van den Abeele, T., Delforge, M. L., Boel, A., Reynders, M., & Padalko, E. (2021). Comparison of 4 commercial enzyme immunoassays for serology testing of human parvovirus B19 infection. *Diagnostic Microbiology and Infectious Disease, 101*(3).

223. De Paschale, M., Pavia, C., Cerulli, T., Cagnin, D., Manco, M. T., Belvisi, L., . . . Clerici, P. (2022). Prevalence of anti-parvovirus B19 IgG and IgM and parvovirus B19 viremia in pregnant women in an urban area of Northern Italy. *Journal of Medical Virology, 94*(11), 5409-5414.

224. Mohd-Azami, S. N. I., Loong, S. K., Khoo, J. J., Sahimin, N., Lim, F. S., Husin, N. A., . . . Abubakar, S. (2022). Molecular evidence of rat bocavirus among rodents in Peninsular Malaysia. *Journal of Veterinary Medical Science, 84*(7), 938-941.

225. Li, Y. P., Gordon, E., Idle, A., Altan, E., Seguin, M. A., Estrada, M., . . . Delwart, E. (2020). Virome of a feline outbreak of diarrhea and vomiting includes Bocaviruses and a novel Chapparvovirus. *Viruses-Basel, 12*(5).

226. Wang, D., Tai, P. W. L., & Gao, G. (2019). Adeno-associated virus vector as a platform for gene therapy delivery. *Nat Rev Drug Discov, 18*(5), 358-378.

227. Canuti, M., Bouchard, E., Rodrigues, B., Whitney, H. G., Hopson, M., Gilroy, C., . . . Verhoeven, J. T. P. (2021). Newlavirus, a novel, highly prevalent, and highly diverse Protoparvovirus of foxes (Vulpes spp.). *Viruses-Basel, 13*(10).

228. Angelova, A., Ferreira, T., Bretscher, C., Rommelaere, J., & Marchini, A. (2021). Parvovirus-based combinatorial immunotherapy: a reinforced therapeutic strategy against poor-prognosis solid cancers. *Cancers, 13*(2).

229. Schildgen, O. (2013). Human bocavirus: lessons learned to date. *Pathogens, 2*(1), 1-12.

230. Kobayashi, H., Shinjoh, M., Sudo, K., Kato, S., Morozumi, M., Koinuma, G., . . . Hasegawa, N. (2019). Nosocomial infection by human bocavirus and human rhinovirus among paediatric patients with respiratory risks. *Journal of Hospital Infection, 103*(3), 341-348.

231. Lee, Q., Padula, M. P., Pinello, N., Williams, S. H., O'Rourke, M. B., Fumagalli, M. J., . . . Jolly, C. J. (2020). Murine and related chapparvoviruses are nephro-tropic and produce novel accessory proteins in infected kidneys. *PLOS Pathogens, 16*(1).

232. Prakash, S., Shukla, S., Bhagat, A. K., Mishra, H., Vangala, R., & Jain, A. (2021). Human parvovirus 4: an emerging etiological agent in cases presenting with influenza like illness. *J Med Virol, 93*(8), 5158-5162.

233. Cao, J., & Zhu, X. Q. (2020). Acute viral encephalitis associated with human parvovirus B19 infection: unexpectedly diagnosed by metagenomic next-generation sequencing. *Journal of Neurovirology, 26*(6), 980-983.

234. Zhu, W. T., Yang, J., Lu, S., Huang, Y. Y., Jin, D., Pu, J., . . . Xu, J. G. (2022). Novel pegiviruses infecting wild birds and rodents. *Virologica Sinica, 37*(2), 208-214.

235. Yu, Y. Q., Wan, Z. Z., Wang, J. H., Yang, X. G., & Zhang, C. Y. (2022). Review of human pegivirus: prevalence, transmission, pathogenesis, and clinical implication. *Virulence, 13*(1), 324-341.

236. Bhattarai, N., & Stapleton, J. T. (2012). GB virus C: the good boy virus? *Trends in Microbiology, 20*(3), 124-130.

237. Baechlein, C., Grundhoff, A., Fischer, N., Alawi, M., Hoeltig, D., Waldmann, K. H., & Becher, P. (2016). Pegivirus infection in domestic pigs, Germany. *Emerg Infect Dis, 22*(7), 1312-1314.

238. Chandriani, S., Skewes-Cox, P., Zhong, W. D., Ganem, D. E., Divers, T. J., Van Blaricum, A. J., . . . Kistler, A. L. (2013). Identification of a previously undescribed divergent virus from the Flaviviridae family in an outbreak of equine serum hepatitis. *Proceedings of the National Academy of Sciences of the United States of America, 110*(15), E1407-E1415.

239. Tuddenham, R., Eden, J. S., Gilbey, T., Dwyer, D. E., Jennings, Z., Holmes, E. C., & Branley, J. M. (2020). Human pegivirus in brain tissue of a patient with encephalitis. *Diagnostic Microbiology and Infectious Disease, 96*(2).

240. Tumbo, A. M., Schindler, T., Dangy, J. P., Orlova-Fink, N., Bieri, J. R., Mpina, M., . . . Daubenberger, C. (2021). Role of human Pegivirus infections in whole Plasmodium falciparum sporozoite vaccination and controlled human malaria infection in African volunteers. *Virology Journal, 18*(1).

241. Kashnikov, A. Y., Epifanova, N. V., & Novikova, N. A. (2020). Picobirnaviruses: prevalence, genetic diversity, detection methods. *Vavilovskii Zhurnal Genetiki I Selektsii, 24*(6), 661-672.

242. Ganesh, B., Banyai, K., Martella, V., Jakab, F., Masachessi, G., & Kobayashi, N. (2012). Picobirnavirus infections: viral persistence and zoonotic potential. *Reviews in Medical Virology, 22*(4), 245-256.

243. Reddy, M. V., Gupta, V., Nayak, A., & Tiwari, S. P. (2023). Picobirnaviruses in animals: a review. *Molecular Biology Reports, 50*(2), 1785-1797.

244. Delmas, B., Attoui, H., Ghosh, S., Malik, Y. S., Mundt, E., Vakharia, V. N., . . . Consortium, I. R. (2019). ICTV virus taxonomy profile: Picobirnaviridae. *Journal of General Virology, 100*(2), 133-134.

245. Karayel-Hacioglu, I., Gul, B., & Alkan, F. (2022). Molecular characterization of picobirnaviruses in small ruminants with diarrhea in Turkey. *Virus Genes, 58*(3), 238-243.

246. Ghosh, S., & Malik, Y. S. (2020). The true host/s of Picobirnaviruses. *Front Vet Sci, 7*, 615293.

247. Conceicao-Neto, N., Mesquita, J. R., Zeller, M., Yinda, C. K., Alvares, F., Roque, S., . . . Matthijnssens, J. (2016). Reassortment among picobirnaviruses found in wolves. *Archives of Virology, 161*(10), 2859-2862.

248. Berg, M. G., Forberg, K., Perez, L. J., Luk, K. C., Meyer, T. V., & Cloherty, G. A. (2021). Emergence of a distinct Picobirnavirus genotype circulating in patients hospitalized with acute respiratory illness. *Viruses, 13*(12).

249. Zell, R., Delwart, E., Gorbalenya, A. E., Hovi, T., King, A. M. Q., Knowles, N. J., . . . Ictv Report, C. (2017). ICTV virus taxonomy profile: Picornaviridae. *J Gen Virol, 98*(10), 2421-2422.

250. Zell, R., Knowles, N. J., & Simmonds, P. (2021). A proposed division of the family Picornaviridae into subfamilies based on phylogenetic relationships and functional genomic organization. *Archives of Virology, 166*(10), 2927-2935.

251. Kuhn, J. H., Sibley, S. D., Chapman, C. A., Knowles, N. J., Lauck, M., Johnson, J. C., . . . Goldberg, T. L. (2020). Discovery of Lanama Virus, a distinct member of species Kunsagivirus C (Picornavirales: Picornaviridae), in wild vervet monkeys (Chlorocebus pygerythrus). *Viruses-Basel, 12*(12).

252. Lim, E. S., Cao, S., Holtz, L. R., Antonio, M., Stine, O. C., & Wang, D. (2014). Discovery of rosavirus 2, a novel variant of a rodent-associated picornavirus, in children from The Gambia. *Virology, 454*, 25-33.

253. Hansen, T. A., Mollerup, S., Nguyen, N. P., White, N. E., Coghlan, M., Alquezar-Planas, D. E., . . . Hansen, A. J. (2016). High diversity of picornaviruses in rats from

different continents revealed by deep sequencing. *Emerging Microbes & Infections, 5*.

254. Yinda, C. K., Zell, R., Deboutte, W., Zeller, M., Conceicao-Neto, N., Heylen, E., . . . Matthijnssens, J. (2017). Highly diverse population of Picornaviridae and other members of the Picornavirales, in Cameroonian fruit bats. *Bmc Genomics, 18*.

255. Boros, A., Kiss, T., Kiss, O., Pankovics, P., Kapusinszky, B., Delwart, E., & Reuter, G. (2013). Genetic characterization of a novel picornavirus distantly related to the marine mammal-infecting aquamaviruses in a long-distance migrant bird species, European roller (Coracias garrulus). *Journal of General Virology, 94*, 2029-2035.

256. Lu, L., Ashworth, J., Nguyen, D., Li, K. J., Smith, D. B., Woolhouse, M., & Consortium, V. (2021). No exchange of Picornaviruses in Vietnam between humans and animals in a high-risk cohort with close contact despite high prevalence and diversity. *Viruses-Basel, 13*(9).

257. Tan, S. Z. K., Tan, M. Z. Y., & Prabakaran, M. (2017). Saffold virus, an emerging human cardiovirus. *Reviews in Medical Virology, 27*(1).

258. Mbani, C. J., Nekoua, M. P., Moukassa, D., & Hober, D. (2023). The fight against Poliovirus is not over. *Microorganisms, 11*(5).

259. Fontana, S., Cimini, D., Marinelli, K., Gori, G., Moroni, V., Bagnarelli, P., . . . Stefanelli, P. (2021). Survey of diagnostic and typing capacity for enterovirus infection in Italy and identification of two echovirus 30 outbreaks. *Journal of Clinical Virology, 137*.

260. Brouwer, L., Benschop, K. S. M., Nguyen, D., Kamau, E., Pajkrt, D., Simmonds, P., & Wolthers, K. C. (2020). Recombination analysis of non-Poliovirus members of the Enterovirus C species: restriction of recombination events to members of the same 3DPol cluster. *Viruses-Basel, 12*(7).

261. Wang, M. Y., He, J., Lu, H. B., Liu, Y. J., Deng, Y. R., Zhu, L. S., . . . Wang, X. P. (2017). A novel enterovirus species identified from severe diarrheal goats. *PLOS ONE, 12*(4).

262. Fieldhouse, J. K., Wang, X., Mallinson, K. A., Tsao, R. W., & Gray, G. C. (2018). A systematic review of evidence that enteroviruses may be zoonotic. *Emerg Microbes Infect, 7*(1), 164.

263. Khan, H., & Khan, A. (2021). Genome-wide population structure inferences of human coxsackievirus-A; insights the genotypes diversity and evolution. *Infection Genetics and Evolution, 95*.

264. Nayak, G., Bhuyan, S. K., Bhuyan, R., Sahu, A., Kar, D., & Kuanar, A. (2022). Global emergence of Enterovirus 71: a systematic review. *Beni-Suef University Journal of Basic and Applied Sciences, 11*(1).

265. Adamu, H. M., Iliyasu, M. Y., Yakubu, M. N., Samaila, A. B., & Umar, A. F. (2020). In-vitro evaluation of antiviral activity of Moringa oleifera extracts against Polio virus. *Journal of Pharmaceutical Research International, 32*(24), 101-109.

266. Li, J., Zhu, R., Huo, D., Du, Y. W., Yan, Y. X., Liang, Z. C., . . . He, Y. (2018). An outbreak of Coxsackievirus A6-associated hand, foot, and mouth disease in a kindergarten in Beijing in 2015. *Bmc Pediatrics, 18*.

267. Ao, Y., Xu, J., & Duan, Z. (2022). A novel cardiovirus species identified in feces of wild Himalayan marmots. *Infection, Genetics and Evolution, 103*, 105347.

268. Kishimoto, M., Hang'ombe, B. M., Hall, W. W., Orba, Y., Sawa, H., & Sasaki, M. (2021). Mastomys natalensis is a possible natural rodent reservoir for encephalomyocarditis virus. *Journal of General Virology, 102*(3).

269. Naeem, A., Hosomi, T., Nishimura, Y., Alam, M. M., Oka, T., Zaidi, S. S. Z., & Shimizu, H. (2014). Genetic diversity of circulating Saffold Viruses in Pakistan and Afghanistan. *Journal of General Virology, 95*, 1945-1957.

270. Gerhauser, I., Hansmann, F., Ciurkiewicz, M., Loscher, W., & Beineke, A. (2019). Facets of Theiler's Murine Encephalomyelitis Virus-induced diseases: an update. *International Journal of Molecular Sciences, 20*(2).

271. Ugai, S., Iwaya, A., Taneichi, H., Hirokawa, C., Aizawa, Y., Hatakeyama, S., & Saitoh, A. (2019). Clinical characteristics of Saffold Virus infection in children. *Pediatric Infectious Disease Journal, 38*(8), 781-785.

272. Ito, H., Miyagaki, S., Sakaue, S., Matsui, F., Katsumi, Y., Otabe, O., . . . Ohara, Y. (2017). Saffold Cardiovirus infection in a 2-year-old boy with acute pancreatitis. *Japanese Journal of Infectious Diseases, 70*(1), 105-107.

273. Kabuga, A. I., Nejati, A., Soheili, P., & Shahmahmoodi, S. (2020). Human parechovirus are emerging pathogens with broad spectrum of clinical syndromes in adults. *Journal of Medical Virology, 92*(12), 2911-2916.

274. Joseph, L., May, M., Thomas, M., Smerdon, C., Tozer, S., Bialasiewicz, S., . . . Clark, J. E. (2019). Human Parechovirus 3 in infants: expanding our knowledge of adverse outcomes. *Pediatric Infectious Disease Journal, 38*(1), 1-5.

275. Suthar, P. P., Hughes, K., Kadam, G., Jhaveri, M., & Gaddikeri, S. (2023). Human parechovirus meningoencephalitis. *Sa Journal of Radiology, 27*(1).

276. Drexler, J. F., Corman, V. M., Lukashev, A. N., van den Brand, J. M. A., Gmyl, A. P., Brunink, S., . . . Hepatovirus Ecology, C. (2015). Evolutionary origins of hepatitis A virus in small mammals. *Proceedings of the National Academy of Sciences of the United States of America, 112*(49), 15190-15195.

277. Gursoy, E., Kilincer, M., Yanmaz, G., & Yagiz, M. (2022). Hepatitis A seroprevalence and factors affecting Hepatitis A vaccination among healthcare workers in a university hospital. *Viral Hepatit Dergisi-Viral Hepatitis Journal, 28*(3), 103-109.

278. Phan, T. G., Vo, N. P., Simmonds, P., Samayoa, E., Naccache, S., Chiu, C. Y., & Delwart, E. (2013). Rosavirus: the prototype of a proposed new genus of the Picornaviridae family. *Virus Genes, 47*(3), 556-558.

279. Buechler, C. R., Bailey, A. L., Lauck, M., Heffron, A., Johnson, J. C., Lawson, C. C., . . . O'Connor, D. H. (2017). Genome sequence of a novel Kunsagivirus (Picornaviridae: Kunsagivirus) from a wild baboon (Papio cynocephalus). *Genome Announcements, 5*(18).

280. Nguyen, K. D., Chamseddin, B. H., Cockerell, C. J., & Wang, R. C. (2019). The biology and clinical features of cutaneous Polyomaviruses. *Journal of Investigative Dermatology, 139*(2), 285-292.

281. Moens, U., Calvignac-Spencer, S., Lauber, C., Ramqvist, T., Feltkamp, M. C. W., Daugherty, M. D., . . . Ictv Report, C. (2017). ICTV virus taxonomy profile: Polyomaviridae. *J Gen Virol, 98*(6), 1159-1160.

282. Ehlers, B., Anoh, A. E., Ben Salem, N., Broll, S., Couacy-Hymann, E., Fischer, D., . . . Calvignac-Spencer, S. (2019). Novel polyomaviruses in mammals from multiple orders and reassessment of Polyomavirus evolution and taxonomy. *Viruses-Basel, 11*(10).

283. Bagasi, A. A., Khandaker, T., Clark, G., Akagha, T., Ball, J. K., Irving, W. L., & McClure, C. P. (2018). Trichodysplasia Spinulosa Polyomavirus in respiratory tract of immunocompromised child. *Emerging Infectious Diseases, 24*(9), 1744-1746.

284. Ciotti, M., Prezioso, C., & Pietropaolo, V. (2019). An overview on human polyomaviruses biology and related diseases. *Future Virology, 14*(7), 487-501.

285. Kamminga, S., van der Meijden, E., Wunderink, H. F., Touze, A., Zaaijer, H. L., & Feltkamp, M. C. W. (2018). Development and evaluation of a broad bead-based multiplex immunoassay to measure IgG seroreactivity against human Polyomaviruses. *Journal of Clinical Microbiology, 56*(4).

286. Wen, K. W., Wang, L. L., Menke, J. R., & Damania, B. (2022). Cancers associated with human gammaherpesviruses. *Febs Journal, 289*(24), 7631-7669.

287. Lubman, O. Y., Cella, M., Wang, X. X., Monte, K., Lenschow, D. J., Huang, Y. H., & Fremont, D. H. (2014). Rodent Herpesvirus Peru encodes a secreted chemokine decoy receptor. *Journal of Virology, 88*(1), 538-546.

288. McKillen, J., Hogg, K., Lagan, P., Ball, C., Doherty, S., Reid, N., . . . Dick, J. T. A. (2017). Detection of a novel gammaherpesvirus (genus Rhadinovirus) in wild muntjac deer in Northern Ireland. *Archives of Virology, 162*(6), 1737-1740.

289. Iftode, N., Radulescu, M. A., Arama, S. S., & Arama, V. (2020). Update on Kaposi sarcoma-associated herpesvirus (KSHV or HHV8) - review. *Romanian Journal of Internal Medicine, 58*(4), 199-208.

290. Fujii, Y., Doan, Y. H., Wahyuni, R. M., Lusida, M. I., Utsumi, T., Shoji, I., & Katayama, K. (2019). Improvement of Rotavirus genotyping method by using the semi-nested multiplex-PCR with new primer set. *Frontiers in Microbiology, 10*.

291. Uprety, T., Wang, D., & Li, F. (2021). Recent advances in rotavirus reverse genetics and its utilization in basic research and vaccine development. *Archives of Virology, 166*(9), 2369-2386.

292. Liu, J., Lurain, K., Sobuz, S. U., Begum, S., Kumburu, H., Gratz, J., . . . Houpt, E. R. (2015). Molecular genotyping and quantitation assay for rotavirus surveillance. *Journal of Virological Methods, 213*, 157-163.

293. Sadiq, A., Bostan, N., Khan, J., & Aziz, A. (2022). Effect of rotavirus genetic diversity on vaccine impact. *Reviews in Medical Virology, 32*(1).

294. Lappe, B. L., Wikswo, M. E., Kambhampati, A. K., Mirza, S. A., Tate, J. E., Kraay, A. N. M., & Lopman, B. (2023). Predicting norovirus and rotavirus resurgence in the United States following the COVID-19 pandemic: a mathematical modelling study. *Bmc Infectious Diseases, 23*(1).

295. Mo, Q. H., Wang, H. B., Tan, H., Wu, B. M., Feng, Z. L., Wang, Q., . . . Yang, Z. (2015). Comparative detection of rotavirus RNA by conventional RT-PCR, TaqMan RT-PCR and real-time nucleic acid sequence-based amplification. *Journal of Virological Methods, 213*, 1-4.

296. Roczo-Farkas, S., Thomas, S., Bogdanovic-Sakran, N., Donato, C. M., Lyons, E. A., Bines, J. E., & Australian Rotavirus Surveillance, G. (2022). Communicable diseases intelligence Australian Rotavirus surveillance program: Annual report, 2021. *Communicable Diseases Intelligence, 46*.

297. Li, K., Lin, X. D., Huang, K. Y., Zhang, B., Shi, M., Guo, W. P., . . . Zhang, Y. Z. (2016). Identification of novel and diverse rotaviruses in rodents and insectivores, and evidence of cross-species transmission into humans. *Virology, 494*, 168-177.

298. Alsudairy, N. M., Aljameely, S. R. S., Alsaihati, F. M. J., Alkhawfi, A. M. A., Alharthi, M. H. M., Alanka, M. A. S., . . . Aljulajil, F. A. M. (2021). Management of gastroenteritis in primary care - a review. *Journal of Pharmaceutical Research International, 33*(43A), 224-231.

299. Gouvea, V., Glass, R. I., Woods, P., Taniguchi, K., Clark, H. F., Forrester, B., & Fang, Z. Y. (1990). Polymerase chain-reaction amplification and typing of rotavirus nucleic-acid from stool specimens. *Journal of Clinical Microbiology, 28*(2), 276-282.

300. Mankertz, A., Chen, M.-H., Goldberg, T. L., Hübschen, J. M., Pfaff, F., Ulrich, R. G., & Consortium, I. R. (2022). ICTV virus taxonomy profile: Matonaviridae 2022. *Journal of General Virology, 103*(12).

301. Bennett, A. J., Paskey, A. C., Ebinger, A., Pfaff, F., Priemer, G., Hoper, D., . . . Goldberg, T. L. (2020). Relatives of rubella virus in diverse mammals. *Nature, 586*(7829), 424-+.

302. Pfaff, F., Breithaupt, A., Rubbenstroth, D., Nippert, S., Baumbach, C., Gerst, S., . . . Beer, M. (2022). Revisiting Rustrela Virus: new cases of encephalitis and a solution to the capsid enigma. *Microbiology Spectrum, 10*(2).

303. Grimwood, R. M., Holmes, E. C., & Geoghegan, J. L. (2021). A novel Rubi-Like virus in the Pacific electric ray (Tetronarce californica) reveals the complex evolutionary history of the Matonaviridae. *Viruses-Basel, 13*(4).

304. George, S., Viswanathan, R., & Sapkal, G. N. (2019). Molecular aspects of the teratogenesis of rubella virus. *Biological Research, 52*(1).

305. Voss, A., Schlieben, P., Gerst, S., Wylezich, C., Pfaff, F., Langner, C., . . . Mundhenk, L. (2022). Rustrela virus infection - an emerging neuropathogen of red-necked wallabies (Macropus rufogriseus). *Transboundary and Emerging Diseases, 69*(6), 4016-4021.

306. Kamada, M., & Kenzaka, T. (2021). A case of Rubella caused by Rubella vaccination. *Vaccines, 9*(9).

307. Coppeta, L., Ferrari, C., Iannuzzi, I., D'Alessandro, I., Balbi, O., Pietroiusti, A., & Aurilio, M. T. (2020). Rubella immunity among Italian female healthcare workers: a serological study. *International Journal of Environmental Research and Public Health, 17*(21).

308. Purnami, N., Rachmadhan, H. F., Moon, I. S., & Sudaryo, M. K. (2023). A study prevalence of congenital Rubella syndrome cases before and after Rubella vaccination campaign. *Indian Journal of Otolaryngology and Head & Neck Surgery*.

309. Fox, C. H., Johnson, F. B., Whiting, J., & Roller, P. P. (1985). Formaldehyde fixation. Journal of Histochemistry & Cytochemistry, 33(8), 845-853.

310. Anthony, S. J., Epstein, J. H., Murray, K. A., Navarrete-Macias, I., Zambrana-Torrelio, C. M., Solovyov, A., . . . Lipkin, W. I. (2013). A strategy to estimate unknown viral diversity in mammals. *Mbio, 4*(5), e00598-00513.

311. Kocher, D. Cyberduck. Retrieved from https://cyberduck.io/,.

312. Bushnell, B., Rood, J., & Singer, E. (2017). BBMerge - accurate paired shotgun read merging via overlap. *PLOS ONE, 12*(10).

313. Minh, B. Q., Schmidt, H. A., Chernomor, O., Schrempf, D., Woodhams, M. D., von Haeseler, A., & Lanfear, R. (2020). IQ-TREE 2: new models and efficient methods for phylogenetic inference in the genomic era. *Molecular Biology and Evolution, 37*(5), 1530-1534.

314. Rambaut, A. FigTree v1.4.4. Retrieved from https://github.com/rambaut/figtree/releases.

315. CZID. CZID. Retrieved from https://czid.org/.

316. Kalantar, K. L., Carvalho, T., de Bourcy, C. F. A., Dimitrov, B., Dingle, G., Egger, R., . . . DeRisi, J. L. (2020). IDseq-an open source cloud-based pipeline and analysis service for metagenomic pathogen detection and monitoring. *Gigascience, 9*(10).

317. Bolger, A. M., Lohse, M., & Usadel, B. (2014). Trimmomatic: a flexible trimmer for Illumina sequence data. *Bioinformatics, 30*(15), 2114-2120.

318. Buchfink, B., Reuter, K., & Drost, H.-G. (2021). Sensitive protein alignments at tree-of-life scale using DIAMOND. *Nature Methods, 18*(4), 366-368.

319. Li, H. (2018). Minimap2: pairwise alignment for nucleotide sequences. *Bioinformatics, 34*(18), 3094-3100.

320. Souvorov, A., Agarwala, R., & Lipman, D. J. (2018). SKESA: strategic k-mer extension for scrupulous assemblies. *Genome Biology, 19*(1), 153.

321. Wood, D. E., Lu, J., & Langmead, B. (2019). Improved metagenomic analysis with Kraken 2. *Genome Biology, 20*(1), 257.

322. International Committee on Taxonomy of Viruses (ICTV). Retrieved from https://ictv.global/taxonomy/.

323. Islam, A., Hossain, M. E., Islam, A., Islam, S., Rahman, M. K., Hasan, R., . . . Rahman, M. Z. (2023). Epidemiology of Group A rotavirus in rodents and shrews in Bangladesh. *Veterinary Research Communications, 47*(1), 29-38.

324. Sajewicz-Krukowska, J., & Domanska-Blicharz, K. (2016). Nearly full-length genome sequence of a novel astrovirus isolated from chickens with 'white chicks' condition. *Arch Virol, 161*(9), 2581-2587.

325. Li, Y. Q., Ghafari, M., Holbrook, A. J., Boonen, I., Amor, N., Catalano, S., . . . Lemey, P. (2023). The evolutionary history of hepaciviruses. *bioRxiv*.

326. Vargiu, L., Rodriguez-Tomé, P., Sperber, G. O., Cadeddu, M., Grandi, N., Blikstad, V., . . . Blomberg, J. (2016). Classification and characterization of human endogenous retroviruses; mosaic forms are common. Retrovirology, 13(1), 7.

327. McDonald, R. A., Webbon, C., & Harris, S. (2000). The diet of stoats (Mustela erminea) and weasels (Mustela nivalis) in Great Britain. Journal of Zoology, 252, 363-371.

328. Laenen, L., Vergote, V., Calisher, C. H., Klempa, B., Klingström, J., Kuhn, J. H., & Maes, P. (2019). Hantaviridae: current classification and future perspectives. *Viruses, 11*(9).

329. Pybus, O. G., & Thézé, J. (2016). Hepacivirus cross-species transmission and the origins of the hepatitis C virus. Current Opinion in Virology, 16, 1-7.

330. Miller, P. J., Boyle, D. B., Eaton, B. T., & Wang, L.-F. (2003). Full-length genome sequence of Mossman virus, a novel paramyxovirus isolated from rodents in Australia. *Virology, 317*(2), 330-344.

331. Mohd-Qawiem, F., Nawal-Amani, A. R., Faranieyza-Afiqah, F., Yasmin, A. R., Arshad, S. S., Norfitriah, M. S., & Nur-Fazila, S. H. (2022). Paramyxoviruses in rodents: s review. *Open Vet J, 12*(6), 868-876.

332. Munks, M. W., Rott, K., Nesterenko, P. A., Smart, S. M., Williams, V., Tatum, A., . . . Hill, A. B. (2023). Latent CMV infection of Lymphatic endothelial cells is sufficient to drive CD8 T cell memory inflation. *PLOS Pathogens, 19*(1).

333. Chanda, B., Shamimuzzaman, M., Gilliard, A., & Ling, K.-S. (2021). Effectiveness of disinfectants against the spread of tobamoviruses: tomato brown rugose fruit virus and cucumber green mottle mosaic virus. *Virology Journal, 18*(1), 7.

334. Booth, J. G., Hanley, B. J., Hodel, F. H., Jennelle, C. S., Guinness, J., Them, C. E., . . . Schuler, K. L. (2023). Sample Size for Estimating Disease Prevalence in Free-

Ranging Wildlife Populations: A Bayesian Modeling Approach. Journal of Agricultural Biological and Environmental Statistics.

335. Occhibove, F., McKeown, N. J., Risley, C., & Ironside, J. E. (2022). Eco-epidemiological screening of multi-host wild rodent communities in the UK reveals pathogen strains of zoonotic interest. International Journal for Parasitology: Parasites and Wildlife, 17, 278-287.

336. Nicolas, V., Mikula, O., Lavrenchenko, L. A., Šumbera, R., Bartáková, V., Bryjová, A., . . . Bryja, J. (2021). Phylogenomics of African radiation of Praomyini (Muridae: Murinae) rodents: first fully resolved phylogeny, evolutionary history and delimitation of extant genera. *Mol Phylogenet Evol, 163*, 107263.

337. Monchatre-Leroy, E., Boue, F., Boucher, J. M., Renault, C., Moutou, F., Gouilh, M. A., & Umhang, G. (2017). Identification of alpha and beta Coronavirus in wildlife species in France: bats, rodents, rabbits, and hedgehogs. *Viruses-Basel, 9*(12).

338. Vincek, V., Nassiri, M., Knowles, J., Nadji, M., & Morales, A. R. (2003). Preservation of tissue RNA in normal saline. *Laboratory Investigation, 83*(1), 137-138.

339. Nakayama, Y., Yamaguchi, H., Einaga, N., & Esumi, M. (2016). Pitfalls of DNA quantification using DNA-binding fluorescent dyes and suggested solutions. *PLOS ONE, 11*(3), e0150528.

340. Simbolo, M., Gottardi, M., Corbo, V., Fassan, M., Mafficini, A., Malpeli, G., . . . Scarpa, A. (2013). DNA qualification workflow for next generation sequencing of histopathological samples. *PLOS ONE, 8*(6).

341. Wong, R. K. Y., MacMahon, M., Woodside, J. V., & Simpson, D. A. (2019). A comparison of RNA extraction and sequencing protocols for detection of small RNAs in plasma. *Bmc Genomics, 20*.

342. Ball, J. K., & Curran, R. (1997). Production of single-stranded DNA using a uracil-N-glycosylase-mediated asymmetric polymerase chain reaction method. *Analytical Biochemistry, 253*(2), 264-267.

343. Fichet-Calvet, E., Lecompte, E., Koivogui, L., Soropogui, B., Doré, A., Kourouma, F., . . . Ter Meulen, J. (2007). Fluctuation of abundance and Lassa virus prevalence in Mastomys natalensis in Guinea, West Africa. *Vector Borne Zoonotic Dis, 7*(2), 119-128.