THE UNIVERSITY
of EDINBURGH

Metagenomic Surveillance of Viruses at the Human-Animal Interface

Carlijn Bogaardt

Doctor of Philosophy

The University of Edinburgh

2020

Acknowledgements

First, I wish to thank my supervisors – Andrew Rambaut and Mark Woolhouse – for their guidance, support, and patience, and for giving me the opportunity to develop as a scientist.

This PhD would not have been possible without the prior work of a great many VIZIONS collaborators in Vietnam, and at the University of Edinburgh, the University of Oxford, the University of Amsterdam Medical Centre, and the Wellcome Trust Sanger Institute. I extend special thanks to Matthew Cotten and My Phan for providing the sequencing data I worked with, and for initial training and discussions on metagenomic analysis. I have appreciated sharing my VIZIONS journey with Margo Chase-Topping, Dung Van Nguyen, Lu Lu, Gail Robertson, and Jordan Ashworth, and thank them for many helpful discussions. I am grateful for funding by the Wellcome Trust and by the European Union's Horizon 2020 research and innovation programme.

I wish to thank Alasdair Ivens, Sujai Kumar, Tim Booth, Lewis Stevens, the Ashworth Code Monkeys, and the Ashworth Bioinformatics Club for technical support and numerous useful discussions about coding and bioinformatics. Similarly, I extend thanks to Bram Van Bunnik, Gail Robertson and Margo Chase-Topping for valuable help on the statistical front.

A heartfelt thanks to current and past members of Epi- and Rambaut groups, with special mention to Dishon and Liam, for providing an environment of dedication, motivation, kindness and fun. Thanks also to my friends, particularly Sophie, Maike, Ana, Sarah, Tasha.

I am grateful to my managers and colleagues at my current work: to Mandy Walsh, Dilys Morgan, and Jake Dunning for the flexibility with regards to leave and time working from home; to Katherine Henderson for so kindly reading and commenting on my thesis; to all, but particularly Mandy, Kat, as well as Bengü Said, for being there for me when I needed it.

My gratitude also goes to Martin Judd, Erin Moorad and Rose Spencer, for helping me build up the toolkit to manage my work and my health during difficult times.

Last but definitely not least, an immense thank you to my family, without whom this work would never have seen an end. Mom and Laurens, thank you for being there, always; Dad, thank you for your patience and for the support from afar.

Abstract

Zoonotic viruses are a major contributor to emerging infectious diseases, and continuously burden public health systems. Early detection and effective response to viral emergence require an overview of what viruses are circulating in animal hosts, which of these can and do infect at-risk human populations, and which pose the greatest risk of further spread. However, knowledge of such epidemiological patterns is generally biased towards known pathogens of humans and of economically important livestock species. With metagenomic sequencing, one can begin to address these biases by generating a more representative picture of what viruses are present in different host species living in a shared environment.

Vietnam is considered a high-risk setting for the emergence of zoonoses, due to its high population and livestock densities and the prevalence of socio-cultural practices involving frequent close contacts between humans, livestock and wildlife. The Vietnam Initiative on Zoonotic Infections (VIZIONS) was established to improve our understanding of zoonotic emergence in this context. Over 2000 faecal samples and rectal swabs were collected from humans and a variety of farmed animals, and subjected to metagenomic sequencing. In this thesis, I use viral taxonomic classification methods to identify and characterise the viruses present in these samples. I investigate any signals for (putative) zoonotic viruses, and assess whether they could represent emerging public health threats. I also evaluate the roles and challenges of metagenomic surveillance for emerging viruses at the human-animal interface.

The first part of this thesis focuses on the development and testing of a viral taxonomic classification pipeline. I describe the basic steps of this pipeline, and the rationale behind the chosen methods. Next, I test the pipeline on a subset of samples and viruses for which diagnostic quantitative PCR (qPCR) data were available for comparison. Receiver operating characteristic (ROC) curve analysis showed that the pipeline accurately distinguishes qPCR-positive from qPCR-negative samples, and read pair counts correlate well with qPCR cycle threshold values. Investigation of samples with discordant qPCR and metagenomic results indicated that taxonomic misclassification by the pipeline plays a minor role in these discrepancies. Additionally, I found that, for each of the tested viruses, negative samples have variable read pair counts (“background noise”) that correlate with the total number of read pairs assigned to the virus across all samples of the same sequencing run. I hypothesise that

this is due to “index switching”, a form of cross-contamination, and model the association. The findings of these investigations allow me to incorporate additional steps into the pipeline to counteract misclassification, and to use signal thresholds that take into account the effect of index switching cross-contamination.

In the second part of this thesis, I focus on the characterisation of viruses identified with the taxonomic classification pipeline. I present an overview of the mammalian viruses found in samples from humans, swine and rats from Dong Thap province. After removing likely contaminants, I categorize the remaining viruses according to their zoonotic potential. Seven of these viruses are known or generally presumed to be zoonotic; three are only found in the animal study populations, but four – *Rotavirus A*, *Picobirnavirus*, *Human associated cyclovirus 8*, and *Mammalian orthoreovirus* – are shared between human and animal populations. Comparison of signals suggests that viral chatter (*Rotavirus A*) and cross-species transmission within a more generalist ecology (*Picobirnavirus*, *Human associated cyclovirus 8*) are plausible in this setting. Additionally, three putative novel zoonoses are identified, but knowledge gaps hinder extensive interpretation. I evaluate the relevance of these 10 zoonotic and putative novel zoonotic viruses as potential emerging public health threats, and highlight the knowledge gaps that need to be addressed before the risks of these viruses can be properly assessed.

Finally, I interpret my findings in the general context of disease emergence, and evaluate the roles and challenges of viral metagenomics as a tool in the surveillance for emerging infectious disease.

Lay summary

The emergence of new infectious diseases is a continuous challenge in global health. Most newly emerging diseases are zoonoses: they originate in animals, but repeatedly jump into humans, where they may cause disease. For example, Middle East respiratory syndrome is caused by a camel virus, and the natural hosts for Ebola virus disease are believed to be bats. Environments where humans and animals are in frequent close contact with each other provide opportunities for the occurrence of such host jumps.

In the Mekong Delta region in Vietnam, close contacts between humans, livestock and wildlife are common, for example in backyard farms and live-animal markets. In recent decades, Vietnam has seen the emergence of multiple zoonoses, including avian influenza H5N1 (bird flu). To improve our understanding of zoonotic emergence in this context, several international universities partnered up with Vietnamese health authorities and hospitals to form the Vietnam Initiative on Zoonotic Infections (VIZIONS). One of the aims of this collaboration was to describe the diversity of viruses in local humans and animals. Identified viruses could then be compared between populations, and analysed further to investigate patterns of emergence.

To address this aim, over 2000 stool samples were collected from humans and a variety of farmed animals. These samples were subjected to viral metagenomics: a set of techniques to sequence all viral genetic material in a sample, allowing one to get an overview of which viruses are present. An important advantage of using viral metagenomics is that it can reveal the presence of previously unknown viruses, and viruses that are unexpected in a setting. Metagenomics involves both laboratory processes, which were performed by my collaborators, and bioinformatic processes, which form the subject of the studies in this thesis.

The first part of this thesis focuses on the development and testing of bioinformatic methods to be used for the identification of viruses. In Chapter 3, I describe the various data processing steps, and the rationale behind the chosen methods. In Chapter 4, I apply these methods to samples from hospital patients with diarrhoea, and compare the results with those of targeted diagnostic tests that were performed on the same samples. I use this comparison to learn about the performance of the viral metagenomic methods, and to identify aspects

that could be improved upon. As a result, I design several additional processing steps, to include before analysis of a larger sample set.

The second part of this thesis (Chapter 5) focuses on the characterisation of mammalian viruses in samples from humans, pigs and rats, as identified with the adapted viral metagenomic methods. I present an overview of the identified viruses and categorize them according to whether they infect humans and/or other animals. Additionally, I perform further investigations for viruses that are or could be zoonotic: I look at which are shared between humans and animals in this study, assess whether any signals could represent host jumps, and evaluate whether any should be considered emerging public health threats. I also highlight knowledge gaps and suggests directions of further studies for each of these potentially zoonotic viruses.

Finally, in the General Discussion (Chapter 6), I interpret my findings from Chapter 5 in the general context of disease emergence, and evaluate the roles and challenges of viral metagenomics as a tool in the surveillance for emerging infectious disease.

Abbreviations

AdV	adenovirus (or the genus *Mastadenovirus*)
AIDS	acquired immune deficiency syndrome
AiV	Aichivirus
AsV	astrovirus (or the genus *Mamastrovirus*)
BLAST	Basic Local Alignment Search Tool
Cat.	category
cDNA	complementary deoxyribonucleic acid
CRESS DNA	circular Rep-encoding single-stranded deoxyribonucleic acid
C_t	threshold cycle
CyV	cyclovirus (or the genus *Cyclovirus)*
DNA	deoxyribonucleic acid
dsDNA	double-stranded deoxyribonucleic acid
DUO	disease of unknown origin
EMCV	encephalomyocarditis virus
ENA	European Nucleotide Archive
HEV	hepatitis E virus
HIV	human immunodeficiency virus
ICTV	International Committee on the Taxonomy of Viruses
KoV	kobuvirus (or the genus *Kobuvirus*)
LCA	lowest common ancestor
LNA	locked nucleic acid
MERS-CoV	Middle East respiratory syndrome-related coronavirus
NBC	Naïve Bayes Classifier
NBV	*Nelson Bay orthoreovirus*

NCBI	National Center for Biotechnology Information
NoV	norovirus (or the genus *Norovirus*)
nt	nucleotide(s)
OTU	operational taxonomic unit
OUCRU	Oxford University Clinical Research Unit
PBV	picobirnavirus (or the genus *Picobirnavirus*)
PCR	polymerase chain reaction
qPCR	quantitative polymerase chain reaction
R_0	basic reproduction number
RNA	ribonucleic acid
RoV	rotavirus (or the genus *Rotavirus*)
rp	read pair(s)
rRNA	ribosomal ribonucleic acid
RVA	*Rotavirus A*
RVC	*Rotavirus C*
S. suis	*Streptococcus suis*
SARS	severe acute respiratory syndrome
SaV	sapovirus (or the genus *Sapovirus*)
SFTS	severe fever with thrombocytopenia syndrome
ssDNA	single-stranded deoxyribonucleic acid
TBE	tick-borne encephalitis
UK	United Kingdom
US	United States
VIDISCA	Virus Discovery by cDNA Amplified Fragment Length Polymorphism
VIZIONS	Vietnam Initiative on Zoonotic Infections
VSBV1	*Variegated squirrel bornavirus 1*

Contents

List of figures

List of tables

Chapter 1. General introduction

I wrote this chapter with minor comments and text edits from Andrew Rambaut and Mark Woolhouse.

1.1 Emerging zoonotic viruses as a public health problem

Despite an important reduction in their global impact on human mortality and morbidity in recent decades, infectious diseases remain a major threat to global public health. In 2016, lower respiratory tract infections, diarrhoeal diseases, malaria and HIV/AIDS were among the ten leading causes of total years of life lost worldwide (G. B. D. Causes of Death Collaborators, 2017). In addition to the constant threat posed by endemic infectious diseases, epidemic-prone infectious diseases can cause large, abrupt increases in morbidity and mortality (G. B. D. Causes of Death Collaborators, 2017). Outbreaks can lead to overstretching of public health infrastructure and ultimately to public health emergencies with far-reaching direct and indirect impacts on the affected areas. For example, the overall economic and social burden of the recent outbreak of Ebola virus disease in West-Africa (2014-2016) has been estimated at $53.19 billion US dollars, largely borne by the directly affected countries. Illustrating the scale of indirect impacts, the mortality from *non*-Ebola causes was estimated to be the largest contributing cost (Huber et al., 2018).

While pathogens have a variety of prevalence patterns and geographical distributions, a common concern is that their epidemiology might change, potentially resulting in increases in their burden of disease. This is known as “disease emergence” and can refer to several different dynamics. First, a pathogen may be emerging in humans as a new host species. An example of this is Middle East respiratory syndrome coronavirus, a camel virus that was first identified in humans in Saudi Arabia in 2012 (Memish et al., 2014, Zaki et al., 2012, Azhar et al., 2014) and has since caused several hundreds of human infections each year (WHO EMRO, 2019). Secondly, a pathogen may be emerging in a new geographical area. This is particularly applicable to vector-borne viruses, for example West Nile and Zika viruses spread to the Americas in 1999 and 2013 respectively (Faria et al., 2016, Zanluca et al., 2015, Garmendia

et al., 2001, Lanciotti et al., 1999). Finally, an emerging pathogen may be significantly expanding in incidence in a known host population, as illustrated by Lyme borreliosis in Europe and North America over the last few decades (Mysterud et al., 2016, Spielman, 1994, Steere et al., 2004). Overall, a landmark study of trends in disease emergence found that, disconcertingly, events associated with disease emergence had increased in frequency over time during the study period (1940-2004) (Jones et al., 2008).

Emerging infectious diseases come with specific challenges. If a pathogen is newly emerging in humans, control efforts are likely to be hindered by limited biological and epidemiological knowledge: effective diagnostics, vaccines, or treatment options may not exist, and it may be unclear how to stop transmission. Additionally, if a human-transmissible pathogen is expanding into a new, immunologically naïve population, susceptibility will be high, and the pathogen may spread rapidly, turning control or elimination into an unattainable goal. To facilitate a rapid response and limit the impact of emergence events, it is important to improve our understanding of what pathogens are likely to emerge, so that targeted preparedness plans can be put in place.

Comparative studies of all known human pathogen species have found that viruses are greatly over-represented among emerging infections. Relative to other pathogens, they are four times as likely to be considered as emerging (Taylor et al., 2001, Woolhouse and Gowtage-Sequeria, 2005, Cleaveland et al., 2001). Their rapid evolvability is generally believed to play an important role: with high mutation rates (Domingo and Holland, 1994) and short generation times, they may be able to adapt to new ecological niches (e.g. host species or tropic tissues) more quickly than bacteria or macroparasites. Furthermore, the scarcity of effective antiviral agents (cf. the wide availability of broad spectrum antibiotics and anthelminthics) makes viral diseases more difficult to treat, and Cleaveland et al. (2001) have suggested that this may contribute to a higher risk of spreading and eventually reaching "emerging" status.

Another group of pathogens that was identified as being disproportionately likely to emerge is the zoonoses, i.e. infections transmissible between humans and animals. Zoonoses have double the risk of non-zoonoses to be labelled as emerging (Taylor et al., 2001, Woolhouse and Gowtage-Sequeria, 2005). In concordance, about 60% of the emergence events studied by Jones et al. (2008) were associated with zoonoses, and the authors particularly noted a marked increase over time in events linked to pathogens originating in wildlife.

This thesis focuses on zoonotic viruses. In particular, it highlights those viruses that are known to repeatedly cross the human-interface, jumping from animals into humans, and those for which this appears a plausible dynamic. This is a very diverse grouping, with varying degrees of pathogenicity in humans and varying relative importance of zoonotic versus human-to-human transmission. However, importantly, it includes a number of high profile pathogens with very high case fatality rates: 25-90% for Ebola virus disease outbreaks (World Health Organization, 2018b), 40-75% for Nipah virus (World Health Organization, 2018b), and nearly 100% for rabies. Additionally, viruses originating in animals have caused pandemics responsible for millions of deaths worldwide: influenza pandemics are caused by novel reassortant viruses that combine genome segments from human, avian and/or swine influenza viruses (Kawaoka et al., 1989, Scholtissek et al., 1978, Smith et al., 2009); and HIV/AIDS, while now transmitted exclusively between humans (and thus not strictly a zoonosis), evolved from non-human primate viruses transferred to humans through contact with bushmeat (Gao et al., 1999, Gao et al., 1992). The public health relevance of zoonotic viruses is illustrated further by their near-exclusive presence on the World Health Organization Research & Development Blueprint list of priority diseases, which highlights "severe emerging diseases with potential to generate a public health emergency, and for which insufficient or no preventive and curative solutions exists" (World Health Organization, 2018a).

1.2 The ecology of emergence of zoonotic pathogens

Given the threat posed by zoonotic viruses, many studies have been conducted to elucidate the drivers and mechanisms behind their emergence. Zoonotic emergence can be considered as a combination of multiple processes: changing dynamics of animal populations and their viruses; introduction of an infection into humans as a novel host species; and further spread within the human population. Comparative studies (Jones et al., 2008, Allen et al., 2017) and investigations of particular emergence events (such as reviewed in Morse (1995), Keesing et al. (2010), and Weiss and McMichael (2004)) have identified ecological drivers that act through changes in these various dynamics.

1.2.1 Dynamics in animals and the "zoonotic pool"

Dynamics of animal populations shape the supply of animal viruses ("zoonotic pool") that could ultimately emerge in humans (Morse, 1993). There are two main aspects to this, the

first of which is abundance: emergence is often the result of increased population sizes of pathogens and their natural hosts. Such increases in abundance have been linked to a variety of ecological factors (Morse, 1995). In wildlife host species, land use changes (e.g. de- and reforestation, shifting agricultural practices) and weather or climate phenomena can affect their habitat and alter the success of their life cycle. The emergence of Argentine haemorrhagic fever was brought about by such environmental drivers: the expansion of maize agriculture in the Argentinian pampa resulted in *Calomys musculinus*, the reservoir host of Junín virus, becoming the dominant rodent in the area (Johnson, 1993). Similarly, expansion of water bodies (e.g. through dam building) and the use of open containers for water storage in cities promote mosquito populations, and the propagation of mosquito-borne diseases. Additionally, today's high volume of international travel and commerce can open up new ecological niches (and thus new opportunities for population growth) for animals and the pathogens they carry. For example, the importation of African rats into the US, and subsequent contacts with and infections in prairie dogs, led to a monkeypox outbreak in the US in 2003 (Centers for Disease Control and Prevention, 2003b). Finally, for livestock species, the combination of intensive (high-density) farming and frequent animal movements means that any infection can rapidly spread within and between herds, further increasing pathogen abundance.

The second aspect is mammal biodiversity, and the opportunities this creates for viral diversification. In spatial models, mammal biodiversity has been directly linked to the emergence of pathogens originating from wildlife (Jones et al., 2008, Allen et al., 2017). There is evidence that interspecies contacts and transmissions play an important part in this: sympatry (overlap of geographical range between multiple host species) is the most important predictor for overall viral richness in mammals (Luis et al., 2013, Olival et al., 2017), zoonotic viral richness in bats and rodents (Luis et al., 2013), and viral pathogen sharing in primates (Davies and Pedersen, 2008). Cross-species transmissions and subsequent emergence of infections in new animal hosts may represent an intermediate stage before their appearance in humans: new animal host species can act as amplifying hosts (generating higher viral loads), or as ecological and/or adaptive bridging hosts. For example, while palm civets in Chinese markets were found to be infected with SARS coronavirus (Kan et al., 2005) and have been speculated to represent the origin of the first human infections, this species is unlikely to be the natural reservoir host. Instead, contact between different animal species at live animal markets probably resulted in an artificial infection cycle, with palm civets

making an ideal bridging host because of their high susceptibility to the virus and their wide distribution in markets and restaurants (reviewed in Wang and Eaton (2007)). Similarly, in the emergence of Nipah virus, pigs played a role as amplifying hosts, having acquired the infection from bats in their farm yards (Chua et al., 2000). Swine have also been proposed to have acted as "mixing vessels" in the generation of pandemic influenza viruses, through reassortment between different human, swine and avian lineages all circulating within porcine hosts (Ito et al., 1998, Scholtissek, 1990). While more recent evidence suggests that this was not the case for the three 20th century pandemics (Nelson and Worobey, 2018), reassortment of diverse lineages in swine was indeed at the origin of the 2009 pandemic (Smith et al., 2009). Altogether, environments and agricultural practices where multiple animals come together pose a particular risk for the emergence of novel viruses.

1.2.2 The human-animal interface: introduction of zoonotic pathogens into the human population

Contact between humans and animals (or animal products) is essential for the introduction of animal viruses into human populations. It follows that environmental and societal changes that promote such contacts are important ecological drivers of emergence. Chief among these is land-use changes (Woolhouse and Gowtage-Sequeria, 2005, Allen et al., 2017): urbanisation, deforestation, agricultural development all involve humans encroaching into animal habitats; additionally, fragmentation of these habitats and easy availability of food in settled areas may result in wildlife species developing more peridomestic behaviours. However, the drivers of zoonotic disease emergence can be more complex, and act in synergy. An interesting example is the upsurge of tick-borne encephalitis (TBE) in Baltic and Central-European states in the early 1990s, after the collapse of the Soviet Union (Sumilo et al., 2007, Sumilo et al., 2008). While the upsurge was widely attributed to climate change, Sumilo et al. (2007) found that in fact climate change did not explain the observed spatial heterogeneities. The authors developed an alternative model in which the shift in political and socio-economic situations resulted in both a conversion of agricultural land to shrub land and forests, favouring ticks and their hosts, and changes in human behaviour, involving more recreational and poverty-driven foraging visits to such forests (Sumilo et al., 2007). In a subsequent study, the authors confirmed a striking correlation between poverty, household expenditure on food, and changes in TBE prevalence, across multiple countries (Sumilo et al., 2008).

The risk of getting infected with an emerging zoonosis is not equally distributed within populations. People with a weakened immune system, such as children, pregnant women, the elderly, and the immunocompromised, generally have an increased risk of infections that are normally easily cleared by the immune system. Children's frequent hand-to-mouth contact and lack of understanding of the role of hygiene in preventing disease also contribute to their increased risk. Additionally, people with occupational or residential exposures, such as butchers, animal health workers and farming households, are at higher risk through their frequent and intense contacts with animals. For example, in a sero-epidemiological study, farmers, veterinarians and meat processing workers were shown to have greatly increased antibody titres against swine influenza viruses, indicating prior infections (Myers et al., 2006). Another risk group is formed by individuals engaging in risky food practices: the preparation or consumption of meat from diseased animals, or dishes containing raw or undercooked meat or blood. Such food-related risk factors have been identified for influenza A H5N1 infections and *Streptococcus suis* meningitis in Vietnam (Dinh et al., 2006, Nghia et al., 2011). A final group includes people exposed to wildlife or bushmeat, through hunting, farming, trading, slaughtering, or consumption. Bushmeat and the wildlife trade have been linked with various outbreaks of zoonotic infections globally, a role which has previously been reviewed in the literature (Daszak et al., 2007, Karesh et al., 2005, Bell et al., 2004, Wolfe et al., 2005a).

1.2.3 Establishment and further spread of emerging zoonotic pathogens

The final element of emergence is the further spread of pathogens recently introduced into a human population. The transmission potential of an infection, traditionally described by the basic reproduction number (R_0, the expected number of secondary cases generated by a primary case in a totally susceptible population), is dependent on three factors: contact rates, the probability of transmission per contact, and duration of infectiousness. Contact rates are high in settings with high population densities, such as urban environments. The connectedness of urban centres, to surrounding rural areas as well as to other urban centres, also facilitates the spread of pathogens between different human populations. Other factors affect both contact rates and transmission probability: high-risk social behaviour (e.g. sex work and intravenous drug use), transmission through medical settings (nosocomial and iatrogenic routes), and breakdown of sanitation or public health infrastructure (Morse, 1995).

The establishment and dissemination of emerging infectious diseases is particularly favoured by simultaneous changes in several of these factors. This is illustrated by the early history of HIV-1 group M (pandemic HIV). While the original zoonotic transfer event likely took place in southeast Cameroon, the virus was inferred to have emerged in epidemic form in Kinshasa in Zaire (now Democratic Republic of Congo) (Sharp and Hahn, 2008, Worobey et al., 2008, Faria et al., 2014). After arrival of the virus in Kinshasa (~1920), the growing urban population and an active transportation network allowed it to spread within the city and to other population centres in the region (Faria et al., 2014, Worobey et al., 2008). But, it is believed that iatrogenic transmission via unsterilized injections in the 1950s and/or changes in commercial sex work in the early 1960s are at the basis of a transition to a much faster growth rate of the virus (Faria et al., 2014). Together, these factors permitted the eventual epidemic and pandemic spread of the virus. Other examples are provided by the recent quick spread of cholera (Khan and Shahidullah, 1982, Centers for Disease Control and Prevention, 2003a, Cowman et al., 2017, The Lancet Infectious Diseases, 2017, Golicha et al., 2018) and Ebola virus (Claude et al., 2018, Dyer, 2018) in regions affected by war and civil unrest, due to densely populated camps with only basic sanitation, and security issues hindering control efforts.

1.2.4 Vietnam as a high-risk setting

In recent years, Vietnam has been affected by a variety of emerging zoonoses: avian influenza A H5N1 (Dinh et al., 2006), severe acute respiratory syndrome (SARS) (Reynolds et al., 2006), porcine-like human rotaviruses (Kaneko et al., 2018, My et al., 2014b), and meningitis due to *S. suis* (Wertheim et al., 2009). Additionally, Vietnam has a substantial burden of disease of which the aetiology cannot be determined (Thompson et al., 2015, Ho Dang Trung et al., 2012). A significant part of this could reflect zoonotic pathogens not yet known to infect humans, or perhaps still unknown to science. Here I briefly discuss some of the factors that make Vietnam a high-risk setting for the emergence of zoonotic infections.

Firstly, approximately two thirds of the Vietnamese population live in rural areas (GSO, 2015), where human-animal contact is part of many people's domestic and professional lives. Farming, including animal husbandry, is the predominant source of livelihood in these areas. A large proportion of this industry is small-scale, with animals kept in the backyard, adjacent to the living quarters; biosecurity measures are limited (Rabaa et al., 2015). Similarly, abattoir workers and meat traders in wet markets operate with minimal basic hygiene, poor cold

chains, and limited meat inspections (Rabaa et al., 2015). In addition, raw or undercooked blood and meat are readily consumed (Rabaa et al., 2015). Such exposures to livestock, via farming, food preparation or high-risk food practices, are known risk factors for cross-species transmission of various zoonotic emerging infections in Southeast Asia (Dinh et al., 2006, Mackenzie, 2005, Nghia et al., 2011). Furthermore, wildlife species are also farmed and consumed in Vietnam (Brooks et al., 2010, Drury, 2009, Wildlife Conservation Society, 2008). With conditions and behaviours as described here, there is ample opportunity for the direct transfer of zoonotic pathogens from either domestic animals or wildlife into human populations.

Secondly, the abundance and diversity of animals in Vietnam suggest that the country harbours a large diversity of animal viruses. Vietnam is rich in mammalian wildlife species (Ceballos and Ehrlich, 2006), but also has one of the highest livestock densities in Southeast Asia (Gerber et al., 2005, FAO). Contacts between distinct animal populations may be common in the wild as well as at wet markets or backyard farms: these often involve multiple different farmed species, and may be open to visits from peridomestic mammals (rodents or bats) or birds. Such opportunities for contact between animal species may favour the exchange of viral genetic material, resulting in novel recombinants or reassortants, of which some might emerge into humans.

Thirdly, Vietnam has a recent history of extensive changes in land use relating to urbanisation, shifts in agricultural crops, and de- and reforestation (Van Dijk et al., 2013). As described above, land use changes can alter the dynamics of animal species, and increase the proximity between animals and humans. They have been found to be strongly associated with emergence of zoonoses from wildlife (Allen et al., 2017, Woolhouse and Gowtage-Sequeria, 2005). In fact, Vietnam is highlighted on maps visualising the risk of such emergence events (Allen et al., 2017, Jones et al., 2008).

Finally, Vietnam has a large human population (91.7 million in 2015), with high population densities particularly in the Red River Delta, South East and Mekong Delta regions (GSO, 2015). Once an emerging pathogen has developed the ability to directly transmit between humans, in densely populated Vietnam this could thus well result in an epidemic.

Altogether these conditions illustrate why Vietnam can be considered a high-risk setting for the emergence of zoonotic infections. More comprehensive reviews of the complex and

interlinked drivers for disease emergence in Southeast Asia can be found elsewhere (Coker et al., 2011, Horby et al., 2013).

1.3 Viral risk factors for emergence

To be able to provide early warning of emergence events, not only is it important to know in what kind of setting to look, but also what viruses to look for. Different viruses come with distinct challenges and preparedness and control measures. Scientists have called for more predictive studies, so that surveillance and preparedness plans can be targeted to specific viruses (Pulliam, 2008, Daszak, 2009, Morse et al., 2012).

Epidemiological studies have identified several ecological and virological traits associated with zoonotic potential or emergence in humans. Overall, a consistently identified risk factor is the breadth of the viral host range (Olival et al., 2017, Woolhouse and Gowtage-Sequeria, 2005), presumably reflecting the use of conserved cellular receptors and replication machinery. Another host-related risk factor is the ability to infect wildlife host species (Cleaveland et al., 2001). This reflects the recently increasing exposure of humans to viruses from wildlife hosts, resulting from human expansion into wildlife territory (Cleaveland et al., 2001, Daszak et al., 2000, Osburn, 1996). With regard to transmission routes, several studies highlighted vector-borne viruses as more likely to be zoonotic and/or emerging (Taylor et al., 2001, Olival et al., 2017, Loh et al., 2015). However, vector-borne RNA viruses were also found to be less likely to transmit efficiently among humans (i.e. exhibit epidemic potential), compared to viruses with other transmission routes (Woolhouse et al., 2013, Woolhouse and Adair, 2013, Geoghegan et al., 2016). Different transmission routes appear to be associated with emergence events related to different ecological drivers (Loh et al., 2015). As regards virological properties, an RNA genome (Luis et al., 2013, Olival et al., 2017) and the ability to replicate in the cytoplasm (Olival et al., 2017, Pulliam and Dushoff, 2009, Luis et al., 2013) are associated with zoonotic potential. These likely reflect high mutation rates allowing for rapid diversification, and the ability to bypass complex and host-specific machinery for transport into the nucleus, respectively. These various risk factors, while providing valuable insights, are rather general in nature, and studies have not been able to highlight specific viral taxa to be alert to. As Woolhouse and Gowtage-Sequeria (2005) noted: “the most striking feature of emerging and re-emerging pathogens is their diversity”.

The usefulness of epidemiological studies has been limited by the complexity of emergence (i.e. different definitions and processes) and the diversity of pathogen types and ecological drivers at play. Studies have sought answers to a multitude of related, but subtly different, questions: they have used different definitions of emergence or transmissibility, different groups of pathogens, different comparison groups, different properties of interest, and/or different comparative methods. As a result, they may have obtained insights that are valid for specific pathogen types (or stages of emergence) but not generalizable, or vice versa.

Another drawback of these studies is that they crucially depend on correct classification of explanatory and outcome variables, but that this information is not always available and it is often biased. For example, studies are necessarily based on *known* viruses only – but these are biased towards pathogens of humans or domestic animals, with an underrepresentation of non-pathogenic viruses and viruses of wildlife. Similarly, among human pathogenic viruses, non-transmissible viruses are less likely to be recognised and classified correctly than human-transmissible viruses, as single cases are more easily missed than outbreaks.

In summary, epidemiological studies have yielded some insights into the properties that make a virus more likely to be zoonotic or deemed as emerging, but due to the complexity of the processes involved in emergence, they cannot predict which specific viruses will emerge next.

1.4 Surveillance

With it being impossible to predict exactly which viruses will emerge next, the best alternative is rapid detection. This requires targeted surveillance, in high-risk populations and settings. Given the predominance of zoonoses among emerging infectious diseases, the human-animal interface should be an important focus for such surveillance.

Surveillance of viruses in settings where humans and animals live closely together has two main forms, each with different functions. First, disease surveillance involves the continuous monitoring and analysis of patterns of disease in humans and animals. This can reveal changes that may reflect the local emergence of pathogens. Early detection can lead to rapid interventions and the prevention of emergence at the global scale. Secondly, screening of high-risk individuals and of key animal host species enables the characterisation of the diversity of pathogens and potential pathogens that humans are exposed to. This includes animal viruses not currently known to infect humans but that could gain zoonotic potential

in future. A knowledge base of what viruses have been found in which host species, in which geographical locations, and whether they were associated with clinical disease, may inform risk assessments and improve the timeliness of pathogen identification in cases of disease of unknown origin (potential zoonotic emergence events).

1.4.1 Disease surveillance in humans and animals

The usefulness of disease surveillance in the investigation of emerging human pathogens is illustrated by the identification of Alongshan virus, a novel flavivirus, in patients with a history of tick bites and febrile illness of unknown origin in China (Wang et al., 2019). After detection of the index case and molecular characterisation of the virus, a heightened surveillance programme for tickborne disease was set up, leading to the identification of additional cases. This allowed a more detailed characterisation of the clinical syndrome, and further pathological and epidemiological investigations.

Disease surveillance in animals similarly provides information about animal diseases, but in the case of zoonotic diseases, this surveillance has additional relevance for human health. Some viruses, including important zoonotic pathogens, do not visibly affect the health of their reservoir hosts but cause significant disease in secondary host species. Secondary hosts can then be used as sentinels to alert public health authorities of the circulation of these zoonotic pathogens. For example, both mass deaths in crows and neurological disease in horses have been used as sentinel systems for the circulation of West Nile virus (Eidson et al., 2001, Saegerman et al., 2016). Ebola virus disease outbreaks in Central Africa have also been preceded by mortality in wildlife (chimpanzees, gorillas, and duikers) (Bermejo et al., 2006, Georges et al., 1999, Leroy et al., 2004, Rouquet et al., 2005), and it has been suggested that monitoring wildlife population indices could provide early warning of virus circulation (Leroy et al., 2004, Rouquet et al., 2005). In the Republic of Congo, conservation scientists have teamed up with the Ministry of Health to implement a low-cost community-based wildlife mortality surveillance network for early detection of Ebola outbreaks, covering ca. 50,000 km^2 in at-risk areas (Kuisma et al., 2019). This network relies on educational outreach visits to local communities; reporting of wildlife carcasses by these communities and national park rangers; and rapid sampling of carcasses by trained teams based across the region (Kuisma et al., 2019). Between 2006 and 2018, 58 carcasses were reported and investigated, with turnaround times as low as 3-4 days after establishment of in-country diagnostic capacity.

The limitations of disease surveillance, as practised in most settings, are two-fold. First, it is biased towards diseases of humans and economically important livestock. Wildlife populations are not easily accessed, and there are often no government agencies dedicated to active disease surveillance in wildlife (Epstein and Anthony, 2017). Secondly, it often does not have the granularity and timeliness to detect and respond to single zoonotic spillover events.

1.4.2 Screening of high-risk individuals

To maximise the probability of detecting zoonotic spillover events, surveillance strategies should also include healthy individuals who have frequent close contact with animals. Samples could also be collected from contact animals. If unexpected animal viruses are identified in these high-risk individuals, comparison of viral sequences with those from sampled animals may provide confirmation of potential spillover events. Additionally, information on specific risk behaviours can facilitate the identification of routes of transmission and suggest potential control measures. However, it is costly, not logistically feasible, and ethically questionable to subject large, ill-defined risk groups to continuous surveillance through public health schemes. Surveillance in these populations is thus limited to shorter-term monitoring of defined cohorts, or to cross-sectional surveys, performed in the context of academic or collaborative initiatives.

An example is provided by the studies by Wolfe et al. (Wolfe et al., 2005a, Wolfe et al., 2005b, Wolfe et al., 2004a), in which bushmeat hunters and butchers from rural communities in Cameroon were screened for simian retroviruses. They found evidence of multiple lineages of simian foamy viruses and primate T-lymphotropic viruses, obtained through independent zoonotic spillover events (Wolfe et al., 2005b, Wolfe et al., 2004a). Based on these findings, the authors hypothesised that cross-species transmissions are widespread in settings where humans have intense contact with animals, but that the great majority of these cases do not result in onwards transmission; they termed this phenomenon “viral chatter” (Wolfe et al., 2004a, Wolfe et al., 2004b, Wolfe et al., 2005a). They posed that high rates of viral chatter lead to humans being exposed to a large genetic diversity of viruses, with this increasing the probability that, ultimately, one variant will be able to successfully transmit between humans and perhaps emerge in a more widespread manner (Wolfe et al., 2004b, Wolfe et al., 2005a).

1.4.3 Screening of key animal species

Many infections do not cause visible disease in animal hosts, and, to provide a more comprehensive overview of viruses that humans are exposed to, surveillance strategies should thus also include healthy animals (Levinson et al., 2013). Cost and logistical considerations preclude the continuous surveillance of a large variety of animals, hence surveillance in healthy animals generally takes the form of one-off studies, targeted towards key host species.

Which animals play these key roles? The great majority of zoonotic pathogens have an origin in mammals, and to a lesser extent birds (Cleaveland et al., 2001). Among these, wildlife species represent the predominant, and an increasingly important, source of emerging viruses (Cleaveland et al., 2001, Jones et al., 2008). Host species that are more closely related to humans have higher proportions of zoonotic viruses (Davies and Pedersen, 2008, Olival et al., 2017), indicating that chimpanzees and other non-human primates may be useful species to monitor. Bats and rodents have also been noted as important hosts for zoonotic viruses (Luis et al., 2013, Olival et al., 2017, Johnson et al., 2015); this may in part relate to their abundance and relatively frequent contacts with humans. They have been the subject of many recent studies searching for novel potential zoonoses (Berto et al., 2017a, Phan et al., 2011, Sachsenroder et al., 2014, Wu et al., 2018, Wu et al., 2016, Yinda et al., 2018). The importance of human-animal contact is also seen in domestic animals, where the abundance of zoonotic pathogens has been associated with time since domestication (Morand et al., 2014). Several studies have highlighted ungulates as having a high viral richness and as hosting the most zoonotic pathogen species and emerging infections (Olival et al., 2017, Woolhouse and Gowtage-Sequeria, 2005). Swine are well studied, because of their role as potential mixing vessels for influenza viruses (Ma et al., 2008) and amplifying hosts for other zoonotic pathogens (Simpson et al., 1976, Chua et al., 2000).

1.4.4 The Wellcome Trust Vietnam Initiative on Zoonotic Infections (VIZIONS)

The studies described in this thesis were performed in the context of surveillance of emerging zoonotic viruses in Vietnam. This was done as part of a larger initiative: the Wellcome Trust Vietnam Initiative on Zoonotic Infections (VIZIONS). The aims of VIZIONS included the characterisation of viral diversity on both sides of the human-animal interface, the

identification of potential zoonotic emergence events, and the investigation of risk factors associated with such events.

VIZIONS homed in on zoonotic viruses with two large surveillance studies, that are described in detail in Chapter 2. Briefly, the first was a disease surveillance study in multiple hospitals, in which patients reporting with any of four common disease syndromes (diarrhoea, respiratory disease, central nervous system disease, and jaundice) were sampled and asked about their potential risk behaviours (including animal contacts) in the previous three months. The second study was a prospective cohort study, in which individuals with frequent residential or occupational contact with animals were followed over three years. Samples were also taken from animals associated with this high-risk cohort.

This thesis describes the search for viruses in these samples, using a viral metagenomics approach.

1.5 Viral metagenomics and its roles in the study of zoonotic emergence

1.5.1 What is metagenomics?

Metagenomics is the application of modern genomics techniques to the study of the entirety of genetic material from environmental samples. It involves the sequence-independent sequencing of any nucleic acids present in a sample, in a (theoretically) unbiased manner, and without requiring prior cultivation of any particular target organism. The term metagenomics also covers subsequent computational analyses, which often involve taxonomic classification procedures and aim to characterise the biodiversity in an environment of interest.

Metagenomics is a relatively young discipline that has its roots in microbial biodiversity studies. It builds on DNA barcoding: the direct sequencing and phylogenetic analysis of universally conserved genes, such as the bacterial 16S rRNA gene, for the characterisation of biodiversity in environmental samples. By sequencing directly from samples, such studies revealed an immense microbial biodiversity that had been missed by traditional, culture-based microbial genomics (Hugenholtz et al., 1998). This insight was paired with nascent shotgun sequencing methods to give birth to the field of metagenomics. Since then, it has

benefitted from the advent of next-generation sequencing technologies and subsequent reductions in costs of large-scale sequencing projects.

The focus in this thesis is on *viral* metagenomics: the application of metagenomic techniques to the viral component of a sample's genetic material. The various roles of viral metagenomics in the study of zoonotic emergence have been described below, starting with its part in surveillance, which is most relevant to this thesis.

1.5.2 Metagenomic surveillance

The aims of viral metagenomic surveillance are to characterise the diversity of viruses present in a setting (here, the human-animal interface), and to screen this information for epidemiological relevance.

The specific advantages of metagenomics as a surveillance tool stem from its ability to capture genetic material from *any and all* viruses, including those that one may not have known about or expected in a sample. It can identify rare or atypical pathogens, viruses with no known association with disease, novel viruses, unculturable viruses, and co-infections. It does not require *a priori* hypotheses about circulating viruses and replaces a battery of tests with a single process. This ability makes metagenomics a useful diagnostic tool, and ideally suited for viral diversity studies – two cornerstones of surveillance.

Viral metagenomics has been used in disease surveillance to gain more comprehensive insights into the diversity of pathogens associated with specific syndromes (Finkbeiner et al., 2008, Yang et al., 2011, Victoria et al., 2009). If applied at a larger scale, metagenomic surveillance could result in more accurate estimations of pathogens' relative contributions to the burden of disease. Similarly, if studies are repeated over time, changes in patterns may highlight the emergence of viruses that may be missed by standard surveillance targeting only known pathogens.

Additionally, as a sequence-based technology, metagenomics allows differentiation of viral variants, and has been used in conjunction with molecular epidemiological and phylogenetic methods to follow the spread of variants through populations in space and time. An interesting example is the investigation of a large increase in Lassa fever cases in Nigeria in 2018: while it was feared that a novel human-transmissible strain had emerged, phylogenetic analysis showed that the viral sequences covered a large genetic diversity and were

interspersed with sequences from previous years, implicating independent spillover events from rodent hosts as the major driver of the outbreak (Kafetzopoulou et al., 2019).

Metagenomics has also been used in the screening of animal host species that are key sources of zoonotic pathogens (Wu et al., 2018, Firth et al., 2014, Sachsenroder et al., 2014, Yinda et al., 2018, Phan et al., 2011, Wu et al., 2016, Shan et al., 2011), arthropod vectors (Bouquet et al., 2017, Brinkmann et al., 2016, Atoni et al., 2018, Xiao et al., 2018), and, to a lesser extent, humans considered at high risk of zoonotic infections (Yinda et al., 2019). The detection of closely related pathogen sequences in humans and geographically linked animal host and vector populations can identify or hint at local zoonotic transmission cycles (Takhampunya et al., 2019, Phan et al., 2016a). Similarly, detection of (near-)identical pathogen sequences in patients and their contact animals is valuable evidence when investigating these animals as potential sources of infection (Hoffmann et al., 2015).

1.5.3 Mystery outbreaks and viral discovery

Another application of viral metagenomics that is particularly relevant to the study of zoonotic emergence and has led to many discoveries with public health implications, is the use as a diagnostic tool in cases and outbreaks of disease of unknown origin (DUO). Through its unbiased nature, metagenomics has identified: novel disease associations of known human pathogens (e.g. astrovirus encephalitis (Naccache et al., 2015, Quan et al., 2010)); cases of human disease associated with viruses that were previously only known to infect other animals (e.g. Borna disease virus 1 (Schlottau et al., 2018)); new species or genotypes of known human pathogens (e.g. a novel human papillomavirus (Mokili et al., 2013)); and highly divergent viruses associated with novel syndromes (e.g. Bas-Congo virus (Grard et al., 2012) and severe fever with thrombocytopenia syndrome (SFTS) virus (Yu et al., 2011)). Such findings may lead to the development of novel specific tests (PCRs and/or serological assays), facilitating further surveillance and epidemiological or pathological investigations. Importantly, they also require risk assessment, possibly followed by risk management and communication (Morgan et al., 2009, Palmer et al., 2005, Human Animal Infections and Risk Surveillance (HAIRS) group, 2018).

An example is provided by the recent discovery of variegated squirrel bornavirus 1 (VSBV1) as a novel zoonotic pathogen, after the deaths by encephalitis of three German squirrel breeders who had traded variegated squirrels with each other. The virus was first identified

through metagenomic analysis of samples from a contact squirrel (Hoffmann et al., 2015). Its genomic characterisation led to the development of a specific PCR system, which was then used to confirm the presence of VSBV1 in the brains of the case patients (Hoffmann et al., 2015), and in squirrels from an additional four species in private collections and zoos in Germany, the Netherlands, and Croatia (Schlottau et al., 2017b, Schlottau et al., 2017a). Further case finding, and mapping of the squirrel trade, identified additional human cases: one confirmed, one probable, and two possible cases (Tappe et al., 2019, Tappe et al., 2018). Overall, these findings represent the first unequivocal evidence of human disease caused by a bornavirus. They resulted in recommendations by German animal health and German and European public health authorities to avoid direct contact with exotic squirrels and to have them tested (European Centre for Disease Prevention and Control (ECDC), 2015, Robert Koch Institute (RKI), 2017, Friedrich Löffler Institute (FLI), 2016, Friedrich Löffler Institute (FLI), 2015). Additionally, they kick-started investigations into the pathology of and immune response to the virus (Tappe et al., 2018).

Laboratory and computational processes involved in viral discovery, and further examples, have been reviewed elsewhere (Chiu, 2013, Jazaeri Farsani et al., 2013, Mokili et al., 2012, Lipkin, 2010).

1.5.4 Outbreak investigations and phylogenetics

In the context of outbreak investigations, metagenomics can not only be used to identify aetiologies, but also to track the spread of viruses and draw epidemiological conclusions. For example, metagenomic sequencing data have been combined with phylogenetic methods to reconstruct the introduction and geographical spread of Zika virus in Central America (Theze et al., 2018), to investigate the relative importance of zoonotic and human-to-human transmission of Lassa virus in a large outbreak in Nigeria (Kafetzopoulou et al., 2019), and to track transmission chains and/or rule out outbreaks in nosocomial settings (Casto et al., 2018, Greninger et al., 2017).

The use of metagenomics for these purposes is particularly relevant for emerging viruses, for which standard diagnostic PCRs have not been developed, and for viruses with a highly variable genome of which not all variants are detectable by standard PCRs (e.g. Lassa virus). Additionally, metagenomics may be the preferred sequencing strategy when researchers are

simultaneously studying multiple co-circulating viruses that cause similar syndromes (e.g. Zika, dengue and chikungunya viruses in the Americas).

1.5.5 Clinical diagnostics

Due to its unbiased diagnostic capacity, (viral) metagenomics is a potentially useful tool in clinical medicine as well as in public health. Depending on the complexity of laboratory procedures (for example the inclusion of a viral enrichment procedure) and computational tools, metagenomics-based diagnostics can have a turnaround time that is comparable to or shorter than many classical, culture-based diagnostics. Additionally, it can detect drug resistance and virulence genes, with potential implications for treatment.

A recent study provides some examples of the real-life clinical usefulness of metagenomics, when used in addition to standard tests, for the diagnosis of infectious meningitis and encephalitis cases in hospitalized patients (Wilson et al., 2019). Participating clinicians noted the following benefits of the metagenomic data: they enabled or supported clinical decisions to replace broad empirical treatment with targeted treatment; helped to rule out co-infections where these would affect patient management; predicted the resistance to antiviral drugs; and provided reassurance to patients that their initial diagnosis was correct. Additionally, metagenomics identified a significant number of aetiologies that had been missed by standard clinical tests. In several cases, these diagnoses guided treatment decisions.

However, several important challenges still stand in the way of an extensive roll-out of metagenomics in the clinic. These include the complexity of metagenomic data, the need for bioinformatics expertise both for the running of computational tools (not user-friendly) and for data interpretation, the high costs, and ethical concerns regarding incidental findings. Metagenomics has thus mainly been used in small studies involving severe cases of disease of unknown origin, or in the context of clinical research.

The opportunities and challenges for the use of metagenomics as clinical diagnostic have been reviewed in more detail elsewhere in the literature (Simner et al., 2018, Goldberg et al., 2015).

1.5.6 Evolutionary biology

Finally, viral metagenomics contributes to our overall scientific knowledge of viral diversity and evolution, which can indirectly advance our understanding of emergence.

The identification of novel viruses, particularly in environmental samples and understudied animal populations, is rapidly filling in the gaps between families in viral taxonomy (Wu et al., 2018, Zhang et al., 2018, Shi et al., 2018, Shi et al., 2016, Li et al., 2015a). This expansion in viral diversity has provided invaluable insights into the diversity and plasticity of genome types, and mechanisms of genomic evolution (Zhang et al., 2018). It also enabled comparisons of viral and host phylogenetic trees over long evolutionary timescales, showing that long-term viral evolution involves frequent cross-species transmissions on a background of co-divergence (Shi et al., 2018, Geoghegan et al., 2017).

1.6 Strategic choices and approaches in viral metagenomics

Viral metagenomics studies involve many strategic choices. Here I highlight some of the choices that need to be made at the bioinformatic processing and data analysis stages, focusing particularly on taxonomic classification as the most relevant to the work in this thesis.

1.6.1 Filtering out non-viral reads before taxonomic classification

One of the main challenges in viral metagenomics is that viral nucleic acids typically represent a small proportion of genetic material in a sample, with the vast majority coming from host organisms and bacteria. To target sequencing resources appropriately, sample processing may include viral enrichment steps prior to sequencing: common approaches include (ultra)centrifugation, size filtration, and selective amplification of viral sequences (reviewed elsewhere, e.g. Hall et al. (2014), Parras-Molto et al. (2018)). However, even with such processes, host and bacterial sequences can still form a significant proportion of the resulting metagenome (Willner et al., 2009a, Cotten et al., 2014a).

To address this issue, a step to computationally filter out host and/or bacterial sequences may be applied before taxonomic classification. This could reduce the misclassification of non-viral sequences to viral taxa, and limit misassembly between viral and non-viral sequences. (Removal of human-derived sequences may also be required for ethical reasons.) Unsurprisingly, a recent review (Nooij et al., 2018) found that metagenomics pipelines with

a filtering step scored high on specificity and precision in benchmark tests with simulated viral data. In contrast, their sensitivity scores were generally lower (Nooij et al., 2018), probably reflecting the accidental removal of viral sequences.

In the VIZIONS study, samples were enriched for viral nucleic acids with an experimental procedure that includes centrifugation and nuclease treatment steps (Cotten et al., 2014a). Additionally, after preliminary exploration of sequence data revealed significant numbers of reads of likely host and prokaryotic origin, a filtering step was included in the bioinformatics pipeline, to prevent misclassification of these sequences to viral taxa.

1.6.2 Read-based analyses versus assembly into contigs

Another important choice is whether to classify individual sequence reads, or whether to assemble them first. Metagenomics pipelines with an assembly step score higher on sensitivity, specificity and precision in benchmarks testing the classification of simulated viral reads (Nooij et al., 2018). Particularly the similarities between very distantly related sequences are more easily detected in longer sequences (Wommack et al., 2008). Assembly also reduces the amount of data and thus the compute-time required for taxonomic classification. Furthermore, longer sequences are more informative in downstream analyses; for example, they result in more robust phylogenies.

However, assembly of viral metagenomic data is not straightforward. The simplest approach, mapping reads to reference genomes, is poorly suited to viral metagenomics, due to the extensive genetic diversity of viruses and their underrepresentation in reference databases. Hence, most studies use *de novo* assembly methods: these combine reads into longer contiguous sequences (contigs) based on overlaps, and can thus assemble sequences from divergent viruses, but they are more resource-intensive. The disadvantage of most standard *de novo* assemblers is that they assume even sequencing depth, whereas metagenomes are characterised by uneven sequencing depth due to the presence of different microbes at different concentrations. As a result, these tools have difficulties forming long, correct contigs. Specific metagenome assemblers (reviewed in Rose et al. (2016)) have been designed to address this issue; these generally outperform standard assemblers in reconstructing multiple genomes at a time, and some have shown success with viral metagenomes (Smits et al., 2014a, Smits et al., 2015). However, various challenges remain, including the high intrapopulation variation of some viruses (discussed in Rose et al. (2016)),

and the formation of chimeric contigs containing sequences from multiple closely related strains (Vazquez-Castellanos et al., 2014).

In the studies described in this thesis, I did not include an assembly step, and instead performed viral taxonomic classification at the read level. My objective was to generate a rapid overview of mammalian viruses circulating in different study populations, and I expected that many detected viral sequences would be of little relevance (e.g. those derived from bacteriophages or dietary viruses). Additionally, metagenomic assembly would be an additional computationally exhaustive process to set up and optimise. I thus considered it preferable to first identify viruses of potential interest (e.g. novel viruses, or potential novel zoonoses), and then, at a later stage, perhaps perform standard genomic assembly before any further analyses.

1.6.3 Taxonomic classification algorithms

Finally, the most important choice involves one's taxonomic classification tool. With sequencing technologies yielding more and more data for analysis, a variety of tools have been developed. These generally employ either a sequence similarity-based approach or a sequence composition-based approach, although hybrid approaches are also found.

Similarity-based tools

The BLAST suite of programmes (Altschul et al., 1990, Camacho et al., 2009) has traditionally been used to identify the closest homologs, and infer the likely origin, of genes or other sequences. It generates local alignments of nucleotide (BLASTn, BLASTx) or protein query sequences (BLASTp, tBLASTn) to reference sequences from a variety of databases. However, scaling this up to taxonomic classification of metagenomic data requires automatised interpretation of outputs. MEGAN (Huson et al., 2007) is a flexible and userfriendly tool that infers a sequence's taxonomic classification from BLAST outputs by identifying the lowest common ancestor (LCA) of all substantial hits. ProViDE (Ghosh et al., 2011) takes a more sophisticated approach, considering orthology (reciprocal similarity) and thresholds on multiple alignment parameters, defined to take into account patterns of viral genetic diversity. A variety of metagenomic pipelines combine specific BLAST searches with diverse taxonomic classification algorithms and additional analyses. For example, the web server MG-RAST (Meyer et al., 2008, Wilke et al., 2016) queries sequences against protein and ribosomal databases, adds functional annotations to protein-encoding genes, and allows

comparison of functional and taxonomic compositions across metagenomes. Several BLAST-based pipelines have been designed specifically for viral metagenomics; these employ different strategies to address the major challenges in the field: the genetic diversity of viruses, and their underrepresentation in reference databases. VirusHunter (Zhao et al., 2013) combines queries to the non-redundant nucleotide database, which is more comprehensive, with queries to the non-redundant protein database, which can detect more remote homology. VIROME (Wommack et al., 2012) queries predicted peptide sequences against different environmental and functional databases, with the aim of providing some biologically meaningful annotation even to sequences for which the taxonomic classification remains unknown or unspecific. Finally, Metavir (Roux et al., 2011, Roux et al., 2014) computes phylogenies of marker genes for various viral families, and comprises multiple analysis tools that facilitate the exploration of the vast unknown fraction of viral diversity, for example through assessment and comparison of gene richness or oligonucleotide frequency biases. While BLAST-based analyses are often considered gold-standard for their high accuracy, their main drawback is that, with metagenomic scale data (several million sequence reads per sample), they are inhibitively slow and resource-intensive.

To address this limitation, several faster aligners were developed, such as UBLAST (Edgar, 2010), RAPSearch2 (Ye et al., 2011, Zhao et al., 2012), Lambda (Hauswedell et al., 2014) and DIAMOND (Buchfink et al., 2015). Particularly the latter has a highly optimised search strategy, with adaptations including alphabet reduction to enhance sensitivity, double indexing to reduce memory requirements, and usage of spaced seeds to increase speed without losing sensitivity; this has resulted in speeds of up to 20,000 times faster than BLASTx, with comparable sensitivity (Buchfink et al., 2015). DIAMOND and similar tools provide output in a similar way to BLAST, allowing substitution in pipelines that incorporate programmes like MEGAN or ProViDE.

Another approach that emerged to speed up similarity searches, is the use of exact *k*-mer matching algorithms: tools that assign query sequences to taxa on the basis of exact matches of subsequences of *k* nucleotides (*k*-mers) to a reference database. These tools typically require a database building step prior to analysis; in this step, reference genomes are split up into overlapping *k*-mers, which are then associated with a taxon, and stored in a hash-based index structure. By keeping the search procedure as simple as looking up *k*-mers from query sequences within the index of reference *k*-mers, classification speed is increased by

several orders of magnitude compared to alignment-based methods. Different programmes have developed different strategies to assign *k*-mers to taxa: LMAT (Ames et al., 2013), Kraken (Wood and Salzberg, 2014), and One Codex (Minot et al., 2015) all use LCA-like approaches, whereas CLARK (Ounit et al., 2015) uses *k*-mers that are only found in reference sequences belonging to a specific taxon at a pre-defined taxonomic rank. The tools also use different methods to scale up from taxon-linked *k*-mers to classification of a full query sequence. Due to their requirement for long exact matches (e.g. Kraken has a default of k = 31), *k*-mer matching algorithms generally have high precision and specificity. However, the same requirement results in these methods suffering from over-specificity, i.e. they have difficulties identifying more divergent members of taxonomic groupings. This can be a problem for viral metagenomics in particular. To increase sensitivity, Kaiju (Menzel et al., 2016) and Taxonomer (Flygare et al., 2016) have adapted *k*-mer matching methodology to work with protein databases. MetLab (Norling et al., 2016) combines Kraken with searches of translated unclassified sequences against hidden Markov model (HMM) profiles of viral proteins; the latter is a probabilistic method (another type of similarity-based approach) to identify remote homologs, reviewed in (Skewes-Cox et al., 2014).

Overall, while sequence similarity-based taxonomic classification methods are very accurate, their success largely depends on the content of the reference database. In the context of viral metagenomics, these methods suffer from the lack of universal marker genes in viruses, their extensive genetic diversity, and their underrepresentation in public databases.

Composition-based tools

Where the methods discussed above use local sequence similarity for taxonomic classification, composition-based methods use global genomic properties, such as GC content, oligonucleotide frequencies and codon usage. These properties are driven by evolutionary forces, and differ sufficiently between organisms to allow discrimination between some species (Karlin and Burge, 1995, Karlin et al., 1997). Several tools have been developed that take advantage of this, by training a taxonomic classifier on a reference database with machine learning techniques. For example, PhyloPythia (McHardy et al., 2007) uses a multiclass support vector machine approach to classify sequences at various taxonomic levels based on relative frequencies of different 4- to 6-mers, whereas NBC (Rosen et al., 2008, Rosen et al., 2011) uses a naïve Bayes classifier, and oligomers of a user-defined length. Phymm (Brady and Salzberg, 2009) uses interpolated Markov models to identify the

most informative length oligomers, and integrates these. In PhymmBL (Brady and Salzberg, 2009, Brady and Salzberg, 2011), Phymm is combined with BLAST, resulting in a hybrid tool that is more accurate than either method on its own. While oligonucleotide frequency biases have been detected in viral genomes (Trifonov and Rabadan, 2010) and metagenomes (Willner et al., 2009b), most composition-based taxonomic classifiers have been designed for and mainly been trained on prokaryotic sequence data. That said, NBC included viral reference genomes in an update (Rosen and Lim, 2012), and, more recently, VirFinder (Ren et al., 2017) was designed specifically for the identification of viral sequences in assembled metagenomic data. Composition-based methods address the main drawback of similarity-based methods: they do not rely on sequence identity at the local level, and are thus better at classifying sequences without close relatives in the reference database. Additionally, once trained, these methods are faster than alignment-based methods (but slower than *k*-mer matching methods) (Bazinet and Cummings, 2012, Wood and Salzberg, 2014). On the other hand, they are generally not as precise or specific as similarity-based methods (Bazinet and Cummings, 2012, Nooij et al., 2018), and, despite their increased sensitivity, they are fundamentally still database-dependent. Most composition-based methods also do not work well on short sequences (<1 kb) (Fancello et al., 2012).

Taxonomic classifier used in this thesis

In the studies in this thesis, I used the exact *k*-mer matching tool Kraken (Wood and Salzberg, 2014) as taxonomic classifier because of a number of advantages over other commonly used tools. With the VIZIONS dataset consisting of several thousand samples with up to multiple millions of sequence read pairs each, alignment-based tools were considered prohibitively slow and resource-intensive. Additionally, the reads were short (ca. 250 nt), disqualifying most composition-based methods. While NBC (Rosen et al., 2008, Rosen et al., 2011) may have been a good option for short reads, it too has a long running time. In comparison, in tests performed on simulated bacterial metagenomes, Kraken did not only have a two-to-four orders of magnitude faster processing speed, but also a better, near-perfect, genus-level classifying precision (Wood and Salzberg, 2014). More recently, Kraken also performed well in benchmark tests on simulated viral sequences (Norling et al., 2016, Nooij et al., 2018).

In addition to speed and accuracy, another consideration in choosing Kraken was its flexibility: the content of its reference database could be customized to match the study's focus on viruses, whereas its size could be reduced to match my workstation's limited RAM.

Finally, as a locally-run tool, Kraken does not suffer from data sensitivity concerns or extended uploading times associated with web-based applications such as VIROME (Wommack et al., 2012) or Metavir (Roux et al., 2011, Roux et al., 2014).

For what concerns the possible over-specificity of *k*-mer matching tools, virus discovery *per se* was not one of the aims of this PhD, and I considered that using a near-comprehensive viral database and adjusting parameters to take a larger sequence diversity into account would likely be sufficient to identify most viruses of interest.

1.7 Aims and outline of this thesis

This thesis has two overarching aims:

1) to contribute to our understanding about the emergence of zoonotic viruses in human populations
2) to explore the roles and limitations of metagenomic surveillance in developing this understanding

More specifically, in the context of surveillance for emerging zoonotic viruses in Dong Thap province, in the densely populated agricultural epicentre of the Mekong Delta region in Vietnam, I aim to answer the following key questions:

> Using metagenomic methods, which mammalian viruses are found in faecal samples from humans and animals in this setting? Which are zoonotic?
>
> Could any of the identified viruses be “at the cusp of emergence” in humans?
>
> Do the findings support the notion of frequent “viral chatter” between humans and animals living in close proximity in this setting?
>
> What lessons can be learned for future metagenomic surveillance studies?

To answer these questions, I built a viral taxonomic classification pipeline, and applied it to metagenomic sequencing data obtained from the Wellcome Trust Vietnam Initiative on Zoonotic Infections (VIZIONS). This initiative homed in on zoonotic viruses with two main sampling strategies: a hospital study, into which patients were recruited if their clinical picture was consistent with viral infection; and a high-risk cohort study, consisting of individuals with residential or occupational exposure to animals, as well as these contact

animals. In this thesis, I include samples from hospital patients, high-risk cohort members, swine, and rats. The focus on these animal hosts was chosen because of their importance as reservoir and/or amplifying hosts for zoonotic viruses, and the significant role they play in the local economy.

In Chapter 2, I introduce the VIZIONS initiative in more detail. I begin with a description of its overall aims, and the design of the hospital and high-risk cohort studies. This is followed by more detailed methods of sample collection and laboratory procedures, as performed by my VIZIONS collaborators to generate the sequencing data.

In Chapter 3, I present the viral taxonomic classification pipeline and its methods. The pipeline consists of two divisions. The first, referred to as the "basic pipeline", cleans, filters, and assigns sequencing data to viral taxa. The second division consists of several adaptations to the outputs of the basic pipeline: grouping of data into customised operational taxonomic units (OTUs), application of signal thresholds, and validation of signals. I designed these adaptations to counteract some limitations identified during testing and preliminary data explorations, as described in Chapter 4.

In Chapter 4, I focus on testing and further development of the pipeline. For this purpose, I use samples from hospital patients for which diagnostic quantitative PCR (qPCR) data are available for comparison. First, I investigate the performance of the basic pipeline. Next, I explore discordant metagenomic and qPCR results, in order to identify processes where the pipeline may be losing sensitivity or specificity. I then turn to investigating and modelling "index switching" contamination of samples. Finally, I suggest a set of post-hoc adaptations to apply to the outputs of the basic pipeline; these are presented in Chapter 3 (section 3.4), and applied before analysis of signals in Chapter 5.

In Chapter 5, I apply the modified pipeline to samples from humans, swine and rats, to identify the viruses circulating in these populations. Considering what is known about the identified viruses and the origin and processing context of the samples, I first remove signals that represent likely contaminants (or non-infectious exposure), and then classify the remaining signals according to their zoonotic potential. For any identified viruses that are known or presumed to be zoonotic, or that are putative novel zoonoses, I investigate whether they are shared between human and animal study populations, and evaluate their

relevance as potential emerging public health threats. I also suggest how further studies may advance our understanding of specific viruses and their significance as potential threats.

In Chapter 6, I discuss how the findings in this thesis fit in with our current understanding of the emergence of zoonotic viruses. I also highlight how my findings illustrate the value and limitations of metagenomic surveillance.

Chapter 2. The Vietnam Initiative on Zoonotic Infections

I wrote this chapter with minor comments and text edits from Andrew Rambaut and Mark Woolhouse.

This chapter details the methods used by my collaborators on the Vietnam Initiative on Zoonotic Infections (VIZIONS) project, to generate the data that I analysed in this thesis. All work described in this chapter was performed by others.

VIZIONS was funded by Wellcome Trust (WT/093724) and led by Jeremy Farrar and Guy Thwaites (Oxford University Clinical Research Unit (OUCRU)), Paul Kellam (Wellcome Trust Sanger Institute), Mark Woolhouse (University of Edinburgh) and Nathan Wolfe (Global Viral).

The hospital study component was a collaboration between OUCRU, Ho Chi Minh City; OUCRU, Ha Noi; the Hospital for Tropical Diseases, Ho Chi Minh City; the National Hospital for Tropical Diseases, Ha Noi; Ba Vi District Hospital, Ha Noi; and Dong Thap, Dak Lak, Khanh Hoa and Hue Provincial Hospitals.

The high-risk sentinel zoonosis cohort study was a collaboration between OUCRU; provincial sub-Departments of Animal Health; Regional Animal Health Office 5, Dak Lak Province; Ba Vi District Veterinary Station, Ha Noi; provincial Preventive Medicine Centers; and Hanoi Medical University.

For the metagenomic study, selected samples were enriched for viral nucleic acids by the Laboratory of Experimental Virology, University of Amsterdam Medical Centre, and libraries were prepared and sequenced by the Viral Genomics team at the Wellcome Trust Sanger Institute.

Parts of this chapter contributed to a published manuscript:

Woolhouse, M., Ashworth, J., Bogaardt, C., Tue, N. T., Baker, S., Thwaites, G. & Phuc, T. M. 2019. Sample descriptors linked to metagenomic sequencing data from human and animal enteric samples from Vietnam. *Sci Data*, 6, 202.

The studies presented in this thesis use data collected in the context of the Vietnam Initiative on Zoonotic Infections (VIZIONS). In this chapter, I provide a general description of the VIZIONS project (partly based on Rabaa et al. (2015)), followed by a more detailed description of the work performed by others in VIZIONS that has particular relevance to the studies in this thesis. The final section describes the samples that were at the origin of the metagenomic data analysed in this thesis.

2.1 Introduction to VIZIONS

In a bid to improve our understanding of the emergence of zoonotic infectious diseases in Vietnam, researchers at the Oxford University Clinical Research Unit (OUCRU) in Ho Chi Minh City, the University of Edinburgh and the Wellcome Trust Sanger Institute established VIZIONS. VIZIONS was set up as a collaborative consortium, incorporating a variety of partners working in human or veterinary health in clinical, academic, and governmental settings. The aims of VIZIONS included the development of these collaborations, of infrastructure and of resources to facilitate integrative, multidisciplinary research on zoonotic diseases, with a focus on viruses (Rabaa et al., 2015). More specific scientific aims of the project are discussed below. The initiative was funded by a Wellcome Trust Strategic Award (grant number WT/093724) from 2011 until 2016.

At the heart of the VIZIONS project were two fundamental components, each targeting zoonotic infections in a different way.

The first component was a hospital disease surveillance study: a retrospective case series of patients presenting with enteric, respiratory or central nervous system disease, or with jaundice. Selection was thus based on outcome (infectious illness). This study particularly homed in on pathogens that cause significant illness and that could potentially be of zoonotic origin. Samples, clinical and epidemiological data were collected from the patients, with the aims of estimating the burden of disease for known pathogens, elucidating the aetiology of diseases of unknown origin (DUOs), and identifying socio-demographic and behavioural risk factors for various infections.

The second component was a high-risk sentinel zoonosis cohort: a prospective longitudinal cohort of people who have frequent residential or occupational contact with animals and are thus perceived to be at high risk of developing zoonotic infections. Hence, selection was based on exposure (animal contact). Outcomes of interest included infections that may

remain asymptomatic or only cause mild illness, which would not be picked up in the hospital study. Samples and epidemiological data were collected from cohort members and their animals at yearly intervals and at any signs of infectious illness. Additionally, social scientists conducted observations of and in-depth interviews with a subset of cohort members with specific exposures to wildlife species. The aims of the study were to characterise the infections (particularly viruses) circulating in high-risk human populations and in animals, to assess the incidence and risk factors for cross-species transmission, and to investigate the socio-cultural context of wildlife consumption and farming.

To address the identification of viruses in samples from these two studies, a viral metagenomic sequencing study was set up. Less biased towards common pathogens and yielding more information than a battery of diagnostic PCRs, metagenomic sequence data would allow an extensive characterisation of the viral flora of various study populations. Such data were also expected to result in the identification of uncommon or novel pathogens, particularly in cases of disease of unknown origin. Subsequent phylogenetic analysis could provide insights into the evolutionary and epidemiological dynamics of the identified viruses, including their spatial and temporal spread through human and animal populations and the inference of any zoonotic transmission events.

The work in this thesis relies on metagenomic data from samples that were collected through the hospital and high-risk cohort studies, and then subjected to metagenomic sequencing. The following three sections therefore contain more detailed descriptions of the methods used in each of these components. All study procedures had been reviewed and approved by the Oxford Tropical Research Ethics Committee (OxTREC; nr. 15-12) in the UK and various national and regional institutional committees for human and veterinary medicine in Vietnam.

2.2 Hospital study methods

2.2.1 Study design and recruitment

The hospital study took place during 2012-2016, at seven hospitals in different locations in Vietnam: the Hospital for Tropical Diseases, Ho Chi Minh City; the National Hospital for Tropical Diseases, Ha Noi; Ba Vi District Hospital, Ha Noi; and Dong Thap, Dak Lak, Khanh Hoa and Hue Provincial Hospitals. Patients were recruited if they presented with any of four clinical syndromes (enteric infection, respiratory infection, central nervous system infection,

or jaundice) considered likely to be due to zoonotic infections. In total, the study included 3616 patients with enteric infections, 4326 patients with respiratory infections, 968 patients with central nervous system infections and 1064 patients with jaundice.

As the work in this thesis focuses on samples from patients with enteric disease, the following paragraphs describe the enrolment, data collection, and diagnostic procedures for this subset of patients.

2.2.2 Inclusion and exclusion criteria

Patients with enteric disease were included in the study if they satisfied the following criteria: a clinical diagnosis of acute diarrhoeal infection (defined as a minimum of three loose stools within 24 hours, or one bloody loose stool); provision of written informed consent; residence within the same province as the attended hospital[1]; and requirement for hospitalisation as decided by the attending physician. Additionally, any member of the high-risk cohort (see section 2.3) would be included, if they presented with enteric disease and were deemed severely ill by the coordinating physician.

Patients were excluded from the study if they had been hospitalised with enteric disease within the previous six months, if they had previously been enrolled to the study with the same syndrome, if the diarrhoea was deemed likely to be due to prior antibiotic treatment, or if they had unrelated medical complications.

2.2.3 Data collection

Upon enrolment, study staff collected a stool sample (on the day of admission, before any antimicrobial treatment) and a variety of demographic, behavioural and other epidemiologically relevant information. Clinical information and results of laboratory tests were additionally recorded during each patient's stay and at time of discharge. Furthermore, the partial address of each patient's residence was collected and mapped with GPS software.

2.2.4 Diagnostics

Stool samples were subjected to a variety of diagnostic tests. First, specimens were cultured at the microbiology department of the participating hospital and subjected to conventional

[1] There were some inconsistencies in the application of this criterion, as multiple patients recruited at Dak Lak and Hue provincial hospitals were actually resident in neighbouring provinces.

biochemical (API 20E) and serological tests to isolate and identify bacterial pathogens (*Enterobacteriaceae* and *Campylobacter*). Residual stool samples were then stored at -80°C and shipped to the central study laboratory at OUCRU. At OUCRU, nucleic acid was extracted from stool samples and used to screen for further pathogens, including viruses and parasites, by PCR and by a Luminex xTAG gastrointestinal pathogen panel assay (Duong et al., 2016).

The work in this thesis uses molecular diagnostics results for sapovirus, astrovirus, Aichivirus, adenovirus, rotavirus, and norovirus. Samples were tested for these viruses by one-step multiplex real-time reverse transcriptase PCR reactions: single-tube combinations of reverse transcription and real-time PCR, set up for the detection of multiple target sequences simultaneously. Three separate assays were set up per sample, to detect the following viruses: 1) rotavirus and norovirus genotype II (GII); 2) norovirus genotype I (GI), Aichivirus and adenovirus; and 3) sapovirus and astrovirus. For each assay, 5 µl total nucleic acid was mixed with target-specific primers and probes (see Table 2.1 for sequences), RNA Master Hydrolysis Probes (Roche Applied Sciences) and associated activator and enhancer solutions, before being subjected to thermal cycling on a LightCycler 480II (Roche Applied Sciences). PCR cycle settings for assays 1 and 2 were as described previously (Dung et al., 2013), except plates were not cooled between reverse transcription and amplification. For assay 3, the reverse transcription phase was extended to 30 min, and the elements of the amplification phase were set to 45 s and 1 min respectively. The assays were validated with positive and internal controls using previously described methods (Dung et al., 2013). Additionally, two sets of negative controls, in which molecular grade water was substituted for a stool specimen and nucleic acid respectively, were included to monitor contamination during nucleic acid extraction and PCR.

Quantitative real-time PCR data were made available in the form of cycle threshold (C_t) values: the time point, expressed as the number of amplification cycles, at which a reaction's fluorescence crosses the signal threshold. The C_t value is thus an inverse proxy for viral load. A reaction was considered PCR-positive and its C_t value was recorded if it was below 39; in contrast, a reaction was considered negative and its C_t value was censored if it came to 39 or above, or in case of total non-reactivity.

Table 2.1 Targets and sequences of primers and probes

SaV, sapovirus; AsV, astrovirus; AiV, Aichivirus; AdV, adenovirus; RoV, rotavirus; NoV, norovirus. Fluorescent dyes – YAK, Yakima Yellow; FAM, 6-carboxyfluorescein; Cyan500/Cy5, indodicarbocyanine. Quenchers – BBQ, BlackBerry Quencher; BHQ1, BlackHole Quencher 1; BHQ3, BlackHole Quencher 3. Sequence ambiguities – W = T, A; M = A, C; R = A, G; Y = C, T; K = G, T; S = C, G; B = C, G, T; N = Any; I = Inosine. +C, +T, +A, +G = LNA bases.

Target		**Primer / probe**		**Reference**
Virus (sp. / genogroup)	**Gene or region**	**Name and direction**	**[Dye-]sequence (5' – 3')[-quencher]**	
SaV (GI, GII, GIV)	ORF1 polymerase-capsid junction	SaV124-F	GAYCASGCTCTCGCYACCTAC	(Oka et al., 2006)
		SaV1245-R	CCCTCCATYTCAAACACTA	
		SaV-Probe	YAK-CCRC+C+T+A+TRAA+C+CA-BBQ	
AsV	ORF2 capsid	AsV-F	TCAACGTGTCCGTAAMATTGTCA	(Logan et al., 2007)
		AsV-R	TGCWGGTTTTGGTCCTGTGA	
		AsV-Probe	FAM-CAACTCAG+G+A+A+A+C+ARG-BBQ	
AiV	3CD (protease-polymerase) junction	AiV-F	CAGGCRTACGGWTACCGC	Designed by OUCRU
		AiV-R	ACGTGGAGKCCCAGRATCTTGA	
		AiV-Probe	Cyan500-CGAAGGYCTGTGCGGWTCCCCGCTTGT-BHQ1	
AdV	Hexon	AdV-F	TCTTACAAAGTGCGCTTTACGC	Designed by OUCRU
		AdV-R	TTAAAGCTGGGRCCACGATC	
		AdV-Probe	Cy5-GACAACCGGGTKTTGGACATGGCCAG-BHQ3	
RoV (A)	NSP3	NVP3-FDeg	ACCATCTWCACRTRACCCTC	(Freeman et al., 2008)
		NVP3-R1	GGTCACATAACGCCCCTATA	
		NVP3-Probe	FAM-ATGAGCACAATAGTTAAAAGCTAACACTGTCAA-BHQ1	
NoV (GII)	ORF1-ORF2 (polymerase-capsid) junction	Cog-2F	CARGARBCNATGTTYAGRTGGATGAG	(Dung et al., 2013, Trujillo et al., 2006)
		Cog-2R	TCGACGCCATCTTCATTCACA	
		Ring-2 (probe)	Cyan500-TGGGAGGGCGATCGCAATCT-BHQ1	
NoV (GI)	ORF1-ORF2 (polymerase-capsid) junction	Cog-1F	CGYTGGATGCGITTYCATGA	(Trujillo et al., 2006)
		Cog-1R	CTTAGACGCCATCATCATTYAC	
		Ring-1C (probe)	FAM-AGATYGCGITCICCTGTCCA-BHQ1	

2.3 High-risk sentinel cohort methods

This section is a summary of the VIZIONS high-risk sentinel cohort methods described in Carrique-Mas et al. (2015) and Saylors et al. (2015).

2.3.1 Study design and recruitment

Recruitment for the high-risk sentinel cohort took place between March 2013 and August 2014 in the provinces of Dong Thap, Dak Lak and Ha Noi, in a collaboration between OUCRU, provincial Sub-Departments of Animal Health, provincial Preventive Medicine Centers and Hanoi Medical University. In total 852 people with frequent occupational and/or domestic contact with animals were recruited.

The groups included were as follows:

- Animal farmers and their relatives, representing people with typical residential exposures to a variety of livestock species. As farming is the most important livelihood in rural Vietnam, it was decided that farmers should form about two thirds of the high-risk cohort. This included both domestic and exotic animal farmers, but excluded farmers of cold blooded animals. Poultry, pig and cattle farmers were originally selected at random from the animal farm census. Exotic animal farmers were selected using convenience sampling to maximise diversity of animals sampled. Up to four family members, including children, were enrolled per farm.
- Pig and poultry slaughterers, representing people with more intense occupational exposures (particularly to blood). These were recruited from the most important abattoirs and slaughter points in each participating location.
- Animal health workers, also representing intense occupational exposures (to sick animals). These were selected through convenience sampling.
- Rat traders and workers in exotic meat restaurants, representing people with exposures to exotic wildlife species. These occupations are not common in all regions of Vietnam but they were enrolled wherever possible. A specific effort was made to identify and recruit these individuals.

All potential cohort members were invited to information meetings about the study prior to enrolment. Upon enrolment, the connection between cohort members and the study was maintained by visits to the participating farm households (once, within the first 3-4 months)

and the abattoirs, slaughter points, veterinary stations and restaurants where members were recruited (all monthly).

2.3.2 Inclusion and exclusion criteria

Potential participants were included in the cohort study if they satisfied the following criteria: involvement with raising, slaughtering or processing livestock or wildlife; provision of informed consent to sampling of themselves and their animals at regular intervals; residence within 40 km of the designated hospital site for clinical presentation in the event of illness.

Potential participants were excluded if they had not reached adulthood (except in case of recruitment at a farming household, where children ≥ 1 year of age were included in recruitment), if they had any immunosuppressive condition or immune deficiency, or in case of ongoing receipt of any immunosuppressive therapy.

2.3.3 Data collection

Included individuals were followed up over up to three years; data were collected at enrolment, yearly thereafter, and when a cohort member reported an episode of illness. During routine visits, a questionnaire was administered to obtain basic demographic and socioeconomic information; medical histories; and details about exposures, through contact with animals, high-risk practices concerning the preparation and consumption of food, or occupational injuries, in the previous year. During disease episodes, individuals were asked about their food exposures in the previous month, and about whether there had been any changes in their animal exposure since the previous questionnaire. Cohort members were also requested to provide a rectal swab, a naso-pharyngeal swab and a blood sample at each visit. In the third year of the study, stool samples were collected instead of (or in addition to) rectal swabs, as it had become apparent that human rectal swabs generally did not have sufficient concentrations of viral nucleic acid for metagenomic sequencing.

The high-risk cohort study also included the sampling of animals. At each visit to participating farm households, up to 15 warm-blooded animals were sampled; this was done according to a scoring system prioritising sick animals over healthy ones, and exotic or large animals over small domestic livestock. Additionally, at regular intervals throughout the study, samples were taken from pigs and poultry at abattoirs and slaughter points where cohort members were recruited, and rats were purchased from market traders. Sample types included faeces

(bats, rats, some pigs) or rectal/cloacal swabs (most pigs and other animals), nasal swabs, blood and (for poultry) feathers.

2.4 Viral metagenomic sequencing methods

A subset of samples from the hospital study and from the high-risk cohort study were selected for viral metagenomic sequencing.

2.4.1 Sample preparation

Selected samples were processed according to a modified Virus Discovery by cDNA Amplified Fragment Length Polymorphism (VIDISCA) method (Cotten et al., 2014a, de Vries et al., 2011, de Vries et al., 2012, van der Hoek et al., 2004). This method enriches samples for viral genetic material, which is otherwise difficult to sequence as it is naturally present at much lower concentrations than host and bacterial nucleic acid. It then converts genetic material from both DNA and RNA viruses to dsDNA to allow for metagenomic library preparation, using a process that is sequence-independent and reduces contamination by host ribosomal RNA.

Detailed procedures were as follows. Faecal samples were suspended in an equal volume of phosphate-buffered saline, and 110 µl were used for nucleic acid extractions. First, viral nucleic acid was enriched for by the removal of cellular debris, mitochondria and bacteria with a centrifugation step (10 minutes at 10,000x g), and subsequent digestion of any residual DNA with DNase (20 units TURBO DNase, Ambion) - viral nucleic acid was presumed to be protected from this treatment by encapsidation within virions. Protected viral nucleic acid was then extracted using the Boom solid phase extraction method (Boom et al., 1990). Viral RNA was converted to a cDNA intermediate using reverse transcriptase (Superscript II, Invitrogen) and a mixture of hexamer primers, designed to efficiently reverse transcribe RNA sequences from all known mammalian viruses but not ribosomal RNA (Endoh et al., 2005). Subsequently, cDNA and viral ssDNA were subjected to second strand synthesis with 5 units Klenow fragment (3'-5' exonuclease defective, New England Biolabs), and the resulting dsDNA was purified by phenol:chloroform extraction and ethanol precipitation.

2.4.2 Library preparation and sequencing

Viral dsDNA as isolated by the VIDISCA method was used to prepare metagenomic sequence libraries for deep sequencing, as described in Phan et al. (2016a).

Library preparation followed standard Illumina protocols, with up to 96 samples being prepared at a time for multiplex sequencing. Nucleic acids were sheared to 400-500 nucleotides (nt) in length, and to distinguish the nucleic acids from different samples, distinct 8 nt indexing barcodes were added to each sample's nucleic acid before pooling. Each pool was sequenced divided over two lanes. Sequencing was done on Illumina HiSeq 2500 machines, yielding up to several million 250 nt paired-end reads per sample (Woolhouse et al., 2019). Sequence reads were de-multiplexed before taken forward for bioinformatic processing as described in Chapter 3.

2.5 Samples in this thesis

In this thesis, I use metagenomic sequencing data from a subset of samples from the VIZIONS hospital and high-risk cohort studies.

It was decided to focus on samples from a single study province where the hospital study and the high-risk cohort study were overlapping: Dong Thap province (Figure 2.1). Located within the highly productive agricultural region of the Mekong Delta in the South of Vietnam, Dong Thap (3,379 km^2) is inhabited by nearly 1.7 million people, about 4.7 million ducks, chickens and other poultry, 233 thousand pigs and 26 thousand cattle and buffaloes (GSO, 2015). Additionally, there is a vigorous rat trade in the Mekong Delta, with an estimated 3300-3600 tonnes of live rats sold for human consumption each year (Khiem et al., 2003).

My study is based on enteric samples from hospital patients with diarrhoea, high-risk cohort members, swine and rats (Table 2.2). The two human subpopulations were targeted based on outcome (infectious illness, possibly zoonotic) and exposure (animal contact) respectively. Swine and rats were chosen as representatives of livestock and wildlife animal populations respectively based on a combination of three factors: 1) the Vietnamese pig industry is one of the largest in the world (FAO) and the rat trade is important locally (Khiem et al., 2003), hence these animals are amongst the most commonly kept/handled domestic and wildlife species; 2) as mammals, they have a higher phylogenetic relatedness to humans than birds and they are thus expected to be more likely sources of novel zoonotic infections; 3) they are known reservoirs of existing zoonotic infections, including viruses (Meerburg et al., 2009, Meerburg, 2010, Ho Dang Trung et al., 2012, My et al., 2014b, Chua et al., 2000).

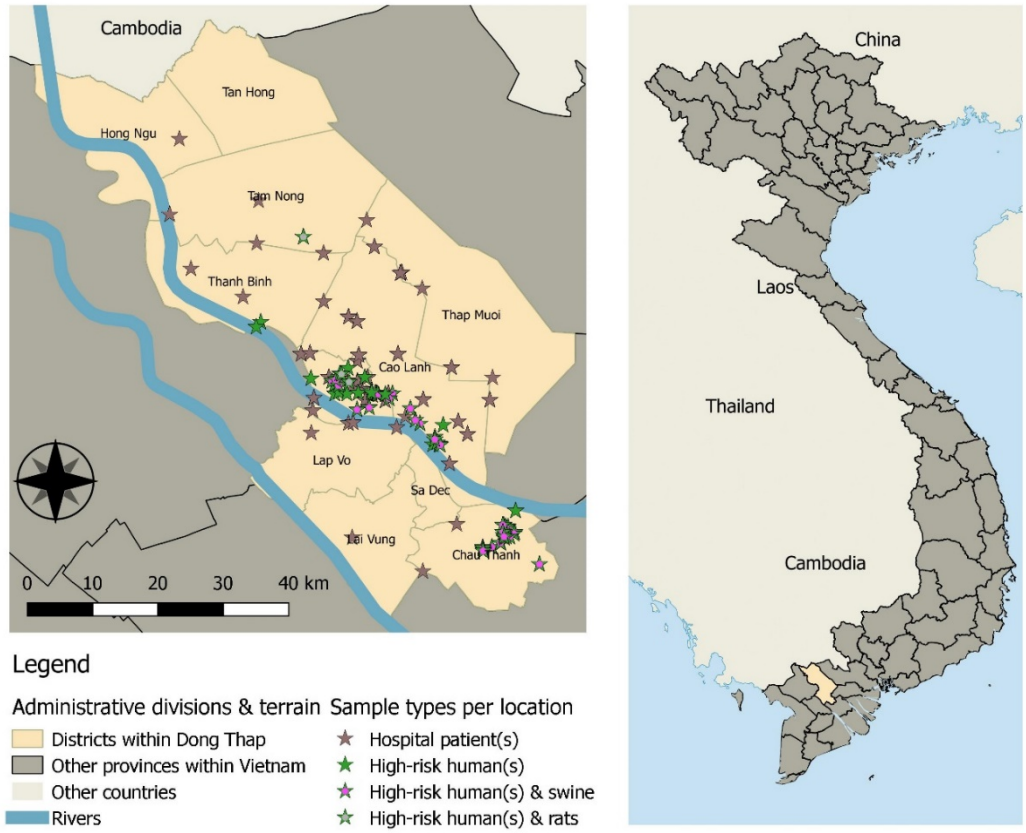


Figure 2.1 Location of sampled individuals in Dong Thap province
The location of hospital patients corresponds to their residence, whereas the location of high-risk cohort members represents their recruitment site (farm household, wet market, abattoir, or slaughtering point).

Table 2.2 Origin and sampling periods for samples used in this thesis

Study population	N° of samples	Sampling dates
Humans	1222	
- Hospital patients	671	November 2012 – May 2016
- High-risk cohort members	551	March – October 2013 & March – May 2016
Swine	285	March – October 2013
Rats	315	June 2013 – November 2014

2.5.1 Human samples

The samples analysed in this study include stools from 671 patients with diarrhoea, hospitalised in Dong Thap General Hospital (Cao Lanh City, Dong Thap province). The selection includes: 50 samples that had tested positive for rotavirus by PCR and that would

act as positive controls during isolation and sequencing of viral nucleic acid; 19 samples that had tested negative for all diagnostic tests, and that were considered of particular interest due to this "disease of unknown origin" status; and 602 samples picked at random from all enteric samples from the study site in Dong Thap that were available at the time.

Additionally, the study includes 551 samples obtained from 281 members of the VIZIONS high-risk sentinel zoonosis cohort. These include farmers and their relatives, abattoir workers and slaughterers, animal health workers and rat traders (Table 2.3). They were recruited in one provincial city and four districts within Dong Thap province: Cao Lanh City, Cao Lanh District, Chau Thanh District, Thanh Binh District and Tam Nong District (Figure 2.1). The samples include rectal swabs taken during baseline data collection (March – October 2013) and clinical episodes early in the study (April – September 2013), as well as faecal samples from the final data collection period (March – May 2016). Rectal swabs collected during the intervening period were not processed for metagenomic sequencing, as it had become apparent that this sampling method did not generally yield sufficient concentrations of viral nucleic acid.

In this thesis, I consider hospital patients and high-risk cohort members as a single study population ("humans"): the focus of this study is on the identification of viruses shared between different host species in this setting, and of putative cross-species transmissions. Viruses shared only between the two human subpopulations are beyond the scope of this work.

Table 2.3 High-risk cohort categories

Numbers of different sampling sites (see Fig. 2.1.), samples taken and individuals sampled. *One rat trading point and one slaughtering point were in the same location (a wet market). Indiv., individuals.

Risk category	Sampling sites	Baseline (2013)		Clinical episodes		End (2016)		Total
		Swabs	Indiv	Swabs	Indiv	Stools	Indiv	
Farm households	61	214	214	24	22	166	166	404
Abattoir workers and slaughterers	3	33	33	19	11	29	28	81
Animal health workers	3	30	30	7	6	23	23	60
Rat traders	3	4	4	0	0	2	2	6
Total	**69***	**281**	**281**	**50**	**39**	**220**	**219**	**551**

2.5.2 Swine samples

Swine samples included in this thesis consist of rectal swabs from 278 domestic pigs and seven farmed wild boars. These are a subset of the samples taken from farm animals during the baseline data collection period of the high-risk cohort study. The samples from domestic pigs were taken at 45 farms in three districts, the wild boar samples were from a single farm (Figure 2.1). The pigs and boars have been considered as a single study population ("swine").

2.5.3 Rat samples

Rat samples in this thesis consist of faecal samples from 315 animals of four different species (Table 2.4), but all considered to be part of a single study population ("rats"). The rats were purchased in seven batches of 15 animals, from rat traders in three different wet markets in Dong Thap province.

Table 2.4 Species of rats and numbers of included samples per species

Species	Nr. of samples
Rattus argentiventer (ricefield rat)	279
Rattus losea (lesser ricefield rat)	20
Rattus tanezumi (oriental house rat)	8
Bandicota indica (greater bandicoot rat)	8
Total	**315**

Chapter 3. The viral taxonomic classification pipeline

I wrote this chapter with minor comments and text edits from Andrew Rambaut and Mark Woolhouse. Figure 3.1 was reproduced from Wood and Salzberg (2014), with permission from the authors.

The basic pipeline (described in section 3.3) is the result of a collaborative effort. Data cleaning was performed by David Jackson (Wellcome Trust Sanger Institute) and Alasdair Ivens (University of Edinburgh). The filtering of host-derived read pairs was an idea from Andrew Rambaut and Mark Woolhouse; the filtering of prokaryote-derived read pairs was an idea from Alasdair Ivens. The filtering procedure was designed and carried out by Alasdair Ivens. Merger of overlapping read pairs was my own idea; the procedure was designed and carried out by Alasdair Ivens. The viral taxonomic classification step was my own work, except that over-represented sequences with unclear organismal origins were identified by Alasdair Ivens.

The three adaptations to the pipeline (described in section 3.4) are completely my own work.

Part of the work described in this chapter (data cleaning and filtering steps of the pipeline) contributed to a published manuscript:

Lu, L., Van Dung, N., Ivens, A., Bogaardt, C., O'Toole, A., Bryant, J. E., Carrique-Mas, J., Van Cuong, N., Anh, P. H., Rabaa, M. A., Tue, N. T., Thwaites, G. E., Baker, S., Simmonds, P., Woolhouse, M. E. & VIZIONS Consortium 2018. Genetic diversity and cross-species transmission of kobuviruses in Vietnam. *Virus Evol,* 4**,** vey002.

Sections 3.3 and 3.4 will be revised for submission as part of further manuscripts with multiple co-authors from the VIZIONS consortium, as appropriate according to their stated contributions.

This thesis focuses on the metagenomic analysis of viruses identified in samples from humans and animals in Vietnam. This chapter describes the viral taxonomic classification pipeline that lies at the basis of this work, for its function of converting millions of raw sequence reads into more useful and meaningful input data. The goal of the pipeline was to identify and classify read pairs derived from viruses – essentially providing me with basic answers to the question "what viruses are found in humans and animals in Vietnam?".

The chapter begins with an introduction to Kraken, the taxonomic classification tool at the heart of the pipeline. This is followed by an overview of the pipeline, and the two divisions it consists of. Finally, a detailed description is given of the methods used in each pipeline division.

3.1 Kraken as taxonomic classification tool

To query the metagenomic sequencing data from the VIZIONS samples for the presence of any viral sequences, a taxonomic classification tool was required. I chose the exact *k*-mer matching tool Kraken (Wood and Salzberg, 2014) for this, as argued in section 1.6. Here, I briefly describe Kraken's algorithm and a number of settings that have particular relevance to the work in this thesis.

3.1.1 Summary of Kraken's exact *k*-mer matching algorithm

Kraken is a similarity-based classifier that assigns query sequences to taxa on the basis of exact matches of subsequences of *k* nucleotides (*k*-mers) to a user-defined reference database. It does this with a two-step process.

The first step is to build the hash index or reference database (also referred to as 'Kraken database' in this thesis). This step requires the user to provide Kraken with a set of reference sequences and an associated taxonomy, for example from the NCBI. The *k*-mer length also needs to be defined at this stage: this will be used as unit for any analyses to be performed with the database. Kraken extracts all *k*-mers contained within the collection of reference sequences. For each *k*-mer, it then identifies all the reference sequences that this *k*-mer is found in, and stores their lowest common ancestor (LCA) taxon in the reference database. The idea is that, for a sufficiently long *k*-mer, the collection of matching sequences is very specific and has a LCA at a low taxonomic level (e.g. species).

The second step is the actual classification of query sequences, e.g. metagenomic sequence reads. To do this, Kraken splits up each query sequence into overlapping *k*-mers, and identifies the LCA taxa associated with these *k*-mers in the reference database. These taxa and their ancestors form a pruned subtree of the general taxonomy: the classification tree (Figure 3.1). Each taxon in the classification tree is weighted according to the number of *k*-mers classified to it. Total scores are obtained for each root-to-leaf path in the tree by summing the weights of the taxa along it. The path with the highest score is assigned as the classification path; the query sequence is then classified to the taxon at the leaf of this path.

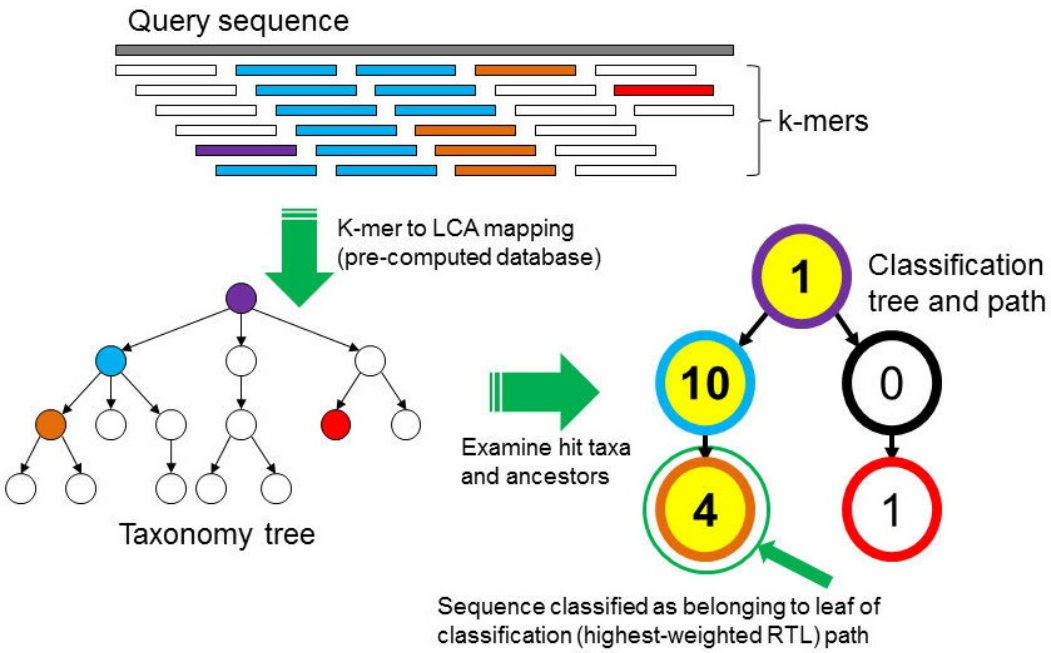


Figure 3.1 Kraken's classification algorithm
The *k*-mers from a query sequence are mapped against a reference database and taxonomy. The pruned subtree formed by the identified taxa and their ancestors is considered the classification tree. In this tree, the root-to-leaf (RTL) path with the highest total number of *k*-mers associated with it is the classification path. The query sequence is assigned to the leaf of the classification path. Figure reproduced from Wood and Salzberg (2014), with permission from the authors.

3.1.2 Relevant settings

k-mer length

The *k*-mer, a sequence of *k* nucleotides, forms the unit over which a query sequence is compared to the reference database. When considering what *k*-mer length to use, one needs to balance two considerations: as the *k*-mer gets longer, the collection of matching taxa (and

thus their LCA) gets more specific; however, this comes with a reduction in sensitivity, as only exact matches are considered. The default *k*-mer length in Kraken is 31 nt.

Maximum database size

To reduce Kraken's memory requirements, it is possible to set a size limit when building a Kraken database. In such cases, the number of *k*-mer-LCA pairs represented in the database is reduced by the minimum factor needed to fulfil the size limit. This causes a reduction in sensitivity compared to Kraken run with a full database, but, in tests run by Kraken's developers on various simulated metagenomes, a 4 GB database consistently showed better precision than the 70 GB default database (Wood and Salzberg, 2014).

Paired reads mode

Kraken can be run so that it classifies paired reads in combination, resulting in increased accuracy. Kraken does not classify *k*-mers with non-ACGT bases, and this can be taken advantage of by concatenating read pairs with a single N between the sequences.

Confidence thresholds

Kraken allows for adjustment of taxonomic classifications to match a required confidence threshold. This user-defined threshold is the minimum proportion of *k*-mers in the query sequence that needs to be assigned to a particular taxon for assignment of the full sequence to this taxon to take place. If the threshold is not met, the suggested classification is shifted up the taxonomic tree, until a taxon is found that does fulfil this confidence criterion.

3.2 Overview of the taxonomic classification pipeline

To apply taxonomic classification to the VIZIONS samples, a bespoke viral taxonomic classification pipeline was designed. This short section describes the overall design of this pipeline, which consists of two divisions (Figure 3.2).

The first division, referred to as the “basic pipeline” throughout this thesis, consists of data cleaning, processing and taxonomic classification procedures. As input, it takes raw sequence read pairs; as output, it gives sets of read pairs associated with viral NCBI taxa. These outputs were used to compare the outcomes of the pipeline with diagnostic quantitative PCR, in studies described in Chapter 4.

The second division consists of a number of adaptations, included after preliminary data explorations (including those in Chapter 4) suggested that the basic pipeline was affected by inaccuracies in the NCBI database and cross-contamination between samples. The three steps in this division redistribute the read pairs to custom operational taxonomic units (OTUs), apply signal thresholds that take into account variable amounts of contamination, and validate potential signals. The outputs are sets of read pairs considered as validated metagenomic signals, associated with OTUs. These were used for the metagenomic overview presented in Chapter 5. The methods associated with each of the divisions have been described in sections 3.3 and 3.4.

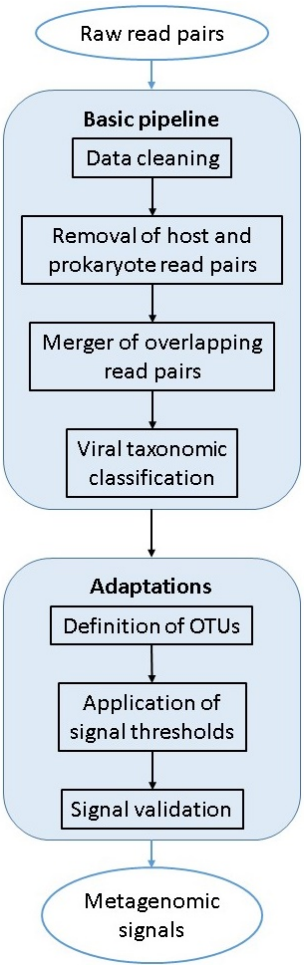


Figure 3.2 Overview of the taxonomic classification pipeline
The shaded rounded rectangles represent the main divisions of the pipeline, as described in sections 3.3 and 3.4. The black rectangles depict the individual steps in each pipeline division, whereas the blue ovals represent overall input and output data types.

3.3 Basic taxonomic classification pipeline

This section describes the basic taxonomic classification pipeline (Figure 3.3), including detailed methods for each step.

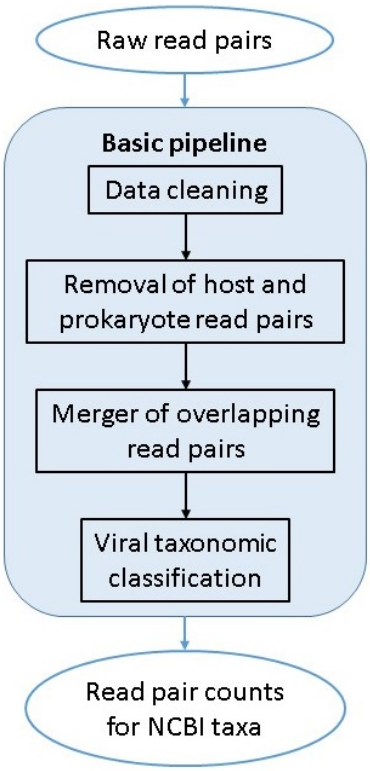


Figure 3.3 The basic taxonomic classification pipeline
In blue ovals, input and output data types. In black rectangles, steps of the pipeline.

3.3.1 Input

The input for the basic pipeline consists of raw sequence data for each sample, derived by metagenomic Illumina HiSeq 2500 sequencing at the Sanger Institute, as described in section 2.4. These are paired-end reads, of a length of approximately 250 nt per read.

3.3.2 Data cleaning

The first step in the pipeline is data cleaning. This took place in two phases. At the Sanger Institute, sequence reads were de-multiplexed and adapters were trimmed off with Biobambam2 (Tischler and Leonard, 2014). For ethical reasons, human-derived sequences were identified by aligning reads to the GRCh38 reference genome (without Eppstein-Barr virus) with BWA-backtrack (Li and Durbin, 2009), and then removed. Total sequencing yields

(used in Chapter 4) were calculated at this stage. The resulting sequence data are publicly available via the European Nucleotide Archive (ENA; study accession numbers PRJEB6505 and PRJEB26687).

After transfer of sequence files to the University of Edinburgh, sequence reads from the two sequencing lanes were merged for each sample, and subjected to further quality control and data cleaning as follows. Reports on each sample's overall sequence quality were generated with FastQC (Andrews, 2010), generally for information only. Any low-quality bases (quality < 20 on the Phred+33 scale) were trimmed from the 3' end of each read, and sequences matching to a database of primers were trimmed from either end of each read, if overlapping with the read by at least 10 nt. The set of sequences considered as primer contaminants was informed by a FastQC report on the first batch of samples. Finally, processed reads shorter than 50 nt were removed from the dataset.

3.3.3 Removal of sequences derived from host organisms and prokaryotes

The second step is a filter to remove prokaryote- and host-derived read pairs, in order to avoid misclassifying non-viral read pairs to viral taxa. This filtering step was included because, during early exploration of the VIZIONS sequencing data, it had become apparent that the data still contained a large amount of bacterial and some host-derived material, despite viral enrichment (described in section 2.4) and an initial filtering of human sequence reads (see above, subsection 3.3.2).

To identify read pairs to be filtered out, Kraken (Wood and Salzberg, 2014) was run on cleaned sequence datasets, with reference databases containing bacterial and archaeal, human, swine, and rat sequences, as described below.

Building of host and prokaryote Kraken databases

Taxonomies and whole genome sequences were downloaded from NCBI for humans (human genome GRCh38.p4, downloaded on 19 August 2015), swine (*Suus scrofa* genome, downloaded on 28 October 2015), and prokaryotes (all bacterial and archaeal genomes in RefSeq, 30 July 2015). Reference sequences for rats were based on a list of rat species sampled during a VIZIONS pilot study (*Rattus argiventer*, *Rattus exulans*, *Rattus norvegicus*, *Rattus tanezumi*, *Rhizomys pruinosus* and *Bandicota indica*). As most of these did not have

full genome sequences available, all nucleotide sequences in Genbank assigned to these six taxa were included (downloaded on 10 November 2015). For each of the three host organisms and for the collection of prokaryotes, a Kraken database was built from the sequences and taxonomies, using a *k*-mer length of 25 nt and a maximum database size of 20 GB. The maximum database size was set with the intention of running Kraken on my work station, which had 24 GB RAM available.

Classification and extraction of non-host, non-prokaryote datasets

For each of the four databases, Kraken's classification algorithm was run on cleaned sequence datasets so as to obtain lists of classified and unclassified read pairs only. Kraken was run in paired read mode and with no confidence threshold, so that classification of a single component *k*-mer would be sufficient to trigger classification of a read pair. For each sample, only the read pairs that remained unclassified with all four databases were deemed to be of further interest. These read pairs were presumed to be of viral origin and extracted from the cleaned sequence datasets.

3.3.4 Merging of overlapping read pairs

The third step tests the cleaned and filtered sequence data for overlap between the individual reads in each pair, and merges them if such overlap is found. This step was included because, during data exploration, a large proportion of read pairs in each sample was found to overlap significantly. Running Kraken on concatenated read pairs (as created by Kraken's paired read mode), while these are in truth overlapping, would cause *k*-mers located within the overlapping section to be counted double. I wanted to avoid this particularly for viral classification, where multiple taxa would be "competing" for *k*-mers and double-counting could affect assignment.

In the non-host, non-prokaryote sequence datasets, read pairs of which the members overlapped were identified with Vsearch (Rognes et al., 2016). A merger was performed if read pairs showed perfect overlap over a minimum of 20 base pairs, with merging of staggered reads (where the 3' end of the reverse read has an overhang to the left of the 5' end of the forward read) permitted. Bases with a quality score below 20 were trimmed and read pairs were not merged if either read was shorter than 50 nt after trimming.

Read pairs that did not match the perfect overlap criterion were concatenated with an N between the read sequences, to mimic the function of Kraken's paired reads mode.

3.3.5 Viral taxonomic classification

Finally, merged and non-merged read pairs were combined and subjected to viral taxonomic classification with Kraken (Wood and Salzberg, 2014).

Building of viral Kraken database

Selection of sequences for the viral database followed two considerations: comprehensiveness and clarity about a sequence's origins. To best cover the large diversity seen in viral sequences, it was decided to use a database based on all viral nucleotide sequences in Genbank, rather than the limited set of genomes in RefSeq. Several steps were then undertaken to remove sequences of which the origin was unclear (e.g. viruses integrated in eukaryotic genomes, or metagenomic whole genome shotgun sequencing projects) and that, if included, could cause erroneous links between *k*-mers and taxa in the Kraken database. Details of the processes are as follows.

All nucleotide sequence records with a virus as primary taxonomic identifier and no associated cellular organism taxonomic identifiers were downloaded from Genbank (4 August 2015). Whole genome shotgun project master records (not containing sequences) were replaced with records of the associated contigs, or removed in case of environmental/metagenomic origins. For segmented records, components without sequences (representing sequence gaps) were removed, and components with the concatenation of all segments' sequences were removed unless these sequences were found to be truly contiguous (as determined by searches in literature, or equivalent). Finally, several specific records were removed after initial data exploration showed that these were over-represented in VIZIONS samples, but had unclear organismal origins: AF191073, AF065755 and AF065756 were labelled as originating from cytomegalovirus cultures but also containing cellular sequences, and AY397620 was labelled as bluetongue virus, but with similarity to mycoplasma rRNA and therefore suspected to be a bacterial contaminant.

The viral Kraken database was generated from these sequences and their associated NCBI taxonomy, using a maximum database size of 20 GB and a *k*-mer length of 20 nt. This short *k*-mer length (relative to the default 31 nt) was chosen to reflect the expectation of finding

few long completely conserved sequences among viruses, due to their higher mutation rate compared to cellular organisms; it was expected to result in a higher sensitivity, but lower specificity. The maximum database size was set with the intention of running Kraken on my work station, which had 24 GB RAM available.

Classification

Kraken was run in single read mode and with a confidence threshold of 0.05, meaning a minimum of five percent of *k*-mers was required to assign a read pair to a taxon. The confidence threshold was set to restore the balance between sensitivity and specificity, by counteracting the drop in specificity due to the short *k*-mer length. I considered classification based on multiple 20 nt stretches a more suitable strategy than that based on one single 31 nt stretch of completely conserved sequence.

3.3.6 Output

The outputs of the basic taxonomic classification pipeline are sets of sequence read pairs assigned to viral NCBI taxa, and summary reports for each sample. These outputs are used to test the basic pipeline in Chapter 4. They are also fed into the "adaptations" division of the pipeline for further processing.

3.4 Adaptations to the taxonomic classification pipeline

After testing the basic pipeline on a subset of samples (Chapter 4), I considered that it could benefit from a number of adaptations, to counteract the effects of sample cross-contamination and taxonomic misclassification. I thus designed several additional steps (Figure 3.4). This section begins with a broad description of each of these steps and why I considered it useful, followed by more detailed methods.

3.4.1 Input

This division of the taxonomic classification pipeline takes as input the sets of read pairs assigned to viral NCBI taxa by the basic pipeline.

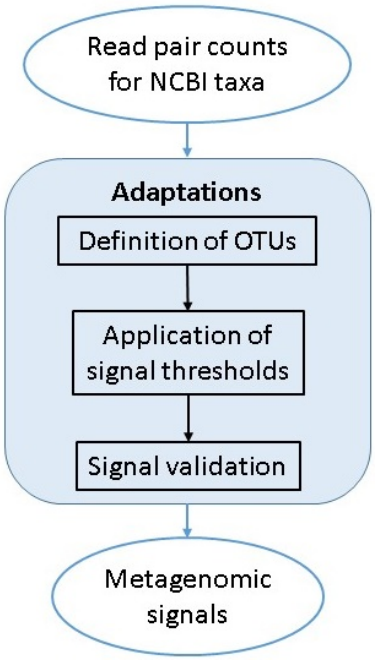


Figure 3.4 Adaptations to the taxonomic classification pipeline
In blue ovals, input and output data types. In black rectangles, individual adaptations.

3.4.2 Definition of OTUs

The first addition to the pipeline is a redistribution of read pair counts to newly defined OTUs. The creation of custom OTUs addressed some issues I had encountered with NCBI taxonomy: several ubiquitous taxonomic groups were unassigned at the genus level in NCBI, and some genera had erroneously assigned subtaxa.

Custom OTUs were created from NCBI viral genera, families, unranked groups, or any of these in combination with related unassigned lineages. The various processes involved in the definition of OTUs have been described here.

Taxon and lineage eligibility criteria for inclusion into OTUs

First, a list was compiled of NCBI viral genera that would form the basis for the OTUs. These genera were selected based on their presence within the samples, and their known or supposed infectivity to mammals. An initial list was generated by processing the basic pipeline outputs for all included samples, and identifying all NCBI virus genera with at least one detection of ≥18 read pairs (the "basic threshold", see section 3.4.3). Infectivity to mammals was inferred from information available in the online resources of the International

Committee on Taxonomy of Viruses (ICTV) (Lefkowitz et al., 2018), ViralZone (Hulo et al., 2011), NCBI databases, other online resources, and/or scientific literature.

The same process and criteria were used to identify mammal-infective viral families with a presence in the included samples. In some cases, these were used as basis for OTUs (described in more detail below).

Additionally, viral lineages that were not assigned to any NCBI genus were identified and considered for inclusion into OTUs. The scope of this list was limited to lineages assigned to mammal-infecting viral families with at least one detection of ≥18 read pairs, and lineages that were also unassigned at the family level and were not bacteriophages. Lineages were considered eligible for inclusion into OTUs if they were specific (at least equivalent to family level); had at least 10 read pairs assigned to them across all included samples; and were known or presumed to infect mammals, or were part of a viral genus (as defined by ICTV but not NCBI) or equivalent group comprising viruses identified in mammals.

Eligible unassigned lineages were combined with each other, or with genera or families, if evidence could be found that they were closely related. This has been detailed below.

OTUs based on NCBI genera

By default, NCBI genera were used as the basis for OTUs. Analysis at the species level would have been more specific, yet rather impractical due to both the large number of existing viral species and the multitude of viral sequences within the NCBI database that had not been assigned to any species. Several genus-based OTUs had one or more eligible lineages added (Table 3.1), on the basis that these lineages truly belonged within the genus, or were closely related. Additionally, some OTUs had one or more subtaxa removed, as during data exploration these were found to be wrongly associated with the genus. In the cases of Protoparvovirus HK-2014 and Sapovirus Hu/Kolkata/J20816, these consisted of mislabelled aveparvoviral and bacterial sequences (see Chapter 4), and were simply removed from all consideration. On the other hand, porcine parvoviruses and TT virus sle1841 were reassigned to different genus-based OTUs. The applied NCBI taxonomy generally corresponded to ICTV taxonomy of 2014 (Adams et al., 2014), but nomenclature of OTUs was adapted to match ICTV taxonomy 2017 (Adams et al., 2017).

Table 3.1 Adapted genus-based OTUs

OTU	NCBI lineages added and removed
Alpharetrovirus	+ Avian endogenous retrovirus EAV-HP
Alphatorquevirus	+ Torque teno virus, SEN virus - TT virus sle1841
Betapapillomavirus	+ Papillomavirus cat/EAA/USA/2001
Betaretrovirus	+ Human endogenous retrovirus K
Betatorquevirus	+ TTV-like mini virus, TT virus sle1841
Copiparvovirus	+ Porcine parvovirus 5, 6
Enterovirus	+ Picornaviridae sp.
Flavivirus	+ Barkedji virus
Gammapapillomavirus	+ Human papillomavirus type 197
Gammaretrovirus	+ Rat retrovirus SC1
Gammatorquevirus	+ Torque teno midi virus, Small anellovirus
Mastadenovirus	+ Titi monkey adenovirus ECC-2011
Pasivirus	+ Parecho-like virus
Percavirus	+ Myotis ricketti herpesvirus 2
Protoparvovirus	+ Bufavirus-1, -2, -3 - Protoparvovirus HK-2014, Porcine parvovirus 2, 5, 6
Sapovirus	- Sapovirus Hu/Kolkata/J20816
Simplexvirus	+ Chimpanzee alpha-1 herpesvirus
Tetraparvovirus	+ Porcine partetravirus, Porcine parvovirus 2

OTUs based on NCBI families

For seven viral NCBI families that, in the applied taxonomy (NCBI 2014) comprised only a single genus known to infect mammals, I used the family as basis for the OTU, to ensure the inclusion of any genus-unassigned lineages within NCBI (Table 3.2).

My assumption was that if any virus from this family were present in any (mammalian) sample, it would belong to the single mammal-infective genus. In the metagenomic analyses (Chapter 5), I have used the names of the presumed genera for these OTUs.

Table 3.2 Family-based OTUs

OTU
Arenaviridae (presumed *Mammarenavirus*)
Arteriviridae (presumed *Arterivirus*)
Asfarviridae (presumed *Asfivirus*)
Astroviridae (presumed *Mamastrovirus*)
Hepadnaviridae (presumed *Orthohepadnavirus*)
Hepeviridae (presumed *Orthohepevirus*)
Polyomaviridae (presumed *Polyomavirus*)

OTUs based on unranked NCBI taxa

I included as OTUs several NCBI species and unranked virus groups (Table 3.3). These represented a recently accepted genus (*Cyclovirus*) and three newly described groups not fitting within current taxonomy. Only groups with at least one eligible lineage were included.

Table 3.3 OTUs based on unranked NCBI taxa

OTU
Cyclovirus
Posavirus 1
Posavirus 3
St-Valerien Swine virus

OTUs based on NCBI lineages

Several OTUs were created from one or more eligible lineages, to represent a newly described group (Po-Circo-like virus) and several recently accepted genera (up until ICTV 2017 (Adams et al., 2017)) that had not yet been implemented within the used NCBI database (Table 3.4).

To form the OTUs for the genera within the new *Genomoviridae* and *Smacoviridae* families, I used proposals available on the ICTV website (Lefkowitz et al., 2018) and a CRESS DNA virus Rep protein phylogeny (Dr Lu Lu and Prof Peter Simmonds, personal communication).

Table 3.4 OTUs based on multiple NCBI lineages

OTU	Included NCBI lineages
Bovismacovirus	Odonata-associated circular virus-21
Gemycircularvirus	Hypericum japonicum associated circular DNA virus, Sewage-associated gemycircularvirus-8
Gemyduguivirus	Dragonfly-associated circular virus 3
Gemygorvirus	Meles meles fecal virus, Sewage-associated gemycircularvirus-5
Gemykibivirus	Gemycircularvirus SL1, Faecal-associated gemycircularvirus 8, Badger feces-associated gemycircularvirus
Gemykrogvirus	Sewage-associated gemycircularvirus-4, Caribou feces-associated gemycircularvirus, HCBI9.212 virus
Huchismacovirus	Sewage-associated circular DNA virus-9, Human smacovirus 1, Pig stool associated circular ssDNA virus
Po-Circo-like virus	Po-Circo-like virus, Po-Circo-like virus 41, Po-Circo-like virus 51
Porprismacovirus	Porcine stool-associated circular virus 1, 2, 3, 7, 8, 9, Gorilla smacovirus, PoSCV Kor J481, Porcine associated stool circular virus, Chimpanzee smacovirus, Chimpanzee stool associated circular ssDNA virus, Black howler monkey smacovirus
Rabovirus	Rabovirus A, Rat picornavirus

3.4.3 Application of signal thresholds

The second addition to the pipeline is the calculation and application of OTU- and sequencing run-specific signal thresholds. Such thresholds were included to separate signals from background noise due to index switching, a form of cross-contamination between samples in the same sequencing run. As the VIZIONS samples were distributed unevenly across sequencing runs (some runs contained mainly samples from hospital patients, with many human enteric pathogens, whereas others contained a mix of animal samples, with a variety of animal viruses), the amount of background noise varied across OTUs and sequencing runs (see Chapter 4). To make signals more comparable across runs, I set signal thresholds that take into account this variability.

In Chapter 4, I derived a model for the relationship between background read pair counts in qPCR-negative samples (excluding samples with an extremely high read pair count) ($rp_{OTU,n}$) and read pair counts summed across all samples of the same sequencing run ($rp_{OTU,run}$). This model is given in equation 3.1.

$$\log_{10}(rp_{OTU,n}+1) = 0.1760(\log_{10}\left(\frac{rp_{OTU,run}}{s_{run}-1}\right))^2 - 0.6792(\log_{10}\left(\frac{rp_{OTU,run}}{s_{run}-1}\right)) + 0.7720$$

Equation 3.1 Quadratic model of "background" read pair counts for a specific OTU on a specific run. $rp_{OTU,n}$, "background" read pair counts in "true negatives" (qPCR-negative samples minus extreme outliers); $rp_{OTU,run}$, read pair counts for a specific OTU summed over all s_{run} samples on the run. Chapter 4 describes the fitting of this model.

To translate this model into signal thresholds, I first applied a "basic threshold" of 18 read pairs, reflecting the back-transformed upper limit of the 99.5% prediction interval at the minimum of the model. This minimum (17.68) occurs at $\log_{10}\left(\frac{rp_{OTU,run}}{s_{run}-1}\right) = 1.93$, with $rp_{OTU,run}$ representing the read pair counts for a specific OTU summed over all s_{run} samples on a run. For all OTUs with any signals above this basic threshold, I determined the total read pair counts for each run (also including all samples not otherwise included in this thesis), and applied the upper limit of the 99.5% prediction interval, rounded up to the next integer, as threshold for the run. To avoid the biologically implausible rise in background levels to the left of the minimum, I used the basic threshold for any combinations of OTU and sequencing run with $\log_{10}\left(\frac{rp_{OTU,run}}{s_{run}-1}\right) < 1.93$.

3.4.4 Signal validation

Finally, I included a signal validation step to identify and eliminate false positive signals resulting from taxonomic misclassification of read pairs by Kraken.

Only signals that could be confirmed by BLAST (Altschul et al., 1990, Camacho et al., 2009) as an independent taxonomic classifier were considered validated. For each potential signal, up to 1000 random read pairs were queried against the non-redundant nucleotide database (downloaded on 31 October 2016), from which sequences obtained through the same VIZIONS metagenomic sequencing procedure (previously assembled and uploaded by collaborators at the Sanger Institute) were removed. The dc-megablast algorithm (version 2.6.0) was used with default parameters. Only top hits (single target sequences and single hsps) were kept for analysis. For each potential signal, the NCBI taxonomic identifiers (txids) of all BLAST top hits were checked against the OTU of interest.

OTU adaptations

As a starting point, OTU definitions described in section 3.4.2 were used, however, several further adaptations were applied. First, I made some adaptations to cover minor changes (taxon names and txids) in NCBI taxonomy since the version used for the Kraken database. Secondly, recognising that novel records and taxa may have been added to NCBI since that time, I reviewed all viral txids encountered in the BLAST top hits but not identified as part of OTUs by my scripts. I consulted online resources and scientific literature to determine whether these should be included as part of OTUs. When such taxa represented heterogeneous record collections (e.g. “Porcine parvovirus” without species number), I considered inclusion at the sequence record level. Finally, I excluded record AB213390.2, probably a mislabelled bacterial sequence, from the *Picobirnavirus* OTU.

Further considerations

To avoid spurious hits, top hits with an alignment length below 50nt were not counted. Additionally, for retroviruses, top hits to mammalian genomes were not considered as matches, despite the possibility that such genomes contained integrated retroviral sequences; this was considered too complicated to investigate.

Application of signal threshold

In considering a signal as verified or not, I applied the appropriate signal threshold from section 3.4.3, multiplied by a factor of 1.4, to the number of queried reads matching the OTU. The factor of 1.4 represented the 2 reads of each pair being queried individually, times a matching percentage of 70%, to allow for some lack in sensitivity. For signals >1000 read pairs that failed to exceed the threshold because it was particularly high (as for rotavirus in some runs), the cut-off was set to 1400 (=70%) queried reads matching the OTU.

3.4.5 Output

The output of the "adaptations" branch of the pipeline consists of sets of read pairs that form validated signals. These data sets form the basis for the metagenomic overview (Chapter 5).

Chapter 4. Validation and further development of the pipeline

I wrote this chapter with minor comments and text edits from Andrew Rambaut and Mark Woolhouse.

A multitude of collaborators from the VIZIONS consortium were involved in generating the metagenomic sequencing data I analysed in this chapter; their contributions have been detailed in Chapter 2 (data generation) and Chapter 3 (bioinformatic processing).

Diagnostic qPCR data (both C_t values and qualitative interpretation) were provided by the Virology team at the Oxford University Clinical Research Unit at the Hospital of Tropical Diseases in Ho Chi Minh City, Vietnam.

The use of receiver operating characteristic (ROC) curves to compare metagenomic and qPCR outcomes was an idea from Mark Woolhouse. The use of scatter plots to visualise and identify samples with discordant metagenomic and qPCR outcomes was an idea from Andrew Rambaut. The application of these ideas to the data, and all other analyses in this chapter, are completely my own work.

This chapter will be revised for submission as one or two manuscripts with multiple co-authors from the VIZIONS consortium, as appropriate according to their stated contributions.

4.1 Introduction

When interested in generating an overview of the diversity of viruses circulating in an ecological setting, metagenomic sequencing is a valuable tool: unlike targeted sequencing, it can alert the investigators to the presence of unsuspected or novel viruses. But such large-scale sequencing studies also have drawbacks. Multiple complex laboratory processes are required to enrich the viral component of samples, so as not to waste sequencing resources on genetic material from organisms that are not of interest. Additionally, post-sequencing, a

suite of bioinformatics programs is needed to classify the generated data, and to extract information of interest. Each step has its own options, biases and potential sources of errors that affect the sensitivity and specificity of the overall process.

It is important to have a certain awareness of these influences: they can be harnessed or neutralised to obtain the balance between sensitivity and specificity that is required for the pipeline to best achieve its goal. Additionally, awareness of these issues facilitates correct interpretation of the resulting metagenomic data. To develop this awareness it is important to test overall pipeline performance, and investigate any unexpected or outlying results.

In this chapter, I validate the basic viral taxonomic classification pipeline (described in section 3.3) by comparing its results with the outcomes of diagnostic quantitative PCR (qPCR), performed for six common enteric pathogens (rotavirus, norovirus, adenovirus, astrovirus, sapovirus, and Aichivirus; referred to as "test viruses") on all samples from hospital patients. The availability of these data allows me to investigate the overall performance of the pipeline, but also to identify and examine samples with discordant metagenomic and qPCR outcomes. These investigations can in turn identify specific vulnerabilities of the pipeline. The overall aim is to formulate a set of recommended pipeline adaptations that address these vulnerabilities, to be applied before metagenomic analysis of the larger VIZIONS sample set.

4.1.1 What balance between sensitivity and specificity is desired?

One of the main goals of the taxonomic classification pipeline described in this thesis is to detect viral infections arising from cross-species transmissions. To achieve this goal, a sensitive approach to detection is essential: cross-species transmissions are expected to be rare events, and viral infections often have a short-lived period of extensive shedding. By choosing sensitivity over specificity, one may have to deal with many false positives, but it is possible to take steps to check and eliminate these. For example, one can use another detection algorithm to validate findings, consider their biological plausibility, or return to the laboratory for additional investigations. In contrast, if choosing specificity, one may just never become aware of that potentially interesting finding. In this pipeline, I therefore seek to apply an initial sensitive approach, combined with additional checks to maintain specificity.

4.1.2 Possible sources of errors in the basic taxonomic classification pipeline

Issues resulting in low read pair counts and/or false negatives

Various laboratory and computational processes can result in true viral infections not being detected by the pipeline, or having lower read pair counts than expected from qPCR results.

Processing failures in the laboratory, whether due to human error, experimental conditions or chance, can result in low concentrations of viral nucleic acid remaining in a sample. This can lead to low overall sequence yields, or low viral read pair counts. Additionally, problems during sequencing may cause reads to be of poor quality, leading to their failure to pass a data cleaning (quality filtering) step. Furthermore, most lab procedures have biases that can result in infections of specific viruses being underrepresented or missed. While the specific biases of the VIDISCA viral enrichment protocol have not been studied, there is some evidence that centrifugation steps may deplete samples of larger viruses (Parras-Molto et al., 2018, Conceicao-Neto et al., 2015). When it comes to sequencing, many methods, including that used by Illumina HiSeq platforms, are less efficient at extreme GC content (Ross et al., 2013, Benjamini and Speed, 2012, Aird et al., 2011). A further discussion of these biases is beyond the scope of this thesis.

On the computational side, viral sequences may be assigned to a taxon that is different to that of their true origin. If the identification of host-derived or bacterial sequences is oversensitive, viral sequences may be mistakenly removed at the filtering stage. Or they can be left unclassified or be misclassified at the viral taxonomic classification stage. This stage is affected by an important limitation: viruses can have a high sequence diversity that is not well covered within reference sequence databases (Fancello et al., 2012, Rose et al., 2016). The result is that taxonomic classification tools relying on sequence similarity, including Kraken, do not perform well when faced with viruses with no close relatives in the reference database. Additionally, errors or uninformatively labelled records in the reference taxonomy can cause misclassification. For example, sequence records that are labelled only as far as the family level can result in related sequences being classified outside the true genus, and thus missed if analysis is performed at the genus level.

In this study, I investigate whether these issues can explain absence of or unexpectedly low read pair counts for test viruses in samples that were qPCR-positive. To check for indications

of the failure of any laboratory processes, I look at whether these samples have particularly low total viral and/or overall read pair counts, and consider the presence or absence of other detections. I also explore whether misclassification by Kraken plays a role, by querying the unfiltered and unclassified sequence data with BLAST to determine whether these contain any reads related to the virus of interest. For any such reads, I then identify whether they were misclassified at the filtering stage, or at the viral classification stage.

Issues resulting in false positives

In addition to missed infections as described above, issues in the lab or with computational processes can also result in false positives.

During field work or in the laboratory, samples may be contaminated with genetic material originating from personnel, lab reagents or from other samples processed in the same location at around the same time (Rosseel et al., 2014, Salter et al., 2014, Ballenghien et al., 2017, de Goffau et al., 2018). As the work in this thesis focuses on viruses, my main concern is contamination with viral genetic material that could be erroneously interpreted as evidence for infection. Described viral contaminants from cell lines, laboratory reagents, or DNA extraction columns include murine leukaemia viruses (Hue et al., 2010, Katzourakis et al., 2011, Lee et al., 2012, Erlwein et al., 2011, Paprotka et al., 2011, Kearney et al., 2012) and parvovirus/circovirus-like hybrid viruses (Xu et al., 2013, Naccache et al., 2013, Naccache et al., 2014, Smuts et al., 2014, Rosseel et al., 2014). But contaminants may also come in the form of viruses expected in the sample(s), in which case they are not easily distinguishable from true infections. In the VIZIONS study, many samples of similar origins were processed together in the various participating laboratories, which could have resulted in cross-contamination.

In the bioinformatics pipeline, taxonomic misclassification of sequences by Kraken (misassignment to the virus of interest) can result in false positives. Index-based classification tools are crucially dependent on a reference sequence database and a linked taxonomy, both of which may contain errors: sequence records labelled with the wrong taxon, or species assigned to the wrong genus. Such errors result in Kraken misclassifying sequences with similarities to these erroneous records or taxa. My use of an extensive but non-curated database for viral taxonomic classification in the VIZIONS metagenomics pipeline means that this kind of misclassification is likely to occur to some extent in this study. Another mechanism by which misclassification can be triggered is chance occurrence of sufficient

sequence identity to the virus of interest. My use of a short *k*-mer favours misclassification by chance, but the application of a confidence threshold counters this effect.

Like for the missed infections, I investigate what could explain unexpectedly high read pair counts for test viruses in samples that tested negative by qPCR. As all six test viruses are common pathogens that are expected in samples from patients with enteric disease, I consider it impossible to distinguish contamination from true infection. I therefore focus on exploring whether misclassification by Kraken may be at the base of these putative false positives, again using BLAST to determine the true origin of reads.

Batch effects

Studies with large numbers of samples, like VIZIONS, often have to resort to processing their samples in different batches. Such studies are invariably affected by "batch effects": effects on subsets of data caused by technical rather than biological factors, for example variations in laboratory conditions, reagent lots and personnel (Leek et al., 2010). In metagenomic studies, differences in efficiency of various lab processes or in the presence of contaminants can result in systematic differences in read counts or taxonomic composition between batches. Additionally, without careful study design, batch effects can confound the comparison of different groups, potentially leading to erroneous biological conclusions. A common way for this to happen is the separation of samples from different study groups into different batches, resulting in confounding by processing group or date (Baggerly et al., 2004, Leek et al., 2010, Akey et al., 2007).

A particularly interesting batch effect, that preliminary studies suggested might affect the VIZIONS data, is a form of cross-contamination known as "index switching". Index switching occurs between samples multiplexed together in an Illumina sequencing run (Kircher et al., 2012, Wright and Vetsigian, 2016b, Bartram et al., 2016, Nelson et al., 2014, D'Amore et al., 2016, Wright and Vetsigian, 2016a, Sinha et al., 2017, Illumina Inc., 2017). In standard Illumina procedures, barcoding indices are sequenced separately from the target sequence, but, when different nucleic acid fragments occupy the same space on a flow cell, this can result in erroneous combinations of index read and sequence read signals. This is seen at high frequencies (up to 5-10% of reads) in the most recent Illumina platforms, which use exclusion amplification chemistry (Sinha et al., 2017, Illumina Inc., 2017), but is also seen at low rates (< 0.5% of reads) in older platforms that use bridge amplification chemistry, like the HiSeq 2500 used in VIZIONS (Kircher et al., 2012, Nelson et al., 2014, Wright and Vetsigian, 2016b,

Wright and Vetsigian, 2016a, D'Amore et al., 2016). When a sequencing run contains several samples with a certain virus in high abundance, even low rates of index switching can result in considerable levels of read pairs from this virus across all samples of the run. If the existence of such "background noise", and its variability across sequencing runs, is not considered, index switching could mistakenly be interpreted as infection. When it involves cross-contamination between samples from different host species, this could further result in the erroneous conclusion that these sequence reads represent a cross-species transmission.

In the final part of this study, I quantify and model index switching contamination in the VIZIONS samples, as a function of the total number of read pairs on the sequencing run that were assigned to the virus. I use this model to inform sequencing run- and virus-specific signal thresholds, to be added to the pipeline (section 3.4.3) and applied in Chapter 5.

4.1.3 Value of this study

By considering the different error sources affecting the VIZIONS pipeline, this study provides me with strategies and valuable context for an improved interpretation of signals in the viral metagenomic overview study described in Chapter 5.

4.2 Methods

4.2.1 Data

In this study I used metagenomic next-generation sequencing data and diagnostic qPCR data from 709 patients with diarrhoea in Vietnam, obtained through the VIZIONS hospital study (see Chapter 2 and Rabaa et al. (2015)). This included 671 hospital patients recruited at Dong Thap General Hospital in Dong Thap Province and an additional 38 patients from Dak Lak General Hospital in Dak Lak Province, Khanh Hoa General Hospital in Khanh Hoa Province, and Hue Central Hospital in Thua Thien Hue Province.

The samples from hospital patients were processed in batches, for which two types of surrogates were available for analysis: (i) lot, corresponding to batches of approximately 300 samples that were processed at the same time at OUCRU and in Amsterdam; and (ii) run, corresponding to smaller batches of 63-96 samples that were multiplexed together in the same sequencing run at the Sanger Institute. Table 4.1 shows the distribution of the included samples over these batches.

Table 4.1 Distribution of samples over batches (sequencing runs and lots)
Number of hospital patient samples, presence of additional samples not included in this study, and the laboratory lots in which the samples were processed, for each of 13 sequencing runs in this study. Additional samples in sequencing runs include samples from high-risk cohort members and animals from the same study setting.

Sequence run	Nr. included samples	Any additional samples?	Lot(s)
16020	96	No	Lot_1
16317	71	No	Lot_1
16318	70	No	Lot_1
16370	63	No	Lot_1
16806	96	No	Lot_2
16845	54	Yes	Lot_2
17668	10	Yes	Lot_4
17819	10	Yes	Lot_3
18923	10	Yes	Lot_9
19344	73	No	Lot_7, Lot_8
19345	73	No	Lot_8
19379	73	No	Lot_8
20745	10	Yes	Lot_10

Sequencing data were cleaned and subjected to a Kraken-based taxonomic classification pipeline that filters out read pairs derived from prokaryotes and host organisms (humans, swine and rats) and assigns remaining read pairs to viral taxa. This pipeline, labelled the "basic taxonomic classification pipeline", has been described in section 3.3. In this study I used data from different stages of this basic pipeline:

- Sequencing yields (total number of read pairs), determined after adapter trimming and filtering of human sequences as performed at the Sanger Institute, but before further data cleaning at the University of Edinburgh (see section 3.3.2)
- "Cleaned but unfiltered" sequence reads, extracted after the pipeline cleaning step but before the filtering step
- Lists of filtered read pairs
- Viral taxonomic classification outcomes for individual read pairs, extracted from Kraken outputs after the pipeline viral classification step
- Sequence reads assigned to the genera of interest, identified through the viral classification step and extracted from the cleaned and filtered datasets
- Read pair counts for different taxa, extracted from Kraken reports after the pipeline viral classification step

I primarily used sequencing data and read pair counts for six genera that contain viruses known to cause diarrhoea and for which diagnostic qPCR data were also available: *Rotavirus*, *Norovirus*, *Mastadenovirus*, *Sapovirus, Mamastrovirus*, and *Kobuvirus*.

Corresponding qPCR outcomes were available for rotaviruses, noroviruses, adenoviruses, sapoviruses, astroviruses and Aichiviruses (human-infective kobuviruses). qPCRs were performed by collaborators on the VIZIONS project, as described in Chapter 2. Separate qPCRs were done for norovirus genogroup I (GI) and genogroup II (GII), but I considered the results in combination to facilitate comparison with the results from the taxonomic classification pipeline. Outcomes were measured as cycle threshold (C_t) values, which negatively correlate with quantity of target nucleic acid, and can thus be used as an inverse proxy for viral load. If a reaction yielded a C_t value below 39, it was considered qPCR-positive and the C_t value was recorded. Any C_t value of 39 or above was censored, and resulted in classification of the reaction as qPCR-negative. A summary of qualitative qPCR results is given in Table 4.2.

Table 4.2 Diagnostic qPCR results for six test viruses
Number and percentage of samples positive by diagnostic qPCR for each of the test viruses, for a total of 709 hospital patients. GI, Genogroup I; GII, Genogroup II.

Virus	Hospital samples positive by qPCR
Rotavirus	244 (34.4%)
Norovirus - GI	0 (0%)
- GII	62 (8.7%)
Adenovirus	22 (3.1%)
Sapovirus	11 (1.6%)
Astrovirus	4 (0.6%)
Aichivirus	3 (0.4%)

4.2.2 ROC curve analysis and normalisation

Receiver operating characteristic (ROC) curve analysis is a tool that can be used to assess the performance of a diagnostic test with a continuous scale, by comparing to a gold-standard, generating two-by-two confusion matrices (numbers of true positives, false positives, true negatives and false negatives) and calculating performance measures for all possible diagnostic thresholds (Obuchowski and Bullen, 2018, Bewick et al., 2004). A traditional ROC curve is obtained by plotting true positive rate (= sensitivity) against false positive rate (= 1-specificity) for all thresholds. The area under the ROC curve (AUC) provides a threshold-

independent measure of accuracy for the diagnostic test: it is an indication of the test's ability to separate truly positive from truly negative samples. If the test performs no better than pure chance, the ROC curve falls on the diagonal line between (0,0) and (1,1), corresponding to an AUC of 0.5. In contrast, if the test is perfect, the curve first raises vertically to (0,1) followed by a horizontal shift to (1,1), and the corresponding AUC is 1.

Here, ROC curve analysis was performed for genera with at least 10 qPCR-positive samples, comparing read pair counts from the basic taxonomic classification pipeline to qualitative outcomes of the corresponding diagnostic qPCRs (gold standard). The R-package *pROC* (Robin et al., 2011) was used to determine AUCs and their 95% confidence intervals. The confidence intervals were generated with the DeLong method (DeLong et al., 1988), using the algorithm by Sun and Xu (2014). The R-package *ROCR* (Sing et al., 2005) was used to plot ROC curves.

To evaluate whether normalisation of read pair counts improved the overall performance of the pipeline, AUCs were also calculated for three normalisation strategies: read pairs assigned to the genus of interest per million read pairs in the sample's overall sequencing yield, per million read pairs taken forward to the viral taxonomic classification stage, and per million read pairs ultimately classified as viral.

4.2.3 Correlation of read pair counts and qPCR C_t values

For each genus with at least 10 qPCR-positive samples, I tested the correlation between read pair counts and C_t values in qPCR-positive samples. For this, Spearman's rank correlation *rho* was used to avoid assumptions of normally distributed data. To illustrate the correlations, read pair count data (rp) were log-transformed using $\log_{10}(rp + 1)$ to allow for inclusion of samples with 0 read pairs, and lines were fitted and 95% prediction intervals calculated by using simple linear models.

I used these models and plots to identify samples with discordant results to investigate further, as well as to define other useful aspects of the data:

- Samples with strong qPCR signals, but low metagenomic read pair counts, were of interest as they could indicate a lack of sensitivity of the taxonomic classification pipeline; they could represent metagenomic false negatives. I defined a sample as a false negative if it was qPCR-positive for a genus of interest, but had a read pair count

of zero or falling below the 95% prediction interval. For genera for which I did not perform statistics (because of few qPCR-positives), I defined a sample as a false negative if it was qPCR-positive but appeared to have an unusually low read pair count at visual inspection of plotted data.

- Background levels of read pairs in qPCR-negative samples were of interest as they could indicate a standard level of (cross-)contamination and/or misclassification in the pipeline. Distributions of these background levels were determined for each sequencing run individually, as I hypothesised that (cross-)contamination levels varied by run. They were defined as the distributions of read pair counts assigned to a genus of interest in qPCR-negative samples, minus values that, when log-transformed, were extremely high (exceeding the third quartile + 3 interquartile ranges; these extreme values were considered false positives, see below).
- I defined a sample as a true negative if it was qPCR-negative for a genus of interest, and had a read pair count of zero or within the background distribution.
- qPCR-negative samples with high metagenomic read pair counts were of interest as they could indicate high amounts of misclassification in the taxonomic classification pipeline; they could represent metagenomic false positives. I defined a sample as a false positive if it was qPCR-negative for a genus of interest, but had a read pair count that, when log-transformed, had an extremely high value with respect to the distribution for all qPCR-negative samples of the same sequencing run.

False positives and false negatives as defined above were used to estimate overall sensitivity (Equation 4.1) and specificity (Equation 4.2) of the pipeline.

$$Sensitivity = \frac{\sum true\ positives}{\sum qPCR\ positives} = \frac{\sum qPCR\ positives - \sum false\ negatives}{\sum qPCR\ positives} \quad \textbf{Equation 4.1}$$

$$Specificity = \frac{\sum true\ negatives}{\sum qPCR\ negatives} = \frac{\sum qPCR\ negatives - \sum false\ positives}{\sum qPCR\ negatives} \quad \textbf{Equation 4.2}$$

4.2.4 False negatives

To investigate whether false negatives may have been due to inefficient or failed laboratory processes, I compared the overall sequencing yield of such samples to those of all other samples included in this study, as well as to all other samples multiplexed in the same sequencing runs (including VIZIONS samples not included in this study). Additionally, for each

sample I considered the percentage of the overall yield that was classified as viral by the taxonomic classification pipeline, the percentage of viral read pairs that was assigned to the genus of interest, and whether any other viral genera had higher read pair counts.

To determine if misclassification by the pipeline contributed to read pair counts being lower than expected, I searched for similarities to the genus of interest in all read pairs from the "cleaned but unfiltered" dataset that were not ultimately assigned to this genus. These read pairs were queried against all nucleotide sequences labelled with the genus of interest in NCBI (downloaded on 26 July 2018), using Magic-BLAST v. 1.3.0 (Boratyn et al., 2019), and defined as "originating from the genus of interest, but misclassified" if at least one member of the read pair had an alignment score ≥ 100. Matches to non-viral sequences in virus-based cloning vectors were excluded. In calculating the percentage of misclassified read pairs, I assumed that the sum of the misclassified read pairs and the genus-level read pair count represented all read pairs truly derived from the genus of interest, and used this quantity as the denominator. The pipeline stage at which the misclassifications occurred was identified by searching for the read pair identifiers in viral classification outputs and in lists of reads that had been classified as host-derived or as bacterial at the filtering stage.

4.2.5 False positives

To investigate if misclassification by the pipeline contributed to false positive detections, I validated the origins of read pairs assigned to the genus of interest in these samples with BLAST (Camacho et al., 2009). For detections of 1,000 or fewer read pairs, I extracted the sequences of all read pairs assigned to the genus of interest; otherwise, I extracted a random selection of 1,000 such read pairs. These read pairs were queried against the NCBI non-redundant nucleotide database (downloaded on 31 October 2016) using the dc-megablast algorithm and default parameters, limiting the output to top hits only (maximum one target sequence and maximum one high-scoring segment pair per target sequence). Any detections with <70% of queried individual reads matching the genus of interest were considered as being affected by taxonomic misclassification by the pipeline. For any detections affected by misclassification but that still had at least one read matching the genus of interest, I performed further BLAST searches of any such matches, to check whether the reference sequences were labelled with an erroneous taxonomy, i.e. whether they showed more similarity to other organisms than to the labelled genus.

4.2.6 Index switching as batch effect and source of background read pairs in true negatives

To investigate how background levels of read pairs varied per run, I performed several statistical analyses on true negative samples. Only data from runs consisting solely of samples from hospital patients were used, to avoid hidden sources of index switching contamination in the form of samples not included in this study. Additionally, for each genus, runs in which <25% of qPCR-negative samples had any read pairs assigned to this genus were excluded. For such combinations, all qPCR-negative samples with any read pairs had been defined as false positives, resulting in the background level of read pairs being set to 0. Unless stated differently, background read pair counts were log-transformed with $\log_{10}(rp + 1)$ before analysis. As 12 out of 13 runs with samples from hospital patients contained samples from only a single lot (Table 4.1), I considered that controlling for run would sufficiently control for any effect of lot.

For each genus, one-way analysis of variance (ANOVA) and the F-test were applied to determine whether the mean transformed background levels of read pairs varied by run. In addition, the correlation between transformed background read pair counts in individual true negative samples and the total number of read pairs assigned to the genus across all samples of the same run was assessed using Spearman's *rho* and the asymptotic *t* approximation.

Quantification of index switching

At the time this study was being carried out, several papers had been published that estimated rates of index switching, using either of two general methodologies. The first involves multiplexing different kind of libraries together (e.g. samples from evolutionarily distant organisms, or different gene libraries), and identifying misassigned reads on the basis that they do not match the known content of each library (Kircher et al., 2012, Mitra et al., 2015, D'Amore et al., 2016, Wright and Vetsigian, 2016a, Nelson et al., 2014). However, as this relies on knowledge of the content of each library, it is not applicable to metagenomics, where the composition of samples is not known in advance.

The second methodology involves counting reads assigned to indices that were included in the sequencing reaction but not associated with sample DNA (e.g. negative control wells). When using unique dual indexing, i.e. when each sample is identifiable by two unique indices

(from both forward and reverse adapters), switching of a single index results in a read being misassigned to an unused index combination; the rate of index switching can then be estimated simply by determining the fraction of reads assigned to such unused index combinations (Bartram et al., 2016, Illumina Inc., 2017)[2]. In contrast, when using single indexing or combinatorial dual indexing, i.e. when each sample is identifiable by a combination of two non-unique indices, switching of a single index can result in a read being misassigned to another sample, rather than an unused index or index combination. If the number of unused indices (negative control wells) or index combinations is large relative to the number of samples, the rate of index switching can still be approximated as above (Bartram et al., 2016, Kircher et al., 2012); however, the proportion of reads misassigned to other samples remains "hidden" and unaccounted for, as noted in several studies (Sinha et al., 2017, Bartram et al., 2016). To solve this, Sinha et al. (2017) estimated the rate of index switching by calculating the percentage of reads in negative control wells relative to the average number of reads seen for each sample, and considering that each sample would also contain a similar percentage of misassigned reads.

In the VIZIONS study, only single indexes had been used, and no read counts had been made available for negative control wells. However, considering each of the test viruses separately, I used true negative samples as equivalent to negative control wells or unused indices.

To account for the "hidden" index switching between positive samples, I needed a conversion factor that could be applied to the number of reads misassigned to true negative samples ("detectable" index switching) to obtain the total amount of index switching. To derive this conversion factor, I devised a conceptual model of index switching contamination (Figure 4.1). In this model, the total number of samples in a run is defined as s, the number of true negative samples for the genus of interest (grey circles in the figure) as n, and the number of presumed positive samples (blue and turquoise circles) as p. The following assumptions are made: (i) in true negatives, all background read pairs are the result of index switching; (ii) each presumed positive sample i contaminates all samples excluding itself, without variation by location on the 96-well plate or between true negative and other samples, with an average of c_i read pairs (arrows in the figure); and (iii) true negative samples do not contribute to contamination. Considering these assumptions, and adding up the number of misassigned

[2] Wright and Vetsigian (2016b) found that this method misses reads misassigned due to switching of the sequence read, rather than either index read, but this is ignored elsewhere.

read pairs for each sample (bold numbers in the figure), the detectable amount of index switching in true negatives comes to $d = n\sum_{i=1}^{p} c_i$ and the hidden index switching in presumed positives to $h = (p-1)\sum_{i=1}^{p} c_i$. Taking these together, the total number of read pairs with switched indices equals $t = (s-1)\sum_{i=1}^{p} c_i$. To state this in terms of the detectable fraction of index switching and a conversion factor, $\frac{d}{n}$ is substituted for $\sum_{i=1}^{p} c_i$, resulting in $t = \frac{s-1}{n} d$.

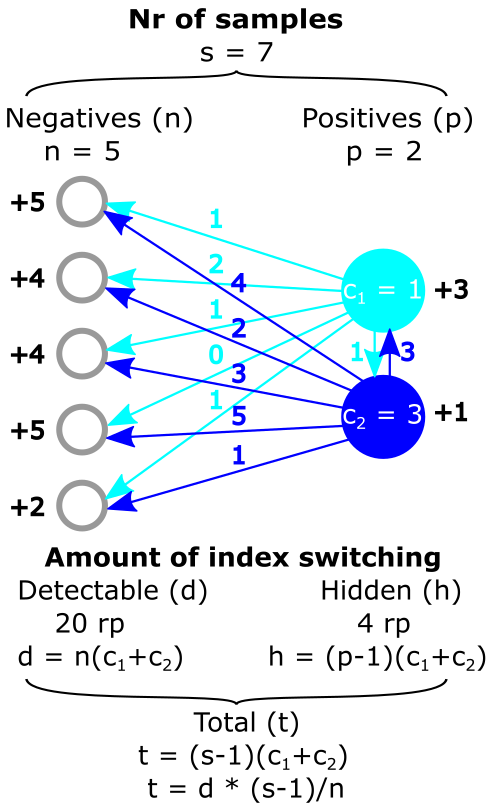


Figure 4.1 Derivation of a conversion factor for the quantification of index switching
Index switching results in the misassignment of read pairs from each positive sample to all other samples; this is represented by arrows, labelled with the number of "contaminating" read pairs. The contamination from a positive sample to each other sample has a mean of c_{sample}, and is assumed not to vary between positive and negative samples, nor by the relative location of the samples on the plate. For each sample, the total number of read pairs misassigned to it is indicated by the bold number next to it. All read pairs in negative samples are considered to result from index switching, thus forming the "detectable" fraction of index switching. In contrast, read pairs misassigned to positive samples cannot be directly observed, and thus form the "hidden" fraction of index switching. Based on the stated assumptions, a conversion factor can be derived to obtain the total number of misassigned read pairs (t) from the detectable amount of index switching (d); this conversion factor depends on the total number of samples (s) and the number of negative samples (n).

For each included genus, the conversion factor $\frac{s-1}{n}$ was applied to the (untransformed) read pair counts assigned to this genus in true negative samples, to yield estimates of the total number of read pairs with switched indices for this genus in the run. The approximate overall proportion of read pairs with switched indices in each run was then obtained by adding up these estimates for all included genera, and dividing by the total number of read pairs assigned to these genera in all samples.

Modelling the association between background read pair counts and total read pair counts in the run

In a bacterial invasion experiment, where each sample came from two strains that were cultured together, Wright and Vetsigian (2016a) applied a simple statistical model of index switching to estimate background levels of read pairs for each strain in each sample. Having used a combinatorial dual indexing scheme, they had found that the extent of index switching for each individual i7 index varied in proportion to the total number of read pairs with this index. They then took into account the contribution of each strain when summed across all other samples with the same i5 index, to obtain the background levels for each strain/sample combination.

To similarly be able to determine and subtract background read pair counts due to index switching in subsequent metagenomic analyses in this PhD (Chapter 5), I modelled the association between background levels of read pairs assigned to a genus in true negative samples, and total read pair counts for the same genus across the run. As qPCR outcomes were only available for a subset of the viruses, samples and sequencing runs that I intended to analyse in Chapter 5, I required a model that was independent of genus or run identifiers. Hence, I combined the data for all included combinations of sequencing run and genus to produce a single model.

Rearranging the equation for the total number of reads with switched indices (see above), $t = \frac{s-1}{n}d$, the average background levels of read pairs in true negative samples can be described as $\frac{d}{n} = \frac{t}{s-1}$. Furthermore, if index switching occurs for a fixed proportion of total reads, as consistent with Wright and Vetsigian (2016a)'s findings, and one considers the read pair count in each true negative sample sample ($rp_{taxon,n}$) as providing a separate estimate,

then the relationship between this and the total number of read pairs for the genus in the run ($rp_{taxon,run}$) can be modelled as $rp_{taxon,n} \sim \frac{rp_{taxon,run}}{s_{run}-1}$. To facilitate plotting, the analysis was performed on the log scale, with $x = \log_{10}\left(\frac{rp_{taxon,run}}{s_{run}-1}\right)$ and $y = \log_{10}(rp_{taxon,n} + 1)$, where 1 was added to avoid $\log_{10}(0)$. A linear regression approach was used, fitting a linear model with an offset term, a linear model without offset, and a quadratic model (Table 4.3). The assumption of linear regression that all data points are independent was not met, but it provided a simple and useful approach to estimate the relation between total read pair counts and background read pair counts due to index switching. I selected the model with the lowest Akaike information criterion (AIC) as best-performing model. Visual inspection of residual plots for these models revealed minor deviations of the linearity and homoscedasticity assumptions (residuals are slightly larger at low fitted values), as well as from normality, but discussions with a statistician (Dr Gail Robertson) did not result in any obvious alternative model structures to consider. Finally, the 99.5% prediction interval for the model was determined, to capture the full distribution of background read pair counts and provide a guide for the setting of future signal thresholds.

Table 4.3 Model structures considered in linear regression modelling of background read pair counts

Models were applied to log-transformed background read pair counts, so that $y = \log_{10}(rp_{taxon,n} + 1)$, whereas total read pair counts for each run were first normalised by the number of samples and then log-transformed, so that $x = \log_{10}\left(\frac{rp_{taxon,run}}{s_{run}-1}\right)$. Df, degrees of freedom; AIC, Akaike Information Criterion.

Model	Model structure	Df	AIC
Linear (offset)	$y = x - 2.8288$	2	5972.02
Linear	$y = 0.5548x - 1.0527$	3	4267.14
Quadratic	$y = 0.1760x^2 - 0.6792x + 0.7720$	4	2600.28

4.3 Results and discussion

4.3.1 Distributions of read pair counts for six “test viruses”

Distributions of numbers of read pairs assigned to the genera *Rotavirus*, *Norovirus*, *Mastadenovirus*, *Sapovirus*, *Mamastrovirus* and *Kobuvirus* (also referred to as “test viruses”) by the taxonomic classification pipeline divide into a few different patterns (Figure 4.2). First, for *Rotavirus* (Figure 4.2A), there are no samples with 0 read pairs classified to this genus;

the bulk of the distribution lies between 100 and 10 million read pairs and is clearly bimodal. This is in sharp contrast with the distribution for *Kobuvirus* (Figure 4.2F): 86.9% of samples have 0 read pairs, and only three samples have 100 or more read pairs classified to this genus. For the other genera (Figure 4.2B-E), the patterns fall in between these two extremes, with considerable proportions of samples having between one and 100 assigned read pairs.

The shapes of these distributions suggest that the taxonomic classification pipeline assigns a certain "background" level of read pairs, including to samples that are truly negative. Comparison with qualitative qPCR data for the same samples further supports this hypothesis. Overlapping read pair count histograms for qPCR-negative and qPCR-positive samples (Figure 4.3) show that the bulky distributions at lower read pair counts are mainly formed by qPCR-negative samples (red). The large numbers of samples involved suggest that these are not just true low-level infections, missed by qPCR due to more variable C_t values at low concentrations of target nucleic acid (Stowers et al., 2010). Instead, a technical explanation seems more likely: the "background noise" in negative samples could be the result of (systematic) misclassification of read pairs by the pipeline, or of contamination of samples.

A second important observation arising from Figure 4.3 is that the distributions of qPCR-negatives and qPCR-positives appear to separate fairly well. Particularly for genera with considerable numbers of qPCR-positive samples (*Rotavirus* and *Norovirus*) it is clear that these have higher read pair counts than the qPCR-negative samples. In addition to being suggestive of a good performance of the pipeline (further investigated in the next section), it also offers the option to use signal thresholds to separate true signals from background levels of read pairs.

Altogether, these findings indicate that, before the output of the basic taxonomic classification pipeline can be used for metagenomic analysis, a signal threshold should be applied to avoid large numbers of false positives due to background noise. The variable extent of this background for the different test genera (notably higher in *Rotavirus* compared to the other genera) suggests that virus-specific thresholds may be the most appropriate. In section 4.3.6, I investigate index switching contamination as a possible source of background noise, and develop a model that can be used to generate virus- and sequencing run-specific signal thresholds.

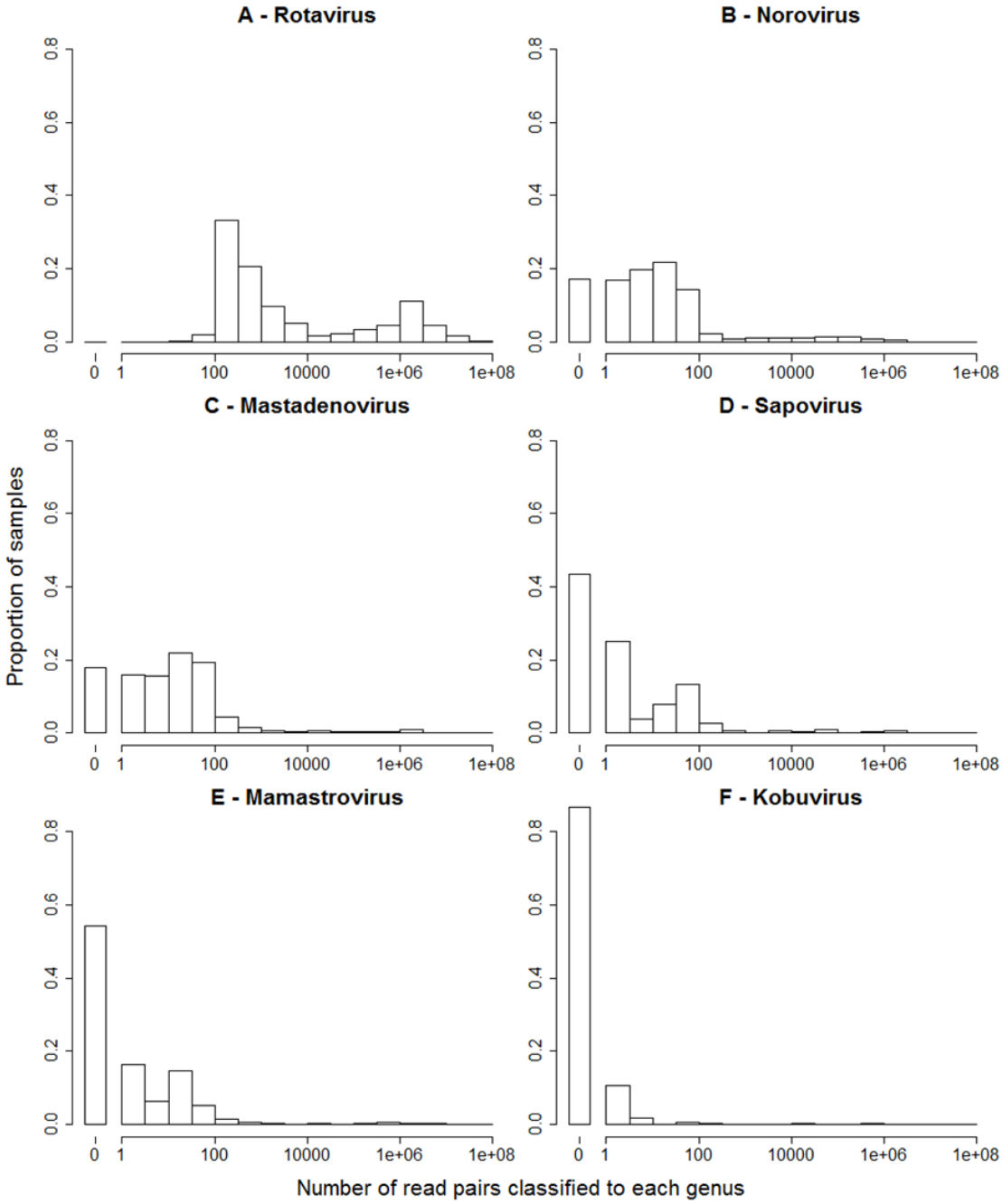


Figure 4.2 Read pair counts per sample, for six test viruses

Numbers of read pairs classified to (A) *Rotavirus*, (B) *Norovirus,* (C) *Mastadenovirus*, (D) *Sapovirus*, (E) *Mamastrovirus* and (F) *Kobuvirus* in enteric samples from 709 hospital patients. Each plot consists of two subplots with a shared y-axis: on the left, a single bar representing samples with 0 read pairs assigned to the genus of interest; on the right, a histogram with a log-scale x-axis, showing the distribution of samples with one or more assigned read pairs.

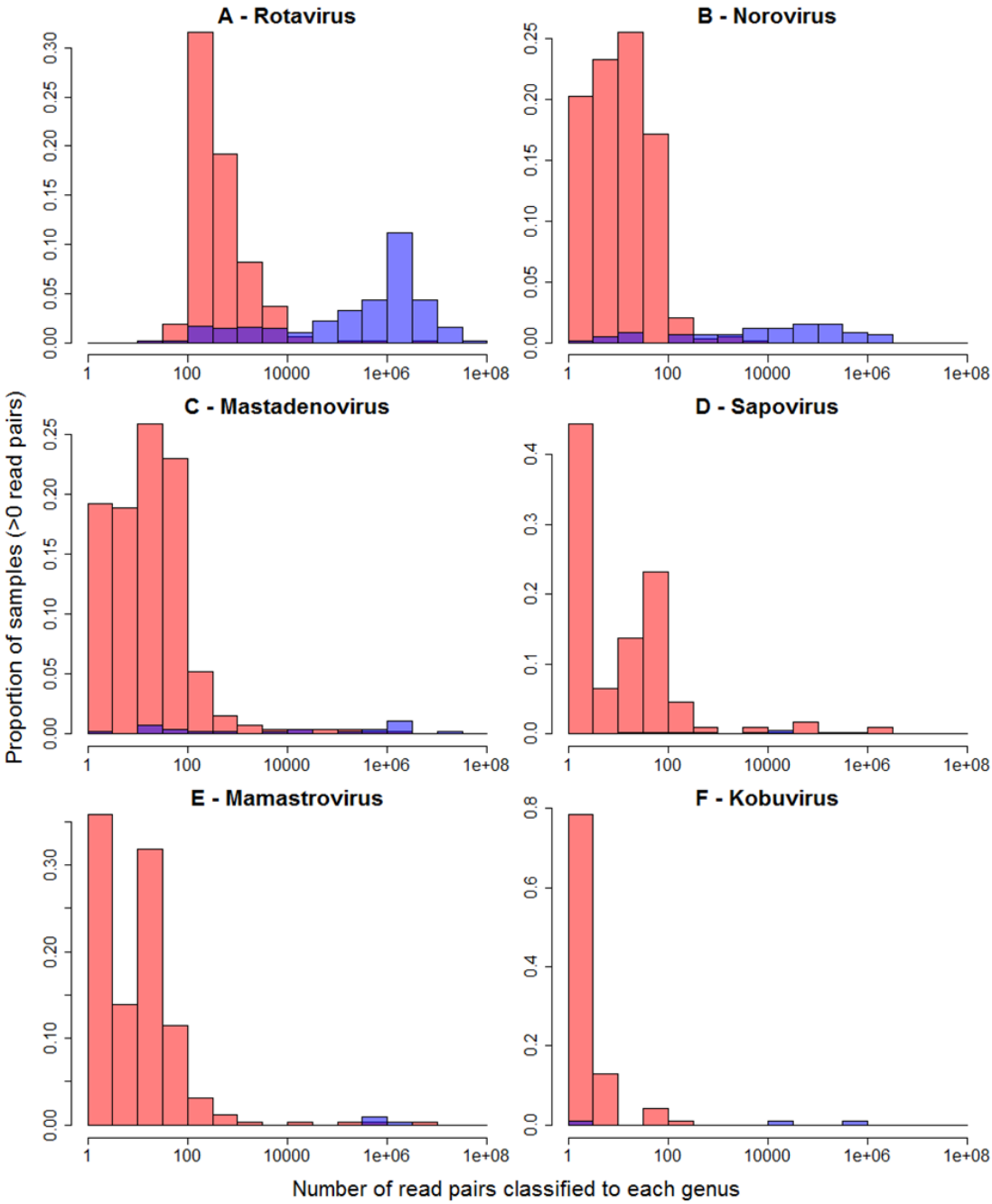


Figure 4.3 Read pair counts for six test viruses, for qPCR-negative and qPCR-positive samples

Read pairs classified to (A) *Rotavirus*, (B) *Norovirus*, (C) *Mastadenovirus*, (D) *Sapovirus*, (E) *Mamastrovirus* and (F) *Kobuvirus*, for samples with at least one read pair assigned to the relevant genus, and plotted separately for qPCR-negative samples, in red, and qPCR-positive samples, in blue.

4.3.2 ROC curve analysis and normalisation

I tested the overall performance of the basic taxonomic classification pipeline by comparing read pair counts and qPCR results for the four test viruses that had at least 10 qPCR-positive samples.

ROC curve analysis on raw read pair counts shows that the pipeline has a good discriminatory ability between samples that tested positive and negative by qPCR (Figure 4.4). The AUC, representing the probability that the pipeline ranks a randomly chosen qPCR-positive sample higher than a randomly chosen qPCR-negative sample, is high or very high (0.82-0.93; Table 4.4) for all four genera. Compared to *Rotavirus* and *Norovirus*, *Mastadenovirus* and *Sapovirus* have lower AUCs and wider confidence intervals, but this reflects the lower numbers of qPCR-positive samples for these genera.

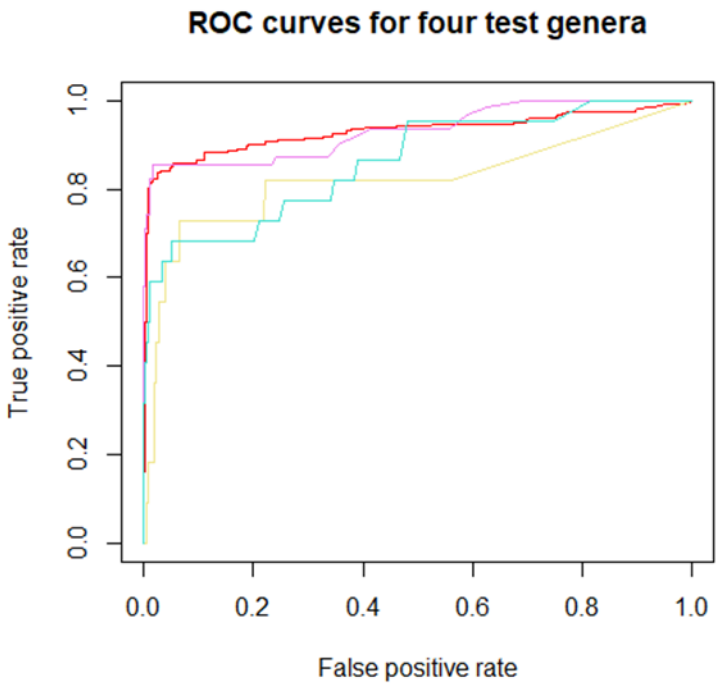


Figure 4.4 Receiver operating characteristic (ROC) curves for four test genera
Rotavirus (red), *Norovirus* (violet), *Mastadenovirus* (turquoise), and *Sapovirus* (khaki). True positive rate = sensitivity; false positive rate = 1 – specificity. A curve nearing (1,1) indicates perfect accuracy.

To check whether pipeline performance could be improved further by using normalised data, I compared AUCs for the raw read pair counts and three different normalisation strategies: taking into account the overall sequencing yield, the number of read pairs taken forward to viral taxonomic classification, and the number of read pairs classified as viral (Table 4.4 and Table 4.5).

Table 4.4 Receiver operating characteristic (ROC) curve analysis
ROC analysis of the basic taxonomic classification pipeline using data for four test genera. Area under the curve (AUC) measures close to 1 indicate a high accuracy of the taxonomic classification pipeline in distinguishing qPCR-positive from qPCR-negative samples. The reported AUCs are based on raw read pair counts.

Genus	AUC (95% CI)
Rotavirus	0.93 (0.90-0.96)
Norovirus	0.93 (0.89-0.97)
Mastadenovirus	0.86 (0.77-0.95)
Sapovirus	0.82 (0.64-1.00)

Table 4.5 ROC curve analysis for different normalisation strategies
ROC analysis of the basic taxonomic classification pipeline using three different normalisation strategies. Area under the curve (AUC) measures are reported for read pair counts normalised by the overall sequencing yield (AUC.yield), by the number of read pairs to which viral taxonomic classification was applied (AUC.taxclass), and by the number of read pairs ultimately classified as viral (AUC.viral).

Genus	AUC.yield (95% CI)	AUC.taxclass (95% CI)	AUC.viral (95% CI)
Rotavirus	0.93 (0.91-0.96)	0.92 (0.90-0.95)	0.91 (0.88-0.94)
Norovirus	0.92 (0.88-0.97)	0.91 (0.86-0.96)	0.91 (0.86-0.96)
Mastadenovirus	0.85 (0.76-0.95)	0.88 (0.79-0.97)	0.88 (0.78-0.97)
Sapovirus	0.82 (0.64-1.00)	0.80 (0.63-0.99)	0.78 (0.60-0.97)

For all four genera, the AUCs for raw read pair counts and for read pair counts normalised by yield are similar, and for three of the four genera, these AUCs are slightly higher than those for the other two strategies. This suggests that normalisation of read pair counts does not significantly improve pipeline performance.

There is an obvious limitation associated with this ROC analysis: the choices for test sample set, test viruses and gold standard were made for practical (data availability) reasons, but they are biased towards the detection of human pathogenic viruses – as is NCBI viral taxonomy. The good performances are thus not representative for all intended targets of the pipeline. The presence of a less-well-defined virus (like many animal or non-pathogenic viruses) would probably result in the pipeline assigning read pairs to various related taxa (perhaps unassigned at the genus level) as well as to less specific groups (e.g. "Viruses", or the virus family); the genus-level signal would thus not be as strong as for a well-defined virus. Pipeline AUCs for such viruses would likely be lower. The presented AUCs are thus best interpreted as reflecting the upper bounds of the pipeline performance.

In summary, the results presented in this section suggest that the overall performance of the pipeline is good and would not benefit from normalisation of read pair counts, but it remains a question whether this is similarly valid for other viruses than the four test genera. Considering practicalities as well as the presented results, I regard raw read pair counts as a suitable data type to use in analyses of the full VIZIONS dataset.

4.3.3 Correlation of read pair counts and qPCR C_t values

Among qPCR-positive samples, read pair counts show a significant negative correlation with C_t values, further supporting the correspondence between the basic pipeline and qPCR outcomes (Table 4.6).

Table 4.6 Correlation between read pair counts and C_t values in qPCR-positive samples
As determined by Spearman's rank correlation test

Genus	Spearman's *rho*	p-value
Rotavirus	-0.75	< 0.001
Norovirus	-0.43	< 0.001
Mastadenovirus	-0.22	< 0.001
Sapovirus	-0.14	< 0.001

While the correlation is weak for *Mastadenovirus* and *Sapovirus*, this is in part a reflection of the sparsity of qPCR-positive samples for these genera. The correlations, with fitted linear models for log-transformed read pair counts, are illustrated in Figure 4.5.

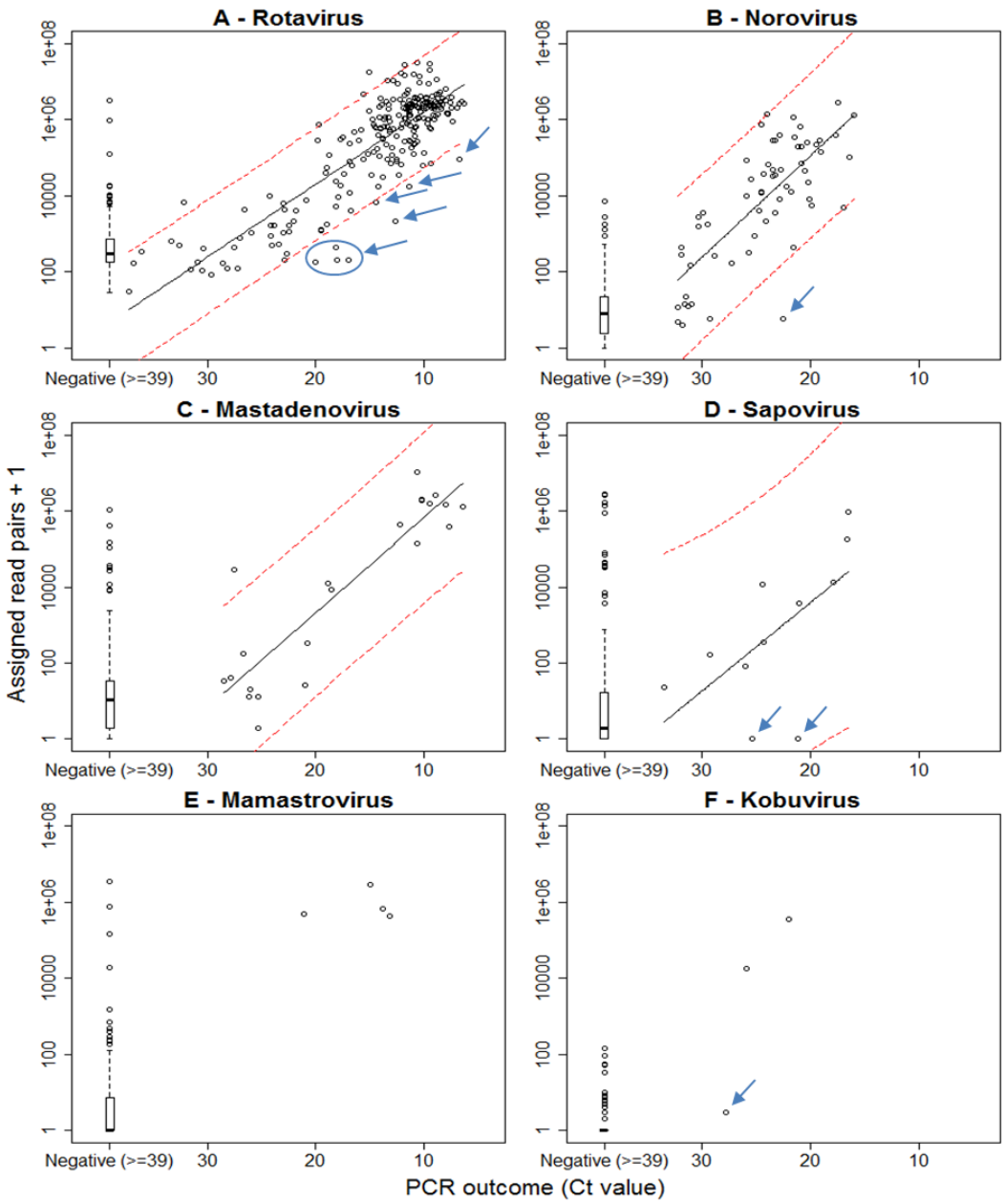


Figure 4.5 Correlation between qPCR outcomes and read pair counts

Correlations for (A) *Rotavirus*, (B) *Norovirus*, (C) *Mastadenovirus*, (D) *Sapovirus*, (E) *Mamastrovirus* and (F) *Kobuvirus*. The fitted line (black) and 95% prediction interval boundaries (red dashed) were derived using a linear model on transformed read pair counts. Blue arrows indicate qPCR-positive samples chosen for further investigation, based on a read pair count that is zero (for *Sapovirus*) or lower than expected (following the 95% prediction interval for *Rotavirus* and *Norovirus*, manually picked for *Kobuvirus* (no statistics performed)). NB: x-axis inverted to correspond to increasing qPCR signal strength.

The models and prediction intervals in Figure 4.5 also reveal the existence of samples for which the metagenomic and qPCR results do not correspond well. The blue arrows in the figure indicate 12 qPCR-positive samples with a read pair count that is either zero, or lower than expected from the corresponding C_t value. These cases could represent true infections missed by the pipeline. Similarly, 152 qPCR-negative samples have read pair counts that are extremely high with respect to the distribution of background read pair counts in their sequencing run (see definitions in section 4.2.3), which could suggest that the taxonomic classification pipeline falsely picks up non-existing infections.

Considering these outlying samples as "false negatives" and "false positives" respectively, I estimate the overall pipeline sensitivity and specificity to be 0.97 and 0.96 respectively. While these estimates are based on informal definitions, they suggest that, if background read pair counts are taken into account in the definition of metagenomic signals, sensitivity and specificity of the pipeline are both very high and well-balanced. However, like for the AUCs, these values have been estimated only for well-defined genera and should thus be considered as reflecting the upper bounds of performance.

4.3.4 False negatives

Twelve samples are positive by qPCR but have low read pair counts for any of the six genera of interest (indicated by blue arrows in Figure 4.5, Table 4.7). I investigated whether these false negatives could be explained by failures or inefficiencies in laboratory processes or by misclassification by the bioinformatics pipeline. My aim was to characterise any issues affecting pipeline sensitivity, so that these could be addressed before metagenomic analysis of the full VIZIONS dataset.

Sequencing yields

To begin with, I questioned whether the identified samples have unusually low total sequencing yields, indicative of low overall DNA concentrations or sequencing inefficiencies. Figure 4.6 shows that the yields of these samples are spread relatively evenly among yields of all samples included in the study. Only for two samples, 18923x72 and 19345x5, do the yields fall below the fifth percentile of samples. A low total yield does thus not appear to be a major cause of false negatives.

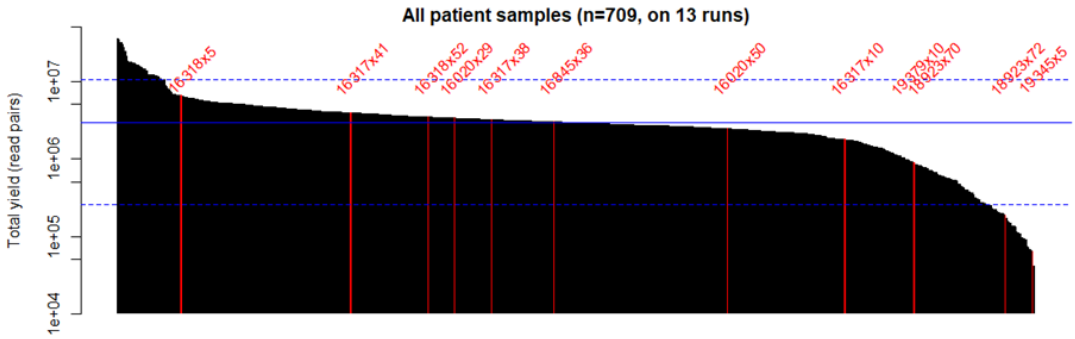


Figure 4.6 Distribution of sequencing yields for all 709 samples included in this study
The blue lines indicate the median (full line) and 5th and 95th percentiles (dashed lines). Samples representing metagenomic false negatives for one of six test viruses have been highlighted in red.

Inefficiencies in viral enrichment

Next, I considered whether the samples of interest have unusually low percentages of read pairs inferred to be of viral origin, potentially reflecting inefficiencies or failures in viral enrichment procedures. The percentage of the total yield that was ultimately classified as viral by the pipeline varied widely (0.02-58.03%) between false negative samples (Table 4.7). For six of twelve samples, this percentage was below 1%, suggesting that viral enrichment inefficiencies or failures could have contributed to low read pair counts. However, due to the complexity of laboratory and bioinformatics processes separating viral enrichment and viral taxonomic classification, the role of any failures in viral enrichment is impossible to determine with certainty. Further investigations into this are therefore considered beyond the scope of this thesis.

Roles for other viruses: sequencing resource depletion, and primer cross-reactivity

In five samples, the read pair count for the genus of interest is 0 or represents a very small percentage of the total viral read pairs. In these same samples, the pipeline does detect other viral genera with higher read pair counts (Table 4.7). This suggests that viral enrichment functioned well for these samples, but it hints at two other explanations for these false negatives.

Table 4.7 Investigations into potential laboratory-based explanations for false negatives

Genera: RoV, *Rotavirus*; NoV, *Norovirus*; SaV, *Sapovirus*; KoV, *Kobuvirus*. Other abbreviations: rp, read pairs; 5%ile, 5th percentile.

Sample, genus	Yield (rp)	Yield <5%ile?	Viral rp (% of yield)	Genus rp (% of viral rp)	Viral genera with higher read pair counts in the sample	Plausible interpretation(s)
16020x29, RoV	3343931	No	1069 (0.03)	443 (41.44)	None	Inefficient viral enrichment
16020x50, RoV	2445050	No	1418846 (58.03)	201 (0.01)	*Norovirus* (1361256 rp), *Parechovirus* (56880 rp)	*Norovirus* dominant infection
16317x10, RoV	1760723	No	99745 (5.67)	94405 (94.65)	None	-
16318x52, RoV	3449793	No	18923 (0.55)	17330 (91.58)	None	Inefficient viral enrichment
16845x36, RoV	2982420	No	625 (0.02)	204 (32.64)	None	Inefficient viral enrichment
18923x70, RoV	886517	No	551 (0.06)	178 (32.3)	*Punalikevirus* (266 rp)	Inefficient viral enrichment
19345x5, RoV	64336	Yes	6692 (10.40)	6618 (98.89)	None	Low nucleic acid concentration loaded on run
19379x10, RoV	904082	No	2207 (0.24)	2130 (96.51)	None	Inefficient viral enrichment
16317x38, NoV	3180037	No	104614 (3.29)	5 (0)	At least 7 genera; highest: Sp6likevirus (56166 rp)	Various more dominant infections
16318x5, SaV	6542279	No	15007 (0.23)	0 (0)	At least 7 genera; highest: *Norovirus* (13222 rp)	Inefficient viral enrichment; *Norovirus/Sapovirus* cross-reactivity of qPCR
18923x72, SaV	189013	Yes	68898 (36.45)	0 (0)	5 genera; highest: *Rotavirus* (67151 rp)	Low nucleic acid concentration loaded on run; *Rotavirus* dominant infection
16317x41, KoV	3931974	No	231979 (5.90)	2 (0)	At least 7 genera; highest: *Sapovirus* (181060 rp); also includes *Enterovirus* (40053 rp)	*Enterovirus/Kobuvirus* cross-reactivity of qPCR

First, nucleic acid from the genus of interest may have been present in a sample, but at a much lower concentration than that from other viruses. In metagenomic sequencing, nucleic acids are amplified and sequenced according to their relative concentrations, meaning that in samples with mixed infections, sequencing reagents are mostly consumed by more dominant infections, resulting in lower read pair counts for lower-level infections. For example, in sample 18923x72, the viral genus with most read pairs is *Rotavirus*, with 97.46% of viral reads; if a sapovirus was present in the sample at a much lower concentration than a rotavirus, it would have been detected by targeted sequencing (like qPCR) but it is likely that in metagenomic sequencing the limited reagents in the sequencing reaction would have been depleted by the rotaviral sequences

Secondly, samples 16318x5 and 16317x41 have relatively high read pair counts for different genera (*Norovirus* and *Enterovirus* respectively) in the same families as the genera of interest (*Sapovirus* and *Kobuvirus*). This suggests that in these cases, the discordant results between qPCR and metagenomics may be related to false positive qPCR signals, due to cross-reactivity of primers and probe.

Misclassification of read pairs at the filtering stage

Finally, I investigated whether any discrepancies in metagenomic and qPCR outcomes can be explained by misclassification of read pairs during the various bioinformatic processes of the pipeline.

Overall, misclassification of read pairs derived from the genus of interest to other taxa appears to play a small role. In samples that are false negatives for *Norovirus*, *Sapovirus* or *Kobuvirus*, read pairs that were classified outside of these genera by the pipeline do not have any good matches to BLAST databases of these genera either (Table 4.8). Thus, it is unlikely that these samples contain any undetected sequences derived from the genera of interest.

In contrast, some misclassification of this type did occur in false negatives for *Rotavirus*. The read pairs affected represent a maximum of 5.12% of all read pairs deemed to be truly derived from rotaviruses (Table 4.8). Thus, misclassifications by the pipeline do not explain any large discrepancies between qPCR and metagenomic outcomes for the investigated samples. With the exception of one read pair, these misclassifications occurred at the filtering stage: the read pairs had been considered as bacterial or host-derived and removed by an oversensitive/unspecific filtering approach. Further investigations revealed that the

single read pair for which this does not apply is a hybrid of individual rotavirus and norovirus reads, that in the viral taxonomic classification step had been assigned to *Norovirus.*

Table 4.8 Investigations into potential pipeline-based explanations for false negatives
Misclassified reads are individual reads not assigned to the genus of interest by the pipeline, but found to be matching the genus in a Magic-BLAST search. The percentage they represent has been calculated by using as denominator the sum of these misclassified read pairs and the read pairs assigned to the genus by the pipeline. Genera: RoV, *Rotavirus*; NoV, *Norovirus*; SaV, *Sapovirus*; KoV, *Kobuvirus*. Other abbreviations: rp, read pairs.

Sample, genus	Reads (part of n rp) misclassified	% of all genus-derived rp	Notes
16020x29, RoV	9 (6)	1.34	6 rp removed at filtering stage
16020x50, RoV	6 (6)	2.90	5 rp removed at filtering stage;
			1 NoV/RoV hybrid rp classified as NoV
16317x10, RoV	576 (359)	0.38	359 rp removed at filtering stage
16318x52, RoV	139 (107)	0.61	107 rp removed at filtering stage
16845x36, RoV	12 (11)	5.12	11 rp removed at filtering stage
18923x70, RoV	5 (3)	1.66	3 rp removed at filtering stage
19345x5, RoV	104 (57)	0.85	57 rp removed at filtering stage
19379x10, RoV	35 (19)	0.88	19 rp removed at filtering stage
16317x38, NoV	0	0	
16318x5, SaV	0	0	
18923x72, SaV	0	0	
16317x41, KoV	0	0	

Summary and implications for further pipeline development

In summary, these investigations show that inefficient laboratory processes and misclassification by the basic taxonomic classification pipeline each could have affected different samples, but that no process on its own explains a majority of the selected false negatives.

The objectives of this part of the study were to identify any issues with the metagenomic procedures that could result in loss of sensitivity, and to consider any post-hoc "fixes" that I may be able to apply to reduce their impact on any analyses. The occasional instances of inefficient viral enrichment, or resource depletion by more dominant viruses in a sample, are inherent to viral metagenomic sequencing and should not be considered issues needing to be resolved. In contrast, inappropriate removal of viral sequences at the filtering stage could

be a problem, but the extent of this happening remains unclear. In this study only *Rotavirus* sequences appear to be affected and only in a minor way, but it is not possible to generalise this finding to other viruses. In principle, the filtering method could be adapted to be more specific, but with no clear objectives for improvement, I consider this beyond the scope of this thesis. Finally, no loss of sensitivity is revealed at the viral taxonomic classification stage of the pipeline. Yet, this finding may not be generalisable either: my investigation only considers detections (or lack thereof) in four viruses that are all well-represented and well-defined in the NCBI database. As discussed in section 4.3.2, I consider it likely that less-well-defined viruses would have their read pairs distributed over multiple taxa, resulting in a lower sensitivity in genus-level analyses. To counteract this, it may be useful to create custom operational taxonomic units (OTUs) that are based on NCBI genera but also include relevant virus groups and lineages that have remained unassigned at the genus level.

4.3.5 False positives

Across genera, between eight and 47 qPCR-negative samples were considered as metagenomic false positives: they have extremely high read pair counts with respect to the background levels identified in other qPCR-negative samples (Table 4.9).

Table 4.9 Properties of false positive detections
Comparison of properties of signals affected by misclassification and those that were found to be correctly classified ("not affected").

	Rota virus	***Noro virus***	***Mastadeno virus***	***Sapo virus***	***Mamastro virus***	***Kobu virus***
False positives	8	15	15	47	28	39
- affected by misclass. (%)	0 (0)	5 (33.3)	3 (20.0)	16 (34.0)	15 (53.6)	13 (33.3)
% reads confirmed as genus	NA	0 – 50.0	0	0 - 46.4	0 – 50.0	0 – 50.0
maximum read pair count	NA	3	28	31,492	4	33
- not affected by misclass. (%)	8 (100)	10 (66.7)	12 (80.0)	31 (67.0)	13 (46.4)	26 (66.7)
% reads confirmed as genus	99.4 - 100	99.5 - 100	99.6 - 100	85.8 - 100	80.0 - 100	98.0 - 100
maximum read pair count	3,335,118	7,296	1,098,475	2,720,401	3,486,363	142

The discrepancies between these high read pair counts and the negative qPCR outcomes could stem from a range of technical issues (e.g. qPCR failure, or contamination of samples before/during sequencing) or be the result of misclassification by Kraken. Here, I focused on investigating the contribution of misclassification, given that qPCR failures and contamination are difficult to detect. My aim was to characterise issues affecting pipeline specificity, so that these too could be addressed before metagenomic analysis of the larger dataset.

Misclassification by the basic pipeline

Seeking confirmation of metagenomic false positives by querying their reads against the non-redundant nucleotide database with BLAST, I found that the pipeline sometimes misassigns reads derived from other taxa to the genera of interest. Fifty-two (34.2%) of 152 false positives are significantly affected by such misclassification, with only up to half the queried reads matching to the genus of interest in BLAST (Table 4.9); for the other 100 detections, considered as validated, this percentage is at least 80%. The extent of misclassification varies considerably by genus: the eight *Rotavirus* detections are all confirmed, whereas for the other genera, 20-54% of false positives are affected by misclassification. For most genera, significant misclassification is limited to detections with few (1-33) read pairs. For *Sapovirus*, on the other hand, significant misclassification is seen in 34% of selected samples, including in those with moderately high read pair counts (up to 31,492 read pairs). However, the strongest metagenomic signals in qPCR-negative samples cannot be explained by misclassification: confirmed by BLAST, these may be true infections undetected by qPCR, perhaps due to sequence variation.

For most genera, the BLAST results for the detections affected by misclassification include top hits to a variety of (mostly bacterial) taxa, suggesting that misclassification is mainly non-systematic. It probably arises from short stretches of sequence identity between a short (and/or overlapping) read pair and the genus of interest: short read pairs consist of a limited number of *k*-mers (overlapping 20-nucleotide stretches of sequence), which means that classification of just a very short stretch of sequence is sufficient to surpass the Kraken confidence threshold (0.05, or 5% of *k*-mers in a read).

For *Sapovirus*, however, I identified systematic misclassification: for 13 out of 16 misclassified detections, including one with 31,492 read pairs, the top-scoring taxa are *Salmonella* phages. For these samples, Genbank record AB212270 (Sapovirus Hu/Kolkata/J20816 pseudogene for

RNA dependent RNA polymerase) is often the only sapovirus record represented in the BLAST top hits. A BLAST search of this record showed that it is most similar to *Salmonella* phages, whereas no similarity with other sapoviruses could be detected. This indicates that this record is erroneously labelled as a sapovirus, and that its inclusion in the Kraken viral database (as well as in the BLAST database for the current analysis) results in the erroneous recruitment of read pairs to this genus. I identified a similar issue for the record L23829, labelled as a norovirus helicase but showing most similarity to *Lactobacillus* sequences; however this record resulted in misclassification of only 1-3 read pairs in two of the investigated norovirus detections.

Implications for further pipeline development

The results presented in this section indicate that the basic pipeline is affected by at least two different types of processes resulting in misclassification of reads. False positive detections due to non-systematic misclassification, where reads are derived from a diversity of sources, are easily identified through validation by BLAST: only a low percentage of reads match the genus of interest. A similar validation procedure could easily be integrated into the taxonomic classification pipeline, allowing the inclusion of only those detections with ≥70-80% of queried reads matching the genus of interest. It may be more complicated to identify cases of systematic misclassification due to mislabelled NCBI records or subtaxa: in the validation process, that also uses the NCBI taxonomy, some proportion of reads may match the same mislabelled record and thus appear as truly derived from the genus. However, adapting the pipeline to use custom OTUs, as suggested in section 4.3.4, could reduce the impact of such records. Mislabelled subtaxa could be removed from genus-based OTUs, and erroneous records directly associated with the genus could be marked as not representing the genus during an adapted validation procedure. It is impossible to identify all erroneously labelled records or taxa in NCBI, but an iterative process of adapting OTUs, running the pipeline with adaptations, and critically evaluating the resulting detections may help remove those with a potential impact on the analysis of the larger VIZIONS dataset.

4.3.6 Index switching as batch effect and source of background read pairs in true negatives

During early explorations of the metagenomic data (not reported here), I observed that sequencing runs with higher total read pair counts for a genus correlated with higher

background levels in presumed negative samples on the same run. I hypothesised that this could be related to contamination between samples through index switching (see section 4.1.2). To investigate this further, I defined background levels of read pairs as the distribution of read pair counts in true negatives (see section 4.2.3), and conducted several statistical analyses to characterise the observed patterns.

In eight runs consisting solely of samples from hospital patients, background levels of read pairs vary significantly by run for five test genera (Table 4.10; *Kobuvirus* was not included because only a single run fulfilled the inclusion criteria (see section 4.2.6)). Additionally, in each of these genera, there is indeed strong evidence for a good positive correlation between background read pair counts and the total read pair counts for that genus on the run (Spearman's *rho* 0.70-0.82, $p<0.001$ for all genera). These findings support my hypothesis that the VIZIONS data are affected by index switching contamination.

Table 4.10 Results of ANOVA analyzing the effect of run on background read pair levels
This analysis was not performed for *Kobuvirus* as only a single run fulfilled the inclusion criteria (see section 4.2.6). Df, degrees of freedom.

Genus	Df (between group, within group)	F statistic	p-value
Rotavirus	7, 404	70.202	< 0.001
Norovirus	7, 539	189.34	< 0.001
Mastadenovirus	7, 575	134.8	< 0.001
Sapovirus	6, 496	475.52	< 0.001
Mamastrovirus	4, 335	111.32	< 0.001

Quantification of index switching

Assuming that all background read pairs are due to index switching contamination, and that this occurs at a fixed rate that does not vary by genus, I estimated the percentage of contaminating read pairs on each run as ranging 0.06-0.22% (Table 4.11). These values are consistent with estimates from other studies performed on Illumina sequencing platforms with bridge amplification chemistry (Kircher et al., 2012, Nelson et al., 2014, Wright and Vetsigian, 2016b, Wright and Vetsigian, 2016a, D'Amore et al., 2016).

Table 4.11 Estimation of percentages of reads due to index switching contamination
Performed for each sequencing run. This assumes that all background read pairs come from this source. Numbers of read pairs with switched indices have been inferred by applying a conversion factor to the background levels of read pairs for each genus, and subsequently summing over all included genera (see section 4.2.6). Rp, read pairs

Sequence run	Number of genera included	Sum of background rp	Inferred rp with switched indices	Total rp in run	% rp with switched indices
16020	4	26734	40501.23	44773076	0.0905
16317	6	22085	35840.85	35141876	0.1020
16318	5	42602	83171.98	65779457	0.1264
16370	5	29963	40594.84	66353266	0.0612
16806	4	29430	31933.18	17396471	0.1836
19344	4	188625	214224.74	98991218	0.2164
19345	5	40426	83395.54	66743400	0.1249
19379	4	30286	40705.08	40315402	0.1010

Modelling the association between background read pair counts and total read pair counts in the run

By modelling the association between background read pair counts and total read pair counts, I can use this to set signal thresholds (for subsequent metagenomic analyses, Chapter 5) that take index switching contamination into account.

I considered all background read pair counts for all included combinations of genus and sequencing run as independent data points, and compared the performance of various general linear model structures on log-transformed data. The best fitting model is a quadratic (Table 4.12, Figure 4.7) that explains a large fraction of the variance in background read pair counts ($F(2,2450)=4970$; $p<0.0001$; adjusted $R^2 = 0.80$).

Table 4.12 Details of the best performing (quadratic) regression model
The model was applied to log-transformed background read pair counts ($rp_{taxon,n}$), so that $y = \log_{10}(rp_{taxon,n} + 1)$, whereas total read pair counts for each run ($rp_{taxon,run}$) were first normalised by the number of samples (s) and then log-transformed, so that $x = \log_{10}\left(\frac{rp_{taxon,run}}{s_{run}-1}\right)$

Variable	Estimate	Std. error	p-value
Intercept	0.7720	0.0460	< 0.001
x^2	0.1760	0.0036	< 0.001
x	-0.6792	0.0260	< 0.001

While this model is purely a statistical approximation of the data, and additionally its assumption of independent data is not actually met, it can be considered useful for practical purposes. The upper limit of a prediction interval (e.g. the 99.5% prediction interval as illustrated in Figure 4.7) can be used to set signal thresholds that consider the predicted distributions of background read pairs for specific taxa on specific runs.

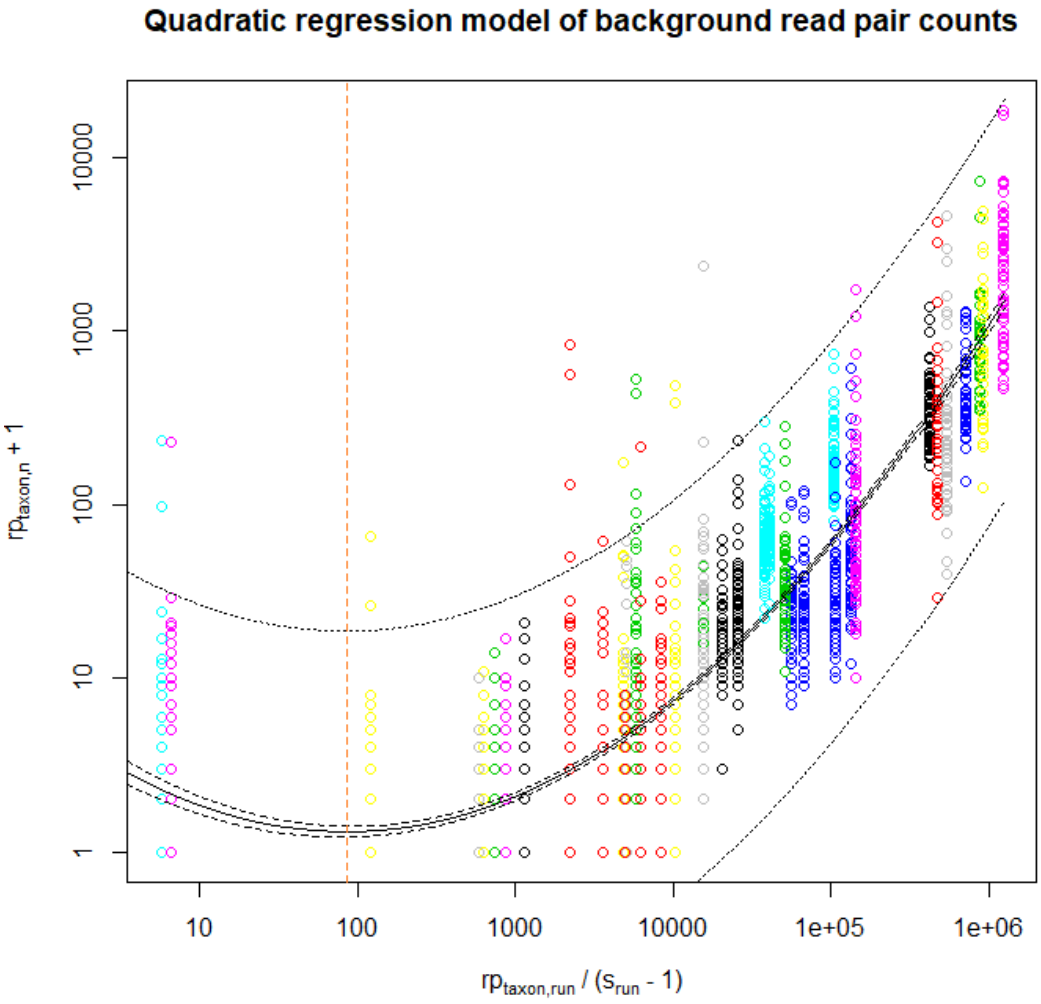


Figure 4.7 Quadratic regression model of background read pair counts
On the x-axis, total read pair counts for each run were first normalised by the number of samples (s). Each column of circles is a set of estimates of background read pair counts for a single test virus and a single run; sets have been coloured by run. The full line represents the quadratic model described in Table 4.12. Dashed black lines are the 95% confidence interval, and dotted lines represent the 99.5% prediction interval suggested for the setting of signal thresholds. The vertical brown dashed line is traced through the minimum of the model. At x values lower than this minimum, the model is not plausible, as it suggests that in runs with a very low overall content of a test virus, index switching contamination is higher.

4.4 Summary and conclusion

This study set out to test the performance of the basic taxonomic classification pipeline, as well as to identify and characterise the processes that could result in loss of sensitivity or specificity. The third aim of the study was to develop ideas for post-hoc adaptations that could be added to the pipeline, to reduce the impact of the identified issues on the metagenomic analysis of the full dataset.

All in all, high AUCs and estimates for sensitivity and specificity, and significant correlations between read pair counts and qPCR results, suggest that the pipeline has a good overall performance. While these findings are subject to an important limitation, namely that they are based on studies of a limited number of viral genera that are well-defined and well-represented in the NCBI database, they are nevertheless reassuring of that the pipeline works as intended.

Investigations of samples with discordant qPCR and metagenomic outcomes inferred a number of ways in which some sensitivity and specificity may be lost, although none of these appeared predominant. Metagenomic false negatives were attributed (in part) to inefficiencies in viral enrichment, low overall sequencing yields, inefficient sequencing of low-level secondary infections, and inappropriate removal of viral sequences at the filtering stage. Additionally, in some false negatives, the presence of sequences derived from viruses closely related to the test viruses suggested that the qPCR signals may actually be false positives, due to cross-reactivity of primers.

Metagenomic false positives could sometimes be explained by misassignment of read pairs to the genera of interest by the pipeline, but this was not the case for the strongest signals, which I then presumed to be qPCR failures (or perhaps contamination). Finally, I considered index switching contamination as the source for background read pair counts in true negative samples, and modelled the relationship between the distribution of this background and total read pair counts in each sequencing run.

In addition to growing my awareness of potential issues affecting the basic pipeline, the value of the studies described in this chapter has been in the opportunities they have generated for further development of the pipeline. Throughout my investigations, I formulated recommendations for post-hoc adaptations I could add to the pipeline, to counteract some

of the identified issues. The application of these recommendations has been described below, and in more detail in section 3.4.

Data normalisation

ROC curve analysis of the pipeline for both raw and normalised read pair counts, described in section 4.3.2, suggested that normalisation procedures were not required for good pipeline performance. I therefore use raw read pair counts in the full analysis of human, swine and rat samples from the VIZIONS study (Chapter 5).

Adapting NCBI taxonomy

In sections 4.3.2 and 4.3.4 I considered that the basic pipeline would probably not perform as well for ill-defined viruses, consisting of multiple unassigned lineages in the NCBI taxonomy, as for well-defined viruses like the six test genera. Creating custom OTUs that, where appropriate, include such unassigned lineages, could limit sensitivity loss for such viruses. Additionally, creating custom OTUs would allow the removal of mislabelled records and taxa, limiting their impact on the specificity of the pipeline (considered in section 4.3.5). I therefore use custom OTUs, instead of standard NCBI genera, in the analysis of the full dataset. How I defined these OTUs has been described in section 3.4.2.

Signal thresholds

Distributions of read pair counts (section 4.3.1) and analysis of background read pair levels (section 4.3.6) indicated that signal thresholds are required to separate signals, representing true infections, from background levels of read pairs, due to index switching. In section 4.3.6, the upper prediction interval limit of a quadratic regression model, predicting background read pair counts based on total read pair counts in all samples of a run, appeared to provide a useful basis for the setting of virus- and run-specific signal thresholds. The findings from this chapter led me to use the 99.5% prediction interval of the model to set OTU- and run-specific signal thresholds for the full metagenomic analysis. The application of these thresholds has been described in section 3.4.3.

Signal validation

In section 4.3.5, I identified misclassification of read pairs as a source of false positives. This led me to add a BLAST-based validation step to the pipeline. The application of this validation step has been described in section 3.4.4.

4.4.1 Conclusion

In conclusion, in this chapter, I found that the basic viral taxonomic classification performs well, with balanced sensitivity and specificity, for several well-studied human pathogens. Whether this performance holds up for ill-defined viruses is, however, uncertain. I also identified several issues that could affect performance of the pipeline, including: the presence of variable background levels of read pairs, presumably due to index switching cross-contamination of samples; unsystematic misclassification of read pairs, leading to false positives; and systematic misclassification of read pairs, due to errors in the NCBI database, also leading to false positives. The findings in this chapter led me to develop three adaptations to the pipeline, to counteract these issues: the creation of custom OTUs; the application of virus- and sequencing run-specific signal thresholds; and the inclusion of a signal validation step. In Chapter 3, section 3.4, I have described the practical application of these adaptations. In Chapter 5, I use the adapted pipeline to describe viral signals in samples from humans, swine, and rats.

Chapter 5. Viruses at the human-animal interface and their relevance to zoonotic emergence

I wrote this chapter with minor comments and text edits from Andrew Rambaut and Mark Woolhouse.

A multitude of collaborators from the VIZIONS consortium were involved in generating the metagenomic sequencing data I analysed in this chapter; their contributions have been detailed in Chapter 2 (data generation) and Chapter 3 (bioinformatic processing).

All analyses in this chapter are completely my own work.

This chapter will be revised for submission as a manuscript with multiple co-authors from the VIZIONS consortium, as appropriate according to their stated contributions.

5.1 Introduction

In Chapter 3 and Chapter 4, I developed and tested a viral taxonomic classification pipeline. In this chapter, I apply the full, adapted pipeline to samples from the VIZIONS hospital and high-risk cohort studies, in order to learn about the zoonotic viruses circulating in the Mekong Delta of Vietnam and to identify any viruses that may be at the cusp of emergence.

This study combines samples from clinical surveillance (hospital study) and from targeted screening of high-risk individuals and animals (high-risk cohort study). As animal hosts species, this study focuses on swine and rats, both for their known roles as hosts of various zoonotic pathogens, and for their importance in local economy and gastronomy. Metagenomic surveillance was chosen as methodology for its ability to detect unexpected and novel viruses as well as known ones.

The specific study goals are three-fold:

First, to contribute to the characterisation of the diversity of human, swine and rat viruses circulating in the study setting. While describing the diversity of non-zoonotic viruses does

not directly contribute to our scientific understanding of zoonoses, development of a knowledge base of locally circulating viruses is nevertheless important in this context. In cases of disease of unknown origin, it could facilitate identification of causative pathogens – including previously unrecognised zoonotic agents. In turn, this would increase the timeliness of the public health response to such putative emergence events.

Secondly, for any detected zoonotic viruses, to assess whether they are shared between human and animal study populations, and, where possible, evaluate their relevance to zoonotic emergence. Knowledge gaps are highlighted, and suggestions are made as to what further studies could be done to improve risk assessment.

Thirdly and finally, to identify and evaluate putative novel zoonotic viruses: viruses that were found in the human study population, but had previously only been found in animals, or vice versa. The identification of putative novel zoonoses in a high-risk setting allows early risk assessment, and, if necessary, the targeting of research and public health measures to avoid emergence at a larger scale.

To address these goals, I apply the pipeline to samples from humans, swine, and rats, and characterise detected viruses to the species, clade or genotype level. After elimination of signals likely due to contamination or non-infectious exposure, I generate an overview of (presumed) viral infections in each population. I categorise the identified viruses according to their zoonotic potential and the population(s) they were detected in. Typical human, ungulate and rodent viruses in their respective hosts are only described by means of population overviews, but known zoonoses and viruses with an unclear origin that may reflect cross-species transmission are investigated and described individually.

5.2 Methods

5.2.1 Data

The samples used in this study have been described in detail in section 2.5 of this thesis. They include 1222 rectal swabs and faecal samples from humans (including both hospital patients and high-risk cohort members), 285 rectal swabs from swine (mostly domestic pigs, but including a few farmed wild boar) and 315 faecal samples from rats, all from Dong Thap province in Vietnam.

5.2.2 Metagenomic sequencing and taxonomic classification

Study samples were enriched for viruses and subjected to metagenomic next-generation sequencing (Illumina HiSeq 2500) by collaborators on the VIZIONS project, as described in section 2.4. To identify viruses present in the resulting sequence data, I used the bespoke taxonomic classification pipeline described in Chapter 3. I applied the basic pipeline and a set of three adaptations designed to improve performance, to obtain sets of sequence read pairs associated with operational taxonomic units (OTUs). These OTUs are mostly based on viral NCBI genera from 2014, but include several newly accepted genera (up until ICTV 2017 (Adams et al., 2017)) and other groups of viruses. Where possible, OTUs have been renamed and contextualised to match a more current state of taxonomy (Adams et al., 2017). Family-level OTUs, where the family contains a single mammal-infective genus, are referred to by the name of this mammal-infective genus rather than the family.

5.2.3 Viral characterisation beyond the OTU

To assess whether signals could be the product of zoonotic transmission, or alternatively represent viruses with zoonotic potential, I characterised signals beyond the OTU level. Where possible, I identified a species, genogroup, genotype or similarly specific phylogenetic clade – whichever grouping was easily identifiable and had the most relevance when considering host range. For this, I used the outputs of the pipeline's validation step: BLAST top hits for up to 1000 read pairs per signal.

I generally considered the reference sequence with the highest total bitscore, summed over matching hits, as proxy ("best-scoring reference sequence") for the detected virus. Information on this reference virus' properties was obtained through literature searches, complemented with further BLAST searches where the literature was sparse. Throughout this chapter, the similarity of a signal to its best-scoring reference sequence is expressed in terms of the average percent identity of all matches to this record, weighted by the length of each match. When this value was below 80%, I considered the virus represented by the signal to be divergent from known viruses, and potentially a novel strain or species.

When a signal had a low identity to its best-scoring reference sequence, or this reference was from a different host species than expected, I additionally investigated the patterns visible in the complete set of BLAST top hits to further interpret the signal. For example, I considered

whether top hits were consistently from the same viral groups or host species, or from a variety of sources.

5.2.4 Likely contaminants and non-infectious exposure

I identified signals representing likely contaminants or non-infectious exposure and removed these from further consideration. To do this, I considered for each signal whether infection is a plausible explanation, based on the biological properties (particularly the known host range) of the virus represented by the best-scoring reference sequence. If infection was deemed implausible, I investigated the context in which the sample was processed and sequenced to identify any putative sources of contamination. Depending on whether a specific source of contamination could be identified, signals were considered probable contaminants, or possible contaminants/non-infectious exposure. I chose this method for the identification and removal of putative contaminants because of its simplicity and its focus on signals that could otherwise be misconstrued as cross-species transmissions.

5.2.5 Categorisation of identified viruses

I categorised the remaining signals according to the zoonotic potential of the viruses they represent. The four categories are:

I. Non-zoonotic human viruses. This includes human viruses that are occasionally detected in animals (likely reverse zoonoses), but that are not generally considered as zoonotic.
II. Non-zoonotic animal viruses. This includes typical ungulate viruses (Category IIa), typical rodent viruses (Category IIb), and animal viruses with an unclear (but probably not human) origin (Category IIc). Category IIc consists of signals that have best-scoring reference sequences from heterologous host species, excluding humans. As an exception, *Cardiovirus* signals in rats that best match to human viruses were included: their overall BLAST top hits are viruses from a variety of host species, and the patterns of these hits match those seen for similar signals with rodent viruses as best-scoring reference sequences.
III. Known or presumed zoonotic viruses. This includes viruses found only in animals in this study (Category IIIa), and viruses found in both animals and humans in this study (Category IIIb).

IV. Putative novel zoonotic viruses. This category consists of viruses detected in human samples in this study, but that had previously only been found in animals, or vice versa.

While I recognise that the cellular tropism of many viruses is not known, I considered previous molecular detections in human and/or animal hosts as sufficient evidence of infection for the purpose of categorisation.

5.3 Results

5.3.1 Likely contaminants and non-infectious exposure

In all samples, there are a total of 22 signals for which I considered infection an implausible explanation, based on the biological properties of the best-scoring reference sequence (Table 5.1). These signals are generally small in size (<100 read pairs assigned to the OTU), suggesting that they represent a low proportion of the total extracted nucleic acid in each sample. They also share a very high (>90%) average percent identity with their best-scoring reference sequences. These findings are consistent with the signals representing contamination with (or non-infectious exposure to) genetic material from a known virus, rather than acute infection with a novel virus variant.

Sixteen of these signals are probable contaminants (Table 5.1), due to the sample metadata or the best-scoring reference sequence suggesting specific sources of contamination. A first such source is other VIZIONS samples that were processed and/or sequenced together with the contaminated samples. For example, bat samples (collected in the context of the VIZIONS project but not included in this thesis) in the same sequencing run are an easily identifiable source for the bat alphacoronavirus signal in a human sample. A second source is provided by non-VIZIONS samples sequenced at the same sequencing facility at around the same time as the contaminated samples. This is most obvious for a Middle East respiratory syndrome-related coronavirus (MERS-CoV) signal: the known geographical distribution of this virus does not include Vietnam, but the involvement of the Sanger Institute in sequencing MERS-CoV outbreak isolates overlapped in time with their work for VIZIONS (Cotten et al., 2014b, Cotten et al., 2013). A third and final source is genetic material contaminating an unidentified laboratory environment or piece of equipment. For example, the murine leukaemia virus-like signal in a human sample is reminiscent of earlier detections of a similar virus, attributed to DNA extraction columns contaminated with murine sequences (Erlwein et al., 2011).

Table 5.1 Description of likely contaminants and non-infectious exposure

Cat., category; N, number of signals; rp, read pairs; % id, average percent identity; Prob cont, probable contaminants; Poss cont/non-inf exp, possible contaminants or non-infectious exposure. Average percent identity to the best-scoring reference sequence was calculated over all reads that had this reference as top hit in the validation step. #Does not include information (accession number and average percent identity) on one best-scoring reference sequence that is related, but labelled as a synthetic construct in NCBI taxonomy

Signal(s)					Best-scoring reference sequence			Evidence that signal(s) due to contamination or exposure	
Cat.	**OTU**	**Sample origin**	**N**	**Size (rp)**	**Accession nr.**	**Species or clade**	**% id**	**Reason(s) infection not plausible**	**Likely source of contamination/exposure**
Prob cont	*Alphacoronavirus*	Human	1	71	DQ648858	*Scotophilus bat coronavirus 512*	95.93	- Virus previously detected only in bats - Sample from a hospital patient who did not report bat contact	Other VIZIONS samples on run/in batch: these included samples from *Scotophilus* bat colonies
	Orthopneumovirus	Rat	1	35	KX765959	*Human orthopneumovirus* (respiratory syncytial virus)	99.55	Host range of virus is limited to humans	Other VIZIONS samples on run/in batch: these included samples from patients with respiratory illness
	Betacoronavirus	Human	1	27	KF294357	*China Rattus coronavirus HKU24* (Longquan Aa mouse coronavirus)	95.46	- Virus previously detected only in rodents - Sample from a hospital patient who did not report contact with rats	Other VIZIONS samples in batch: these included samples from rats with similar viruses
	Betacoronavirus	Human	1	679	KT861628	*Middle East respiratory syndrome-related coronavirus* (MERS-CoV)	99.66	Geographical range of virus not known to include Vietnam	Sequencing facility: same facility sequenced MERS-CoV isolates at around same time
	Lentivirus	Swine	5	28-75	various	*Human immunodeficiency virus 1* (HIV-1)	90.52-98.79	- Host range of virus is limited to humans - Best reference seqs are African strains	Sequencing facility: same facility sequenced HIV strains from Africa

	Lentivirus	Human	1	23	DQ826727	*Human immunodeficiency virus 1* (HIV-1)	91.41	- Best reference seq is African strain - Similar finding in swine samples above	Sequencing facility: same facility sequenced HIV strains from Africa
	Lentivirus	Human	1	33	KJ372918	*Human immunodeficiency virus 1* (HIV-1)	97.33	- Best reference seq is African strain - Similar finding in swine samples above	Sequencing facility: same facility sequenced HIV strains from Africa
	Gammaretrovirus	Human	4	161 - 247	XM_011248564	*Murine leukemia virus* (MLV)-like sequences	99.81-99.91	MLV-related viruses in human samples have previously been attributed to contaminated laboratory environments and equipment (Hue et al., 2010, Erlwein et al., 2011, Katzourakis et al., 2011, Paprotka et al., 2011, Kearney et al., 2012, Lee et al., 2012)	Unidentified laboratory environment or equipment contaminated with murine genetic material
	Influenzavirus A	Human	1	30	HE802059	*Influenza A virus*	99.65	Best reference seq is laboratory strain derived from an isolate from the 1930s, and previously implicated in lab contamination (Enserink, 2005, Worobey, 2008)	Unidentified laboratory environment where experiments with influenzavirus were undertaken
Poss cont/ non-inf exp	*Gammacoronavirus*	Swine	3	18-68	KP118894, KT946798#	*Avian coronavirus* (infectious bronchitis virus)	98.24-98.56 #	Host range of virus is limited to birds	- Chickens at farm - Unidentified source in lab
	Betapapillomavirus	Rat	1	27	AY382779	*Human papillomavirus 96*	93.16	Host range of virus is limited to humans	- Rat traders/humans at market - Unidentified source in lab
	Gammapapillomavirus	Swine	1	52	GU117632	*Human papillomavirus 132*	83.48	Host range of virus is limited to humans	- Humans at farm - Unidentified source in lab
	Mastadenovirus	Swine	1	63	EU128937	*Human mastadenovirus C*	99.87	Host range of virus is limited to humans	- Humans at farm - Unidentified source in lab

For six signals, a specific source of contamination could not be identified (Table 5.1). They are consistent with either sample contamination from an unidentified source (e.g. any of the laboratories involved in processing the sample) or non-infectious exposure of the host to the virus (e.g. through ingestion of contaminated soil or food.

5.3.2 Metagenomic overview

After exclusion of putative contaminants, 3212 signals remain which I assume are due to infection. These are part of 59 OTUs, representing 52 genera, two abolished genera (split up in recent years), and an unclassified group in 22 viral families, as well as four groups not fitting within current taxonomy (Table 5.2). The majority of OTUs (33) are RNA viruses; altogether these contain 2195 signals (68.34%). The 26 DNA virus OTUs (including reverse transcribing DNA viruses) contain 1017 signals.

Many of the detected OTUs consist of viruses that are known to infect mammals, through years of microscopic observations of clinical samples and cultivation in mammalian cell lines. However, other OTUs (or their component viruses) have only been characterised based on genetic sequences, and their true cellular hosts are not known. Several of these (e.g. cosaviruses and novel astroviruses) are part of *bona fide* mammal-infecting virus families and can be assumed to have mammalian hosts. But for others, questions of tropism and host range remain wide open. Examples are the posaviruses within the order *Picornavirales*, and the circular Rep-encoding single-stranded (CRESS) DNA viruses, represented in this thesis by the *Circoviridae*, Po-Circo-like viruses, and novel families *Genomoviridae* and *Smacoviridae*. When categorising such viruses, I assumed that they are mammal-infective, but in the following sections, when discussing specific signals and their implication in the context of zoonotic emergence, I return to these questions of tropism.

Population-specific overviews, including information about the best-scoring reference sequence and categorisation for each signal, are given in the Appendix.

5.3.3 Category I: non-zoonotic human viruses

Non-zoonotic human viruses are represented by 547 (17.03%) signals (Table 5.3).

The human viruses show great diversity: they include major diarrhoeal pathogens (noroviruses, adenoviruses, sapoviruses, and astroviruses), other acutely-infecting pathogens (e.g. measles virus, human respiratory syncytial virus), persistent viruses with

various disease associations (e.g. hepatitis B virus), viruses with unknown pathogenic significance (e.g. cosaviruses) and viruses believed to be commensal (e.g. torque teno viruses). They also include human associated gemykibiviruses and porprismacoviruses, members of two newly-recognised CRESS DNA virus families about which very little is known, but for which human-derived sequences cluster together in phylogenies (Varsani and Krupovic, 2017, Varsani and Krupovic, 2018).

The signals in this category have a generally high average identity (>80%) between the best-scoring reference sequence and the reads matching it (Table 5.3), suggesting that, in absence of further investigations, the biological properties of the detected viruses can be assumed to be similar to those of the reference viruses. However, this identity is lower for multiple signals assigned to OTUs within the *Anelloviridae* family (alpha-, beta-, and gammatorqueviruses). This reflects the large diversity of these viruses (Biagini, 2009), as well as that they are relatively understudied due to difficulties supporting the viral cycle in cell culture and, presumably, the lack of association with disease (Okamoto, 2009).

Interestingly, not all included signals are due to natural infections: four enterovirus signals have polioviruses as best-scoring references, probably related to recent vaccination. Vietnam eliminated polio in 2000, but continued using the trivalent oral polio vaccine, with live but weakened virus, until May 2016 (Gurung et al., 2017).

5.3.4 Category II: non-zoonotic animal viruses

Animal viruses with no indication of zoonotic potential are represented by 1742 (54.23%) signals assumed to be due to infections.

Category IIa: typical ungulate viruses

Typical ungulate viruses, found in the swine samples, are represented by 1298 signals (Table 5.4). They include important pathogens (e.g. porcine reproductive and respiratory syndrome virus, porcine teschoviruses), viruses generally associated with asymptomatic or subclinical infections (e.g. porcine toroviruses), and newly-described viruses with unknown pathogenic potential (e.g. posaviruses and po-Circo-like viruses).

Table 5.2 Viral OTUs detected in the three study populations

With numbers of signals per population. OTUs labelled as "presumed <genus>" were defined at the family level, but are considered to correspond to the indicated genus (being the only mammalian genus within the family). This table uses ICTV 2017 taxonomy (Adams et al., 2017), except in the case of "old genera", which were left as in the pipeline viral database (dating from 2014). Relevant updates in taxonomical status since 2014 have been indicated.

Genome	Family (*-viridae*)	Genus/OTU	Updated ICTV status (2017)	Human	Swine	Rat	Total
dsDNA	*Adeno*	*Mastadenovirus*	Genus	47	51	27	125
dsDNA	*Herpes*	*Cytomegalovirus*	Genus	23	0	0	23
dsDNA	*Herpes*	*Lymphocryptovirus*	Genus	2	0	0	2
dsDNA	*Papilloma*	*Alphapapillomavirus*	Genus	1	0	0	1
dsDNA	*Papilloma*	*Betapapillomavirus*	Genus	11	0	0	11
dsDNA	*Papilloma*	*Pipapillomavirus*	Genus	0	0	1	1
dsDNA	*Polyoma*	presumed *Polyomavirus*	Old genus (split up, 2015)	14	0	0	14
dsDNA	*Pox*	*Molluscipoxvirus*	Genus	1	0	0	1
dsDNA	*Pox*	*Suipoxvirus*	Genus	0	5	0	5
ssDNA	*Anello*	*Alphatorquevirus*	Genus	113	0	0	113
ssDNA	*Anello*	*Betatorquevirus*	Genus	26	0	0	26
ssDNA	*Anello*	*Gammatorquevirus*	Genus	18	0	0	18
ssDNA	*Anello*	*Kappatorquevirus*	Genus	0	1	0	1
ssDNA	*Circo*	*Circovirus*	Genus	1	16	0	17
ssDNA	*Circo*	*Cyclovirus*	New genus (2015)	13	17	1	31
ssDNA	*Genomo*	*Gemykibivirus*	New genus (2016)	2	1	0	3
ssDNA	*Genomo*	*Gemykrogvirus*	New genus (2016)	5	0	0	5
ssDNA	*Parvo*	*Bocaparvovirus*	Genus	6	166	0	172
ssDNA	*Parvo*	*Copiparvovirus*	Genus	0	22	0	22
ssDNA	*Parvo*	*Dependoparvovirus*	Genus	0	27	11	38
ssDNA	*Parvo*	*Protoparvovirus*	Genus	0	10	58	68
ssDNA	*Parvo*	*Tetraparvovirus*	Genus	0	11	0	11
ssDNA	*Smaco*	*Porprismacovirus*	New genus (2017)	3	149	0	152
ssDNA	NA	*Huchismacovirus*-like	Unclassified	0	61	0	61
ssDNA	NA	Po-Circo-like virus	Unclassified	0	88	0	88
dsDNA-RT	*Hepadna*	presumed *Orthohepadnavirus*	Genus	8	0	0	8
dsRNA	*Picobirna*	*Picobirnavirus*	Genus	105	259	66	430

Genome	Family (*-viridae*)	Genus/OTU	Updated ICTV status (2017)	Human	Swine	Rat	Total
dsRNA	*Reo*	*Orthoreovirus*	Genus	1	4	0	5
dsRNA	*Reo*	*Rotavirus*	Genus	384	22	85	491
ssRNA (-)	*Paramyxo*	*Morbillivirus*	Genus	2	0	0	2
ssRNA (-)	*Paramyxo*	*Rubulavirus*	Genus	2	0	0	2
ssRNA (-)	*Pneumo*	*Orthopneumovirus*	Genus (renamed, in new family, 2015)	1	0	0	1
ssRNA (+)	*Arteri*	presumed *Arterivirus*	Old genus (split up, 2016)	0	1	0	1
ssRNA (+)	*Astro*	presumed *Mamastrovirus*	Genus	22	251	59	332
ssRNA (+)	*Calici*	*Norovirus*	Genus	64	2	12	78
ssRNA (+)	*Calici*	*Sapovirus*	Genus	32	23	4	59
ssRNA (+)	*Calici*	St-Valerien swine virus	Unclassified	0	6	0	6
ssRNA (+)	*Corona*	*Betacoronavirus*	Genus	1	0	19	20
ssRNA (+)	*Corona*	*Torovirus*	Genus	0	12	0	12
ssRNA (+)	*Flavi*	*Flavivirus*	Genus	1	0	0	1
ssRNA (+)	*Flavi*	*Hepacivirus*	Genus	2	0	0	2
ssRNA (+)	*Flavi*	*Pegivirus*	Genus	1	0	0	1
ssRNA (+)	*Flavi*	*Pestivirus*	Genus	0	3	0	3
ssRNA (+)	*Hepe*	presumed *Orthohepevirus*	Genus	1	20	23	44
ssRNA (+)	*Picorna*	*Cardiovirus*	Genus	7	0	22	29
ssRNA (+)	*Picorna*	*Cosavirus*	Genus	9	0	0	9
ssRNA (+)	*Picorna*	*Enterovirus*	Genus	69	107	9	185
ssRNA (+)	*Picorna*	*Hunnivirus*	Genus	0	0	41	41
ssRNA (+)	*Picorna*	*Kobuvirus*	Genus	8	18	42	68
ssRNA (+)	*Picorna*	*Mosavirus*	Genus	0	0	1	1
ssRNA (+)	*Picorna*	*Parechovirus*	Genus	37	0	1	38
ssRNA (+)	*Picorna*	*Pasivirus*	Genus	0	69	0	69
ssRNA (+)	*Picorna*	*Rabovirus*	New genus (2016)	0	0	8	8
ssRNA (+)	*Picorna*	*Rosavirus*	Genus	0	0	12	12
ssRNA (+)	*Picorna*	*Salivirus*	Genus	12	0	0	12
ssRNA (+)	*Picorna*	*Sapelovirus*	Genus	0	82	0	82
ssRNA (+)	*Picorna*	*Teschovirus*	Genus	0	91	0	91
ssRNA (+)	NA	Posavirus 1	Unclassified	0	31	0	31
ssRNA (+)	NA	Posavirus 3	Unclassified	0	29	0	29
Total number of valid signals, assumed to be infections				**1055**	**1655**	**502**	**3212**

Table 5.3 Primate viruses found in human samples (Category I)

With lowest identities and signal count. Signals were assigned to a viral species and/or clade based on information about the best-scoring reference sequence acquired through literature searches. The "average percent identity (% id)" to the best-scoring reference sequence was calculated for each signal, as calculated over all reads that had this reference as top hit in the validation step. Given here is the lowest value for each set of signals. This does not take into account signals with best-scoring reference sequences outside the OTU. #These signals were not assigned to viral species, as the correspondence between reference sequences and species was unclear from the literature.

OTU	Species name (synonym or clade name)	Lowest "average % id"	Nº signals Sp.	OTU
Alphatorquevirus	Various torque teno viruses[#]	79.62		113
Enterovirus	*Enterovirus B*	91.78	27	69
	Enterovirus A	89.88	17	
	Enterovirus C	89.86	11	
	Rhinovirus A	88.18	7	
	Rhinovirus C	94.44	5	
	Rhinovirus B	91.11	2	
Norovirus	*Norwalk virus* (GI, GII (human clades))	90.75		64
Mastadenovirus	*Human mastadenovirus F*	99.19	19	47
	Human mastadenovirus C	98.71	9	
	Human mastadenovirus B	99.56	7	
	Human mastadenovirus D	97.72	7	
	Human mastadenovirus A	98.34	5	
Parechovirus	*Parechovirus A*	89.00		37
Sapovirus	*Sapporo virus* (GI, GII (human clades), GV)	83.78		32
Betatorquevirus	Various torque teno mini viruses[#]	73.74		26
Cytomegalovirus	*Human betaherpesvirus 5* (human cytomegalovirus)	97.90		23
Mamastrovirus	*Mamastrovirus 1* (classical human astroviruses (HAstV))	97.88	17	22
	Mamastrovirus 6 (HAstV-MLB strains)	98.10	3	
	Mamastrovirus 9 (HAstV-VA1/HMO-C, VA3/HMO-B and PS)	97.33	2	
Gammatorquevirus	Various torque teno midi viruses[#]	77.94		18
Polyomavirus	*Human polyomavirus 5* (Merkel cell polyomavirus)	99.46	7	14
	Human polyomavirus 2 (JC polyomavirus)	99.71	5	
	Human polyomavirus 4 (WU polyomavirus)	99.49	2	
Salivirus	*Salivirus A* (human klassevirus)	92.58		12

OTU	Species name (synonym or clade name)	Lowest "average % id"	Nº signals Sp.	Nº signals OTU
Betapapillomavirus	*Betapapillomavirus 2*	90.39	6	11
	Betapapillomavirus 1	97.33	4	
	Betapapillomavirus 5	99.75	1	
Cosavirus	*Cosavirus D*	88.97	4	9
	Cosavirus A	91.24	3	
	Cosavirus B	89.91	1	
	Cosavirus E	89.69	1	
Kobuvirus	*Aichivirus A* (Aichi virus 1)	95.60		8
Orthohepadnavirus	*Hepatitis B virus*	98.48		8
Cardiovirus	*Cardiovirus B* (Saffold virus)	90.64		7
Bocaparvovirus	*Primate bocaparvovirus 1*	99.67	5	6
	Primate bocaparvovirus 2	99.85	1	
Porprismacovirus	*Human associated porprismacovirus 2*	88.58		3
Gemykibivirus	*Human associated gemykibivirus 2*	98.78		2
Hepacivirus	*Hepacivirus C* (hepatitis C virus)	97.43		2
Lymphocryptovirus	*Human gammaherpesvirus 4* (Epstein–Barr virus)	99.69		2
Morbillivirus	*Measles morbillivirus* (measles virus)	99.32		2
Rubulavirus	*Human rubulavirus 4* (human parainfluenza virus 4)	96.41		2
Alphapapillomavirus	*Alphapapillomavirus 4*	99.62		1
Betacoronavirus	*Betacoronavirus 1* (human coronavirus OC43)	99.57		1
Cyclovirus	*Human associated cyclovirus 9*	91.32		1
Flavivirus	*Dengue virus* (serotype 2, genotype Asian I (Pickett et al., 2012))	99.48		1
Molluscipoxvirus	*Molluscum contagiosum virus*	99.67		1
Orthohepevirus	*Orthohepevirus A* (serotype 1)	97.90		1
Orthopneumovirus	*Human orthopneumovirus* (human respiratory syncytial virus)	99.60		1
Pegivirus	*Pegivirus C* (human pegivirus)	92.13		1
Total				**547**

Table 5.4 Ungulate viruses found in swine samples (Category IIa)

With lowest identities and signal count. Signals were assigned to a viral species and/or clade based on information about the best-scoring reference sequence acquired through literature searches. The "average percent identity (% id)" to the best-scoring reference sequence was calculated for each signal, as calculated over all reads that had this reference as top hit in the validation step. Given here is the lowest value for each set of signals.

OTU	Species name (synonym or clade name)	Lowest "average % id"	Nº signals Sp.	Nº signals OTU
Mamastrovirus	Unassigned (porcine astrovirus 4-like)	84.64	146	251
	Unassigned (porcine astrovirus 2-like)	85.14	95	
	Unassigned (porcine astrovirus 5-like)	90.47	4	
	Mamastrovirus 3 (porcine astrovirus)	88.94	3	
	Unassigned (porcine astrovirus 3-like)	86.90	3	
Bocaparvovirus	*Ungulate bocaparvovirus 5*	88.44	129	166
	Ungulate bocaparvovirus 3	99.43	17	
	Ungulate bocaparvovirus 2	95.61	13	
	Ungulate bocaparvovirus 4	98.59	7	
Porprismacovirus	*Porcine associated porprismacovirus 4*	86.76	35	113
	Porcine associated porprismacovirus 5	81.57	24	
	Porcine associated porprismacovirus 9	86.94	17	
	Porcine associated porprismacovirus 2	85.75	14	
	Porcine associated porprismacovirus 3	86.62	13	
	Porcine associated porprismacovirus 7	83.98	9	
	Porcine associated porprismacovirus 1	91.85	1	
Enterovirus	*Enterovirus G*	81.19		105
Teschovirus	*Teschovirus A* (porcine teschovirus)	77.39		91
Po-Circo-like virus	Unclassified (Po-Circo-like viruses 21 & 22, 41, 51)	84.25		88
Sapelovirus	*Sapelovirus A* (porcine sapelovirus)	86.17		82
Pasivirus	*Pasivirus A*	83.09		69
***Huchismacovirus*-like**	Unclassified	80.66		61
Mastadenovirus	*Porcine mastadenovirus A*	91.61	35	51
	Porcine mastadenovirus C	97.32	16	
Posavirus 1	Unclassified	91.22		31
Posavirus 3	Unclassified	83.90		29
Dependoparvovirus	Unassigned or taxonomy unclear	88.63	19	27
	Adeno-associated dependoparvovirus B	86.46	8	
Sapovirus	*Sapporo virus* (GIII, GVI, GVII, GVIII, GXI)	81.05		23
Copiparvovirus	Unclassified (porcine parvovirus 6)	97.44	13	22
	Ungulate copiparvovirus 2 (porcine parvoviruses 4 and 5)	99.04	9	

OTU	Species name (synonym or clade name)	Lowest "average % id"	Nº signals Sp.	Nº signals OTU
Kobuvirus	*Aichivirus C* (porcine kobuvirus)	87.43		18
Circovirus	*Porcine circovirus 2*	98.86		16
Torovirus	*Porcine torovirus*	74.92		12
Tetraparvovirus	*Ungulate tetraparvovirus 2* (porcine parvovirus 3; hokovirus) *Ungulate tetraparvovirus 3* (porcine parvovirus 2; Cnvirus)	99.00 98.20	65	11
Protoparvovirus	Unassigned (porcine bufavirus) *Ungulate protoparvovirus 1*	96.47 98.55	91	10
St-Valerien swine virus	Unclassified	89.40		6
Suipoxvirus	*Swinepox virus*	97.05		5
Rotavirus	*Rotavirus B* (porcine genotypes) *Rotavirus H* (porcine genotypes)	96.87 97.99	22	4
Pestivirus	Unclassified (atypical porcine pestivirus)	85.57		3
Norovirus	*Norwalk virus* (GII (porcine clades))	85.49		2
Arterivirus	*Porcine reproductive and respiratory syndrome virus 2* (North American genotype)	99.47		1
Kappatorquevirus	*Torque teno sus virus k2b*	95.78		1
Total				**1298**

Like for the human viruses found in human samples, most signals have good matches to their best-scoring reference sequences. However, signals in two OTUs form exceptions to this finding. Firstly, among 91 *Teschovirus* signals, three have an average identity that is lower than 80%. This suggests that the diversity of these viruses has not fully been explored and that these samples may contain novel, divergent strains. This hypothesis is supported by recent findings of two novel teschovirus serotypes in piglets from China (Sun et al., 2015).

Secondly, 10 of 12 *Torovirus* signals have matches to best-scoring references that not only are of relatively low identity, but also cover just a small portion of the torovirus genome. Further investigations indicated that these signals probably do not represent torovirus infections, but infections with a recombinant enterovirus strain that acquired a torovirus-like cysteine protease, as recently reported from Belgium and the USA (Conceicao-Neto et al., 2017, Shang et al., 2017, Knutson et al., 2017). This is supported by the close similarity between these recombinants and enterovirus sequences previously reported from Vietnam through VIZIONS (Van Dung et al., 2016, Shang et al., 2017, Knutson et al., 2017). In the

current study, all samples with these odd *Torovirus* signals do indeed also have *Enterovirus* signals.

Interestingly, in addition to viral genera or equivalent groups included in this study as OTUs, the pipeline reveals evidence for the presence of picornavirus-like viruses that were discovered too recently to be incorporated into OTUs. In the swine samples, two small "signals" were assigned to the genus *Hunnivirus* by Kraken but showed very high average identities (96.83-97.63%) to a related unassigned virus, porcine picornavirus Japan (Naoi et al., 2016), in the validation step. Similarly, two swine signals were assigned to posavirus 1 by Kraken, but upon validation they showed more similarity to recently discovered husavirus-like viruses (posaviruses Bu-1 and 6282) (Hause et al., 2016, Sano et al., 2016). These "signals" were not validated as part of any included OTUs and are thus not included in any tables or discussed any further in this thesis.

Category IIb: typical rodent viruses

Typical rodent viruses, found in the rat samples, are represented by 370 signals (Table 5.5). They are mostly viruses that were discovered relatively recently through metagenomic studies (for example (Sachsenroder et al., 2014, Firth et al., 2014, Phan et al., 2011)). Their pathogenicity remains unknown, and their sequence diversity is in many cases unexplored.

This also explains the larger number of OTUs with signals matching less well to their best-scoring reference sequences: *Norovirus*, *Sapovirus*, *Mosavirus* and *Parechovirus* signals all have average identities below 80% (Table 5.5). Publicly available sequences for these OTUs are sparse: at the time the Kraken and BLAST databases were created, the NCBI database contained a single Sebokele virus sequence (Joffret et al., 2013), two mosavirus sequences (Phan et al., 2011, Reuter et al., 2014) and six rodent sapovirus sequences (Firth et al., 2014, Sachsenroder et al., 2014). Additionally, some of these viruses have been characterised in rodent populations (canyon mice for mosavirus, African wood mice for Sebokele virus) that are very different from the rats included in this study. It is thus not unexpected to find evidence of rather distantly related viruses in the VIZIONS rat samples.

The signal suggesting the presence of a novel Sebokele virus-like parechovirus is consistent with recent descriptions of similar novel viruses from rats in the US (Firth et al., 2014) and voles in China (Wu et al., 2018). The *Mosavirus* signal represents the first time a virus of this genus is found in rodents of the family Muridae.

Table 5.5 Rodent viruses found in rat samples (Category IIb)
with lowest identities and signal count. Signals were assigned to a viral species and/or clade based on information about the best-scoring reference sequence acquired through literature searches. The "average percent identity (% id)" to the best-scoring reference sequence was calculated for each signal, as calculated over all reads that had this reference as top hit in the validation step. Given here is the lowest value for each set of signals.

OTU	Species name (synonym or clade name)	Lowest "average % id"	N° signals Sp.	N° signals OTU
Rotavirus	*Rotavirus B* (infectious diarrhoea of infant rats (IDIR) agent)	81.63		78
Protoparvovirus	*Rodent protoparvovirus 1*	86.98		58
Kobuvirus	*Aichivirus A* (murine kobuvirus)	87.21		42
Hunnivirus	Unclassified (hunnivirus 83GR-70-RAT106)	95.53		41
Mamastrovirus	Unassigned (rat astrovirus, proposed MAstV25)	87.48		39
Mastadenovirus	*Murine mastadenovirus B*	83.34		27
Betacoronavirus	Unassigned (China *Rattus* coronavirus HKU24)	94.76	15	19
	Murine coronavirus	88.99	4	
Cardiovirus	*Cardiovirus A* (EMCV-2)	88.40	12	16
	Cardiovirus C (Boone cardiovirus)	86.19	4	
Norovirus	*Norwalk virus* (GV)	74.28		12
Rosavirus	Unclassified (rosavirus B)	83.91		12
Dependoparvovirus	Unassigned or taxonomy unclear	82.88		11
Rabovirus	*Rabovirus A*	84.97		8
Sapovirus	*Sapporo virus* (GII (rodent sapovirus 2))	79.30		4
Mosavirus	*Mosavirus A*	77.24		1
Parechovirus	*Parechovirus C* (Sebokele virus)	77.02		1
Pipapillomavirus	*Pipapillomavirus 2*	85.31		1
Total				**370**

Category IIc: animal viruses with unclear origin

Animal viruses with an unclear (but probably not human) origin are represented by 74 signals (Table 5.6). These signals mostly belong to families of which the known diversity is rapidly expanding, and the many signals with low identities (below 80%) to reference sequences suggest that they represent novel viruses. Here, I describe my interpretation of the origin of these viruses, based on their best-scoring reference sequences and other evidence, including signal sizes and any patterns in the full list of top hits from the BLAST signal validation process.

Table 5.6 Signals representing animal viruses with unclear host range (Category IIc)

The "average percent identity" (% id) to the best-scoring reference sequence for each signal was calculated over all reads that had this reference as top hit in the validation step.

Signal(s)				Best-scoring reference sequence				Other evidence	Interpretation
OTU	**Host**	**N**	**Size (rp)**	**Accession nr.**	**Species or clade**	**Host**	**% id**		
Porprismacovirus	Swine	36	19-315	KP233192	*Gorilla associated porprismacovirus 1*	Gorilla	84.30-89.79	Species closely related to *Porcine associated porprismacovirus 7* (Ng et al., 2015)	Swine virus within clade of gorilla and swine viruses (Ng et al., 2015) - possibly infection of an (archaeal) symbiont
Mamastrovirus	Rat	16	31-99069	KT599569 (15), KT599570 (1)	Unassigned (unclear)	Macaque	81.21-88.01	- Large numbers of BLAST top hits to rodent and porcine astroviruses - Sequence similarity between best-scoring reference and rodent viruses	Astroviruses from newly-identified rodent virus clades (To et al., 2017, Hu et al., 2014, Wu et al., 2018) AND/OR recombinants that include sequences from other animal viruses
		3	86-580	KJ571449 (2), KJ571470 (1)	*E. cachinus* AstV Group 1	Kachin red-backed vole	76.58-77.82	Large numbers of BLAST top hits to vole, canine and some other (incl. bovine) astroviruses	
		1	82	KP404149	*Mamastrovirus 5* (canine astrovirus)	Dog	76.97		
Enterovirus	Rat	4	176-1544	KJ641693	Unassigned (Bat picornavirus)	Bat	77.61-79.09	- BLAST top hits to various entero-, sapelo- and rabovirus-like records - Additional validated rabovirus signals in some samples	Novel virus within entero-/sapelo-/rabovirus supergroup AND/OR divergent rabovirus
		4	93-2234	KJ950883	*Rabovirus A*	Norway rat	79.80-81.89		
		1	544	JX627573	*Sapelovirus B*	Macaque	76.39		
	Swine	2	52-91	KJ641696	Unassigned (Bat picornavirus)	Bat	76.60-76.90	BLAST top hits to various entero- and sapelovirus-like records	Novel virus within entero-/sapelo-/rabovirus supergroup

Cardiovirus	Rat	3	21-51147	AB090161 (2), KJ950912 (1)	*Cardiovirus B* (Thera virus)	Norway rat	76.28-78.16	BLAST top hits to various cardiovirus, senecavirus and mischivirus records	Novel cardiovirus or cardiovirus-like virus
		2	66-135	EU723237	*Cardiovirus B* (Vilyuisk human encephalomyelitis virus)	Human or mouse (controversial)	74.92-75.28		
		1	3462	AB747253	*Cardiovirus B* (Saffold virus)	Human	79.72		
Gemykibivirus	Swine	1	47	KF371634	*Black robin associated gemykibivirus 1*	Black robin	76.04	NA	Uninterpretable – novel swine virus OR cross-species transmission OR other

Gorilla associated porprismacoviruses in swine samples

Gorilla associated porprismacovirus 1 (genus *Porprismacovirus*, family *Smacoviridae*) is a newly recognised CRESS DNA virus species, defined on the basis of two genomes detected in faeces from gorillas in a zoo in the United States (Ng et al., 2015). Other porprismacoviruses have been detected in faeces from various birds and mammals, including swine and humans (including in this study, see Table 5.3 and Table 5.4). However, smacoviruses have only been detected with molecular methods, and it is unclear whether mammals and birds represent their true replication hosts (Varsani and Krupovic, 2018).

In this study, thirty-six swine samples have signals matching a gorilla associated smacovirus as best-scoring reference sequence (Table 5.6). At first it may appear as if these signals originated from cross-species transmissions, but published smacovirus phylogenies show that gorilla associated smacoviruses fall within a clade of porcine associated viruses, of the species *Porcine associated porprismacovirus 7* (Ng et al., 2015, Varsani and Krupovic, 2018). The most parsimonious explanation is that the signals in the VIZIONS samples represent smacoviruses circulating in swine, rather than cross-species transmissions from primates.

Alternatively, as the signals are small, they could represent non-infectious exposure, or viruses infecting porcine symbionts. Smacovirus-like sequences were recently identified as CRISPR spacers in archaea, leading to the suggestion that, rather than infecting animals directly, smacoviruses infect archaea living in animal gastrointestinal systems (Diez-Villasenor and Rodriguez-Valera, 2019)

Astroviruses in rat samples

The genus *Mamastrovirus* (family *Astroviridae*) contains viruses that infect and cause disease in a broad range of mammals, including humans, swine and rodents (Donato and Vijaykrishna, 2017). In recent years, the known diversity of these viruses has expanded significantly through large scale sequencing initiatives. For example, a variety of novel, still unassigned, astroviruses have recently been described in different rodent species in China (Hu et al., 2014, Wu et al., 2018, To et al., 2017).

Here, sixteen astrovirus signals in rats have two sequences from macaques as best-scoring references (Table 5.6). In the BLAST validation step, reads from the signals match to a variety of mamastroviruses; in addition to the macaque-derived sequences, murine, rat and porcine viruses have the highest overall bitscores. BLAST queries of the macaque-derived sequences

reveal that they too share high identities with rodent astroviruses, within Lineage 1 from (Wu et al., 2018) and Cluster A from (To et al., 2017). Together, these results suggest that the investigated signals likely represent novel rodent astroviruses. Sequence regions with similarity to viruses from other hosts may have been obtained through recombination and/or they could represent overall phylogenetic relatedness (murine astroviruses are closely related to porcine astroviruses).

Additionally, one astrovirus signal in a rat has a canine astrovirus (species *Mamastrovirus 5*) as best-scoring reference sequence, and large numbers of top hits matching vole astroviruses (*Eothenomys cachinus* AstV Group 1 from (Hu et al., 2014), corresponding to Lineage 2 from (Wu et al., 2018)). A further three signals with vole astroviruses as best-scoring reference sequences have large numbers of top hits matching canine astroviruses. These results suggest that, despite their different best-scoring references, the four signals represent closely related astroviruses. They may be members of a novel rat astrovirus clade with similarities to the identified vole clade, or recombinants of rodent (rat/vole) and non-rodent (canine) viruses.

Enterovirus-like viruses in rat and swine samples

The genera *Enterovirus, Sapelovirus* and *Rabovirus*, and multiple unassigned viruses, form a "supergroup" in the family *Picornaviridae* (supergroup 3 on www.picornaviridae.com). Enteroviruses infect a variety of mammals, including humans, swine and rodents; sapeloviruses infect swine, non-human primates, and birds; rabovirus sequences have been detected in rodents; and related unassigned viruses have been detected in a variety of mammals, including bats (Zell et al., 2017).

In this study, nine rat samples contain signals assigned to the genus *Enterovirus*, with sufficient matches to this OTU to pass the signal validation stage. However, these signals all have other viruses from supergroup 3 as best-scoring reference sequences (Table 5.6): bat picornaviruses (part of Clade 1 in (Wu et al., 2016)), raboviruses, and simian sapeloviruses. Within each investigated signal, top hits for individual reads also match a variety of sequences from throughout this supergroup. These results suggest that the samples contain viruses that fall within this supergroup, but are not easily classifiable beyond that. Interestingly, in five samples, *Rabovirus* signals (Category IIb) were also detected and validated, and it is unclear whether these and the enterovirus-like signals represent one and the same virus (perhaps a divergent rabovirus strain), or two related viruses in the same

samples. The suggested presence of potentially novel viruses in the supergroup is consistent with the recent discovery of enteroviruses and sapelovirus-like viruses in rodents in China (Du et al., 2016, Wu et al., 2018).

Two swine samples also contain validated signals for *Enterovirus* with low-identity matches to a best-scoring reference sequence of a bat picornavirus (part of Clade 2 in Wu et al. (2016); Table 5.6). Like in the rat signals described above, top hits within each signal match to a variety of sapeloviruses, enteroviruses and bat picornaviruses, again suggesting the presence of a related but unclassifiable virus in the samples. The hypothesis that the signals represent a novel virus in supergroup 3 is supported by a recent description of a new porcine picornavirus, sapelo-like porcine picornavirus Japan (Masuda et al., 2018), with extensive sequence similarity to the same bat picornavirus BtVs-PicoV/SC2013.

From these read-based metagenomic investigations it is impossible to determine whether the putative novel enterovirus-like viruses represent recent cross-species transmissions (possibly from bats) into rats and/or swine, or viruses that have already been circulating for some time in the relevant host populations.

Cardiovirus-like viruses in rat samples

Cardiovirus B (genus *Cardiovirus*, family *Picornaviridae*) is a diverse species. It encompasses Theiler's murine encephalomyelitis virus (a murine pathogen), Vilyuisk human encephalomyelitis virus (a controversial human pathogen - possibly a contaminant of murine origin (Lipton, 2008)), thera virus (detected in healthy rats), genet faecal theilovirus (detected in a clinically normal genet), and Saffold viruses (human pathogens).

In this study, six rat signals have best-scoring reference sequences corresponding to different viruses within the species *Cardiovirus B* (Table 5.6). The top hits for individual reads show similar patterns for all six signals: they include the same cardioviruses, as well as viruses from related genera, e.g. *Senecavirus* and *Mischivirus*. These results suggest the presence of a novel cardiovirus or cardiovirus-like virus in these samples. This is supported by previous VIZIONS-based studies finding new cardiovirus sequences in Vietnamese rats (Nguyen, 2015), as well as by recent reports of novel cardioviruses in multiple rodent species from China (Wang et al., 2018, Wu et al., 2018).

While these signals have a potential link to human viruses, illustrated by their BLAST matches to Saffold viruses and Vilyuisk human encephalitis virus, the variety of the BLAST top hits and

the low identities suggest that these matches reflect remote shared ancestry of viruses rather than any recent cross-species transmission or recombination events. As most other cardioviruses have rodents as reservoir hosts, it is likely that this putatively novel cardiovirus is a rodent virus too.

Gemykibivirus in a swine sample

Black robin associated gemykibivirus 1 (genus *Gemykibivirus*, family *Genomoviridae*) is a newly recognised CRESS DNA virus species consisting of a single virus, the genome of which was detected in the faeces of a passerine bird (Sikorski et al., 2013). Other gemykibiviruses have been identified in a dragonfly, and multiple birds and mammals, including humans and cattle, but not swine (Varsani and Krupovic, 2017). However, they have only been detected with molecular methods and their replication hosts are not known (the only genomovirus with known tropism, a gemycircularvirus, infects fungi (Yu et al., 2010)).

Here, one swine sample has a signal for *Gemykibivirus* (Table 5.6), with faecal-associated gemycircularvirus 8 (species *Black robin associated gemykibivirus 1*) as best-scoring reference. The match has a low average identity, indicating that the virus in the sample could be a new gemykibivirus species. It is unclear whether the signal represents a recent cross-species transmission event, the discovery of an established virus-host relationship, or a different situation altogether (e.g. non-infectious exposure, or infection of a member of the swine microbiome).

5.3.5 Category III: known and presumed zoonotic viruses

Viruses with an established zoonotic potential, or for which this is widely presumed on the basis of phylogenies, are represented by 914 (28.46%) signals. Here I introduce these viruses, summarising existing evidence of their zoonotic status and any disease associations, before describing the signals detected in my study. For zoonotic viruses found in both animal and humans in this study, I compare the characterisations across study populations to clarify whether the identified viruses are truly shared, beyond the genus or species level.

Category IIIa: zoonotic viruses found only in animal populations

Fifty-two signals represent zoonotic viruses that, within this study, have only been found in animal species (Table 5.7).

Table 5.7 Signals for zoonotic viruses in animal populations only (Category IIIa)
OTU, operational taxonomic unit; N, number of signals; % id, average percent identity. The "average percent identity (% id)" to the best-scoring reference sequence was calculated for each signal, as calculated over all reads that had this reference as top hit in the validation step. Given here is the lowest value for each set of signals.

Signals			Best-scoring reference sequences					
OTU	Host	N	Accession nr.	Species (genotype)	Host	Country	Reference(s)	Lowest "average % id"
Orthohepevirus		15	JQ679014, AF060668-9, AB089824	*Orthohepevirus A* (HEV-A3)	Human	UK, US, Japan	(Shukla et al., 2012, Erker et al., 1999, Schlauder et al., 1998, Tokita et al., 2003)	89.26
	Swine	5	AB740232, KJ507956		Swine	Japan, Canada	(Shiota et al., 2013, Ward et al., 2015)	91.21
Orthohepe	Rats	23	AB847305-7, LC145329, LC145332	*Orthohepevirus C*	Rats (*Rattus rattus*)	Indonesia	(Mulyanto et al., 2014, Takahashi et al., 2016)	84.65
Rotavirus	Swine	9	KX362470, KX362475, KX362492, KX362502, KX362503, KX362505	*Rotavirus C*	Swine	Vietnam	(Phan et al., 2016a)	94.63

Orthohepevirus A genotype 3, in swine samples

The species *Orthohepevirus A* (HEV-A; genus *Orthohepevirus*, family *Hepeviridae*) contains hepatitis E viruses (HEV) isolated from a wide range of mammals. The genotype detected in this study, genotype 3, has a broad host range, with swine considered to be the main reservoir host (Purdy et al., 2017, Lu et al., 2006, Doceul et al., 2016). Genotype 3 is also commonly associated with infections in humans, which are presumed to be zoonotic, with transmission occurring via the consumption of raw or undercooked animal products (Takahashi et al., 2004, Tei et al., 2003, Colson et al., 2010) and possibly via direct contact with animals (Renou et al., 2007). Molecular evidence suggests a (near-) absence of transmission barriers between human and swine (Bouquet et al., 2011, Bouquet et al., 2012). Human HEV infections are generally asymptomatic or associated with self-limiting acute hepatitis, but they can become chronic and result in severe disease or death in immunocompromised patients. In swine, HEV infections are asymptomatic.

Here, 20 swine samples have signals for HEV-A3. The signals have high-identity matches to a variety of best-scoring reference sequences from porcine strains (Shiota et al., 2013, Ward et al., 2015), human clinical isolates (Erker et al., 1999, Schlauder et al., 1998, Tokita et al., 2003) and a human-origin laboratory strain (Shukla et al., 2012) (Table 5.7).

Orthohepevirus C in rat samples

Orthohepevirus C (HEV-C; genus *Orthohepevirus*, family *Hepeviridae*) contains HEV isolates from rodents, eulipotyphlids and mustelids (Purdy et al., 2017). In 2018 and 2019, the first human HEV-C infections were detected. Human cases known so far involve six patients with co-morbidities presenting with persistent hepatitis or liver function derangement in Hong Kong (Coston, 2019, Siddharth et al., 2018), and a previously healthy patient presenting with severe acute hepatitis in Canada (Andonov et al., 2019). Whether HEV-C causes disease in rodents is unknown.

In this study, 23 rat samples have HEV-C signals. The signals all match with high identity to reference sequences from rats sampled in Indonesia (Mulyanto et al., 2014, Takahashi et al., 2016) (Table 5.7). These reference sequences are part of two major phylogenetic clades that also include sequences from Vietnamese rats and the first human patient in Hong Kong, but not the patient from Canada (Takahashi et al., 2016, Andonov et al., 2019).

Rotavirus C in swine samples

Members of species *Rotavirus C* (RVC; genus *Rotavirus*, family *Reoviridae*) are pathogens of humans and animals. In humans, RVC has been associated with both sporadic cases and large outbreaks of diarrhoea in all age groups. In addition to humans, RVC has been detected in pigs, cattle, ferrets and dogs. Human and animal RVC sequences generally cluster separately in phylogenies (Suzuki et al., 2015, Yamamoto et al., 2011, Phan et al., 2016a), but there is evidence that RVC strains are capable of zoonotic transmission: porcine RVC sequences have been found in human samples (Gabbay et al., 2008); a human-origin RVC segment was found in porcine samples (Kattoor et al., 2017); and farming communities in England and Wales have higher RVC seroprevalences than urban populations (Iturriza-Gomara et al., 2004).

Here, nine swine samples have RVC signals (Table 5.7). The best-scoring reference sequences are all from domestic pigs, sampled during an earlier VIZIONS study in Dong Thap province, Vietnam (Phan et al., 2016a). Unsurprisingly, the identities of the matches are very high.

However, the identification and characterisation of RVC signals is impacted by two limitations of this study, associated with the use of a single best-scoring reference per signal per OTU (here the genus *Rotavirus*). First, rotaviruses have segmented genomes, but each best-scoring reference sequence only describes one segment, meaning that the similarity of other segments to porcine or other rotaviruses remains unexplored. As a result, it may be that reassortants with segments originating from cross-species transmissions have been missed in this analysis. Additionally, in the case of infections with multiple different rotaviruses, the signal gets assigned to a single species, most likely representing the dominant infection. In a previous study in Vietnam, RVC infections were found in humans, but only as part of co-infections with *Rotavirus A* (Nguyen et al., 2007b). It is thus possible that there are additional RVC infections in this study, including in the human samples, that remain "hidden" as part of other *Rotavirus* (particularly *Rotavirus* A) signals.

Category IIIb: zoonotic viruses found in animal and human populations

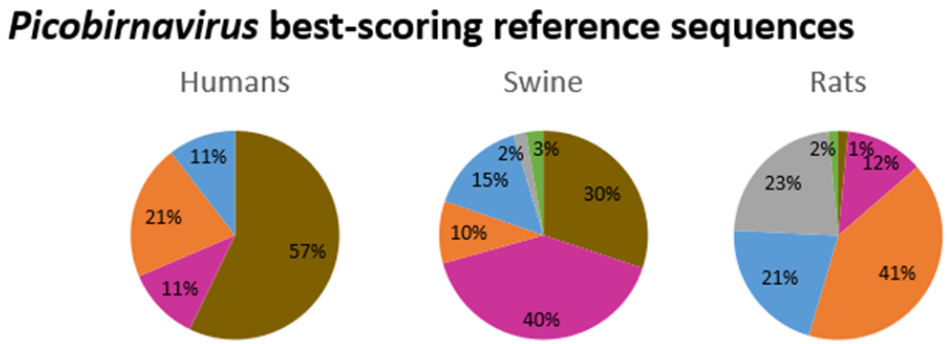
Eight hundred sixty-two signals represent zoonotic viruses that, within this study, are shared between animal and human host populations.

Picobirnavirus

Picobirnaviruses (PBV; genus *Picobirnavirus*, family *Picobirnaviridae*) have been detected in humans and other animals, including a wide variety of mammals, birds, reptiles and invertebrates (Delmas et al., 2019, Bodewes et al., 2014, Fregolente et al., 2009, Smits et al., 2012). PBV sequences do not cluster by host species in published phylogenies (Banyai et al., 2008, Smits et al., 2011, Ganesh et al., 2012, Malik et al., 2014, Gonzalez et al., 2017, Woo et al., 2016), and on this basis it is often considered to be zoonotic. However, it has never been cultured and its natural host remains unknown. PBVs have been found in stools from both diarrhoeic and healthy humans. It has been suggested that they are opportunistic pathogens in immunocompromised individuals, but the evidence is inconclusive (Ganesh et al., 2012).

In this study, I detected 430 PBV signals: 105 in human samples, 259 in swine samples and 66 in rat samples. For each study population, best-scoring reference sequences come from a wide range of sources, covering four mammalian taxonomic orders, birds, and sewage (Figure 5.1). Given the lack of clustering by hosts in phylogenies, it is unclear how well the hosts of these reference sequences reflect the immediate sources of PBVs in the current study. However, the majority of PBV signals in human samples have primate-derived reference

sequences, swine signals most often match ungulate-derived reference sequences, and rat signals have a significant proportion of rodent-derived reference sequences (Figure 5.1).



Host of best-scoring reference sequence	Humans (n = 105)			Swine (n = 259)			Rats (n = 66)		
	N	Accession nrs.	Countries	N	Accession nrs.	Countries	N	Accession nrs.	Countries
Primates									
Human	55	21 records, incl. KR827418 (12)	Various	71	13 records, incl. KU892530 (23)	Various	0	N/A	N/A
Macaque	5	5 records, incl. KT334936 (1)	Bangladesh, China	7	KT335010 (5), KT334917 (1), KT334992 (1)	Bangladesh	1	KT335056	Bangladesh
Ungulates									
Camel	4	KM573806 (2), KM573802 (1), KM573807 (1)	United Arab Emirates	65	7 records, incl. KM573802 (25)	United Arab Emirates	3	KM573798 (1), KM573801 (1), KM573807 (1)	United Arab Emirates
Swine	8	KF861773 (4), KJ650571 (3), HM070240 (1)	Italy, India, China	22	13 records, incl. KC841459 (3)	China, Italy, India	5	HM070240 (3), KF861773 (2)	China, Italy
Horse	0	N/A	N/A	13	KR902505	USA	0	N/A	N/A
Cow	0	N/A	N/A	5	AB828072 (4), KP843617 (1)	Japan, Brazil	0	N/A	N/A
Carnivores									
Fox	14	KP941111	Croatia	16	4 records, incl. KF823810 (10)	Spain, Netherlands, Croatia	25	4 records, incl. KF823810 (11)	Spain, Netherlands, Croatia
California sea lion	1	JQ776552	China	7	JQ776552	China	0	N/A	N/A
Cat	6	KF792838	Portugal	0	N/A	N/A	0	N/A	N/A
Wolf	1	KT934307	Portugal	1	KT934308	Portugal	2	KT934307 (1), KT934308 (1)	Portugal
Genet	0	N/A	N/A	1	KF823812	Spain	0	N/A	N/A
Birds									
Chicken	10	4 records, incl. KM254165 (7)	South Korea, Brazil	31	4 records, incl. KM254164 (16)	South Korea, Brazil	11	4 records, incl. KM254161 (6)	South Korea
Turkey	1	KJ476132	USA	8	KJ476133 (6), KJ476129 (1), KJ476131 (1)	USA	3	KJ476125 (1), KJ476126 (1), KJ476133 (1)	USA
Rodents									
Mouse	0	N/A	N/A	1	JF755419	USA	12	JF755419	USA
Meadow vole	0	N/A	N/A	4	JF755420	USA	3	JF755420	USA
Environment									
Wastewater/ sewage	0	N/A	N/A	7	5 records, incl. KJ135810 (2), KJ135877 (2)	China, USA	1	KJ135916	China

Figure 5.1 Origins of *Picobirnavirus* best-scoring reference sequences
N, number of signals.

Despite this apparent structure, the signals from the human, swine and rat study populations share several of their best-scoring reference sequences. Fifty-four percent of human signals match sequences that are also best-scoring reference sequences for swine signals, and 21% are shared with rats (Figure 5.2).

All three study populations contain signals matching to dromedary camel strain c5128 (KM573807), fox strain 55590 (KP941111), and porcine strains 221/04-16/ITA/2004 (KF861773) and pig/SD (HM070240). This indicates that relatively closely related PBVs circulate in multiple different host populations in Dong Thap province.

The range of average identities to the best-scoring reference sequences includes both values above 99% and below 80% (see the Appendix), suggesting that the study samples contain both known and divergent PBVs.

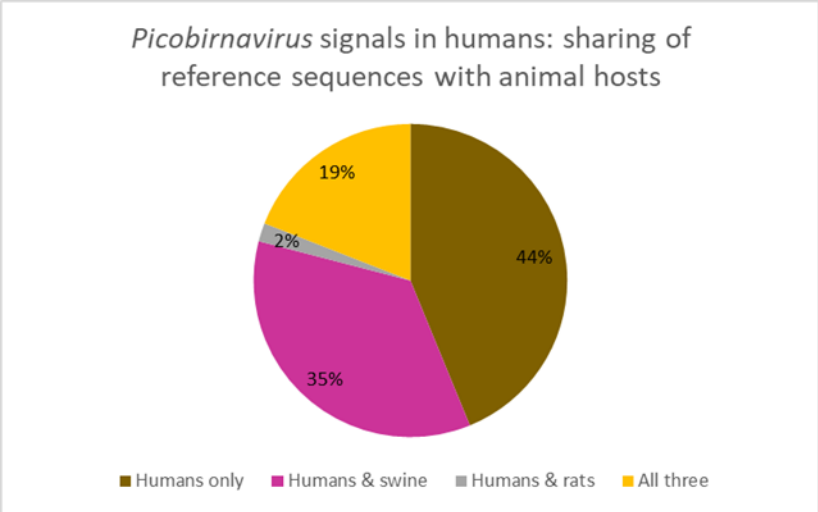


Figure 5.2 Sharing of *Picobirnavirus* best-scoring reference sequences
Proportions of human *Picobirnavirus* signals that share a best-scoring reference sequence with animal study populations.

Rotavirus A

Members of the species *Rotavirus A* (RVA; genus *Rotavirus*, family *Reoviridae*) are important diarrhoeic pathogens of mammals and birds. In humans, RVA is the predominant cause of diarrhoea worldwide, leading to about 200,000 deaths per year in children under the age of five. RVA is an established zoonosis, and its history of and potential for cross-species transmissions have been extensively reviewed (Cook et al., 2004, Martella et al., 2010, Ghosh and Kobayashi, 2014). Briefly, different segment constellations appear to be common in

different animals, but the continuous detection of unusual constellation/host combinations suggests that host species boundaries are leaky, and that cross-species transmission and reassortment contribute to the maintenance of genetic diversity of RVA.

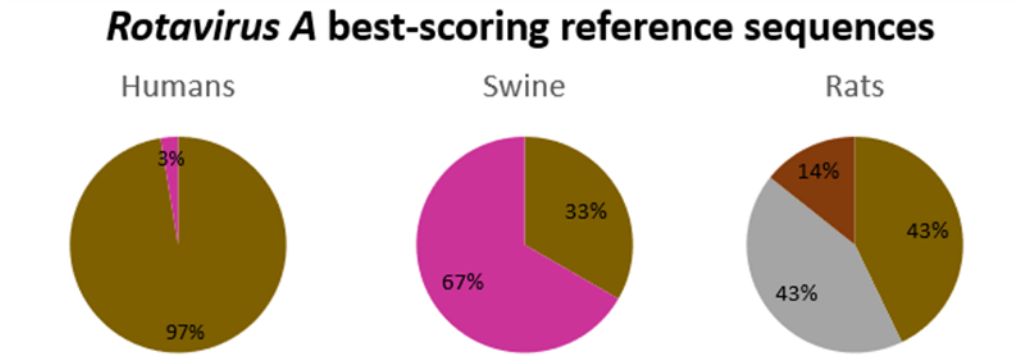
Here, 400 samples have signals with best-scoring reference sequences falling within species *Rotavirus A*. These are 384 signals in human samples, nine in swine samples and seven in rat samples.

The great majority of signals (374 human, 6 swine, and 3 rat signals) have best-scoring references from host species that match the relevant study populations (Figure 5.3). For the human and swine signals, the identified references include human and porcine sequences from an overlapping study by the VIZIONS consortium[3] (Phan et al., 2016a). Other human signals match to human sequences from Thailand (Tacharoenmuang et al., 2016) and a variety of countries in other continents. The rat signals all have the same best-scoring reference, from a rat in Germany (Sachsenroder et al., 2014). These results suggest that these viruses were obtained through transmission from within each population, rather than from other host species. However, by considering just a single best-scoring reference sequence per signal, one cannot exclude that the signals actually represent reassortants that also have segments normally associated with viruses infecting other host species.

Of the remaining RVA signals, ten human and three swine signals have best-scoring reference sequences from porcine and human rotaviruses respectively, suggesting possible origins in cross-species transmission (Figure 5.3). The best-scoring references for the human signals are all from the previously mentioned overlapping VIZIONS study (Phan et al., 2016a). One of these signals is in a sample that was also included in this earlier study, where the identified virus was explicitly described as having a porcine-like constellation of genotypes, shared with pig samples in the same geographical area. It is plausible that the other nine signals represent similar porcine-like viruses, or they could be reassortants with at least one porcine-origin segment. The three swine signals have best-scoring references from the same VIZIONS study and from a Thai study (Tacharoenmuang et al., 2016). One could interpret these the same way as the porcine-like viruses in human samples, but a different reading is also plausible: the signals are all very small (Figure 5.3), and could thus represent non-infectious exposure

[3] In the signal validation BLAST database, I excluded virus sequences that were identified in the same next-generation sequence data and the same samples as I used as input (see Chapter 2). The overlap with this study consisted of 96 human samples; pig samples did not overlap.

to human virus, or contamination from the human samples. However, altogether these results indicate exchange of rotaviruses between humans and swine in Dong Thap province. This is illustrated further by 5% of human RVA signals sharing best reference sequences with swine RVA signals in this study (Figure 5.4).



Humans (n = 384)			Swine (n = 9)			Rats (n = 7)		
N	Accession nrs.	Countries	N	Accession nrs.	Countries	N	Accession nrs.	Countries
Best-scoring reference sequences from humans								
374	56 records, incl. KU550279 (65)	Various	3	KX363301 (1), LC169985 (1), LC169990 (1)	Vietnam, Thailand	3	KF541282	USA
Best-scoring reference sequences from swine								
10	7 records, incl. KX363412 (3)	Vietnam	6	5 records, incl. KX363352 (2)	Vietnam	0	N/A	N/A
Best-scoring reference sequences from rats								
0	N/A	N/A	0	N/A	N/A	3	KJ879450	Germany
Best-scoring reference sequences from bats								
0	N/A	N/A	0	N/A	N/A	1	KF649188	China

Figure 5.3 Origins of *Rotavirus A* best-scoring reference sequences
N, number of signals.

Finally, four rat RVA signals have best-scoring reference sequences from non-rodent host species, with further investigations indicating that the interpretation of these signals is complicated. Three of the signals match to a VP2 sequence from human-derived strain 200972711 (KF541282) from the USA (Mijatovic-Rustempasic et al., 2016); one matches to a VP7 sequence from bat-derived strain MYAS33 (KF649188) from China (Figure 5.3) (Xia et al., 2014). Interestingly, these sequences are both within segment genotypes (C3 and G3, respectively) that are also often found in Chinese rodents (Li et al., 2016a), indicating that these matches need not represent cross-species origins. When considering BLAST top hits at the individual read level, the four signals show patterns that are similar to those of the three

signals with best references from rodents, discussed above: they all include large numbers of hits to the same (human) VP2, (bat) VP7 and (rat) VP3 sequences, as well as a variety of matches to other segments from rat strains KS/11/0573, RA116 and WC179 (Sachsenroder et al., 2014, Li et al., 2016a). The various rat signals could thus represent reassortants of rodent and non-rodent (bat and/or human) viruses, or, alternatively, novel "typical" rodent viruses, with the non-rodent top hits reflecting the very limited representation of rodent RVA sequences in the NCBI database.

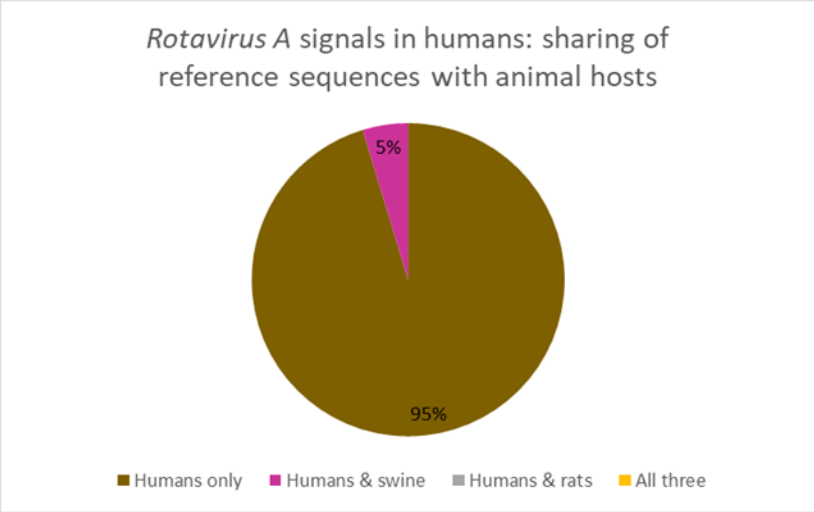


Figure 5.4 Sharing of *Rotavirus A* best-scoring reference sequences
Proportions of human *Rotavirus A* signals that share a best-scoring reference sequence with animal study populations.

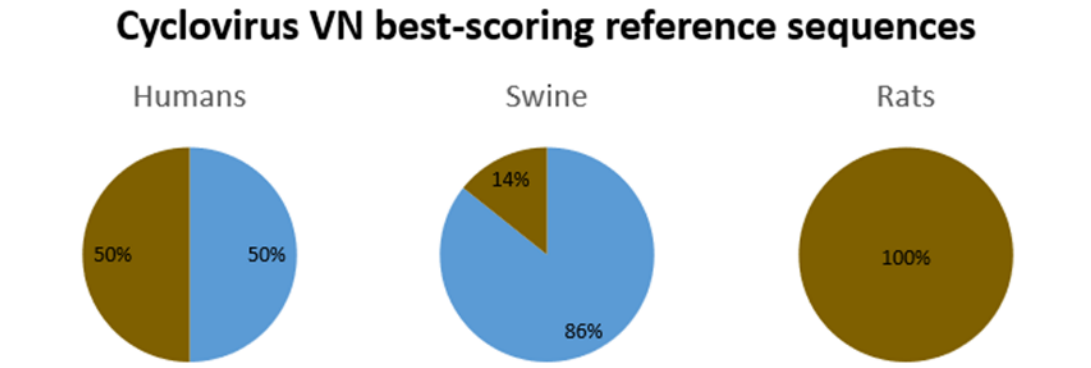
For the rotavirus signals in this study, the average identities to the best-scoring reference sequences are very high (>97%) for all human and swine signals, and slightly lower (85-95%) for all rat signals. This is consistent with the NCBI database covering human and porcine RVA strains much better than rodent RVA isolates. However, for RVA, like for RVC, these specific values have minimal meaning as each best-scoring reference sequence represents only one of 11 genome segments.

Human associated cyclovirus 8 (Cyclovirus VN)

Viruses of the species *Human associated cyclovirus 8* (Cyclovirus VN (CyV-VN); genus *Cyclovirus*, family *Circoviridae*) were discovered in 2013 in patients with acute central nervous system infections, and subsequently also found in healthy humans and in pigs, poultry, rodents and shrews (Tan le et al., 2013, Sasaki et al., 2015). CyV-VN has only been

described on the basis of sequences, and its natural host remains unknown. Similarly, its role in human or animal disease is unclear.

Twelve signals in human samples, 14 in swine samples and one in a rat sample have best-scoring reference sequences within the species *Human associated cyclovirus 8* (Figure 5.5). All but one of the signals match best to sequences previously identified in human and poultry samples from Vietnam (Tan le et al., 2013). The final signal, from a swine sample, matches to a human-derived sequence from Madagascar (Garigliany et al., 2014).



Humans (n = 12)			Swine (n = 14)			Rats (n = 1)		
N	Accession nrs.	Countries	N	Accession nrs.	Countries	N	Accession nr.	Country
Best-scoring reference sequences from chickens								
6	KF031471	Vietnam	12	KF031471	Vietnam	0	N/A	N/A
Best-scoring reference sequences from humans								
6	KF031466 (3), KF031468 (3)	Vietnam	2	KF031468 (1), KM392289 (1)	Vietnam, Madagascar	1	KF031468	Vietnam

Figure 5.5 Origins of cyclovirus VN best-scoring reference sequences
N, number of signals.

There are strong indications that the signals in this study represent a population of viruses that is shared between humans, swine and rats as well as other animals in Vietnam. Two of the four best reference sequences are shared between different host populations in the study: isolate hcf4 (KF031468, from a Vietnamese CNS patient) between humans, swine and rats, and isolate cs1 (KF031471, from a Vietnamese chicken) between humans and swine. Seventy-five percent of human signals match to these shared reference sequences (Figure 5.6). Additionally, all signals have very high average identities (96.80-99.44%; see the Appendix) to their reference sequences, and these were also reported to share pairwise identities of >97% between themselves (Tan le et al., 2013, Garigliany et al., 2014). Altogether these results show that CyV-VN circulates in multiple host populations in the study setting.

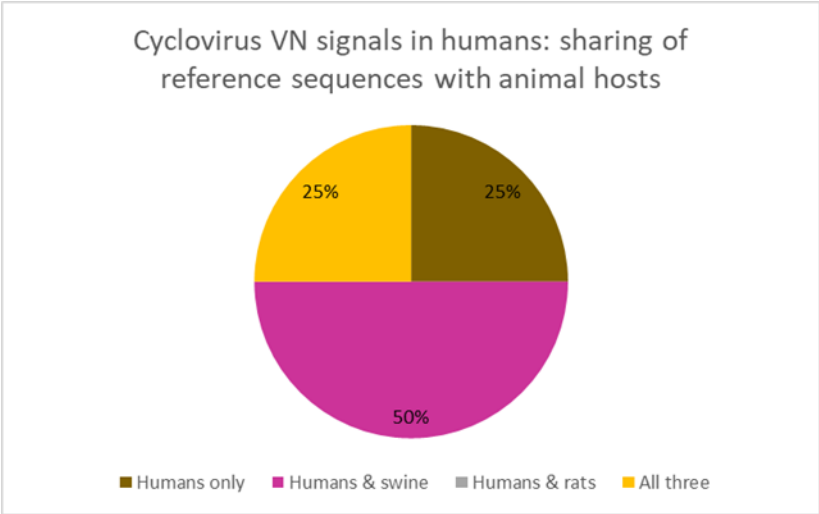


Figure 5.6 Sharing of cyclovirus VN best-scoring reference sequences
Proportions of human cyclovirus VN signals that share a best-scoring reference sequence with animal study populations.

Mammalian orthoreovirus

The species *Mammalian orthoreovirus* (MRV; genus *Orthoreovirus*, family *Reoviridae*) contains viruses that are known to infect humans and a variety of other mammals, including swine, mice and bats. In swine, MRVs are associated with diarrhoea and respiratory disease (Thimmasandra Narayanappa et al., 2015). In humans, MRV infections are mostly asymptomatic, but they can also cause mild enteric or respiratory disease and have been implicated in several cases of severe disease, including central nervous system involvement (Ouattara et al., 2011, Jiang et al., 2006, Tyler et al., 2004, Hermann et al., 2004, Johansson et al., 1996). The lack of clustering by host in phylogenies and the detection of closely related MRVs in European bats and diseased humans have led to hypotheses about a role for zoonotic transmission originating in bats (Steyer et al., 2013, Lewandowska et al., 2018, Lelli et al., 2013, Kohl et al., 2012), paralleling the epidemiology of sister species *Nelson Bay orthoreovirus* in Southeast Asia (Chua et al., 2011, Tan et al., 2017). MRVs can also be transmitted indirectly: infective MRV has been found in environmental samples (Lodder et al., 2010, Lodder and de Roda Husman, 2005, Spinner and Di Giovanni, 2001, Irving and Smith, 1981).

Here, one human sample and four swine samples have MRV signals (Figure 5.7). The best-scoring reference sequence for the human signal is a sequence for segment M1 from a bat isolate from China (Yang et al., 2015) (Figure 5.7). It is unclear whether this indicates a bat

origin for the virus in the sample, as the best-scoring reference only represents one of 10 segments of the MRV genome, and, additionally, MRV sequences do not neatly separate into host-specific clades in phylogenies (Thimmasandra Narayanappa et al., 2015, Lelli et al., 2016, Lewandowska et al., 2018). The identified M1 reference sequence clusters with other bat sequences and a mink sequence, also from China (Yang et al., 2015). The best-scoring reference sequences for the swine signals are various isolates from pig products and diarrhoeic pigs from the USA, China, and Italy (Thimmasandra Narayanappa et al., 2015, Dai et al., 2012, Lelli et al., 2016) (Figure 5.7). Phylogenies show that the identified segment sequences are part of clades that include other porcine sequences, but also sequences from humans, mink, and cattle (Thimmasandra Narayanappa et al., 2015, Lelli et al., 2016).

***Mammalian orthoreovirus* best-scoring reference sequences**

Humans (n = 1)			Swine (n = 4)		
N	Accession nr.	Country	N	Accession nrs.	Countries
Best-scoring reference sequences from swine					
0	N/A	N/A	4	KM820746 (1), KM820761 (1), JX486065 (1), KX343209 (1)	USA, China, Italy
Best-scoring reference sequences from bats					
1	KT444545	China	0	N/A	N/A

Figure 5.7 Origins of *Mammalian orthoreovirus* best-scoring reference sequences for signals in the human and swine study populations. N, number of signals.

Based on single best-scoring reference sequences, it is difficult to determine whether the MRVs in human and pig samples are closely related: the sequences represent different segments. BLAST top hits for individual reads show distinct patterns for the human and swine signals. In addition to their best references, the human signal has many hits to mink isolate SD-14, whereas the swine signals have large numbers of hits to porcine strains, bovine-origin lab strain clone18 and several different bat isolates. These results suggest that the MRVs in human and pig samples are distinct.

Overall, the matches to the best-scoring reference sequences have very high average identities (>90%). However, like for RVC and RVA signals, the meaning of this measure is limited for MRV as it only considers one of 10 genomic segments.

5.3.6 Category IV: putative novel zoonotic viruses

Nine (0.28%) signals represent viruses for which their detection in this study is the first (limited) evidence of putative zoonotic potential (Table 5.8). These are two viruses detected in human samples but previously only found in animals, and one virus detected in swine samples but previously only found in humans. Here, I investigate these signals further to gain insights into whether they are likely to represent recent zoonotic transmissions and/or viruses with zoonotic potential.

Gemykrogviruses in human samples

Caribou associated gemykrogvirus 1 (genus *Gemykrogvirus*, family *Genomoviridae*) is a newly recognised CRESS DNA virus species consisting of a single virus, the genome of which was detected in caribou faeces (Ng et al., 2014). Other gemykrogviruses have been found in sewage and in cattle serum (Varsani and Krupovic, 2017). With no non-genetic evidence of infection in caribou, cattle or other organisms, the natural hosts of these viruses remain unknown. So far, gemykrogviruses have shown no link with disease: the health status of the caribou was unknown, and the cattle serum was from healthy cattle.

In this study, five human samples have signals for the genus *Gemykrogvirus* (Table 5.8), with caribou faeces-associated gemycircularvirus (species *Caribou associated gemykrogvirus 1)* as best-scoring reference sequence. The signals are all in patients hospitalised with diarrhoea.

Exposure to caribou in the Vietnamese Mekong delta is improbable, but the signals could represent infections from a different zoonotic source. The patients did not report any animal exposures in common, however, four out of five (80%) reported eating, cooking, or handling raw pig meat in the two weeks prior to onset. In other patients with diarrhoea living in the same two areas as the *Gemykrogvirus* signals (Cao Lanh City and Cao Lanh District), only 241 of 942 (25.58%) reported a connection with raw pig meat. Although the numbers are very small, they provide strong evidence against the null hypothesis that the pattern was observed by chance ($\chi^2(1, N=947) = 7.7$; $p<0.01$). Interestingly, the same virus was not found in swine in my study, and the practical significance of the association thus remains unclear.

Table 5.8 Signals representing viruses with unclear zoonotic potential (Category IV)

The "average percent identity" (% id) to the best-scoring reference sequence for each signal was calculated over all reads that had this reference as top hit in the validation step

Signal(s)				**Best-scoring reference sequence**				**Other evidence**	**Interpretation**
OTU	**Host**	**N**	**Size (rp)**	**Accession nr.**	**Species or clade**	**Host**	**% id**		
Gemykrogvirus	Human	5	21-425	KJ938717	*Caribou associated gemykrogvirus 1*	Caribou	86.88-89.90	NA	Uninterpretable – novel virus OR cross-species transmission OR exposure
Cyclovirus	Swine	3	19-23	GQ404890 (2), GQ404855 (1)	*Human associated cyclovirus 7*	Human	82.83-83.70	NA	Uninterpretable – cross-species transmission OR exposure
Circovirus	Human	1	46	KC241982	*Canine circovirus*	Dog	88.82	Sample from animal health worker	Uninterpretable – cross-species transmission OR exposure

Adding to the uncertainty, the signals are rather small (21-425 read pairs), which makes it impossible to distinguish between true infection and contamination or non-infectious exposure. It thus remains unclear whether these signals represent recent zoonotic transmission events, the discovery of an established virus-host relationship, non-infectious exposure, or infections of a member of the human microbiome.

Human associated cyclovirus 7 in swine samples

The species *Human associated cyclovirus 7* (genus *Cyclovirus*, family *Circoviridae*) contains a single virus, with one full genome and one partial sequence detected in stool samples of Nigerian children with non-polio acute flaccid paralysis (Li et al., 2010). It is unclear whether these detections represent infections, or whether the virus contributed to the disease. Other cycloviruses have been found in a variety of mammals and birds, and in arthropods (Breitbart et al., 2017), but the only confirmed host association is for an endogenous cyclovirus sequence integrated into an ant genome (Dennis et al., 2018).

Here, three swine samples have signals assigned to the genus *Cyclovirus*, with as best-scoring references the two *Human associated cyclovirus 7* sequences (Table 5.8).

The detection of these signals in swine raises questions as to whether *Human associated cyclovirus 7* is a virus that infects both humans and swine, and if so, whether cross-species transmissions occur. However, the sparsity of closely related sequences in the NCBI sequence database, the very little knowledge about the biology of cycloviruses, and the lack of apparent similar signals in the human samples in this study hinder any further interpretation of these signals.

Finally, the signals are very small (19-23 read pairs) and it is thus also possible that they represent non-infectious exposure, for example dietary exposure to insect viruses as suggested in (Dennis et al., 2018).

Canine circovirus in a human sample

The species *Canine circovirus* (genus *Circovirus*, family *Circoviridae*) consists of viruses that infect dogs and wild carnivores (Zaccaria et al., 2016). It has been implicated in diarrhoea, vasculitis, granulomatous lymphadenitis, and respiratory disease, although its role, as causative pathogen or more likely as complicating factor in disease caused by co-infecting pathogens, remains unclear (Anderson et al., 2017, Decaro et al., 2014, Dowgier et al., 2017, Hsu et al., 2016, Li et al., 2013a, Piewbang et al., 2018, Thaiwong et al., 2016). Other

circoviruses have been detected in various mammals, birds and fish (Breitbart et al., 2017, Rosario et al., 2017), and are recognised pathogens of swine and birds (Maclachlan et al., 2017). However, little is known about the biology (including the host range) of mammalian circoviruses other than porcine circovirus (Breitbart et al., 2017). It is unclear whether any circoviruses infect or cause disease in humans (Fields et al., 2013); they have been detected in human stool samples, but some of these are likely to represent dietary contamination (Li et al., 2010).

In this study, one human sample contains a signal with as best-scoring reference sequence a canine circovirus (Table 5.8). This is a routine sample obtained from a member of the high-risk cohort, who is an animal health worker and could thus have been exposed to infected dogs. Unfortunately no metadata are available regarding the specific exposures or health status of the cohort member around the time of sampling.

Until now, there has been no indication that canine circoviruses are zoonotic. There is an anecdotal report of illness in dog owners and veterinary staff during a mystery outbreak of canine illness in the US, attributed by some to canine circovirus – but the virus was ruled out as primary cause of the outbreak in dogs, and to date there is no scientific evidence for infections in humans (Scheidegger, 2013). The signal in the current study does not change this: the small signal size (46 read pairs) precludes differentiation between low-level infection and non-infectious exposure. Further studies are needed to confirm or exclude infection.

5.4 Discussion

This is the first large scale viral metagenomic investigation of humans, swine and rats from the same setting: the province of Dong Thap, in the Mekong Delta in Vietnam.

5.4.1 Overall diversity and novelties

The samples contain a large diversity of mammalian viruses. This includes known pathogens of humans (e.g. rotaviruses, enteroviruses) and swine (e.g. teschoviruses, porcine circovirus 2), some of which have previously been identified in clinical (Nguyen et al., 2007b, Duong et al., 2016, Thompson et al., 2015, Do et al., 2016, Khanh et al., 2012, Phan et al., 2005, Tan le et al., 2014, Le et al., 2010, Pham et al., 2014b, Pham et al., 2003, Tran et al., 2003) or veterinary research (Feng et al., 2008, Thuy et al., 2013, Huynh et al., 2014) settings in Vietnam. Other viruses detected in this study were discovered relatively recently through

sequence-based investigations; their associations with disease remain largely unclear. Among these are known or presumed mammal-infective viruses (e.g. cosaviruses, copiparvoviruses) as well as viruses of which the tropism is unknown (e.g. gemykibiviruses, porprismacoviruses). Many of these viruses have been reported in academic studies in comparable human (Dai et al., 2010, Khamrin et al., 2012, Shan et al., 2010, Yu et al., 2015), swine (Zhang et al., 2014, Yu et al., 2013a, Yu et al., 2013b, Cheng et al., 2010) and rat (http://www.mgc.ac.cn/DRodVir/ (Chen et al., 2017) and (Wang et al., 2015b, Li et al., 2016a, To et al., 2017, Wu et al., 2018, Du et al., 2016)) populations in Southeast Asia or China whereas others (e.g. rat sapovirus (Sachsenroder et al., 2014, Firth et al., 2014)) have only been described in single or few studies worldwide.

Below are some examples of how the findings from this study have contributed to the knowledge base of locally circulating viruses.

Novel virus variants

The metagenomic data contain evidence for the presence of novel viruses, indicated by low average identities of signals to their best-scoring reference sequences.

The majority of signals representing putative novel viruses are part of virus families that have recently seen an expansion in diversity, related to the increase in sequence-based studies in humans and animals. Many signals suggest the presence of novel picornaviruses in rats (parecho-, mosa-, cardio-, and entero-/sapelovirus-like viruses) and swine (tescho- and entero-/sapelovirus-like viruses). These findings fit within the recent rapid expansion of known diversity of the *Picornaviridae* family, as illustrated by its growth from 12 genera (containing 29 species) in 2011 (King et al., 2011) to 35 genera (containing 80 species) in 2017 (Zell et al., 2017). Similarly, signals suggesting the presence of a novel gemykibivirus (in swine) and new picobirnaviruses (in all three study populations) are consistent with other metagenomic studies reporting new CRESS DNA viruses (Krupovic et al., 2016, Phan et al., 2015, Kim et al., 2014, Ng et al., 2015, Phan et al., 2016b, Sachsenroder et al., 2012, Yinda et al., 2019) and picobirnaviruses (Luo et al., 2018, Yinda et al., 2018, Shi et al., 2016, Li et al., 2015b, Woo et al., 2014).

In addition to these rapidly growing virus families, rat samples in this study have divergent signals for several other viruses: sapo-, noro-, astro- and rotaviruses. In contrast, in humans, in addition to picobirnaviruses, only viruses in the *Anelloviridae* family (alpha-, beta-, and

gammatorqueviruses) – all non-pathogenic – have signals with low average identities. These findings reflect the historical bias of sequencing efforts towards human pathogens, resulting in a relative lack of coverage of wildlife viruses and non-pathogenic viruses in the NCBI database.

Novel virus-host associations

Among viruses with no zoonotic connotations (Categories I and II), there are two novel virus-host associations when considering the virus as the genus level. The *Gemykibivirus* signal in a swine sample represents the first detection of this genus in this host species. Other gemykibiviruses have previously been found in other mammals and birds (Varsani and Krupovic, 2017), but the significance of these associations remains unclear. Similarly, the *Mosavirus* signal in a rat sample represents the first time a virus of this genus is found in rodents of the family Muridae, having previously only been detected in a canyon mouse (family Cricetidae) (Phan et al., 2011), a marmot (family Sciuridae) (Luo et al., 2018), and a bird (probably diet-related) (Reuter et al., 2014).

A third novel virus-host association at the genus level, but with suggestions of zoonotic potential, is the *Gemykrogvirus* in humans, discussed in section 5.4.3.

5.4.2 Non-zoonotic viruses

An extensive discussion of non-zoonotic viruses, beyond their contribution to the knowledge base of locally circulating viruses as described above, is not relevant to this thesis.

However, it is important to consider that some of these viruses, while not currently known as zoonotic, may yet have or be able to develop zoonotic potential. It is of interest to be aware of viruses circulating in local animal populations, so that, if spillover infections occur in future, these can be promptly identified and responded to. Basic risk assessments should be considered, including evaluations of the possibility of human-infectivity, putative routes of exposure, and our ability to detect any putative human infections (Palmer et al., 2005).

5.4.3 Known and presumed zoonotic viruses

Six virus species and one genus with known or presumed zoonotic potential were identified in the samples. Here I place the findings for each of these viruses into context with what is known about their circulation in humans and animals in Vietnam, China or Southeast Asia. I

also suggest further studies and actions that could advance our understanding of their significance as potentially emerging pathogens.

Orthohepevirus A, genotype 3

The detection of HEV-A3 in swine is consistent with many reports from domestic swine herds and wild boars throughout the western hemisphere and in Asia (Purdy et al., 2017, Doceul et al., 2016). Zoonotic hepatitis E is emerging in humans in these regions, with HEV-A3 a major agent, including in Japan and Singapore (Teo et al., 2017, Wong et al., 2019, Doceul et al., 2016, Adlhoch et al., 2016).

In Vietnam, its circulation among domestic pigs in Dong Thap province was previously shown in a PCR-based study by the VIZIONS consortium (Berto et al., 2017b). In the human population, the seroprevalence of HEV is high (Berto et al., 2017b, Tran et al., 2003), but the relative contribution of genotypes is unknown as few molecular studies have been carried out. With its dominance in local swine herds and the importance of pigs and pork products in daily life, HEV-A3 is believed to make up a significant part of these infections (Berto et al., 2017b). The occurrence of zoonotic HEV infections is supported by the previous identification of another porcine-origin genotype, HEV-A4, in patients with acute sporadic hepatitis in Hanoi (Hijikata et al., 2002). However, the single human HEV signal in the current study (in Category I) was HEV-A1, a human-specific, water-borne, genotype that is endemic in large parts of Asia and that has also been detected in Vietnam (Lu et al., 2006).

Generally, with development, other countries have seen large water-borne outbreaks being replaced with sporadic zoonotic cases. More molecular epidemiological studies are needed to understand whether this is also happening in Vietnam, and what the relative contribution of HEV-A3 is to human infections and disease.

Orthohepevirus C

HEV-C was detected in rats in this study, in accordance with previous identifications in similar rat populations in Vietnam (Obana et al., 2017, Li et al., 2013b, Van Nguyen et al., 2018).

Interestingly, the virus from the first recognised human case, in Hong Kong, was most closely related to one of these earlier Vietnamese sequences (Siddharth et al., 2018). Based on their best-scoring reference sequences, several of the HEV-C viruses detected in this study also appear to belong to the same clade as these strains (Takahashi et al., 2016, Andonov et al., 2019). Whether these viruses cause zoonotic infections or disease in humans in Vietnam is

unknown, although there is serological evidence for such cases in febrile patients in Hanoi (Shimizu et al., 2016). The lack of human HEV-C signals in the current study is not surprising and should not be seen as evidence of absence: human HEV-C infections are likely to be rare, and, in this study, samples came from healthy humans and diarrhoeic patients, rather than patients with febrile illness or jaundice.

Physicians in Vietnam should be aware that HEV-C strains circulating in local rats may contribute to unexplained febrile illness or hepatitis cases in the region. A pan-*Orthohepevirus* or HEV-C-specific PCR system should be used to screen such patients (standard HEV diagnostics do not detect HEV-C), as well as individuals with frequent exposure to rats, to learn more about whether spillover HEV-C infections occur in Vietnam.

Rotavirus C

The RVC signals in swine in this study are in line with the recent first detection of porcine RVC in Vietnamese swine by Phan et al. (2016a), and reports from several other countries worldwide (Suzuki et al., 2015).

Although human RVC signals were not found in the current study, they have previously been reported in Vietnam, in children hospitalised with gastroenteritis – but at a very low prevalence and only in combination with RVA (Nguyen et al., 2007b). Other countries in the wider region in which human RVC has been detected include Thailand, Malaysia and China (Penaranda et al., 1989, Rasool et al., 1994, Yamamoto et al., 2011). Recent molecular studies in India found an unexpectedly high prevalence of RVC in diarrhoeic children, with viruses closely related to a human/porcine reassortant previously detected in swine, suggesting a role for cross-species transmission (Kattoor et al., 2017, Bhat et al., 2018). In Vietnam, the prevalence of RVC in humans and animals and the relative contribution of zoonotic infections remains unknown, as studies have largely focused on species *Rotavirus A* due to its greater public health relevance.

To shed further light on the epidemiology of RVC in Vietnam and in general, further molecular epidemiological studies with a broadly sensitive or an RVC-specific PCR system are required.

Picobirnavirus

The detection of PBV in all three study populations is consistent with previous findings of PBVs in faecal samples from humans (Pereira et al., 1988, Smits et al., 2014b, Ganesh et al., 2011b), pigs (Zhang et al., 2014, Sachsenroder et al., 2012, Banyai et al., 2008), rodents (Phan

et al., 2011, Fregolente et al., 2009) and a wide variety of other mammals across the globe (Yinda et al., 2018, Bodewes et al., 2014, Woo et al., 2014, Woo et al., 2016). To my knowledge, the signals in this study represent the first PBVs identified in Vietnam, but they have been detected in the region, including China and Thailand (Luo et al., 2018, Woo et al., 2016, Zhang et al., 2014, Wilburn et al., 2017, Wakuda et al., 2005).

The sharing of best-scoring reference sequences between multiple study populations, suggesting that closely related PBVs circulate in different species in the same setting, is in line with previous reports of high pairwise identities between human and animal PBV sequences from the same geographical region (Banyai et al., 2008, Giordano et al., 2011, Ganesh et al., 2011a). This seems to indicate a generalist ecology with a likely role for cross-species transmissions.

However, this interpretation is clouded by a lack of knowledge about the basic biology (including the cellular host) of PBVs. Interestingly, recent genomic studies have resulted in theories that PBVs may actually be prokaryotic viruses (perhaps infecting our gut flora) (Krishnamurthy and Wang, 2018), or that a subset may infect mitochondria (Yinda et al., 2018). It is thus unclear whether the similarities between PBV sequences detected in different host species actually represent cross-species transmissions of these viruses, or of the bacteria that may be their true hosts, or whether a different situation applies altogether.

While this study does not address these biological questions, it does indicate opportunities for further investigations that may elucidate some aspects of PBV ecology, if PBVs are indeed animal viruses. The PBV signals detected in this study, could, when properly assembled, become the first collection of PBV genomes from a large number of individuals belonging to multiple host species all living in the same setting. A phylogenetic study incorporating these sequences could shed more light on the relative ecological roles of different host species in this environment, compared to previous studies that have either been conducted in single populations, or been limited to few individuals per host species.

Rotavirus A

RVA signals were present in all three host populations. This is in line with many previous findings in humans and swine in Vietnam (Do et al., 2017, Nguyen et al., 2007b, Phan et al., 2016a, Thompson et al., 2015, Pham et al., 2014a), and recent reports in rats in China (Li et al., 2016a). RVA is a major cause of morbidity in children in Vietnam, responsible for circa

50% of severe diarrhoea cases in those aged under five, and for several thousand deaths a year (Thompson et al., 2015, Van Man et al., 2005).

Here, in accordance with the study by Phan et al. (2016a), I found evidence for sharing of RVAs between humans and swine in the study setting. There have been multiple other recent reports of porcine-like, possibly bovine-like, and animal/human reassortant RVA strains in human gastroenteritis patients in Vietnam (Ahmed et al., 2007, Nguyen et al., 2007a, My et al., 2014a, Do et al., 2017, Kaneko et al., 2018, Hoa-Tran et al., 2016). Although I did not detect any sharing of RVAs between humans and rats, the recent identification of closely related RVAs in rodents and a human in China indicates that rodent RVAs also have zoonotic potential (Li et al., 2016a).

The frequency of zoonotic transmission of RVAs in Vietnam is unknown, as there is only limited genomic surveillance of RVA in human and animal populations. Phan et al. (2016a) found one porcine-like constellation among 146 human RVA infections, but in my study, 3% of human RVA signals have a best-scoring reference from swine, suggesting that the frequency of zoonotic infections (or reassortants with porcine-like segments) could be even higher.

From the current study, it is not clear whether the identified putative porcine-like RVA signals in humans are the result of one or multiple zoonotic spillover events. Most previously identified zoonotic RVA infections in Vietnam appeared to represent isolated cases, but other reports suggested that porcine-like P[19] strains had become established in the Vietnamese human population (My et al., 2014a), and that a human/bovine-like reassortant had emerged as a predominant RVA strain (Hoa-Tran et al., 2016). Phylogenetic studies on assembled sequences from this study could bring to light whether the putative zoonotic signals are related and/or represent further human-to-human transmission of zoonotic strains. Additionally, continuous genomic surveillance of RVA would be useful to develop a better understanding of the dynamics of such strains, and their contribution to the burden of disease.

Human associated cyclovirus 8 (Cyclovirus VN)

The finding of CyV-VN in humans, swine and rats in this study is in accordance with its original discovery in Vietnam, in a patient with acute central nervous system infection and subsequently in pigs and poultry (Tan le et al., 2013). Similar viruses have also been reported

in healthy humans in Madagascar (Garigliany et al., 2014, Sauvage et al., 2018), HIV-, HBV- or HCV-infected humans in Italy (Macera et al., 2016), pigs in Cameroon (Garigliany et al., 2014) and a rodent (multimammate mouse) and shrews in Zambia (Sasaki et al., 2015). As CyV-VN is a newly discovered virus (from 2013), its prevalence patterns in human and animals in Vietnam are unknown, and its role in disease is still undetermined.

As with PBVs, our results and those of others suggest that CyV-VN may be a generalist virus: our three host populations share several best-scoring references, and the initial paper also reports high pairwise identities of sequences from different host species (Tan le et al., 2013). But, like for PBVs, very little is known about cycloviruses and this hinders drawing firm conclusions. So far, knowledge about cycloviruses is limited to what can be derived from sequences; they have never been isolated and their cellular host is unknown (Breitbart et al., 2017). A recent phylogenetic study of viruses in the *Circoviridae* family suggests that cycloviruses are insect viruses, and that their detection in samples from other animals is the result of contamination (or non-infectious exposure) (Dennis et al., 2018). However, in the same study it is noted that CyV-VN is part of a basal lineage that, unlike other cycloviruses, has been identified exclusively in vertebrate samples and could thus truly be vertebrate-infective (Dennis et al., 2018). Thus, if one assumes CyV-VN is a genuine vertebrate virus that infects the humans, swine and rats in our study populations, it is likely to be a generalist and cross-species transmissions may be important; but it remains unclear whether this assumption is valid.

Further studies should focus on identifying the tropism of CyV-VN. In the meantime, phylogenetic analyses with assembled data from the current study, and ideally from additional local human and animal populations and environmental samples, could help generate hypotheses.

Mammalian orthoreovirus

MRV signals were present in human and swine samples. To my knowledge, this represents the first molecular evidence of MRV in these populations in Vietnam, with the only previous sequences from the country being from dogs (Dung, 2017). However, MRV is believed to circulate globally, and multiple reports exist from humans (Duan et al., 2003, Song et al., 2008), pigs (Dai et al., 2012, Zhang et al., 2011) and bats (Yang et al., 2015, Li et al., 2016b, Wang et al., 2015a) in China. As human MRV infections are largely asymptomatic, the virus is only of limited academic interest, and little is known about its prevalence in Vietnam or the

relative contribution of any zoonotic infections. This is in contrast to sister species *Nelson Bay orthoreovirus* (NBV), which has been implicated in several recent zoonotic emergence events, originating in pteropine bats and resulting in acute respiratory tract infections in humans in Southeast Asia (Chua et al., 2011, Tan et al., 2017).

The humans and swine signals in this study do not share best-scoring reference sequences and have different patterns in their BLAST top hits, strongly suggesting that they represent different virus variants. The swine signals match other porcine MRVs from multiple different continents, suggesting that, for this virus variant, transmission between swine herds, even across continents, has a more important role than cross-species transmission on a relatively local scale. On the other hand, the similarity of the human signal to a segment from a Chinese bat MRV lends some support to the hypothesis that bats could be donor hosts of zoonotic MRV infections in addition to NBV infections. However, these suppositions are based on similarity across single segments, and there are no data from local populations available for comparison.

Further studies on assembled sequences for all segments, and including more samples from a greater diversity of local populations – particularly including bats – are needed to gain a better understanding of MRV ecology and epidemiology in Vietnam.

5.4.4 Putative novel zoonotic viruses

In this study, three viruses were identified that are putative novel zoonoses.

The detection of *Canine circovirus* in this study represents the first report of this animal pathogen in a human sample. However, it is unclear whether the single, small signal in an animal health worker represents an infection, or is a marker of professional exposure to infected dogs. The virus had not previously been found in Vietnam, but its circulation in dogs in the country is plausible, considering findings in China (Sun et al., 2019), Thailand (Piewbang et al., 2018) and Taiwan (Hsu et al., 2016). Further investigations should prioritise genetic characterisation of the virus and verification of infection in the cohort member (e.g. through serology). If infection is confirmed, the incidence of spillover infections could be investigated with targeted surveillance studies in individuals exposed to dogs.

The *Caribou associated gemykrogvirus 1* and *Human associated cyclovirus 7* signals in this study, in humans and in swine respectively, also represent novel virus/host combinations.

They had previously only been detected in single studies, in caribou faeces from Canada (Ng et al., 2014) and in human stools from Nigeria (Li et al., 2010), respectively. Due to the lack of further detections, or of any other, non-genetical evidence of infections, the significance of the signals in the current study remains unclear. More studies, both epidemiological and experimental, are needed to determine the host range of these viruses, and whether they are pathogens, "commensal" viruses, or perhaps viruses that infect our symbionts.

In summary, in this study I have not found significant evidence of novel zoonotic viruses emerging in Vietnam. With regards to the viruses I did consider as putative novel zoonoses, the label "emerging" perhaps applies more to our knowledge of these viruses than to the viruses themselves.

5.4.5 Limitations of this study

The set of viruses described here is not an exhaustive list of viruses present in the study samples. The detected viruses are biased towards known, well-studied viruses, with novel or ill-defined viruses more likely to have been missed. This is because the pipeline relies on a similarity-based classification tool, and these are not very sensitive in detecting viruses without close relatives in their reference database. Additionally, I only considered genera and easily identifiable groups of viruses (e.g. proposed genera) and did not record less specific (e.g. family-level) signals. I did this to keep the study manageable, and because viral discovery was not considered a major aim. However, as viral sequencing efforts have traditionally focused on human and economically important veterinary pathogens, it is likely that this bias towards known viruses translates into an under-detection of non-pathogenic and wildlife viruses.

Further, the signal descriptions beyond the OTU level are the result of a simple methodology and should be interpreted with the aims of the study in mind: they are not meant to be detailed characterisations, but to highlight interesting viruses for further study or assessment. These descriptions are based on sets of short sequence reads, and cannot be read in the same way as one would read the comparison of a single long sequence to a reference database. Importantly, the best-scoring reference sequence does not necessarily represent the closest relative of, or the sequence with the highest overall nucleotide identity to, the virus in the sample. Its definition depends on both the bitscores and the numbers of matching sequence reads, and is thus sensitive to disproportionally amplified sequences.

Additionally, in case of segmented viruses, recombinants or mixed infections, it reflects just one of multiple segments or strains contributing to the signal. Furthermore, the "average percent identity" measure used to describe the similarity of a signal to its best-scoring reference is not a very meaningful summary statistic: it is only based on reads that match to this particular sequence, the proportion of which is very variable. Best-scoring references and average identities should thus be interpreted with caution. Identified signals should be assembled for more detailed characterisation of viruses of interest.

Generally, it is impossible to determine based on metagenomic sequence data alone whether a signal represents an infection of mammalian host cells, an infection of a symbiont, non-infectious exposure, or contamination. In this study, I distinguished between these options by using prior knowledge or common assumptions about a virus' host and geographical ranges. Signals for which infection was considered biologically implausible were dismissed as likely contaminants or exposure. While this method is easy to apply, it can lead to the inappropriate exclusion of true infections arising from cross-species transmissions, particularly of viruses mistakenly thought to be host-specific. Such infections would be of particular interest to this study, and it is thus important to consider multiple strands of evidence in the dismissal of these signals. Here, all dismissed signals are relatively small in size, and for the majority I could identify an obvious source of contamination, supporting interpretation of these signals as contaminants but not excluding other explanations. For signals with a virus-host combination that is unexpected, but that would carry important implications if true, it may be appropriately cautious to exclude the remote possibility that they are true infections by performing further virus-specific diagnostic tests in the laboratory.

Finally, while one of the aims of this study was to identify potential zoonotic cross-species transmissions, this could only be partially fulfilled. Signals for putative novel zoonoses (Category IV) could not be interpreted with confidence because current knowledge of their reference virus' host range is limited, and they were found in only one study population. In such cases, testing of other implicated potential host species should be considered, to identify additional sequences for comparison. On the other hand, signals for known zoonotic viruses shared between human and animal study populations (Category IIIb) could not be fully interpreted due to the limited resolution of taxonomy, the limited significance of best-scoring reference sequences for segmented viruses, and a lack of direct comparison across study populations. Signals assigned to both categories should be assembled and investigated

with phylogenetics, allowing a more detailed characterisation, and direct comparison between different study populations, while also taking into account evolutionary processes.

5.4.6 Conclusion

This study provides a metagenomic overview of viruses circulating in humans, swine and rats in Dong Thap province in the Vietnamese Mekong Delta. The vast majority of viruses were found only in single host species, in a manner that matched with known host ranges and previous detections. Interestingly, the findings suggest the presence of a variety of novel animal viruses in the study samples, particularly those from rats. Of more relevance is the identification of several zoonotic viruses, including some for which signals in human and animal study populations resemble the same reference sequences. The findings support the occurrence of occasional cross-species transmissions of rotavirus A between humans and swine in this setting, and the notion that picobirnaviruses and cyclovirus VN have a generalist ecology that includes humans, swine and rats. Additionally, multiple viruses were detected that could represent putative novel zoonoses; but the signals were small, suggesting that contamination or non-infectious exposure could also be plausible sources.

In addition to increasing our awareness of viruses of potential public health relevance in Vietnam, the study findings provide some insight regarding viral emergence in general. In particular, while they are consistent with “viral chatter” for a few specific viruses – all already known or presumed to be zoonotic –, the lack of clear signals for other animal viruses in a human study population with extensive and intensive animal contacts, suggests that viral chatter occurs only rarely, or not at all, for most viruses. Barriers to spillover infections in human hosts could relate to human genetics (Warren and Sawyer, 2019), or to any of a number of ecological, epidemiological and behavioural factors (Plowright et al., 2017).

The study findings also illustrate the advantages and limitations of metagenomics as a tool in the surveillance for zoonotic emergence. For example, the detection of lesser-known, non-pathogenic and apparently novel viruses illustrates how metagenomics can yield unexpected findings, and is less biased towards veterinary or zoonotic pathogens than standard targeted surveillance. This is an important advantage when a project aims to characterise the diversity of animal viruses that local humans are exposed to, or to identify newly emerging zoonotic viruses. On the other hand, interpretation of signals, including the evaluation of whether detected (putative) zoonoses represent emerging public health threats, has demonstrated to

be challenging, due to knowledge gaps concerning the tropism, pathogenic potential, and/or local epidemiology of many viruses. This shows how metagenomic surveillance is useful as a hypothesis-generating exercise, but that it needs to be combined with more targeted virological and epidemiological studies to allow proper risk assessment and to generate actionable knowledge.

These broader aspects are explored further in Chapter 6, the General discussion.

Chapter 6. General discussion

I wrote this chapter with minor text edits from Andrew Rambaut.

This study explored the use of metagenomics as a tool for surveillance of emerging zoonotic viruses. It was embedded within VIZIONS, a collaborative, multidisciplinary project of which the overarching aim was to understand the emergence of zoonotic viruses in Vietnam. Considering these two aims, the work in this thesis was guided by four key research questions:

> Using metagenomic methods, which mammalian viruses are found in faecal samples from humans and animals in the study setting? Which are zoonotic?
>
> Do the findings support the notion of frequent "viral chatter" between humans and animals living in close proximity in this setting?
>
> Could any of the identified viruses be "at the cusp of emergence" in humans?
>
> What lessons can be learned for future metagenomic surveillance studies?

To address these questions, I developed a viral taxonomic classification pipeline (Chapter 3). I then tested its performance by comparing its outcomes with diagnostic qPCR data (Chapter 4). To address some limitations identified during this process, I built several additional steps into the pipeline. I then applied the adapted pipeline to faecal samples from humans, swine, and rats from Dong Thap province in the Vietnamese Mekong Delta, generating an overview of viruses circulating in these populations (Chapter 5). I categorised these viruses as non-zoonotic or as known, presumed or putative novel zoonoses. Several viruses were shared between the human and animal study populations. Where possible, I evaluated the relevance of these viruses as potential emerging threats to public health.

This chapter summarises and discusses the main findings from these studies, with regards to the design and performance of the pipeline and the four research questions above.

6.1 Pipeline design and performance

To survey the diversity of mammalian viruses circulating in humans, swine and rats in the study setting, a metagenomics pipeline was required that could detect well-known viruses as well as understudied viruses and novel variants. The detection of very divergent viruses (viral discovery) was not considered a goal *per se*.

In accordance with these goals, I designed a viral taxonomic classification pipeline that was meant to be broadly sensitive, without a focus on viral discovery. The pipeline (described in detail in Chapter 3) was built around the taxonomic classification tool Kraken, which uses an exact *k*-mer matching algorithm to infer the likely taxonomic origin of query sequences. I deemed a nucleotide identity-based method sufficiently sensitive for our purposes, and Kraken appeared to provide a good balance of speed, precision and flexibility. To ensure an appropriate balance between sensitivity and specificity, I created a comprehensive viral reference database based on all viral nucleotide sequences in Genbank; used a short *k*-mer of 20 nt; and applied a confidence threshold that required matches over at least five percent of a read pair's *k*-mers before assigning it to a taxon. In addition to the main viral taxonomic classification process, the "basic pipeline" also consisted of data cleaning and filtering steps, and the merger of overlapping read pairs.

In Chapter 4, I tested the performance of this basic pipeline for several well-studied viruses (human enteric pathogens), by comparing read pair counts with corresponding diagnostic qPCR data available for samples from hospital patients. ROC curve AUCs of 0.82-0.93 indicated that, for these viruses, the accuracy of the pipeline was high. Estimates of overall sensitivity and specificity were also high – 0.97 and 0.96 respectively – and well-balanced. Furthermore, discordant metagenomic and qPCR results were investigated, but could not be attributed disproportionately to any one source of error. Altogether, this suggested that the basic pipeline worked well.

However, the estimated performance measures probably represent the upper bounds of the basic pipeline's performance. My investigations in Chapter 4 revealed that some false positive metagenomic signals were due to erroneously labelled sequence records in the NCBI database. Similarly, early data exploration identified instances where inaccuracies in NCBI taxonomy could lead to loss of sensitivity, for example due to the existence of multiple synonymous taxa. Such errors in NCBI databases are likely to be more common for less well-

defined viruses, than for the well-studied viruses the pipeline was tested on. To limit the impacts of these errors on further analyses, I added two post-hoc adaptations to the pipeline: I redistributed read pairs to custom OTUs, modified from NCBI taxa to better reflect ICTV taxonomy, and added a BLAST-based signal validation step to facilitate detection of misclassification.

Another issue I addressed in Chapter 4 is the presence of cross-contamination between samples in the same sequencing run, assumed to be due to index switching. Background levels of read pairs in truly negative samples varied significantly by sequencing run, and strongly correlated with read pair counts for the same virus as summed over all samples in the run. If unaccounted for, this variation in background levels can result in false positive signals. In turn, this can lead to wrong conclusions in runs with samples from different host species – particularly if a virus has a high prevalence and high viral loads in one host type (e.g. *Rotavirus A* in patients with diarrhoea, various viruses in bats), but is unexpected in another. To be able to account for this variation in background levels in further analyses, I modelled the overall relation between background read pair counts and total read pair counts across each run. Using the resulting model, I defined sets of OTU- and run-specific signal thresholds, and incorporated these as a third adaptation into the pipeline.

After the inclusion of adaptations to mitigate these issues, the pipeline detected a wide variety of viruses when applied to the full sample set in Chapter 5. These included viral taxa with single or few genomes in the database (e.g. *Mosavirus*) and newly-formed taxa that I had manually created as custom OTUs (e.g. *Porprismacovirus*). Further characterisation of signals, using summary measures from the validation step, revealed the presence of putative novel variants of known viruses (e.g. a novel teschovirus) and also, when combined with manual review of unexpected validated signals, putative novel viruses that are closely related to, but perhaps not part of, the taxa they were assigned to (e.g. enterovirus-like viruses). These various detections, in addition to the signals for well-studied viruses, suggest that the designed pipeline was fit for purpose.

6.2 Detected viruses

In Chapter 5, more than 1800 stools and rectal swabs from humans and animals from Dong Thap province were investigated through viral metagenomic analysis. The human study population included 671 hospital patients with diarrhoea and 281 individuals with frequent

residential or occupational contact with animals (the latter sampled multiple times), whereas the animals included 285 swine (mainly domestic pigs) and 315 rats (intended for human consumption).

Application of the bespoke viral taxonomic classification pipeline to metagenomic sequence data from these samples yielded validated signals for a wide variety of viruses. After excluding likely contaminants, the 3212 remaining signals covered 59 viral genera or equivalent groupings (OTUs) within 22 viral families. Characterisation beyond the genus level revealed the presence of well-known human and animal viruses, including many pathogens; viruses identified more recently through sequencing studies, including some for which the natural hosts are unclear; and putative novel viruses.

Many of the detected viruses are non-zoonotic, or not currently known to be zoonotic. Human, swine and rat samples all contained signals for viruses that are typical for, or had previously mainly been identified in, these hosts. Swine and rat samples also contained a variety of putatively novel viruses with suggested origins in cross-species transmission from other (non-human) animals. While the description of non-zoonotic viruses has not been a major focus of this thesis, these identifications are not irrelevant to surveillance for zoonotic emergence: characterisation of the local "zoonotic pool" – the collection of animal viruses that humans are exposed to, and that could develop zoonotic potential in future – may lead to faster pathogen identification in case zoonotic transmission eventually occurs. The characterisation of locally circulating human and animal pathogens and potential pathogens, whether zoonotic or not, also provides insights of value to public and/or veterinary health.

Seven of the detected viruses are known or presumed to be zoonotic. Five of these are bona fide mammal-infective viruses, associated with disease in humans and/or other mammals. **Rat hepatitis E virus** (HEV-C), recently recognised as a cause of sporadic hepatitis cases in humans, was found in several of the rat samples. **Hepatitis E virus genotype 3** (HEV-A3), endemic in swine herds worldwide and a common cause of zoonotic hepatitis E infections in humans, and ***Rotavirus C*** (RVC), a diarrhoeic pathogen of humans and animals but for which the role of cross-species transmissions is not quite clear, were each found in several swine samples. The major diarrhoeic pathogen ***Rotavirus A*** (RVA), for which zoonotic infections are well-documented and believed to play an important role in the maintenance of viral diversity, was the most commonly identified virus in the human samples, and was also found in several swine and rat samples. Finally, ***Mammalian orthoreovirus*** (MRV), which has a broad host

range but an unclear or variable pathogenic potential in humans, was found in several swine samples and one human sample.

The two remaining "zoonotic" viruses have an unknown tropism and pathogenicity, but are generally considered as zoonoses because of previous detections in human and animal samples and the lack of clustering by host species in phylogenies. **Cyclovirus VN** (CyV-VN) was found in several human, swine and rat samples in this study, and ***Picobirnavirus*** (PBV) was among the most commonly detected viruses in all three study populations. Recent research suggests that PBV infects prokaryotes (Krishnamurthy and Wang, 2018), and the applicability of the label "zoonosis" thus remains questionable.

Most of these seven viruses had previously been identified in similar human and animal populations in Vietnam: HEV-C in rats (Obana et al., 2017, Li et al., 2013b, Van Nguyen et al., 2018); HEV-A3 in swine (Berto et al., 2017b); RVC, RVA, and CyV-VN in both humans and swine (Phan et al., 2016a, Nguyen et al., 2007b, Tan le et al., 2013). The detections of RVA and CyV-VN in rats are, to my knowledge, the first to be reported from Vietnam, but they are not unexpected given the detection of these viruses in rodents elsewhere (Li et al., 2016a, Sasaki et al., 2015). Finally, PBV and MRV have not been studied (much) in Vietnam – the only previous MRV sequences from the country are from dogs (Dung, 2017), whereas none exist for PBV – but their detections are in line with previous findings from similar populations in neighbouring countries (Song et al., 2008, Dai et al., 2012, Zhang et al., 2014, Luo et al., 2018, Wakuda et al., 2005).

In addition to known and presumed zoonoses, the samples from humans and swine contained signals for three putative novel zoonoses. The species ***Canine circovirus*** (CaCV) was detected in one human sample, and ***Caribou associated gemykrogvirus 1*** in five, when these viruses had previously only been found in animal samples. Similarly, the species ***Human associated cyclovirus 7*** was detected in three swine samples, when this virus had only ever been detected in human samples. CaCV is known to truly infect mammals and is associated with disease in carnivores. However, the small size of the single signal in a human in this study does not allow for a clear identification as zoonotic infection; it could also represent non-infectious exposure. The other two viruses have been defined on the basis of very limited sequence data, and little is known of their biology. It thus remains unclear whether any of these three viruses are truly zoonotic.

6.3 Viral chatter

Wolfe et al. (2004b) coined the term "viral chatter" to describe the hypothesised widespread cross-species transmission of animal viruses to exposed humans. In their studies of primate retroviruses in rural Cameroon, they identified multiple independent zoonotic spillover infections in bushmeat hunters (Wolfe et al., 2004a, Wolfe et al., 2005b). This was in stark contrast to the few retroviruses that have seen global spread within human populations, and suggested that not cross-species transmission, but viral adaptation to humans, formed the bottleneck in retroviral emergence (Wolfe et al., 2005b). They hypothesised that similar patterns of repeated zoonotic transmissions without onward human-to-human spread are a common stage in viral emergence, and that the greater the frequency and diversity of such viral chatter, the higher the probability is that, eventually, one zoonotic infection will be able to successfully transmit between humans (Wolfe et al., 2004b, Wolfe et al., 2005a).

To investigate whether viral chatter occurs across the human-animal interface in the Vietnamese Mekong Delta, I considered the likely origins of any signals for zoonotic viruses, and explored whether there were any overlaps between human and animal study populations. Of the seven zoonotic viruses, three were detected only in animal samples, but four – MRV, RVA, PBV, and CyV-VN – were also found in human samples. For each of these four viruses, at least one human signal had a best-scoring reference sequence of animal origin, suggesting it could represent a zoonotic infection. For three viruses, human and animal signals had overlapping best-scoring reference sequences, indicating the presence of closely related virus lineages in the human and swine and/or rat study populations. Altogether, these results are strongly suggestive of cross-species viral chatter in the study setting.

For MRV, similarities to different reference sequences indicated that the viruses in human and swine samples were not closely related. The single human signal had a best-scoring reference sequence derived from a bat, but due to the lack of sequences from local populations for comparison and the limited clustering by host in phylogenies, it is not possible to infer whether the signal results from cross-species viral chatter.

For RVA, the study results are suggestive of occasional viral chatter between swine and humans, but not between rats and humans. Ten (3%) human RVA signals had a porcine best-scoring reference sequence, and conversely three (33%) swine signals had a human best-

scoring reference sequence. Overall, 5% of human RVA signals shared a best-scoring reference with swine signals. In contrast, their best-scoring reference sequences did not overlap with those of rat signals.

For PBV and CyV-VN, the study results are in accordance with the generalist ecology of these viruses as suggested by published phylogenies. Approximately half of the human PBV (47%) and CyV-VN (50%) signals had best-scoring reference sequences from animals (a wide range of animals for PBV, chickens for CyV-VN), and more than half of the human PBV (56%) and CyV-VN (75%) signals shared best-scoring reference sequences with the signals in swine and/or rat samples. However, the significance of these findings is clouded by the unclear tropism of these viruses.

While the results of this study are consistent with a role for viral chatter for some viruses, they do not show widespread zoonotic transmissions for mammalian viruses overall. The great majority of viruses were found only in a single population. Additionally, the human samples, originating in part from farmers and slaughterers with frequent contacts with swine, rats, dogs and/or poultry, did not contain signals for many viruses known to occur in these animal species. It is likely that many of these viruses cannot infect humans: human infectivity is the major bottleneck in viral emergence (Warren and Sawyer, 2019). To infect a human, a virus needs to successfully interact with human proteins it requires to enter cells, replicate itself and transmit, all the while escaping the various parts of the human immune system (Warren and Sawyer, 2019). Differences in relevant host proteins form barriers to viral replication in a new host species, and may cancel the effect of viral immune escape mechanisms. The magnitude of these barriers is different for different viruses, but overall, viruses from mammals that are more closely related to humans, such as non-human primates, are more likely to be zoonotic (Olival et al., 2017). Viruses from more remotely related host species, such as the swine and rats focused on in this study, have larger barriers to overcome, and human infective variants are thus likely to be rare.

However, detection of viral chatter could also have been limited due to our study scope and sampling frame. In Wolfe et al. (2005b, 2004a), sampling was targeted to maximize the detection of retrovirus spillover transmissions from non-human primates, but given the broader interests of the VIZIONS study, our sampling frame was more diverse, incorporating people exposed to different animal species, as well as hospital patients selected for their probable viral infections rather than any particular exposures. Hence, compared to large

targeted studies, our study had less power to detect zoonotic transmissions of any one specific virus. Particularly the study of the human-rat interface could have benefitted from a more targeted approach: with only six samples from rat sellers, and few among the human study subjects recalling recent contact with or consumption of rats, the human side of this interface appears undersampled for the detection of rare viral chatter. Additionally, the inclusion of patients with diarrhoea, but not other syndromes, biased the detections towards diarrhoeal pathogens and reduced the probability of detecting viral chatter resulting in different symptomatology. These elements could, at least in part, explain the absence of signals for HEV-A3 and HEV-C in the human samples, despite their zoonotic potential and their circulation in local animal populations.

Finally, I wish to return to the notion that viral chatter is relatively common, but that only on rare occasions it results in further emergence in the human population. In this study, my characterisation of viral signals was too rough to investigate the validity of this notion. By using unassembled sequence data and comparing signals in different study populations in an indirect manner, I could rapidly identify zoonotic viruses that were shared between humans and animals in the study setting, but I was not able to accurately determine a particular infection's origin, nor to distinguish between direct cross-species transmissions and infections obtained via onward human-to-human transmission. Assembly of the VIZIONS viral metagenomic data, and direct comparison of sequences from different study populations via phylogenetic analyses, could shed further light on the relative frequencies of viral chatter and onward human-to-human transmission of zoonotic viruses in the Vietnamese Mekong Delta.

6.4 Emergence

Two major aims of surveillance programmes at the human-animal interface are (1) to identify current and potential future zoonotic pathogens at risk of local emergence, and (2) to detect local emergence events as early as possible, so that they can be attended to before emergence at a larger scale. In accordance with these aims, I evaluated whether any zoonotic or putative novel zoonotic viruses identified in this study could represent emerging public health threats.

Of the seven zoonotic and presumed zoonotic viruses identified in this study, one is currently emerging in several human populations globally, including in Southeast Asia: HEV-A3. In the

last decade, western European countries and Singapore have seen important increases in incidence of zoonotic hepatitis E, predominantly caused by HEV-A3 (Adlhoch et al., 2016, Wong et al., 2019). The drivers of these changes are not entirely clear. It is not known whether HEV-A3 is also emerging in humans in Vietnam, but the predominance of this genotype in local swine herds, seen in this study and also by Berto et al. (2018), and the importance of pork in local cuisine suggest that the country may be at risk.

Another zoonotic virus, HEV-C, may be at the cusp of emergence in humans. Newly recognised as infecting humans in 2018 (Siddharth et al., 2018), it has since been identified in multiple patients in Hong Kong (Coston, 2019), and one in Canada with a travel history to Africa (Andonov et al., 2019). The presence of this virus in rats in Vietnam as seen in this study, and alsoy described elsewhere (Obana et al., 2017, Li et al., 2013b, Van Nguyen et al., 2018), the role of rats in local cuisine, and previous serological evidence for HEV-C infections in febrile patients in Hanoi (Shimizu et al., 2016), suggest that similar spillover cases could also occur in Vietnam. While the occurrence of occasional zoonotic infections does not imply further emergence in human populations, the fact that several patients did not recall any contact with or signs of the presence of rodents (Coston, 2019, Andonov et al., 2019) suggests that this virus should be targeted for enhanced surveillance.

Additionally, the signal for CaCV in a human sample could reflect a new zoonotic emergence event. The small signal size, however, is suggestive of contamination or non-infectious exposure. If considered of interest, further investigations of the sample or study subject should be undertaken to confirm or reject the finding. A risk assessment of the zoonotic potential of CaCV should also be considered (see, for example, the algorithms used by the UK intergovernmental Human Animal Infections and Risk Surveillance group (Human Animal Infections and Risk Surveillance (HAIRS) group, 2018, Morgan et al., 2009, Palmer et al., 2005)).

One zoonotic virus is *not* emerging when considered at the species level as in this study: RVA is well-established in humans, in Vietnam as well as globally. However, specific variants, including some of zoonotic origin, may be emerging. For example, (Hoa-Tran et al., 2016) describe the sudden emergence and subsequent predominance of a RVA reassortant with a bovine-like segment in children hospitalised with diarrhoea in Vietnam. In this study, I did not attempt to assign genotypes to genome segments in the detected RVA signals; this hindered the identification of unusual variants, and any evaluation of detected signals. A

more in-depth characterisation, as well as more targeted longitudinal studies, are needed to assess whether the RVA signals in humans that showed similarity to swine viruses represent any form of variant emergence: do these viruses have full porcine-like segment constellations, or are they reassortants? Overall, are they more closely related to local porcine RVAs, or to porcine-like RVAs previously found in human populations? Are they the result of multiple separate cross-species transmissions, or a single event followed by chains of human-to-human transmission?

For the remaining four zoonotic and presumed zoonotic viruses – RVC, MRV, PBV and CyV-VN – and two putative novel zoonoses – *Caribou associated gemykrogvirus 1* and *Human associated cyclovirus 7* –, knowledge gaps hinder an evaluation as to whether they are emerging. The epidemiology of these viruses is not well studied, because of a limited public health relevance and/or because they have only recently been discovered through sequencing studies. Consequently, overall incidence and prevalence patterns in human and animal populations, as well as relative roles of anthroponotic and zoonotic transmission, remain unclear. Similar is valid for the pathogenicity of most of these viruses. More basic epidemiological and/or virological studies are needed before these viruses can be classified as emerging public health threats, occasional pathogens, or innocent bystanders.

Finally, while in this thesis I focus on viruses that are already known or presumed to have zoonotic potential, it is worth noting that viruses "at the cusp of emergence" can also be found among animal viruses that are not currently known to be zoonotic. Recent studies investigating the evolution of transmissibility and pandemic potential in a variety of viruses showed that development of these traits rarely passes through the human-infective (but non-transmissible) stage (Lu et al., 2019). It is estimated that there are approximately 40,000 virus species in mammals, of which 10,000 have zoonotic potential (Carlson et al., 2019); yet, we currently only know about 300 viruses that are human-infective (King et al., 2011). The recognition that the great majority of mammal viruses are still unknown, and that our understanding of the drivers of viral emergence is still limited, has led some to propose a large-scale survey of viral diversity in animals – the Global Virome Project (Carroll et al., 2018). Carroll et al. (2018) claim that, while costly, the Global Virome Project would result in an abundance of data that could improve predictive studies on emergence. However, the project is controversial, with others suggesting that prediction of emergence is impossible – emergence events are affected by too many variables, and too few have occurred so far to

train predictive models on – and that money is better spent on surveillance and rapid response (Holmes et al., 2018).

6.5 Lessons for future metagenomic surveillance studies

Because of their ability to detect a wide variety of agents, including novel and unexpected viruses, metagenomic methods appear ideally suited to surveillance studies at the human-animal interface. They can contribute to the characterisation of the local zoonotic pool, identification of novel zoonoses in high-risk populations, description of the molecular epidemiology of known pathogens, and elucidation of aetiologies of disease of unknown origin – perhaps even all at once. However, metagenomic sequencing remains costly, and "metagenomic surveillance" is thus often limited in scope and time, in turn restricting the power of such studies to detect rare events and investigate complex situations. An additional important drawback is that the large amount of generated data can be noisy and difficult to interpret. In future metagenomic surveillance studies, it may be beneficial to focus one's limited sequencing power on a specific question, and to carefully consider study design to eliminate as much noise and ambiguity as possible. Here, I discuss some considerations that follow from the limitations of the current study.

6.5.1 Increasing relevant signals through more targeted sampling

A major aim of the studies in this thesis was to learn about viral chatter and subsequent emergence of zoonotic viruses in Vietnam. Most aspects of this, including estimating the relative frequencies of chatter and onwards transmission, require the detection of multiple spillover events. However, zoonotic infections are rare, and here, few were detected – especially if one considers *bona fide* mammal-infective pathogens. As most human samples came from hospital patients with diarrhoea, human diarrhoeal pathogens were in abundance, but, with the exception of RVA, these were non-zoonotic. The samples from the high-risk cohort came from multiple risk groups with different animal contacts, which meant that the study had relatively low power to detect spillover across any one particular human-animal interface. As a result, while the VIZIONS metagenomics study was well-suited to provide insights into which viruses circulate in humans, swine and rats in the Vietnamese Mekong Delta, it was not sufficiently powered to investigate the epidemiology or risk of emergence of any such viruses (with the exception, perhaps, of RVA).

In future studies with similar aims, power can be increased by a more targeted sampling frame. To reduce irrelevant detections of common human pathogens in hospital patients, one can apply metagenomics only in case a diagnostic screen does not reveal an aetiology. In complement to the above, to increase the likelihood of detecting spillover events, one can focus on a single animal host species, ideally the most relevant in the study setting, and mainly target humans with contacts with this particular animal species. If one is specifically interested in zoonotic *pathogens*, one can focus on clinical episodes in these individuals and/or their animals, rather than sample at set time points. Studies can be targeted further by matching sampling procedures to the tropism and disease presentations of any expected viruses. However, an overly narrow sampling frame can also result in missing elements of interest, such as onwards transmission after spillover events if one only focuses on animal contacts. Overall, one's sampling frame should match the main question of the study, with an appropriate balance between focus and breadth.

6.5.2 Minimizing noise from contamination and exposure

Metagenomic surveillance studies can reveal the presence of a variety of viruses of which the relevance to human or animal health is unknown. While novel and unexpected viruses can be of particular interest in the context of emerging infections, this interest is generally conditional upon the ability of these viruses to infect humans or animals. In this respect, an important limitation of metagenomic methods is that they cannot distinguish between direct infections, and viruses from other sources, such as infections of symbionts, dietary viruses and contaminants.

In the studies in this thesis, I used a combination of biological plausibility and small signal size to identify signals that were unlikely to be the result of infections. But these are rough and unreliable filters, and, additionally, application of a plausibility criterion depends on prior knowledge about the detected viruses or their relatives. As a result, signals for a variety of viruses remained difficult to interpret. This was particularly the case for PBV and CyV-VN, which appear zoonotic on the basis of molecular detections in humans and animals, but of which it remains unknown whether they truly infect (and are thus able to cause disease in) mammals. Another example is the single small CaCV signal in a human sample, which could represent a novel zoonosis: current knowledge about circoviruses was too little to exclude human infection, but non-infectious exposure and contamination appeared more likely explanations.

In metagenomics studies, two complementary strategies can be used to address such situations: the minimization of contamination, and the provision of additional evidence in support of a signal representing an infection.

There are a variety of actions one can take to minimize contamination. Index switching, a likely cause of cross-contamination between samples in the VIZIONS study, can be eliminated by two minor modifications of sequencing protocols: the use of two separate indexes for each sample, and the application of a quality filter to the index reads (Kircher et al., 2012, Wright and Vetsigian, 2016b). Additionally, multiple different index sets can be used to avoid carry-over of index sequences from one run to the next. Recommendations to minimize and deal with other types of contamination in sequencing studies have been proposed elsewhere (Salter et al., 2014, Strong et al., 2014, de Goffau et al., 2018). In general, contamination can be limited by using stringent protocols and dedicated laboratory space, kit and staff. To facilitate detection and removal of residual contamination, negative controls should be included for all sample processing steps, and records should be kept of all kit batches. Finally, at the computational stage, any identified taxa can be checked against databases of known contaminants.

To demonstrate that a metagenomic signal represents an infection, evidence is required that supports viral replication and excludes non-infectious exposure or contamination. One way in which this can be achieved is by combining metagenomics with serological methods. An autologous antibody capture process, using convalescent serum from diseased subjects (rich in virus-specific antibodies), can be used to enrich faecal or respiratory samples for immunogenic viruses prior to metagenomic sequencing (Oude Munnink et al., 2013). Serological assays that detect virus-specific antibodies can also be used to confirm any metagenomic findings. Alternatively, detection of viruses in internal compartments, such as blood, tissues or CSF, may be considered sufficient evidence, as non-infectious viruses do not pass through these compartments in large numbers (Li and Delwart, 2011). Altogether, it may thus be useful to take multiple sample types, and to consider the need for confirmation by other methods when designing and costing a metagenomic study.

6.5.3 Demonstrating associations with disease or animal exposure: comparative studies

Another major challenge regarding the interpretation of metagenomic signals in the context of surveillance is that, for many novel or recently discovered viruses, it is not known whether they are pathogenic, or whether their transmission dynamics have a zoonotic component.

Traditionally, in microbiology, causation of disease is demonstrated through fulfilment of Koch's postulates[4] (Koch, 1890); however, their reliance on the absence of asymptomatic infections, the ability to culture the pathogen, and the availability of an animal model, means that they are not easily applied to viruses. Various alternative guidelines have been proposed that also allow evidence from observational rather than experimental studies – for example, for viruses (Rivers, 1937), epidemiological studies (Hill, 1965), and molecular detections (Fredricks and Relman, 1996, Falkow, 1988). While different strategies can be used to investigate the pathogenicity of novel viruses discovered by metagenomics (reviewed in (Oude Munnink and van der Hoek, 2016, Lipkin, 2010, Li and Delwart, 2011, Mokili et al., 2012)), a straightforward first step is testing whether there is an association between viral detection and disease status.

Metagenomic surveillance studies can be designed so that disease associations can be tested directly. This requires a set up as a case-control study, with sampling of healthy controls, or individuals with unrelated symptoms, alongside hospital patients or high-risk cohort members with disease. Controls should be matched according to age, location, and other variables affecting exposure to viruses, such as occupation and animal contacts, to minimize differences between comparison groups.

Similarly, metagenomic and epidemiological methods can be paired to test whether any virus detected in humans is associated with specific animal exposures. To this end, high-risk cohorts should be set up in a more structured way, so that viral detections can be compared between groups with different exposures.

In designing such studies, it is important to not only carefully consider one's comparison groups, but also to think of batch effects and the impact they could have on analyses. To

[4]As summarised from Koch's speech: (1) a pathogen is present in all individuals with the disease, but not in healthy individuals; (2) it must be isolated from a diseased individual and grown in pure culture; and (3) upon inoculation of this culture into a healthy individual, it should cause similar disease.

prevent batch effects from causing false patterns, samples from different comparison groups should be equally distributed among any processing batches. Further practical suggestions to prevent, detect and adjust for batch effects have been reviewed by Leek et al. (2010).

6.6 Future directions

An obvious continuation of the studies in this thesis would be to assemble the sequence data and characterize the full genomes of any identified viruses considered of interest, such as putative novel viruses and viruses with established or putative zoonotic potential. For assembly, a variety of methods can be used, as reviewed in (Rose et al., 2016, Smits et al., 2015). Closest relatives can then be identified with BLAST (Camacho et al., 2009, Altschul et al., 1990), and identities assessed across the genomes. Sequences should also be confirmed by PCR, both to validate assembly and to exclude the original detections being due to contamination during sample processing. With assembled sequences, metagenomic sequence data from all samples can also be searched directly for related viruses that may have been missed by the taxonomic classification pipeline.

After assembly, viral sequences from different study populations can be compared to infer their evolutionary relationships. Resulting phylogenies can be used to further explore some epidemiological and ecological questions that remained unanswered in this thesis. For example, they allow identification of zoonotic spillover events and likely transmission chains. For RVA, this may bring insights into the relative frequencies of zoonotic spillover infections and onwards transmission of such viruses in human populations. Additionally, ecological traits, like host species and geographical location, can be mapped onto phylogenies, and their clustering quantified, to investigate to what extent they structure viral transmission (e.g. Faria et al. (2014)). When applied to PBV and CyV-VN, this may shed further light onto the ecology of these viruses. Furthermore, phylogenies can confirm the position of a viral sequence as falling within the diversity of a clade with known features, or instead highlight it as divergent, in which case its features may be different from those of known viruses. This may help formulate further hypotheses about the nature and/or origin of novel viruses or unexpected findings, such as the CaCV signal in a human sample.

This thesis also raised questions that are of public health relevance, but that require more data to answer. For example, what are the relative contributions of zoonotic hepatitis E and rotavirus infections to human disease in Vietnam, and is this changing over time? Extending

the use of broad PCR systems and whole genome sequencing methods in disease surveillance programmes would improve our ability to detect unusual infections and to distinguish between anthroponotic and zoonotic strains. With such capacity in place, surveillance programmes can track changes in genotype composition over time, and also provide more data for epidemiological studies seeking to characterise disease associations or to identify risk factors.

6.7 Conclusion

In conclusion, the studies in this thesis have focused on the exploration of viral metagenomics as a surveillance tool for emerging zoonotic infections, and on the characterization of mammalian viruses at the human-animal interface in Vietnam. A viral taxonomic classification pipeline was built and tested by comparison with diagnostic qPCR, revealing an apparent high discriminatory accuracy. While cross-contamination of samples in a sequencing run was identified as a limitation of the pipeline, the extent of this was modelled, so that it could be taken into account in the setting of signal thresholds. Application of the pipeline to samples from human, swine and rats identified a variety of viruses: viruses typical for these hosts, putative novel variants, viruses known or presumed to be zoonotic, and putative novel zoonoses. Comparison of signals for viruses shared between human and animal study populations suggested that viral chatter (RVA) and cross-species transmission within a more generalist ecology (PBV and CyV-VN) are plausible in this setting. However, assessment of the risks that these and other viruses pose as potential emerging public health threats was hampered by relatively few signals representing cross-species transmissions, or by uncertainty about viral tropism. More extensive genomic surveillance of viruses with a known zoonotic component (such as rotaviruses and orthohepeviruses), and investigations into the basic biology and ecology of understudied and novel viruses (such as PBV and cycloviruses) are needed to shed further light on the risks posed by these viruses.

References

Database of Rodent-associated Viruses [Online]. Available: http://www.mgc.ac.cn/DRodVir/ [Accessed].

Picornaviridae.com [Online]. The Pirbright Institute. Available: www.picornaviridae.com [Accessed 2019].

Adams, M. J., Lefkowitz, E. J., King, A. M. & Carstens, E. B. 2014. Ratification vote on taxonomic proposals to the International Committee on Taxonomy of Viruses (2014). *Arch Virol,* 159**,** 2831-41.

Adams, M. J., Lefkowitz, E. J., King, A. M. Q., Harrach, B., Harrison, R. L., Knowles, N. J., Kropinski, A. M., Krupovic, M., Kuhn, J. H., Mushegian, A. R., Nibert, M., Sabanadzovic, S., Sanfacon, H., Siddell, S. G., Simmonds, P., Varsani, A., Zerbini, F. M., Gorbalenya, A. E. & Davison, A. J. 2017. Changes to taxonomy and the International Code of Virus Classification and Nomenclature ratified by the International Committee on Taxonomy of Viruses (2017). *Arch Virol,* 162**,** 2505-2538.

Adlhoch, C., Avellon, A., Baylis, S. A., Ciccaglione, A. R., Couturier, E., De Sousa, R., Epstein, J., Ethelberg, S., Faber, M., Feher, A., Ijaz, S., Lange, H., Mandakova, Z., Mellou, K., Mozalevskis, A., Rimhanen-Finne, R., Rizzi, V., Said, B., Sundqvist, L., Thornton, L., Tosti, M. E., Van Pelt, W., Aspinall, E., Domanovic, D., Severi, E., Takkinen, J. & Dalton, H. R. 2016. Hepatitis E virus: Assessment of the epidemiological situation in humans in Europe, 2014/15. *J Clin Virol,* 82**,** 9-16.

Ahmed, K., Anh, D. D. & Nakagomi, O. 2007. Rotavirus G5P[6] in child with diarrhea, Vietnam. *Emerg Infect Dis,* 13**,** 1232-5.

Aird, D., Ross, M. G., Chen, W. S., Danielsson, M., Fennell, T., Russ, C., Jaffe, D. B., Nusbaum, C. & Gnirke, A. 2011. Analyzing and minimizing PCR amplification bias in Illumina sequencing libraries. *Genome Biol,* 12**,** R18.

Akey, J. M., Biswas, S., Leek, J. T. & Storey, J. D. 2007. On the design and analysis of gene expression studies in human populations. *Nat Genet,* 39**,** 807-8; author reply 808-9.

Allen, T., Murray, K. A., Zambrana-Torrelio, C., Morse, S. S., Rondinini, C., Di Marco, M., Breit, N., Olival, K. J. & Daszak, P. 2017. Global hotspots and correlates of emerging zoonotic diseases. *Nat Commun,* 8**,** 1124.

Altschul, S. F., Gish, W., Miller, W., Myers, E. W. & Lipman, D. J. 1990. Basic local alignment search tool. *J Mol Biol,* 215**,** 403-10.

Ames, S. K., Hysom, D. A., Gardner, S. N., Lloyd, G. S., Gokhale, M. B. & Allen, J. E. 2013. Scalable metagenomic taxonomy classification using a reference genome database. *Bioinformatics,* 29**,** 2253-60.

Anderson, A., Hartmann, K., Leutenegger, C. M., Proksch, A. L., Mueller, R. S. & Unterer, S. 2017. Role of canine circovirus in dogs with acute haemorrhagic diarrhoea. *Vet Rec,* 180**,** 542.

Andonov, A., Robbins, M., Borlang, J., Cao, J., Hattchete, T., Stueck, A., Deschaumbault, Y., Murnaghan, K., Varga, J. & Johnston, B. 2019. Rat hepatitis E virus linked to severe acute hepatitis in an immunocompetent patient. *J Infect Dis*.

Andrews, S. 2010. FastQC: A Quality Control Tool for High Throughput Sequence Data.

Atoni, E., Wang, Y., Karungu, S., Waruhiu, C., Zohaib, A., Obanda, V., Agwanda, B., Mutua, M., Xia, H. & Yuan, Z. 2018. Metagenomic Virome Analysis of Culex Mosquitoes from Kenya and China. *Viruses,* 10.

Azhar, E. I., El-Kafrawy, S. A., Farraj, S. A., Hassan, A. M., Al-Saeed, M. S., Hashem, A. M. & Madani, T. A. 2014. Evidence for camel-to-human transmission of MERS coronavirus. *N Engl J Med,* 370**,** 2499-505.

Baggerly, K. A., Edmonson, S. R., Morris, J. S. & Coombes, K. R. 2004. High-resolution serum proteomic patterns for ovarian cancer detection. *Endocr Relat Cancer,* 11**,** 583-4; author reply 585-7.

Ballenghien, M., Faivre, N. & Galtier, N. 2017. Patterns of cross-contamination in a multispecies population genomic project: detection, quantification, impact, and solutions. *BMC Biology,* 15**,** 25.

Banyai, K., Martella, V., Bogdan, A., Forgach, P., Jakab, F., Meleg, E., Biro, H., Melegh, B. & Szucs, G. 2008. Genogroup I picobirnaviruses in pigs: evidence for genetic diversity and relatedness to human strains. *J Gen Virol,* 89**,** 534-9.

Bartram, J., Mountjoy, E., Brooks, T., Hancock, J., Williamson, H., Wright, G., Moppett, J., Goulden, N. & Hubank, M. 2016. Accurate Sample Assignment in a Multiplexed, Ultrasensitive, High-Throughput Sequencing Assay for Minimal Residual Disease. *J Mol Diagn,* 18**,** 494-506.

Bazinet, A. L. & Cummings, M. P. 2012. A comparative evaluation of sequence classification programs. *BMC Bioinformatics,* 13**,** 92.

Bell, D., Roberton, S. & Hunter, P. R. 2004. Animal origins of SARS coronavirus: possible links with the international trade in small carnivores. *Philos Trans R Soc Lond B Biol Sci,* 359**,** 1107-14.

Benjamini, Y. & Speed, T. P. 2012. Summarizing and correcting the GC content bias in high-throughput sequencing. *Nucleic Acids Res,* 40**,** e72.

Bermejo, M., Rodriguez-Teijeiro, J. D., Illera, G., Barroso, A., Vila, C. & Walsh, P. D. 2006. Ebola outbreak killed 5000 gorillas. *Science,* 314**,** 1564.

Berto, A., Anh, P. H., Carrique-Mas, J. J., Simmonds, P., Van Cuong, N., Tue, N. T., Van Dung, N., Woolhouse, M. E., Smith, I., Marsh, G. A., Bryant, J. E., Thwaites, G. E., Baker, S., Rabaa, M. A. & Consortium, V. 2017a. Detection of potentially novel paramyxovirus and coronavirus viral RNA in bats and rats in the Mekong Delta region of southern Viet Nam. *Zoonoses Public Health*.

Berto, A., Pham, H. A., Thao, T. T. N., Vy, N. H. T., Caddy, S. L., Hiraide, R., Tue, N. T., Goodfellow, I., Carrique-Mas, J. J., Thwaites, G. E., Baker, S., Boni, M. F. & Consortium, V. 2017b. Hepatitis E in southern Vietnam: Seroepidemiology in humans and molecular epidemiology in pigs. *Zoonoses Public Health*.

Berto, A., Pham, H. A., Thao, T. T. N., Vy, N. H. T., Caddy, S. L., Hiraide, R., Tue, N. T., Goodfellow, I., Carrique-Mas, J. J., Thwaites, G. E., Baker, S., Boni, M. F. & Consortium, V. 2018. Hepatitis E in southern Vietnam: Seroepidemiology in humans and molecular epidemiology in pigs. *Zoonoses Public Health,* 65**,** 43-50.

Bewick, V., Cheek, L. & Ball, J. 2004. Statistics review 13: receiver operating characteristic curves. *Crit Care,* 8**,** 508-12.

Bhat, S., Kattoor, J. J., Malik, Y. S., Sircar, S., Deol, P., Rawat, V., Rakholia, R., Ghosh, S., Vlasova, A. N., Nadia, T., Dhama, K. & Kobayashi, N. 2018. Species C Rotaviruses in Children with Diarrhea in India, 2010-2013: A Potentially Neglected Cause of Acute Gastroenteritis. *Pathogens,* 7.

Biagini, P. 2009. Classification of TTV and related viruses (anelloviruses). *Curr Top Microbiol Immunol,* 331**,** 21-33.

Bodewes, R., Ruiz-Gonzalez, A., Schapendonk, C. M., Van Den Brand, J. M., Osterhaus, A. D. & Smits, S. L. 2014. Viral metagenomic analysis of feces of wild small carnivores. *Virol J,* 11**,** 89.

Boom, R., Sol, C. J., Salimans, M. M., Jansen, C. L., Wertheim-Van Dillen, P. M. & Van Der Noordaa, J. 1990. Rapid and simple method for purification of nucleic acids. *J Clin Microbiol,* 28**,** 495-503.

Boratyn, G. M., Thierry-Mieg, J., Thierry-Mieg, D., Busby, B. & Madden, T. L. 2019. Magic-BLAST, an accurate RNA-seq aligner for long and short reads. *BMC Bioinformatics,* 20**,** 405.

Bouquet, J., Cheval, J., Rogee, S., Pavio, N. & Eloit, M. 2012. Identical consensus sequence and conserved genomic polymorphism of hepatitis E virus during controlled interspecies transmission. *J Virol,* 86**,** 6238-45.

Bouquet, J., Melgar, M., Swei, A., Delwart, E., Lane, R. S. & Chiu, C. Y. 2017. Metagenomic-based Surveillance of Pacific Coast tick Dermacentor occidentalis Identifies Two Novel Bunyaviruses and an Emerging Human Ricksettsial Pathogen. *Sci Rep,* 7**,** 12234.

Bouquet, J., Tesse, S., Lunazzi, A., Eloit, M., Rose, N., Nicand, E. & Pavio, N. 2011. Close similarity between sequences of hepatitis E virus recovered from humans and swine, France, 2008-2009. *Emerg Infect Dis,* 17**,** 2018-25.

Brady, A. & Salzberg, S. 2011. PhymmBL expanded: confidence scores, custom databases, parallelization and more. *Nat Methods,* 8**,** 367.

Brady, A. & Salzberg, S. L. 2009. Phymm and PhymmBL: metagenomic phylogenetic classification with interpolated Markov models. *Nat Methods,* 6**,** 673-6.

Breitbart, M., Delwart, E., Rosario, K., Segales, J., Varsani, A. & Consortium, I. R. 2017. ICTV Virus Taxonomy Profile: Circoviridae. *Journal of General Virology,* 98**,** 1997-1998.

Brinkmann, A., Nitsche, A. & Kohl, C. 2016. Viral Metagenomics on Blood-Feeding Arthropods as a Tool for Human Disease Surveillance. *Int J Mol Sci,* 17.

Brooks, E. G. E., Roberton, S. I. & Bell, D. J. 2010. The conservation impact of commercial wildlife farming of porcupines in Vietnam. *Biological Conservation,* 143**,** 2808-2814.

Buchfink, B., Xie, C. & Huson, D. H. 2015. Fast and sensitive protein alignment using DIAMOND. *Nat Methods,* 12**,** 59-60.

Camacho, C., Coulouris, G., Avagyan, V., Ma, N., Papadopoulos, J., Bealer, K. & Madden, T. L. 2009. BLAST+: architecture and applications. *BMC Bioinformatics,* 10**,** 421.

Carlson, C. J., Zipfel, C. M., Garnier, R. & Bansal, S. 2019. Global estimates of mammalian viral diversity accounting for host sharing. *Nat Ecol Evol,* 3**,** 1070-1075.

Carrique-Mas, J. J., Tue, N. T., Bryant, J. E., Saylors, K., Cuong, N. V., Hoa, N. T., An, N. N., Hien, V. B., Lao, P. V., Tu, N. C., Chuyen, N. K., Chuc, N. T., Tan, D. V., Duong, H. V., Toan, T. K., Chi, N. T., Campbell, J., Rabaa, M. A., Nadjm, B., Woolhouse, M., Wertheim, H., Thwaites, G. & Baker, S. 2015. The baseline characteristics and interim analyses of the high-risk sentinel cohort of the Vietnam Initiative on Zoonotic InfectiONS (VIZIONS). *Sci Rep,* 5**,** 17965.

Carroll, D., Daszak, P., Wolfe, N. D., Gao, G. F., Morel, C. M., Morzaria, S., Pablos-Mendez, A., Tomori, O. & Mazet, J. a. K. 2018. The Global Virome Project. *Science,* 359**,** 872-874.

Casto, A. M., Adler, A. L., Makhsous, N., Crawford, K., Qin, X., Kuypers, J. M., Huang, M. L., Zerr, D. M. & Greninger, A. L. 2018. Prospective real-time metagenomic sequencing during norovirus outbreak reveals discrete transmission clusters. *Clin Infect Dis*.

Ceballos, G. & Ehrlich, P. R. 2006. Global mammal distributions, biodiversity hotspots, and conservation. *Proc Natl Acad Sci U S A,* 103**,** 19374-9.

Centers for Disease Control and Prevention 2003a. Cholera epidemic after increased civil conflict--Monrovia, Liberia, June-September 2003. *MMWR Morb Mortal Wkly Rep,* 52**,** 1093-5.

Centers for Disease Control and Prevention 2003b. Multistate outbreak of monkeypox--Illinois, Indiana, and Wisconsin, 2003. *MMWR Morb Mortal Wkly Rep,* 52**,** 537-40.

Chen, L., Liu, B., Wu, Z., Jin, Q. & Yang, J. 2017. DRodVir: A resource for exploring the virome diversity in rodents. *J Genet Genomics,* 44**,** 259-264.

Cheng, W. X., Li, J. S., Huang, C. P., Yao, D. P., Liu, N., Cui, S. X., Jin, Y. & Duan, Z. J. 2010. Identification and nearly full-length genome characterization of novel porcine bocaviruses. *PLoS One,* 5**,** e13583.

Chiu, C. Y. 2013. Viral pathogen discovery. *Curr Opin Microbiol,* 16**,** 468-78.

Chua, K. B., Bellini, W. J., Rota, P. A., Harcourt, B. H., Tamin, A., Lam, S. K., Ksiazek, T. G., Rollin, P. E., Zaki, S. R., Shieh, W., Goldsmith, C. S., Gubler, D. J., Roehrig, J. T., Eaton, B., Gould, A. R., Olson, J., Field, H., Daniels, P., Ling, A. E., Peters, C. J., Anderson, L. J. & Mahy, B. W. 2000. Nipah virus: a recently emergent deadly paramyxovirus. *Science,* 288**,** 1432-5.

Chua, K. B., Voon, K., Yu, M., Keniscope, C., Abdul Rasid, K. & Wang, L. F. 2011. Investigation of a potential zoonotic transmission of orthoreovirus associated with acute influenza-like illness in an adult patient. *PLoS One,* 6**,** e25434.

Claude, K. M., Underschultz, J. & Hawkes, M. T. 2018. Ebola virus epidemic in war-torn eastern DR Congo. *Lancet,* 392**,** 1399-1401.

Cleaveland, S., Laurenson, M. K. & Taylor, L. H. 2001. Diseases of humans and their domestic mammals: pathogen characteristics, host range and the risk of emergence. *Philos Trans R Soc Lond B Biol Sci,* 356**,** 991-9.

Coker, R. J., Hunter, B. M., Rudge, J. W., Liverani, M. & Hanvoravongchai, P. 2011. Emerging infectious diseases in southeast Asia: regional challenges to control. *Lancet,* 377**,** 599-609.

Colson, P., Borentain, P., Queyriaux, B., Kaba, M., Moal, V., Gallian, P., Heyries, L., Raoult, D. & Gerolami, R. 2010. Pig liver sausage as a source of hepatitis E virus transmission to humans. *J Infect Dis,* 202**,** 825-34.

Conceicao-Neto, N., Theuns, S., Cui, T., Zeller, M., Yinda, C. K., Christiaens, I., Heylen, E., Van Ranst, M., Carpentier, S., Nauwynck, H. J. & Matthijnssens, J. 2017. Identification of an enterovirus recombinant with a torovirus-like gene insertion during a diarrhea outbreak in fattening pigs. *Virus Evol,* 3**,** vex024.

Conceicao-Neto, N., Zeller, M., Lefrere, H., De Bruyn, P., Beller, L., Deboutte, W., Yinda, C. K., Lavigne, R., Maes, P., Van Ranst, M., Heylen, E. & Matthijnssens, J. 2015. Modular approach to customise sample preparation procedures for viral metagenomics: a reproducible protocol for virome analysis. *Sci Rep,* 5**,** 16532.

Cook, N., Bridger, J., Kendall, K., Gomara, M. I., El-Attar, L. & Gray, J. 2004. The zoonotic potential of rotavirus. *J Infect,* 48**,** 289-302.

Coston, M. 2019. *Hong Kong Reports 6th Human Infection With Rat Hepatitis E Virus* [Online]. Available: http://afludiary.blogspot.com/2019/06/hong-kong-reports-6th-human-infection.html [Accessed 16/07/2019 2019].

Cotten, M., Oude Munnink, B. B., Canuti, M., Deijs, M., Watson, S. J., Kellam, P. & Van Der Hoek, L. 2014a. Full genome virus detection in fecal samples using sensitive nucleic acid preparation, deep sequencing, and a novel iterative sequence classification algorithm. *PLoS One,* 9**,** e93269.

Cotten, M., Watson, S. J., Kellam, P., Al-Rabeeah, A. A., Makhdoom, H. Q., Assiri, A., Al-Tawfiq, J. A., Alhakeem, R. F., Madani, H., Alrabiah, F. A., Al Hajjar, S., Al-Nassir, W. N., Albarrak, A., Flemban, H., Balkhy, H. H., Alsubaie, S., Palser, A. L., Gall, A., Bashford-Rogers, R., Rambaut, A., Zumla, A. I. & Memish, Z. A. 2013. Transmission and evolution of the Middle East respiratory syndrome coronavirus in Saudi Arabia: a descriptive genomic study. *Lancet,* 382**,** 1993-2002.

Cotten, M., Watson, S. J., Zumla, A. I., Makhdoom, H. Q., Palser, A. L., Ong, S. H., Al Rabeeah, A. A., Alhakeem, R. F., Assiri, A., Al-Tawfiq, J. A., Albarrak, A., Barry, M., Shibl, A.,

Alrabiah, F. A., Hajjar, S., Balkhy, H. H., Flemban, H., Rambaut, A., Kellam, P. & Memish, Z. A. 2014b. Spread, Circulation, and Evolution of the Middle East Respiratory Syndrome Coronavirus. *mBio,* 5.

Cowman, G., Otipo, S., Njeru, I., Achia, T., Thirumurthy, H., Bartram, J. & Kioko, J. 2017. Factors associated with cholera in Kenya, 2008-2013. *Pan Afr Med J,* 28**,** 101.

D'amore, R., Ijaz, U. Z., Schirmer, M., Kenny, J. G., Gregory, R., Darby, A. C., Shakya, M., Podar, M., Quince, C. & Hall, N. 2016. A comprehensive benchmarking study of protocols and sequencing platforms for 16S rRNA community profiling. *BMC Genomics,* 17**,** 55.

Dai, X. Q., Hua, X. G., Shan, T. L., Delwart, E. & Zhao, W. 2010. Human cosavirus infections in children in China. *J Clin Virol,* 48**,** 228-9.

Dai, Y., Zhou, Q., Zhang, C., Song, Y., Tian, X., Zhang, X., Xue, C., Xu, S., Bi, Y. & Cao, Y. 2012. Complete genome sequence of a porcine orthoreovirus from southern China. *J Virol,* 86**,** 12456.

Daszak, P. 2009. A call for "Smart Surveillance": a lesson learned from H1N1. *Ecohealth,* 6**,** 1-2.

Daszak, P., Cunningham, A. A. & Hyatt, A. D. 2000. Emerging infectious diseases of wildlife--threats to biodiversity and human health. *Science,* 287**,** 443-9.

Daszak, P., Epstein, J. H., Kilpatrick, A. M., Aguirre, A. A., Karesh, W. B. & Cunningham, A. A. 2007. Collaborative research approaches to the role of wildlife in zoonotic disease emergence. *Curr Top Microbiol Immunol,* 315**,** 463-75.

Davies, T. J. & Pedersen, A. B. 2008. Phylogeny and geography predict pathogen community similarity in wild primates and humans. *Proc Biol Sci,* 275**,** 1695-701.

De Goffau, M. C., Lager, S., Salter, S. J., Wagner, J., Kronbichler, A., Charnock-Jones, D. S., Peacock, S. J., Smith, G. C. S. & Parkhill, J. 2018. Recognizing the reagent microbiome. *Nat Microbiol,* 3**,** 851-853.

De Vries, M., Deijs, M., Canuti, M., Van Schaik, B. D., Faria, N. R., Van De Garde, M. D., Jachimowski, L. C., Jebbink, M. F., Jakobs, M., Luyf, A. C., Coenjaerts, F. E., Claas, E. C., Molenkamp, R., Koekkoek, S. M., Lammens, C., Leus, F., Goossens, H., Ieven, M., Baas, F. & Van Der Hoek, L. 2011. A sensitive assay for virus discovery in respiratory clinical samples. *PLoS One,* 6**,** e16118.

De Vries, M., Oude Munnink, B. B., Deijs, M., Canuti, M., Koekkoek, S. M., Molenkamp, R., Bakker, M., Jurriaans, S., Van Schaik, B. D., Luyf, A. C., Olabarriaga, S. D., Van Kampen, A. H. & Van Der Hoek, L. 2012. Performance of VIDISCA-454 in feces-suspensions and serum. *Viruses,* 4**,** 1328-34.

Decaro, N., Martella, V., Desario, C., Lanave, G., Circella, E., Cavalli, A., Elia, G., Camero, M. & Buonavoglia, C. 2014. Genomic characterization of a circovirus associated with fatal hemorrhagic enteritis in dog, Italy. *PLoS One,* 9**,** e105909.

Delmas, B., Attoui, H., Ghosh, S., Malik, Y. S., Mundt, E., Vakharia, V. N. & Ictv Report, C. 2019. ICTV virus taxonomy profile: Picobirnaviridae. *J Gen Virol,* 100**,** 133-134.

Delong, E. R., Delong, D. M. & Clarke-Pearson, D. L. 1988. Comparing the areas under two or more correlated receiver operating characteristic curves: a nonparametric approach. *Biometrics,* 44**,** 837-45.

Dennis, T. P. W., Flynn, P. J., De Souza, W. M., Singer, J. B., Moreau, C. S., Wilson, S. J. & Gifford, R. J. 2018. Insights into Circovirus Host Range from the Genomic Fossil Record. *J Virol,* 92.

Diez-Villasenor, C. & Rodriguez-Valera, F. 2019. CRISPR analysis suggests that small circular single-stranded DNA smacoviruses infect Archaea instead of humans. *Nat Commun,* 10**,** 294.

Dinh, P. N., Long, H. T., Tien, N. T. K., Hien, N. T., Mai, L. T. Q., Phong, L. H., Tuan, L. V., Tan, H. V., Nguyen, N. B., Tu, P. V. & Phuong, N. T. M. 2006. Risk Factors for Human

Infection with Avian Influenza A H5N1, Vietnam, 2004. *Emerging Infectious Disease journal,* 12**,** 1841.

Do, L. A., Bryant, J. E., Tran, A. T., Nguyen, B. H., Tran, T. T., Tran, Q. H., Vo, Q. B., Tran Dac, N. A., Trinh, H. N., Nguyen, T. T., Le Binh, B. T., Le, K., Nguyen, M. T., Thai, Q. T., Vo, T. V., Ngo, N. Q., Dang, T. K., Cao, N. H., Tran, T. V., Ho, L. V., Farrar, J., De Jong, M. & Van Doorn, H. R. 2016. Respiratory Syncytial Virus and Other Viral Infections among Children under Two Years Old in Southern Vietnam 2009-2010: Clinical Characteristics and Disease Severity. *PLoS One,* 11**,** e0160606.

Do, L. P., Kaneko, M., Nakagomi, T., Gauchan, P., Agbemabiese, C. A., Dang, A. D. & Nakagomi, O. 2017. Molecular epidemiology of Rotavirus A, causing acute gastroenteritis hospitalizations among children in Nha Trang, Vietnam, 2007-2008: Identification of rare G9P[19] and G10P[14] strains. *J Med Virol,* 89**,** 621-631.

Doceul, V., Bagdassarian, E., Demange, A. & Pavio, N. 2016. Zoonotic Hepatitis E Virus: Classi fi cation, Animal Reservoirs and Transmission Routes. *Viruses,* 8.

Domingo, E. & Holland, J. J. 1994. Mutation rates and rapid evolution of RNA viruses. *In:* MORSE, S. S. (ed.) *The Evolutionary Biology of Viruses.* New York: Raven Press.

Donato, C. & Vijaykrishna, D. 2017. The Broad Host Range and Genetic Diversity of Mammalian and Avian Astroviruses. *Viruses,* 9.

Dowgier, G., Lorusso, E., Decaro, N., Desario, C., Mari, V., Lucente, M. S., Lanave, G., Buonavoglia, C. & Elia, G. 2017. A molecular survey for selected viral enteropathogens revealed a limited role of Canine circovirus in the development of canine acute gastroenteritis. *Vet Microbiol,* 204**,** 54-58.

Drury, R. 2009. Reducing urban demand for wild animals in Vietnam: examining the potential of wildlife farming as a conservation tool. *Conservation Letters,* 2**,** 263-270.

Du, J., Lu, L., Liu, F., Su, H., Dong, J., Sun, L., Zhu, Y., Ren, X., Yang, F., Guo, F., Liu, Q., Wu, Z. & Jin, Q. 2016. Distribution and characteristics of rodent picornaviruses in China. *Sci Rep,* 6**,** 34381.

Duan, Q., Zhu, H., Yang, Y., Li, W. H., Zhou, T., Song, L. H., Gan, Y. H., Tan, H., Jin, B. F., Li, H. Y., Zuo, T. T., Chen, D. H. & Zhang, X. M. 2003. Reovirus, isolated from SARS patients. *Chinese Science Bulletin,* 48**,** 1293-1296.

Dung, N. V. 2017. *Characterization of canine enteric viruses in Vietnam.* PhD, Yamaguchi University.

Dung, T. T., Phat, V. V., Nga, T. V., My, P. V., Duy, P. T., Campbell, J. I., Thuy, C. T., Hoang, N. V., Van Minh, P., Le Phuc, H., Tuyet, P. T., Vinh, H., Kien, D. T., Huy Hle, A., Vinh, N. T., Nga, T. T., Hau, N. T., Chinh, N. T., Thuong, T. C., Tuan, H. M., Simmons, C., Farrar, J. J. & Baker, S. 2013. The validation and utility of a quantitative one-step multiplex RT real-time PCR targeting rotavirus A and norovirus. *J Virol Methods,* 187**,** 138-43.

Duong, V. T., Phat, V. V., Tuyen, H. T., Dung, T. T., Trung, P. D., Minh, P. V., Tu Le, T. P., Campbell, J. I., Le Phuc, H., Ha, T. T., Ngoc, N. M., Huong, N. T., Tam, P. T., Huong, D. T., Xang, N. V., Dong, N., Phuong Le, T., Hung, N. V., Phu, B. D., Phuc, T. M., Thwaites, G. E., Vi, L. L., Rabaa, M. A., Thompson, C. N. & Baker, S. 2016. Evaluation of Luminex xTAG Gastrointestinal Pathogen Panel Assay for Detection of Multiple Diarrheal Pathogens in Fecal Samples in Vietnam. *J Clin Microbiol,* 54**,** 1094-100.

Dyer, O. 2018. Ebola: new Congo epidemic grows as conflict hampers aid efforts. *BMJ,* 362**,** k3599.

Edgar, R. C. 2010. Search and clustering orders of magnitude faster than BLAST. *Bioinformatics,* 26**,** 2460-1.

Eidson, M., Komar, N., Sorhage, F., Nelson, R., Talbot, T., Mostashari, F., Mclean, R. & West Nile Virus Avian Mortality Surveillance, G. 2001. Crow deaths as a sentinel

surveillance system for West Nile virus in the northeastern United States, 1999. *Emerg Infect Dis,* 7**,** 615-20.

Endoh, D., Mizutani, T., Kirisawa, R., Maki, Y., Saito, H., Kon, Y., Morikawa, S. & Hayashi, M. 2005. Species-independent detection of RNA virus by representational difference analysis using non-ribosomal hexanucleotides for reverse transcription. *Nucleic Acids Res,* 33**,** e65.

Enserink, M. 2005. Infectious diseases. Experts dismiss pig flu scare as nonsense. *Science,* 307**,** 1392.

Epstein, J. H. & Anthony, S. J. 2017. Viral discovery as a tool for pandemic preparedness. *Rev Sci Tech,* 36**,** 499-512.

Erker, J. C., Desai, S. M., Schlauder, G. G., Dawson, G. J. & Mushahwar, I. K. 1999. A hepatitis E virus variant from the United States: molecular characterization and transmission in cynomolgus macaques. *J Gen Virol,* 80 (Pt 3)**,** 681-90.

Erlwein, O., Robinson, M. J., Dustan, S., Weber, J., Kaye, S. & Mcclure, M. O. 2011. DNA extraction columns contaminated with murine sequences. *PLoS One,* 6**,** e23484.

European Centre for Disease Prevention and Control (Ecdc) 2015. Rapid Risk Assessment: Novel zoonotic Borna disease virus associated with severe disease in breeders of variegated squirrels in Germany – first update, 5 May 2015. Stockholm: ECDC.

Falkow, S. 1988. Molecular Koch's postulates applied to microbial pathogenicity. *Rev Infect Dis,* 10 Suppl 2**,** S274-6.

Fancello, L., Raoult, D. & Desnues, C. 2012. Computational tools for viral metagenomics and their application in clinical research. *Virology,* 434**,** 162-74.

Fao. *FAOSTAT data* [Online]. Available: http://faostat.fao.org/ [Accessed].

Faria, N. R., Azevedo, R., Kraemer, M. U. G., Souza, R., Cunha, M. S., Hill, S. C., Theze, J., Bonsall, M. B., Bowden, T. A., Rissanen, I., Rocco, I. M., Nogueira, J. S., Maeda, A. Y., Vasami, F., Macedo, F. L. L., Suzuki, A., Rodrigues, S. G., Cruz, A. C. R., Nunes, B. T., Medeiros, D. B. A., Rodrigues, D. S. G., Queiroz, A. L. N., Da Silva, E. V. P., Henriques, D. F., Da Rosa, E. S. T., De Oliveira, C. S., Martins, L. C., Vasconcelos, H. B., Casseb, L. M. N., Simith, D. B., Messina, J. P., Abade, L., Lourenco, J., Alcantara, L. C. J., De Lima, M. M., Giovanetti, M., Hay, S. I., De Oliveira, R. S., Lemos, P. D. S., De Oliveira, L. F., De Lima, C. P. S., Da Silva, S. P., De Vasconcelos, J. M., Franco, L., Cardoso, J. F., Vianez-Junior, J., Mir, D., Bello, G., Delatorre, E., Khan, K., Creatore, M., Coelho, G. E., De Oliveira, W. K., Tesh, R., Pybus, O. G., Nunes, M. R. T. & Vasconcelos, P. F. C. 2016. Zika virus in the Americas: Early epidemiological and genetic findings. *Science,* 352**,** 345-349.

Faria, N. R., Rambaut, A., Suchard, M. A., Baele, G., Bedford, T., Ward, M. J., Tatem, A. J., Sousa, J. D., Arinaminpathy, N., Pépin, J., Posada, D., Peeters, M., Pybus, O. G. & Lemey, P. 2014. The early spread and epidemic ignition of HIV-1 in human populations. *Science,* 346**,** 56-61.

Feng, Y., Zhao, T., Nguyen, T., Inui, K., Ma, Y., Nguyen, T. H., Nguyen, V. C., Liu, D., Bui, Q. A., To, L. T., Wang, C., Tian, K. & Gao, G. F. 2008. Porcine respiratory and reproductive syndrome virus variants, Vietnam and China, 2007. *Emerg Infect Dis,* 14**,** 1774-6.

Fields, B. N., Knipe, D. M. & Howley, P. M. 2013. *Fields Virology,* Philadelphia, PA, USA, Wolters Kluwer Health/Lippincott Williams & Wilkins.

Finkbeiner, S. R., Allred, A. F., Tarr, P. I., Klein, E. J., Kirkwood, C. D. & Wang, D. 2008. Metagenomic analysis of human diarrhea: viral detection and discovery. *PLoS Pathog,* 4**,** e1000011.

Firth, C., Bhat, M., Firth, M. A., Williams, S. H., Frye, M. J., Simmonds, P., Conte, J. M., Ng, J., Garcia, J., Bhuva, N. P., Lee, B., Che, X., Quan, P. L. & Lipkin, W. I. 2014. Detection of

zoonotic pathogens and characterization of novel viruses carried by commensal Rattus norvegicus in New York City. *MBio,* 5**,** e01933-14.

Flygare, S., Simmon, K., Miller, C., Qiao, Y., Kennedy, B., Di Sera, T., Graf, E. H., Tardif, K. D., Kapusta, A., Rynearson, S., Stockmann, C., Queen, K., Tong, S., Voelkerding, K. V., Blaschke, A., Byington, C. L., Jain, S., Pavia, A., Ampofo, K., Eilbeck, K., Marth, G., Yandell, M. & Schlaberg, R. 2016. Taxonomer: an interactive metagenomics analysis portal for universal pathogen detection and host mRNA expression profiling. *Genome Biol,* 17**,** 111.

Fredricks, D. N. & Relman, D. A. 1996. Sequence-based identification of microbial pathogens: a reconsideration of Koch's postulates. *Clin Microbiol Rev,* 9**,** 18-33.

Freeman, M. M., Kerin, T., Hull, J., Mccaustland, K. & Gentsch, J. 2008. Enhancement of detection and quantification of rotavirus in stool using a modified real-time RT-PCR assay. *Journal of Medical Virology,* 80**,** 1489-1496.

Fregolente, M. C., De Castro-Dias, E., Martins, S. S., Spilki, F. R., Allegretti, S. M. & Gatti, M. S. 2009. Molecular characterization of picobirnaviruses from new hosts. *Virus Res,* 143**,** 134-6.

Friedrich Löffler Institute (Fli). 2015. *Neuer Bornavirus bei Bunthörnchen entdeckt – möglicher Zusammenhang mit Infektionen bei Menschen [New bornavirus discovered in variegated squirrels – possible connection with infections in humans]* [Online]. Greifswald: FLI. [Accessed 21/06/2019].

Friedrich Löffler Institute (Fli). 2016. *Weitere Fälle von Bunthörnchen-Bornavirus 1 festgestellt [Further cases of variegated squirrel bornavirus identified]* [Online]. Greifswald: FLI. Available: https://www.openagrar.de/servlets/MCRFileNodeServlet/Document_derivate_00014078/Kurznachricht-2016-03-01.pdf [Accessed 21/06/2019].

G. B. D. Causes of Death Collaborators 2017. Global, regional, and national age-sex specific mortality for 264 causes of death, 1980-2016: a systematic analysis for the Global Burden of Disease Study 2016. *Lancet,* 390**,** 1151-1210.

Gabbay, Y. B., Borges, A. A., Oliveira, D. S., Linhares, A. C., Mascarenhas, J. D., Barardi, C. R., Simoes, C. M., Wang, Y., Glass, R. I. & Jiang, B. 2008. Evidence for zoonotic transmission of group C rotaviruses among children in Belem, Brazil. *J Med Virol,* 80**,** 1666-74.

Ganesh, B., Banyai, K., Martella, V., Jakab, F., Masachessi, G. & Kobayashi, N. 2012. Picobirnavirus infections: viral persistence and zoonotic potential. *Rev Med Virol,* 22**,** 245-56.

Ganesh, B., Banyai, K., Masachessi, G., Mladenova, Z., Nagashima, S., Ghosh, S., Nataraju, S. M., Pativada, M., Kumar, R. & Kobayashi, N. 2011a. Genogroup I picobirnavirus in diarrhoeic foals: can the horse serve as a natural reservoir for human infection? *Vet Res,* 42**,** 52.

Ganesh, B., Nagashima, S., Ghosh, S., Nataraju, S. M., Rajendran, K., Manna, B., Ramamurthy, T., Niyogi, S. K., Kanungo, S., Sur, D., Kobayashi, N. & Krishnan, T. 2011b. Detection and molecular characterization of multiple strains of Picobirnavirus causing mixed infection in a diarrhoeic child: Emergence of prototype Genogroup II-like strain in Kolkata, India. *Int J Mol Epidemiol Genet,* 2**,** 61-72.

Gao, F., Bailes, E., Robertson, D. L., Chen, Y., Rodenburg, C. M., Michael, S. F., Cummins, L. B., Arthur, L. O., Peeters, M., Shaw, G. M., Sharp, P. M. & Hahn, B. H. 1999. Origin of HIV-1 in the chimpanzee Pan troglodytes troglodytes. *Nature,* 397**,** 436-41.

Gao, F., Yue, L., White, A. T., Pappas, P. G., Barchue, J., Hanson, A. P., Greene, B. M., Sharp, P. M., Shaw, G. M. & Hahn, B. H. 1992. Human infection by genetically diverse SIVSM-related HIV-2 in west Africa. *Nature,* 358**,** 495-9.

Garigliany, M. M., Hagen, R. M., Frickmann, H., May, J., Schwarz, N. G., Perse, A., Jost, H., Borstler, J., Shahhosseini, N., Desmecht, D., Mbunkah, H. A., Daniel, A. M., Kingsley, M. T., Campos Rde, M., De Paula, V. S., Randriamampionona, N., Poppert, S., Tannich, E., Rakotozandrindrainy, R., Cadar, D. & Schmidt-Chanasit, J. 2014. Cyclovirus CyCV-VN species distribution is not limited to Vietnam and extends to Africa. *Sci Rep,* 4**,** 7552.

Garmendia, A. E., Van Kruiningen, H. J. & French, R. A. 2001. The West Nile virus: its recent emergence in North America. *Microbes Infect,* 3**,** 223-9.

Geoghegan, J. L., Duchene, S. & Holmes, E. C. 2017. Comparative analysis estimates the relative frequencies of co-divergence and cross-species transmission within viral families. *PLoS Pathog,* 13**,** e1006215.

Geoghegan, J. L., Senior, A. M., Di Giallonardo, F. & Holmes, E. C. 2016. Virological factors that increase the transmissibility of emerging human viruses. *Proc Natl Acad Sci U S A,* 113**,** 4170-5.

Georges, A. J., Leroy, E. M., Renaut, A. A., Benissan, C. T., Nabias, R. J., Ngoc, M. T., Obiang, P. I., Lepage, J. P., Bertherat, E. J., Benoni, D. D., Wickings, E. J., Amblard, J. P., Lansoud-Soukate, J. M., Milleliri, J. M., Baize, S. & Georges-Courbot, M. C. 1999. Ebola hemorrhagic fever outbreaks in Gabon, 1994-1997: epidemiologic and health control issues. *J Infect Dis,* 179 Suppl 1**,** S65-75.

Gerber, P., Chilonda, P., Franceschini, G. & Menzi, H. 2005. Geographical determinants and environmental implications of livestock production intensification in Asia. *Bioresour Technol,* 96**,** 263-76.

Ghosh, S. & Kobayashi, N. 2014. Exotic rotaviruses in animals and rotaviruses in exotic animals. *Virusdisease,* 25**,** 158-72.

Ghosh, T. S., Mohammed, M. H., Komanduri, D. & Mande, S. S. 2011. ProViDE: A software tool for accurate estimation of viral diversity in metagenomic samples. *Bioinformation,* 6**,** 91-4.

Giordano, M. O., Martinez, L. C., Masachessi, G., Barril, P. A., Ferreyra, L. J., Isa, M. B., Valle, M. C., Massari, P. U. & Nates, S. V. 2011. Evidence of closely related picobirnavirus strains circulating in humans and pigs in Argentina. *J Infect,* 62**,** 45-51.

Goldberg, B., Sichtig, H., Geyer, C., Ledeboer, N. & Weinstock, G. M. 2015. Making the Leap from Research Laboratory to Clinic: Challenges and Opportunities for Next-Generation Sequencing in Infectious Disease Diagnostics. *MBio,* 6**,** e01888-15.

Golicha, Q., Shetty, S., Nasiblov, O., Hussein, A., Wainaina, E., Obonyo, M., Macharia, D., Musyoka, R. N., Abdille, H., Ope, M., Joseph, R., Kabugi, W., Kiogora, J., Said, M., Boru, W., Galgalo, T., Lowther, S. A., Juma, B., Mugoh, R., Wamola, N., Onyango, C., Gura, Z., Widdowson, M. A., Decock, K. M. & Burton, J. W. 2018. Cholera Outbreak in Dadaab Refugee Camp, Kenya - November 2015-June 2016. *MMWR Morb Mortal Wkly Rep,* 67**,** 958-961.

Gonzalez, G., Sasaki, M., Burkitt-Gray, L., Kamiya, T., Tsuji, N. M., Sawa, H. & Ito, K. 2017. An optimistic protein assembly from sequence reads salvaged an uncharacterized segment of mouse picobirnavirus. *Sci Rep,* 7**,** 40447.

Grard, G., Fair, J. N., Lee, D., Slikas, E., Steffen, I., Muyembe, J. J., Sittler, T., Veeraraghavan, N., Ruby, J. G., Wang, C., Makuwa, M., Mulembakani, P., Tesh, R. B., Mazet, J., Rimoin, A. W., Taylor, T., Schneider, B. S., Simmons, G., Delwart, E., Wolfe, N. D., Chiu, C. Y. & Leroy, E. M. 2012. A novel rhabdovirus associated with acute hemorrhagic fever in central Africa. *PLoS Pathog,* 8**,** e1002924.

Greninger, A. L., Waghmare, A., Adler, A., Qin, X., Crowley, J. L., Englund, J. A., Kuypers, J. M., Jerome, K. R. & Zerr, D. M. 2017. Rule-Out Outbreak: 24-Hour Metagenomic Next-

Generation Sequencing for Characterizing Respiratory Virus Source for Infection Prevention. *J Pediatric Infect Dis Soc,* 6**,** 168-172.

Gso 2015. *Statistical Yearbook of Vietnam 2015,* Ha Noi, Vietnam.

Gurung, S., Harris, J. B., Eltayeb, A. O., Hampton, L. M., Diorditsa, S., Avagyan, T. & Schluter, W. W. 2017. Experience With Inactivated Polio Vaccine Introduction and the "Switch" From Trivalent to Bivalent Oral Polio Vaccine in the World Health Organization's Western Pacific Region. *J Infect Dis,* 216**,** S101-S108.

Hall, R. J., Wang, J., Todd, A. K., Bissielo, A. B., Yen, S., Strydom, H., Moore, N. E., Ren, X., Huang, Q. S., Carter, P. E. & Peacey, M. 2014. Evaluation of rapid and simple techniques for the enrichment of viruses prior to metagenomic virus discovery. *J Virol Methods,* 195**,** 194-204.

Hause, B. M., Palinski, R., Hesse, R. & Anderson, G. 2016. Highly diverse posaviruses in swine faeces are aquatic in origin. *J Gen Virol,* 97**,** 1362-7.

Hauswedell, H., Singer, J. & Reinert, K. 2014. Lambda: the local aligner for massive biological data. *Bioinformatics,* 30**,** i349-55.

Hermann, L., Embree, J., Hazelton, P., Wells, B. & Coombs, R. T. 2004. Reovirus type 2 isolated from cerebrospinal fluid. *Pediatr Infect Dis J,* 23**,** 373-5.

Hijikata, M., Hayashi, S., Trinh, N. T., Ha Le, D., Ohara, H., Shimizu, Y. K., Keicho, N. & Yoshikura, H. 2002. Genotyping of hepatitis E virus from Vietnam. *Intervirology,* 45**,** 101-4.

Hill, A. B. 1965. The Environment and Disease: Association or Causation? *Proc R Soc Med,* 58**,** 295-300.

Ho Dang Trung, N., Le Thi Phuong, T., Wolbers, M., Nguyen Van Minh, H., Nguyen Thanh, V., Van, M. P., Thieu, N. T., Van, T. L., Song, D. T., Thi, P. L., Thi Phuong, T. N., Van, C. B., Tang, V., Ngoc Anh, T. H., Nguyen, D., Trung, T. P., Thi Nam, L. N., Kiem, H. T., Thi Thanh, T. N., Campbell, J., Caws, M., Day, J., De Jong, M. D., Van Vinh, C. N., Van Doorn, H. R., Tinh, H. T., Farrar, J., Schultsz, C. & Network, V. C. I. 2012. Aetiologies of central nervous system infection in Viet Nam: a prospective provincial hospital-based descriptive surveillance study. *PLoS One,* 7**,** e37825.

Hoa-Tran, T. N., Nakagomi, T., Vu, H. M., Do, L. P., Gauchan, P., Agbemabiese, C. A., Nguyen, T. T., Nakagomi, O. & Thanh, N. T. 2016. Abrupt emergence and predominance in Vietnam of rotavirus A strains possessing a bovine-like G8 on a DS-1-like background. *Arch Virol,* 161**,** 479-82.

Hoffmann, B., Tappe, D., Hoper, D., Herden, C., Boldt, A., Mawrin, C., Niederstrasser, O., Muller, T., Jenckel, M., Van Der Grinten, E., Lutter, C., Abendroth, B., Teifke, J. P., Cadar, D., Schmidt-Chanasit, J., Ulrich, R. G. & Beer, M. 2015. A Variegated Squirrel Bornavirus Associated with Fatal Human Encephalitis. *N Engl J Med,* 373**,** 154-62.

Holmes, E. C., Rambaut, A. & Andersen, K. G. 2018. Pandemics: spend on surveillance, not prediction. *Nature,* 558**,** 180-182.

Horby, P. W., Pfeiffer, D. & Oshitani, H. 2013. Prospects for Emerging Infections in East and Southeast Asia 10 Years after Severe Acute Respiratory Syndrome. *Emerging Infectious Diseases,* 19**,** 853-860.

Hsu, H. S., Lin, T. H., Wu, H. Y., Lin, L. S., Chung, C. S., Chiou, M. T. & Lin, C. N. 2016. High detection rate of dog circovirus in diarrheal dogs. *BMC Vet Res,* 12**,** 116.

Hu, B., Chmura, A. A., Li, J., Zhu, G., Desmond, J. S., Zhang, Y., Zhang, W., Epstein, J. H., Daszak, P. & Shi, Z. 2014. Detection of diverse novel astroviruses from small mammals in China. *J Gen Virol,* 95**,** 2442-9.

Huber, C., Finelli, L. & Stevens, W. 2018. The Economic and Social Burden of the 2014 Ebola Outbreak in West Africa. *J Infect Dis,* 218**,** S698-S704.

Hue, S., Gray, E. R., Gall, A., Katzourakis, A., Tan, C. P., Houldcroft, C. J., Mclaren, S., Pillay, D., Futreal, A., Garson, J. A., Pybus, O. G., Kellam, P. & Towers, G. J. 2010. Disease-associated XMRV sequences are consistent with laboratory contamination. *Retrovirology,* 7.

Hugenholtz, P., Goebel, B. M. & Pace, N. R. 1998. Impact of culture-independent studies on the emerging phylogenetic view of bacterial diversity. *J Bacteriol,* 180**,** 4765-74.

Hulo, C., De Castro, E., Masson, P., Bougueleret, L., Bairoch, A., Xenarios, I. & Le Mercier, P. 2011. ViralZone: a knowledge resource to understand virus diversity. *Nucleic Acids Res,* 39**,** D576-82.

Human Animal Infections and Risk Surveillance (Hairs) Group 2018. Processes of risk assessment. London: Public Health England (PHE).

Huson, D. H., Auch, A. F., Qi, J. & Schuster, S. C. 2007. MEGAN analysis of metagenomic data. *Genome Res,* 17**,** 377-86.

Huynh, T. M., Nguyen, B. H., Nguyen, V. G., Dang, H. A., Mai, T. N., Tran, T. H., Ngo, M. H., Le, V. T., Vu, T. N., Ta, T. K., Vo, V. H., Kim, H. K. & Park, B. K. 2014. Phylogenetic and phylogeographic analyses of porcine circovirus type 2 among pig farms in Vietnam. *Transbound Emerg Dis,* 61**,** e25-34.

Illumina Inc. 2017. *Effects of Index Misassignment on Multiplexing and Downstream Analysis* [Online]. Available: https://www.illumina.com/content/dam/illumina-marketing/documents/products/whitepapers/index-hopping-white-paper-770-2017-004.pdf?linkId=36607862 [Accessed 22/08/2017].

Irving, L. G. & Smith, F. A. 1981. One-year survey of enteroviruses, adenoviruses, and reoviruses isolated from effluent at an activated-sludge purification plant. *Appl Environ Microbiol,* 41**,** 51-9.

Ito, T., Couceiro, J. N., Kelm, S., Baum, L. G., Krauss, S., Castrucci, M. R., Donatelli, I., Kida, H., Paulson, J. C., Webster, R. G. & Kawaoka, Y. 1998. Molecular basis for the generation in pigs of influenza A viruses with pandemic potential. *J Virol,* 72**,** 7367-73.

Iturriza-Gomara, M., Clarke, I., Desselberger, U., Brown, D., Thomas, D. & Gray, J. 2004. Seroepidemiology of group C rotavirus infection in England and Wales. *Eur J Epidemiol,* 19**,** 589-95.

Jazaeri Farsani, S. M., Oude Munnink, B. B., Deijs, M., Canuti, M. & Van Der Hoek, L. 2013. Metagenomics in virus discovery. *ISBT Science Series,* 8**,** 193-194.

Jiang, J., Hermann, L. & Coombs, K. M. 2006. Genetic characterization of a new mammalian reovirus, type 2 Winnipeg (T2W). *Virus Genes,* 33**,** 193-204.

Joffret, M. L., Bouchier, C., Grandadam, M., Zeller, H., Maufrais, C., Bourhy, H., Despres, P., Delpeyroux, F. & Dacheux, L. 2013. Genomic characterization of Sebokele virus 1 (SEBV1) reveals a new candidate species among the genus Parechovirus. *J Gen Virol,* 94**,** 1547-53.

Johansson, P. J., Sveger, T., Ahlfors, K., Ekstrand, J. & Svensson, L. 1996. Reovirus type 1 associated with meningitis. *Scand J Infect Dis,* 28**,** 117-20.

Johnson, C. K., Hitchens, P. L., Evans, T. S., Goldstein, T., Thomas, K., Clements, A., Joly, D. O., Wolfe, N. D., Daszak, P., Karesh, W. B. & Mazet, J. K. 2015. Spillover and pandemic properties of zoonotic viruses with high host plasticity. *Scientific Reports,* 5.

Johnson, K. M. 1993. Emerging viruses in context: an overview of viral hemorrhagic fevers. *In:* MORSE, S. S. (ed.) *Emerging viruses.* New York: Oxford University Press.

Jones, K. E., Patel, N. G., Levy, M. A., Storeygard, A., Balk, D., Gittleman, J. L. & Daszak, P. 2008. Global trends in emerging infectious diseases. *Nature,* 451**,** 990-3.

Kafetzopoulou, L. E., Pullan, S. T., Lemey, P., Suchard, M. A., Ehichioya, D. U., Pahlmann, M., Thielebein, A., Hinzmann, J., Oestereich, L., Wozniak, D. M., Efthymiadis, K., Schachten, D., Koenig, F., Matjeschk, J., Lorenzen, S., Lumley, S., Ighodalo, Y.,

Adomeh, D. I., Olokor, T., Omomoh, E., Omiunu, R., Agbukor, J., Ebo, B., Aiyepada, J., Ebhodaghe, P., Osiemi, B., Ehikhametalor, S., Akhilomen, P., Airende, M., Esumeh, R., Muoebonam, E., Giwa, R., Ekanem, A., Igenegbale, G., Odigie, G., Okonofua, G., Enigbe, R., Oyakhilome, J., Yerumoh, E. O., Odia, I., Aire, C., Okonofua, M., Atafo, R., Tobin, E., Asogun, D., Akpede, N., Okokhere, P. O., Rafiu, M. O., Iraoyah, K. O., Iruolagbe, C. O., Akhideno, P., Erameh, C., Akpede, G., Isibor, E., Naidoo, D., Hewson, R., Hiscox, J. A., Vipond, R., Carroll, M. W., Ihekweazu, C., Formenty, P., Okogbenin, S., Ogbaini-Emovon, E., Gunther, S. & Duraffour, S. 2019. Metagenomic sequencing at the epicenter of the Nigeria 2018 Lassa fever outbreak. *Science,* 363**,** 74-77.

Kan, B., Wang, M., Jing, H., Xu, H., Jiang, X., Yan, M., Liang, W., Zheng, H., Wan, K., Liu, Q., Cui, B., Xu, Y., Zhang, E., Wang, H., Ye, J., Li, G., Li, M., Cui, Z., Qi, X., Chen, K., Du, L., Gao, K., Zhao, Y. T., Zou, X. Z., Feng, Y. J., Gao, Y. F., Hai, R., Yu, D., Guan, Y. & Xu, J. 2005. Molecular evolution analysis and geographic investigation of severe acute respiratory syndrome coronavirus-like virus in palm civets at an animal market and on farms. *J Virol,* 79**,** 11892-900.

Kaneko, M., Do, L. P., Doan, Y. H., Nakagomi, T., Gauchan, P., Agbemabiese, C. A., Dang, A. D. & Nakagomi, O. 2018. Porcine-like G3P[6] and G4P[6] rotavirus A strains detected from children with diarrhoea in Vietnam. *Arch Virol*.

Karesh, W. B., Cook, R. A., Bennett, E. L. & Newcomb, J. 2005. Wildlife trade and global disease emergence. *Emerg Infect Dis,* 11**,** 1000-2.

Karlin, S. & Burge, C. 1995. Dinucleotide relative abundance extremes: a genomic signature. *Trends Genet,* 11**,** 283-90.

Karlin, S., Mrazek, J. & Campbell, A. M. 1997. Compositional biases of bacterial genomes and evolutionary implications. *J Bacteriol,* 179**,** 3899-913.

Kattoor, J. J., Saurabh, S., Malik, Y. S., Sircar, S., Dhama, K., Ghosh, S., Banyai, K., Kobayashi, N. & Singh, R. K. 2017. Unexpected detection of porcine rotavirus C strains carrying human origin VP6 gene. *Vet Q,* 37**,** 252-261.

Katzourakis, A., Hue, S., Kellam, P. & Towers, G. J. 2011. Phylogenetic Analysis of Murine Leukemia Virus Sequences from Longitudinally Sampled Chronic Fatigue Syndrome Patients Suggests PCR Contamination Rather than Viral Evolution. *Journal of Virology,* 85**,** 10909-10913.

Kawaoka, Y., Krauss, S. & Webster, R. G. 1989. Avian-to-human transmission of the PB1 gene of influenza A viruses in the 1957 and 1968 pandemics. *J Virol,* 63**,** 4603-8.

Kearney, M. F., Spindler, J., Wiegand, A., Shao, W., Anderson, E. M., Maldarelli, F., Ruscetti, F. W., Mellors, J. W., Hughes, S. H., Le Grice, S. F. & Coffin, J. M. 2012. Multiple sources of contamination in samples from patients reported to have XMRV infection. *PLoS One,* 7**,** e30889.

Keesing, F., Belden, L. K., Daszak, P., Dobson, A., Harvell, C. D., Holt, R. D., Hudson, P., Jolles, A., Jones, K. E., Mitchell, C. E., Myers, S. S., Bogich, T. & Ostfeld, R. S. 2010. Impacts of biodiversity on the emergence and transmission of infectious diseases. *Nature,* 468**,** 647-52.

Khamrin, P., Chaimongkol, N., Malasao, R., Suantai, B., Saikhruang, W., Kongsricharoern, T., Ukarapol, N., Okitsu, S., Shimizu, H., Hayakawa, S., Ushijima, H. & Maneekarn, N. 2012. Detection and molecular characterization of cosavirus in adults with diarrhea, Thailand. *Virus Genes,* 44**,** 244-6.

Khan, M. U. & Shahidullah, M. 1982. Role of water and sanitation in the incidence of cholera in refugee camps. *Trans R Soc Trop Med Hyg,* 76**,** 373-7.

Khanh, T. H., Sabanathan, S., Thanh, T. T., Thoa Le, P. K., Thuong, T. C., Hang, V., Farrar, J., Hien, T. T., Chau, N. & Van Doorn, H. R. 2012. Enterovirus 71-associated hand, foot, and mouth disease, Southern Vietnam, 2011. *Emerg Infect Dis,* 18**,** 2002-5.

Khiem, N. T., Cuong, L. Q. & Chien, H. V. 2003. Market study of meat from field rats in the Mekong delta. *In:* SINGLETON, G. R., HINDS, L. A., KREBS, C. J. & SPRATT , D. M. (eds.) *Rats, mice and people: rodent biology and management.* Canberra: Australian Centre for International Agricultural Research.

Kim, A. R., Chung, H. C., Kim, H. K., Kim, E. O., Nguyen, V. G., Choi, M. G., Yang, H. J., Kim, J. A. & Park, B. K. 2014. Characterization of a complete genome of a circular single-stranded DNA virus from porcine stools in Korea. *Virus Genes,* 48**,** 81-8.

King, A. M. Q., Adams, M. J., Carstens, E. B. & Lefkowitz, E. J. 2011. *Virus taxonomy. Ninth report of the international committee on taxonomy of viruses.,* London, San Diego, Elsevier Academic Press.

Kircher, M., Sawyer, S. & Meyer, M. 2012. Double indexing overcomes inaccuracies in multiplex sequencing on the Illumina platform. *Nucleic Acids Res,* 40**,** e3.

Knutson, T. P., Velayudhan, B. T. & Marthaler, D. G. 2017. A porcine enterovirus G associated with enteric disease contains a novel papain-like cysteine protease. *J Gen Virol,* 98**,** 1305-1310.

Koch, R. 1890. An Address on Bacteriological Research. *Br Med J,* 2**,** 380-3.

Kohl, C., Lesnik, R., Brinkmann, A., Ebinger, A., Radonic, A., Nitsche, A., Muhldorfer, K., Wibbelt, G. & Kurth, A. 2012. Isolation and characterization of three mammalian orthoreoviruses from European bats. *PLoS One,* 7**,** e43106.

Krishnamurthy, S. R. & Wang, D. 2018. Extensive conservation of prokaryotic ribosomal binding sites in known and novel picobirnaviruses. *Virology,* 516**,** 108-114.

Krupovic, M., Ghabrial, S. A., Jiang, D. & Varsani, A. 2016. Genomoviridae: a new family of widespread single-stranded DNA viruses. *Arch Virol,* 161**,** 2633-43.

Kuisma, E., Olson, S. H., Cameron, K. N., Reed, P. E., Karesh, W. B., Ondzie, A. I., Akongo, M. J., Kaba, S. D., Fischer, R. J., Seifert, S. N., Munoz-Fontela, C., Becker-Ziaja, B., Escudero-Perez, B., Goma-Nkoua, C., Munster, V. J. & Mombouli, J. V. 2019. Long-term wildlife mortality surveillance in northern Congo: a model for the detection of Ebola virus disease epizootics. *Philos Trans R Soc Lond B Biol Sci,* 374**,** 20180339.

Lanciotti, R. S., Roehrig, J. T., Deubel, V., Smith, J., Parker, M., Steele, K., Crise, B., Volpe, K. E., Crabtree, M. B., Scherret, J. H., Hall, R. A., Mackenzie, J. S., Cropp, C. B., Panigrahy, B., Ostlund, E., Schmitt, B., Malkinson, M., Banet, C., Weissman, J., Komar, N., Savage, H. M., Stone, W., Mcnamara, T. & Gubler, D. J. 1999. Origin of the West Nile virus responsible for an outbreak of encephalitis in the northeastern United States. *Science,* 286**,** 2333-7.

Le, V. T., Phan, T. Q., Do, Q. H., Nguyen, B. H., Lam, Q. B., Bach, V., Truong, H., Tran, T. H., Nguyen, V., Tran, T., Vo, M., Tran, V. T., Schultsz, C., Farrar, J., Van Doorn, H. R. & De Jong, M. D. 2010. Viral etiology of encephalitis in children in southern Vietnam: results of a one-year prospective descriptive study. *PLoS Negl Trop Dis,* 4**,** e854.

Lee, D., Das Gupta, J., Gaughan, C., Steffen, I., Tang, N., Luk, K. C., Qiu, X. X., Urisman, A., Fischer, N., Molinaro, R., Broz, M., Schochetman, G., Klein, E. A., Ganem, D., Derisi, J. L., Simmons, G., Hackett, J., Silverman, R. H. & Chiu, C. Y. 2012. In-Depth Investigation of Archival and Prospectively Collected Samples Reveals No Evidence for XMRV Infection in Prostate Cancer. *Plos One,* 7.

Leek, J. T., Scharpf, R. B., Bravo, H. C., Simcha, D., Langmead, B., Johnson, W. E., Geman, D., Baggerly, K. & Irizarry, R. A. 2010. Tackling the widespread and critical impact of batch effects in high-throughput data. *Nat Rev Genet,* 11**,** 733-9.

Lefkowitz, E. J., Dempsey, D. M., Hendrickson, R. C., Orton, R. J., Siddell, S. G. & Smith, D. B. 2018. Virus taxonomy: the database of the International Committee on Taxonomy of Viruses (ICTV). *Nucleic Acids Res,* 46**,** D708-D717.

Lelli, D., Beato, M. S., Cavicchio, L., Lavazza, A., Chiapponi, C., Leopardi, S., Baioni, L., De Benedictis, P. & Moreno, A. 2016. First identification of mammalian orthoreovirus type 3 in diarrheic pigs in Europe. *Virol J,* 13**,** 139.

Lelli, D., Moreno, A., Lavazza, A., Bresaola, M., Canelli, E., Boniotti, M. B. & Cordioli, P. 2013. Identification of Mammalian orthoreovirus type 3 in Italian bats. *Zoonoses Public Health,* 60**,** 84-92.

Leroy, E. M., Rouquet, P., Formenty, P., Souquiere, S., Kilbourne, A., Froment, J. M., Bermejo, M., Smit, S., Karesh, W., Swanepoel, R., Zaki, S. R. & Rollin, P. E. 2004. Multiple Ebola virus transmission events and rapid decline of central African wildlife. *Science,* 303**,** 387-90.

Levinson, J., Bogich, T. L., Olival, K. J., Epstein, J. H., Johnson, C. K., Karesh, W. & Daszak, P. 2013. Targeting surveillance for zoonotic virus discovery. *Emerg Infect Dis,* 19**,** 743-7.

Lewandowska, D. W., Capaul, R., Prader, S., Zagordi, O., Geissberger, F. D., Kugler, M., Knorr, M., Berger, C., Gungor, T., Reichenbach, J., Shah, C., Boni, J., Zbinden, A., Trkola, A., Pachlopnik Schmid, J. & Huber, M. 2018. Persistent mammalian orthoreovirus, coxsackievirus and adenovirus co-infection in a child with a primary immunodeficiency detected by metagenomic sequencing: a case report. *BMC Infect Dis,* 18**,** 33.

Li, C. X., Shi, M., Tian, J. H., Lin, X. D., Kang, Y. J., Chen, L. J., Qin, X. C., Xu, J., Holmes, E. C. & Zhang, Y. Z. 2015a. Unprecedented genomic diversity of RNA viruses in arthropods reveals the ancestry of negative-sense RNA viruses. *Elife,* 4.

Li, H. & Durbin, R. 2009. Fast and accurate short read alignment with Burrows-Wheeler transform. *Bioinformatics,* 25**,** 1754-60.

Li, K., Lin, X. D., Huang, K. Y., Zhang, B., Shi, M., Guo, W. P., Wang, M. R., Wang, W., Xing, J. G., Li, M. H., Hong, W. S., Holmes, E. C. & Zhang, Y. Z. 2016a. Identification of novel and diverse rotaviruses in rodents and insectivores, and evidence of cross-species transmission into humans. *Virology,* 494**,** 168-77.

Li, L. & Delwart, E. 2011. From orphan virus to pathogen: the path to the clinical lab. *Curr Opin Virol,* 1**,** 282-8.

Li, L., Giannitti, F., Low, J., Keyes, C., Ullmann, L. S., Deng, X., Aleman, M., Pesavento, P. A., Pusterla, N. & Delwart, E. 2015b. Exploring the virome of diseased horses. *J Gen Virol,* 96**,** 2721-33.

Li, L., Kapoor, A., Slikas, B., Bamidele, O. S., Wang, C., Shaukat, S., Masroor, M. A., Wilson, M. L., Ndjango, J. B., Peeters, M., Gross-Camp, N. D., Muller, M. N., Hahn, B. H., Wolfe, N. D., Triki, H., Bartkus, J., Zaidi, S. Z. & Delwart, E. 2010. Multiple diverse circoviruses infect farm animals and are commonly found in human and chimpanzee feces. *J Virol,* 84**,** 1674-82.

Li, L., Mcgraw, S., Zhu, K., Leutenegger, C. M., Marks, S. L., Kubiski, S., Gaffney, P., Dela Cruz, F. N., Jr., Wang, C., Delwart, E. & Pesavento, P. A. 2013a. Circovirus in tissues of dogs with vasculitis and hemorrhage. *Emerg Infect Dis,* 19**,** 534-41.

Li, T. C., Ami, Y., Suzaki, Y., Yasuda, S. P., Yoshimatsu, K., Arikawa, J., Takeda, N. & Takaji, W. 2013b. Characterization of full genome of rat hepatitis E virus strain from Vietnam. *Emerg Infect Dis,* 19**,** 115-8.

Li, Z., Liu, D., Ran, X., Liu, C., Guo, D., Hu, X., Tian, J., Zhang, X., Shao, Y., Liu, S. & Qu, L. 2016b. Characterization and pathogenicity of a novel mammalian orthoreovirus from wild short-nosed fruit bats. *Infect Genet Evol,* 43**,** 347-53.

Lipkin, W. I. 2010. Microbe hunting. *Microbiol Mol Biol Rev,* 74**,** 363-77.

Lipton, H. L. 2008. Human Vilyuisk encephalitis. *Rev Med Virol,* 18**,** 347-52.

Lodder, W. J. & De Roda Husman, A. M. 2005. Presence of noroviruses and other enteric viruses in sewage and surface waters in The Netherlands. *Appl Environ Microbiol,* 71**,** 1453-61.

Lodder, W. J., Van Den Berg, H. H., Rutjes, S. A. & De Roda Husman, A. M. 2010. Presence of enteric viruses in source waters for drinking water production in The Netherlands. *Appl Environ Microbiol,* 76**,** 5965-71.

Logan, C., O'leary, J. J. & O'sullivan, N. 2007. Real-time reverse transcription PCR detection of norovirus, sapovirus and astrovirus as causative agents of acute viral gastroenteritis. *J Virol Methods,* 146**,** 36-44.

Loh, E. H., Zambrana-Torrelio, C., Olival, K. J., Bogich, T. L., Johnson, C. K., Mazet, J. A., Karesh, W. & Daszak, P. 2015. Targeting Transmission Pathways for Emerging Zoonotic Disease Surveillance and Control. *Vector Borne Zoonotic Dis,* 15**,** 432-7.

Lu, L., Brierley, L., Robertson, G., Zhang, F., Lycett, S., Smith, D., Chase-Topping, M., Simmonds, P. & Woolhouse, M. 2019. Evolutionary origins of epidemic potential among human RNA viruses. *bioRxiv*.

Lu, L., Li, C. & Hagedorn, C. H. 2006. Phylogenetic analysis of global hepatitis E virus sequences: genetic diversity, subtypes and zoonosis. *Rev Med Virol,* 16**,** 5-36.

Luis, A. D., Hayman, D. T., O'shea, T. J., Cryan, P. M., Gilbert, A. T., Pulliam, J. R., Mills, J. N., Timonin, M. E., Willis, C. K., Cunningham, A. A., Fooks, A. R., Rupprecht, C. E., Wood, J. L. & Webb, C. T. 2013. A comparison of bats and rodents as reservoirs of zoonotic viruses: are bats special? *Proc Biol Sci,* 280**,** 20122753.

Luo, X. L., Lu, S., Jin, D., Yang, J., Wu, S. S. & Xu, J. 2018. Marmota himalayana in the Qinghai-Tibetan plateau as a special host for bi-segmented and unsegmented picobirnaviruses. *Emerg Microbes Infect,* 7**,** 20.

Ma, W., Kahn, R. E. & Richt, J. A. 2008. The pig as a mixing vessel for influenza viruses: Human and veterinary implications. *J Mol Genet Med,* 3**,** 158-66.

Macera, L., Focosi, D., Vatteroni, M. L., Manzin, A., Antonelli, G., Pistello, M. & Maggi, F. 2016. Cyclovirus Vietnam DNA in immunodeficient patients. *J Clin Virol,* 81**,** 12-5.

Mackenzie, J. S. 2005. Emerging zoonotic encephalitis viruses: lessons from Southeast Asia and Oceania. *J Neurovirol,* 11**,** 434-40.

Maclachlan, N. J., Dubovi, E. J. & Fenner, F. 2017. *Fenner's Veterinary Virology, 5th Edition,* London, Elsevier.

Malik, Y. S., Kumar, N., Sharma, K., Dhama, K., Shabbir, M. Z., Ganesh, B., Kobayashi, N. & Banyai, K. 2014. Epidemiology, phylogeny, and evolution of emerging enteric Picobirnaviruses of animal origin and their relationship to human strains. *Biomed Res Int,* 2014**,** 780752.

Martella, V., Banyai, K., Matthijnssens, J., Buonavoglia, C. & Ciarlet, M. 2010. Zoonotic aspects of rotaviruses. *Vet Microbiol,* 140**,** 246-55.

Masuda, T., Sunaga, F., Naoi, Y., Ito, M., Takagi, H., Katayama, Y., Omatsu, T., Oba, M., Sakaguchi, S., Furuya, T., Yamasato, H., Shirai, J., Makino, S., Mizutani, T. & Nagai, M. 2018. Whole genome analysis of a novel picornavirus related to the Enterovirus/Sapelovirus supergroup from porcine feces in Japan. *Virus Res,* 257**,** 68-73.

Mchardy, A. C., Martin, H. G., Tsirigos, A., Hugenholtz, P. & Rigoutsos, I. 2007. Accurate phylogenetic classification of variable-length DNA fragments. *Nat Methods,* 4**,** 63-72.

Meerburg, B. G. 2010. Rodents are a risk factor for the spreading of pathogens on farms. *Vet Microbiol,* 142**,** 464-5; author reply 466.

Meerburg, B. G., Singleton, G. R. & Kijlstra, A. 2009. Rodent-borne diseases and their risks for public health. *Crit Rev Microbiol,* 35**,** 221-70.

Memish, Z. A., Cotten, M., Meyer, B., Watson, S. J., Alsahafi, A. J., Al Rabeeah, A. A., Corman, V. M., Sieberg, A., Makhdoom, H. Q., Assiri, A., Al Masri, M., Aldabbagh, S., Bosch, B. J., Beer, M., Muller, M. A., Kellam, P. & Drosten, C. 2014. Human infection with MERS coronavirus after exposure to infected camels, Saudi Arabia, 2013. *Emerg Infect Dis,* 20**,** 1012-5.

Menzel, P., Ng, K. L. & Krogh, A. 2016. Fast and sensitive taxonomic classification for metagenomics with Kaiju. *Nat Commun,* 7**,** 11257.

Meyer, F., Paarmann, D., D'souza, M., Olson, R., Glass, E. M., Kubal, M., Paczian, T., Rodriguez, A., Stevens, R., Wilke, A., Wilkening, J. & Edwards, R. A. 2008. The metagenomics RAST server - a public resource for the automatic phylogenetic and functional analysis of metagenomes. *BMC Bioinformatics,* 9**,** 386.

Mijatovic-Rustempasic, S., Roy, S., Teel, E. N., Weinberg, G. A., Payne, D. C., Parashar, U. D. & Bowen, M. D. 2016. Full genome characterization of the first G3P[24] rotavirus strain detected in humans provides evidence of interspecies reassortment and mutational saturation in the VP7 gene. *J Gen Virol,* 97**,** 389-402.

Minot, S. S., Krumm, N. & Greenfield, N. B. 2015. One Codex: A Sensitive and Accurate Data Platform for Genomic Microbial Identification. *bioRxiv*.

Mitra, A., Skrzypczak, M., Ginalski, K. & Rowicka, M. 2015. Strategies for achieving high sequencing accuracy for low diversity samples and avoiding sample bleeding using illumina platform. *PLoS One,* 10**,** e0120520.

Mokili, J. L., Dutilh, B. E., Lim, Y. W., Schneider, B. S., Taylor, T., Haynes, M. R., Metzgar, D., Myers, C. A., Blair, P. J., Nosrat, B., Wolfe, N. D. & Rohwer, F. 2013. Identification of a novel human papillomavirus by metagenomic analysis of samples from patients with febrile respiratory illness. *PLoS One,* 8**,** e58404.

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

Morand, S., Mcintyre, K. M. & Baylis, M. 2014. Domesticated animals and human infectious diseases of zoonotic origins: domestication time matters. *Infect Genet Evol,* 24**,** 76-81.

Morgan, D., Kirkbride, H., Hewitt, K., Said, B. & Walsh, A. L. 2009. Assessing the risk from emerging infections. *Epidemiol Infect,* 137**,** 1521-30.

Morse, S. S. 1993. Examining the origins of emerging viruses. *In:* MORSE, S. S. (ed.) *Emerging viruses.* New York: Oxford University Press.

Morse, S. S. 1995. Factors in the emergence of infectious diseases. *Emerg Infect Dis,* 1**,** 7-15.

Morse, S. S., Mazet, J. a. K., Woolhouse, M., Parrish, C. R., Carroll, D., Karesh, W. B., Zambrana-Torrelio, C., Lipkin, W. I. & Daszak, P. 2012. Prediction and prevention of the next pandemic zoonosis. *The Lancet,* 380**,** 1956-1965.

Mulyanto, Suparyatmo, J. B., Andayani, I. G., Khalid, Takahashi, M., Ohnishi, H., Jirintai, S., Nagashima, S., Nishizawa, T. & Okamoto, H. 2014. Marked genomic heterogeneity of rat hepatitis E virus strains in Indonesia demonstrated on a full-length genome analysis. *Virus Res,* 179**,** 102-12.

My, P. V., Rabaa, M. A., Donato, C., Cowley, D., Phat, V. V., Dung, T. T., Anh, P. H., Vinh, H., Bryant, J. E., Kellam, P., Thwaites, G., Woolhouse, M. E., Kirkwood, C. D. & Baker, S. 2014a. Novel porcine-like human G26P[19] rotavirus identified in hospitalized paediatric diarrhoea patients in Ho Chi Minh City, Vietnam. *J Gen Virol,* 95**,** 2727-33.

My, P. V. T., Rabaa, M. A., Donato, C., Cowley, D., Phat, V. V., Dung, T. T. N., Anh, P. H., Vinh, H., Bryant, J. E., Kellam, P., Thwaites, G., Woolhouse, M. E. J., Kirkwood, C. D. & Baker, S. 2014b. Novel porcine-like human G26P[19] rotavirus identified in hospitalized paediatric diarrhoea patients in Ho Chi Minh City, Vietnam. *The Journal of General Virology,* 95**,** 2727-2733.

Myers, K. P., Olsen, C. W., Setterquist, S. F., Capuano, A. W., Donham, K. J., Thacker, E. L., Merchant, J. A. & Gray, G. C. 2006. Are swine workers in the United States at increased risk of infection with zoonotic influenza virus? *Clin Infect Dis,* 42**,** 14-20.

Mysterud, A., Easterday, W. R., Stigum, V. M., Aas, A. B., Meisingset, E. L. & Viljugrein, H. 2016. Contrasting emergence of Lyme disease across ecosystems. *Nat Commun,* 7**,** 11882.

Naccache, S. N., Greninger, A. L., Lee, D., Coffey, L. L., Phan, T., Rein-Weston, A., Aronsohn, A., Hackett, J., Jr., Delwart, E. L. & Chiu, C. Y. 2013. The perils of pathogen discovery: origin of a novel parvovirus-like hybrid genome traced to nucleic acid extraction spin columns. *J Virol,* 87**,** 11966-77.

Naccache, S. N., Hackett, J., Jr., Delwart, E. L. & Chiu, C. Y. 2014. Concerns over the origin of NIH-CQV, a novel virus discovered in Chinese patients with seronegative hepatitis. *Proc Natl Acad Sci U S A,* 111**,** E976.

Naccache, S. N., Peggs, K. S., Mattes, F. M., Phadke, R., Garson, J. A., Grant, P., Samayoa, E., Federman, S., Miller, S., Lunn, M. P., Gant, V. & Chiu, C. Y. 2015. Diagnosis of neuroinvasive astrovirus infection in an immunocompromised adult with encephalitis by unbiased next-generation sequencing. *Clin Infect Dis,* 60**,** 919-23.

Naoi, Y., Kishimoto, M., Masuda, T., Ito, M., Tsuchiaka, S., Sano, K., Yamasato, H., Omatsu, T., Aoki, H., Furuya, T., Katayama, Y., Oba, M., Okada, T., Shirai, J., Mizutani, T. & Nagai, M. 2016. Characterization and phylogenetic analysis of a novel picornavirus from swine feces in Japan. *Arch Virol,* 161**,** 1685-90.

Nelson, M. C., Morrison, H. G., Benjamino, J., Grim, S. L. & Graf, J. 2014. Analysis, optimization and verification of Illumina-generated 16S rRNA gene amplicon surveys. *PLoS One,* 9**,** e94249.

Nelson, M. I. & Worobey, M. 2018. Origins of the 1918 Pandemic: Revisiting the Swine "Mixing Vessel" Hypothesis. *Am J Epidemiol,* 187**,** 2498-2502.

Ng, T. F., Chen, L. F., Zhou, Y., Shapiro, B., Stiller, M., Heintzman, P. D., Varsani, A., Kondov, N. O., Wong, W., Deng, X., Andrews, T. D., Moorman, B. J., Meulendyk, T., Mackay, G., Gilbertson, R. L. & Delwart, E. 2014. Preservation of viral genomes in 700-y-old caribou feces from a subarctic ice patch. *Proc Natl Acad Sci U S A,* 111**,** 16842-7.

Ng, T. F., Zhang, W., Sachsenroder, J., Kondov, N. O., Da Costa, A. C., Vega, E., Holtz, L. R., Wu, G., Wang, D., Stine, C. O., Antonio, M., Mulvaney, U. S., Muench, M. O., Deng, X., Ambert-Balay, K., Pothier, P., Vinje, J. & Delwart, E. 2015. A diverse group of small circular ssDNA viral genomes in human and non-human primate stools. *Virus Evol,* 1**,** vev017.

Nghia, H. D. T., Tu, L. T. P., Wolbers, M., Thai, C. Q., Hoang, N. V. M., Nga, T. V. T., Thao, L. T. P., Phu, N. H., Chau, T. T. H., Sinh, D. X., Diep, T. S., Hang, H. T. T., Truong, H., Campbell, J., Chau, N. V. V., Chinh, N. T., Dung, N. V., Hoa, N. T., Spratt, B. G., Hien, T. T., Farrar, J. & Schultsz, C. 2011. Risk factors of Streptococcus suis infection in Vietnam. A case-control study. *PLoS One,* 6**,** e17604.

Nguyen, D. V. 2015. *The contribution of newly discovered and emerging viruses to human disease.* PhD, University of Edinburgh.

Nguyen, T. A., Khamrin, P., Trinh, Q. D., Phan, T. G., Pham Le, D., Hoang Le, P., Hoang, K. T., Yagyu, F., Okitsu, S. & Ushijima, H. 2007a. Sequence analysis of Vietnamese P[6] rotavirus strains suggests evidence of interspecies transmission. *J Med Virol,* 79**,** 1959-65.

Nguyen, T. A., Yagyu, F., Okame, M., Phan, T. G., Trinh, Q. D., Yan, H., Hoang, K. T., Cao, A. T., Le Hoang, P., Okitsu, S. & Ushijima, H. 2007b. Diversity of viruses associated with acute gastroenteritis in children hospitalized with diarrhea in Ho Chi Minh City, Vietnam. *J Med Virol,* 79**,** 582-90.

Nooij, S., Schmitz, D., Vennema, H., Kroneman, A. & Koopmans, M. P. G. 2018. Overview of Virus Metagenomic Classification Methods and Their Biological Applications. *Frontiers in Microbiology,* 9.

Norling, M., Karlsson-Lindsjo, O. E., Gourle, H., Bongcam-Rudloff, E. & Hayer, J. 2016. MetLab: An In Silico Experimental Design, Simulation and Analysis Tool for Viral Metagenomics Studies. *PLoS One,* 11**,** e0160334.

Obana, S., Shimizu, K., Yoshimatsu, K., Hasebe, F., Hotta, K., Isozumi, R., Nguyen, H. T., Le, M. Q., Yamashiro, T., Tsuda, Y. & Arikawa, J. 2017. Epizootiological study of rodent-borne hepatitis E virus HEV-C1 in small mammals in Hanoi, Vietnam. *J Vet Med Sci,* 79**,** 76-81.

Obuchowski, N. A. & Bullen, J. A. 2018. Receiver operating characteristic (ROC) curves: review of methods with applications in diagnostic medicine. *Phys Med Biol,* 63**,** 07TR01.

Oka, T., Katayama, K., Hansman, G. S., Kageyama, T., Ogawa, S., Wu, F. T., White, P. A. & Takeda, N. 2006. Detection of human sapovirus by real-time reverse transcription-polymerase chain reaction. *J Med Virol,* 78**,** 1347-53.

Okamoto, H. 2009. History of discoveries and pathogenicity of TT viruses. *Curr Top Microbiol Immunol,* 331**,** 1-20.

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,* advance online publication.

Osburn, B. I. 1996. Emerging diseases with a worldwide impact and the consequences for veterinary curricula. *Vet Q,* 18 Suppl 3**,** S124-6.

Ouattara, L. A., Barin, F., Barthez, M. A., Bonnaud, B., Roingeard, P., Goudeau, A., Castelnau, P., Vernet, G., Paranhos-Baccala, G. & Komurian-Pradel, F. 2011. Novel human reovirus isolated from children with acute necrotizing encephalopathy. *Emerg Infect Dis,* 17**,** 1436-44.

Oude Munnink, B. B., Jazaeri Farsani, S. M., Deijs, M., Jonkers, J., Verhoeven, J. T., Ieven, M., Goossens, H., De Jong, M. D., Berkhout, B., Loens, K., Kellam, P., Bakker, M., Canuti, M., Cotten, M. & Van Der Hoek, L. 2013. Autologous antibody capture to enrich immunogenic viruses for viral discovery. *PLoS One,* 8**,** e78454.

Oude Munnink, B. B. & Van Der Hoek, L. 2016. Viruses Causing Gastroenteritis: The Known, The New and Those Beyond. *Viruses,* 8.

Ounit, R., Wanamaker, S., Close, T. J. & Lonardi, S. 2015. CLARK: fast and accurate classification of metagenomic and genomic sequences using discriminative k-mers. *BMC Genomics,* 16**,** 236.

Palmer, S., Brown, D. & Morgan, D. 2005. Early qualitative risk assessment of the emerging zoonotic potential of animal diseases. *BMJ,* 331**,** 1256-60.

Paprotka, T., Delviks-Frankenberry, K. A., Cingoz, O., Martinez, A., Kung, H. J., Tepper, C. G., Hu, W. S., Fivash, M. J., Jr., Coffin, J. M. & Pathak, V. K. 2011. Recombinant origin of the retrovirus XMRV. *Science,* 333**,** 97-101.

Parras-Molto, M., Rodriguez-Galet, A., Suarez-Rodriguez, P. & Lopez-Bueno, A. 2018. Evaluation of bias induced by viral enrichment and random amplification protocols in metagenomic surveys of saliva DNA viruses. *Microbiome,* 6**,** 119.

Penaranda, M. E., Cubitt, W. D., Sinarachatanant, P., Taylor, D. N., Likanonsakul, S., Saif, L. & Glass, R. I. 1989. Group C rotavirus infections in patients with diarrhea in Thailand, Nepal, and England. *J Infect Dis,* 160**,** 392-7.

Pereira, H. G., Fialho, A. M., Flewett, T. H., Teixeira, J. M. & Andrade, Z. P. 1988. Novel viruses in human faeces. *Lancet,* 2**,** 103-4.

Pham, H. A., Carrique-Mas, J. J., Nguyen, V. C., Ngo, T. H., Nguyet, L. A., Do, T. D., Vo, B. H., Phan, V. T., Rabaa, M. A., Farrar, J., Baker, S. & Bryant, J. E. 2014a. The prevalence

and genetic diversity of group A rotaviruses on pig farms in the Mekong Delta region of Vietnam. *Vet Microbiol,* 170**,** 258-65.

Pham, T. H., Nguyen, T. H., Herrero, R., Vaccarella, S., Smith, J. S., Nguyen Thuy, T. T., Nguyen, H. N., Nguyen, B. D., Ashley, R., Snijders, P. J., Meijer, C. J., Munoz, N., Parkin, D. M. & Franceschi, S. 2003. Human papillomavirus infection among women in South and North Vietnam. *Int J Cancer,* 104**,** 213-20.

Pham, V. H., Nguyet, D. P., Mai, K. N., Truong, K. H., Huynh, L. V., Pham, T. H. & Abe, K. 2014b. Measles Epidemics Among Children in Vietnam: Genomic Characterization of Virus Responsible for Measles Outbreak in Ho Chi Minh City, 2014. *EBioMedicine,* 1**,** 133-40.

Phan, M. V. T., Anh, P. H., Cuong, N. V., Munnink, B. B. O., Van Der Hoek, L., My, P. T., Tri, T. N., Bryant, J. E., Baker, S., Thwaites, G., Woolhouse, M., Kellam, P., Rabaa, M. A. & Cotten, M. 2016a. Unbiased whole-genome deep sequencing of human and porcine stool samples reveals circulation of multiple groups of rotaviruses and a putative zoonotic infection. *Virus Evolution,* 2**,** vew027-vew027.

Phan, T. G., Da Costa, A. C., Del Valle Mendoza, J., Bucardo-Rivera, F., Nordgren, J., O'ryan, M., Deng, X. & Delwart, E. 2016b. The fecal virome of South and Central American children with diarrhea includes small circular DNA viral genomes of unknown origin. *Arch Virol,* 161**,** 959-66.

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 Pathog,* 7**,** e1002218.

Phan, T. G., Mori, D., Deng, X., Rajindrajith, S., Ranawaka, U., Fan Ng, T. F., Bucardo-Rivera, F., Orlandi, P., Ahmed, K. & Delwart, E. 2015. Small circular single stranded DNA viral genomes in unexplained cases of human encephalitis, diarrhea, and in untreated sewage. *Virology,* 482**,** 98-104.

Phan, T. G., Nguyen, T. A., Shimizu, H., Yagyu, F., Okitsu, S., Muller, W. E. & Ushijima, H. 2005. Identification of enteroviral infection among infants and children admitted to hospital with acute gastroentritis in Ho Chi Minh City, Vietnam. *J Med Virol,* 77**,** 257-64.

Pickett, B. E., Sadat, E. L., Zhang, Y., Noronha, J. M., Squires, R. B., Hunt, V., Liu, M., Kumar, S., Zaremba, S., Gu, Z., Zhou, L., Larson, C. N., Dietrich, J., Klem, E. B. & Scheuermann, R. H. 2012. ViPR: an open bioinformatics database and analysis resource for virology research. *Nucleic Acids Res,* 40**,** D593-8.

Piewbang, C., Jo, W. K., Puff, C., Van Der Vries, E., Kesdangsakonwut, S., Rungsipipat, A., Kruppa, J., Jung, K., Baumgartner, W., Techangamsuwan, S., Ludlow, M. & Osterhaus, A. 2018. Novel canine circovirus strains from Thailand: Evidence for genetic recombination. *Sci Rep,* 8**,** 7524.

Plowright, R. K., Parrish, C. R., Mccallum, H., Hudson, P. J., Ko, A. I., Graham, A. L. & Lloyd-Smith, J. O. 2017. Pathways to zoonotic spillover. *Nat Rev Microbiol,* 15**,** 502-510.

Pulliam, J. R. 2008. Viral host jumps: moving toward a predictive framework. *Ecohealth,* 5**,** 80-91.

Pulliam, J. R. & Dushoff, J. 2009. Ability to replicate in the cytoplasm predicts zoonotic transmission of livestock viruses. *J Infect Dis,* 199**,** 565-8.

Purdy, M. A., Harrison, T. J., Jameel, S., Meng, X. J., Okamoto, H., Van Der Poel, W. H. M., Smith, D. B. & Ictv Report, C. 2017. ICTV Virus Taxonomy Profile: Hepeviridae. *J Gen Virol,* 98**,** 2645-2646.

Quan, P. L., Wagner, T. A., Briese, T., Torgerson, T. R., Hornig, M., Tashmukhamedova, A., Firth, C., Palacios, G., Baisre-De-Leon, A., Paddock, C. D., Hutchison, S. K., Egholm, M., Zaki, S. R., Goldman, J. E., Ochs, H. D. & Lipkin, W. I. 2010. Astrovirus encephalitis in boy with X-linked agammaglobulinemia. *Emerg Infect Dis,* 16**,** 918-25.

Rabaa, M. A., Tue, N. T., Phuc, T. M., Carrique-Mas, J., Saylors, K., Cotten, M., Bryant, J. E., Nghia, H. D., Cuong, N. V., Pham, H. A., Berto, A., Phat, V. V., Dung, T. T., Bao, L. H., Hoa, N. T., Wertheim, H., Nadjm, B., Monagin, C., Van Doorn, H. R., Rahman, M., Tra, M. P., Campbell, J. I., Boni, M. F., Tam, P. T., Van Der Hoek, L., Simmonds, P., Rambaut, A., Toan, T. K., Van Vinh Chau, N., Hien, T. T., Wolfe, N., Farrar, J. J., Thwaites, G., Kellam, P., Woolhouse, M. E. & Baker, S. 2015. The Vietnam Initiative on Zoonotic Infections (VIZIONS): A Strategic Approach to Studying Emerging Zoonotic Infectious Diseases. *Ecohealth,* 12**,** 726-35.

Rasool, N. B., Hamzah, M., Jegathesan, M., Wong, Y. H., Qian, Y. & Green, K. Y. 1994. Identification of a human group C rotavirus in Malaysia. *J Med Virol,* 43**,** 209-11.

Ren, J., Ahlgren, N. A., Lu, Y. Y., Fuhrman, J. A. & Sun, F. 2017. VirFinder: a novel k-mer based tool for identifying viral sequences from assembled metagenomic data. *Microbiome,* 5**,** 69.

Renou, C., Cadranel, J. F., Bourliere, M., Halfon, P., Ouzan, D., Rifflet, H., Carenco, P., Harafa, A., Bertrand, J. J., Boutrouille, A., Muller, P., Igual, J. P., Decoppet, A., Eloit, M. & Pavio, N. 2007. Possible zoonotic transmission of hepatitis E from pet pig to its owner. *Emerg Infect Dis,* 13**,** 1094-6.

Reuter, G., Boros, A., Kiss, T., Delwart, E. & Pankovics, P. 2014. Complete genome characterization of mosavirus (family Picornaviridae) identified in droppings of a European roller (Coracias garrulus) in Hungary. *Arch Virol,* 159**,** 2723-9.

Reynolds, M. G., Anh, B. H., Thu, V. H., Montgomery, J. M., Bausch, D. G., Shah, J. J., Maloney, S., Leitmeyer, K. C., Huy, V. Q., Horby, P., Plant, A. J. & Uyeki, T. M. 2006. Factors associated with nosocomial SARS-CoV transmission among healthcare workers in Hanoi, Vietnam, 2003. *BMC Public Health,* 6**,** 207.

Rivers, T. M. 1937. Viruses and Koch's Postulates. *J Bacteriol,* 33**,** 1-12.

Robert Koch Institute (Rki). 2017. *Neues Bornavirus bei Bunt- und Schönhörnchen entdeckt – wahrscheinlicher Zusammenhang mit Infektionen bei Menschen [New bornavirus discovered in variegated squirrels and Callosciurinae – probable connection with infections in humans]* [Online]. Berlin: RKI. Available: https://www.rki.de/DE/Content/InfAZ/Z/Zoonosen/Bornavirus_Bunthoernchen.html [Accessed 21/06/2019].

Robin, X., Turck, N., Hainard, A., Tiberti, N., Lisacek, F., Sanchez, J. C. & Muller, M. 2011. pROC: an open-source package for R and S+ to analyze and compare ROC curves. *BMC Bioinformatics,* 12**,** 77.

Rognes, T., Flouri, T., Nichols, B., Quince, C. & Mahe, F. 2016. VSEARCH: a versatile open source tool for metagenomics. *PeerJ,* 4**,** e2584.

Rosario, K., Breitbart, M., Harrach, B., Segales, J., Delwart, E., Biagini, P. & Varsani, A. 2017. Revisiting the taxonomy of the family Circoviridae: establishment of the genus Cyclovirus and removal of the genus Gyrovirus. *Arch Virol,* 162**,** 1447-1463.

Rose, R., Constantinides, B., Tapinos, A., Robertson, D. L. & Prosperi, M. 2016. Challenges in the analysis of viral metagenomes. *Virus Evolution,* 2**,** vew022-vew022.

Rosen, G., Garbarine, E., Caseiro, D., Polikar, R. & Sokhansanj, B. 2008. Metagenome fragment classification using N-mer frequency profiles. *Adv Bioinformatics,* 2008**,** 205969.

Rosen, G. L. & Lim, T. Y. 2012. NBC update: The addition of viral and fungal databases to the Naive Bayes classification tool. *BMC Res Notes,* 5**,** 81.

Rosen, G. L., Reichenberger, E. R. & Rosenfeld, A. M. 2011. NBC: the Naive Bayes Classification tool webserver for taxonomic classification of metagenomic reads. *Bioinformatics,* 27**,** 127-9.

Ross, M. G., Russ, C., Costello, M., Hollinger, A., Lennon, N. J., Hegarty, R., Nusbaum, C. & Jaffe, D. B. 2013. Characterizing and measuring bias in sequence data. *Genome Biol,* 14**,** R51.

Rosseel, T., Pardon, B., De Clercq, K., Ozhelvaci, O. & Van Borm, S. 2014. False-positive results in metagenomic virus discovery: a strong case for follow-up diagnosis. *Transbound Emerg Dis,* 61**,** 293-9.

Rouquet, P., Froment, J. M., Bermejo, M., Kilbourn, A., Karesh, W., Reed, P., Kumulungui, B., Yaba, P., Delicat, A., Rollin, P. E. & Leroy, E. M. 2005. Wild animal mortality monitoring and human Ebola outbreaks, Gabon and Republic of Congo, 2001-2003. *Emerg Infect Dis,* 11**,** 283-90.

Roux, S., Faubladier, M., Mahul, A., Paulhe, N., Bernard, A., Debroas, D. & Enault, F. 2011. Metavir: a web server dedicated to virome analysis. *Bioinformatics,* 27**,** 3074-5.

Roux, S., Tournayre, J., Mahul, A., Debroas, D. & Enault, F. 2014. Metavir 2: new tools for viral metagenome comparison and assembled virome analysis. *Bmc Bioinformatics,* 15.

Sachsenroder, J., Braun, A., Machnowska, P., Ng, T. F., Deng, X., Guenther, S., Bernstein, S., Ulrich, R. G., Delwart, E. & Johne, R. 2014. Metagenomic identification of novel enteric viruses in urban wild rats and genome characterization of a group A rotavirus. *J Gen Virol,* 95**,** 2734-47.

Sachsenroder, J., Twardziok, S., Hammerl, J. A., Janczyk, P., Wrede, P., Hertwig, S. & Johne, R. 2012. Simultaneous identification of DNA and RNA viruses present in pig faeces using process-controlled deep sequencing. *PLoS One,* 7**,** e34631.

Saegerman, C., Alba-Casals, A., Garcia-Bocanegra, I., Dal Pozzo, F. & Van Galen, G. 2016. Clinical Sentinel Surveillance of Equine West Nile Fever, Spain. *Transbound Emerg Dis,* 63**,** 184-93.

Salter, S. J., Cox, M. J., Turek, E. M., Calus, S. T., Cookson, W. O., Moffatt, M. F., Turner, P., Parkhill, J., Loman, N. J. & Walker, A. W. 2014. Reagent and laboratory contamination can critically impact sequence-based microbiome analyses. *BMC Biol,* 12**,** 87.

Sano, K., Naoi, Y., Kishimoto, M., Masuda, T., Tanabe, H., Ito, M., Niira, K., Haga, K., Asano, K., Tsuchiaka, S., Omatsu, T., Furuya, T., Katayama, Y., Oba, M., Ouchi, Y., Yamasato, H., Ishida, M., Shirai, J., Katayama, K., Mizutani, T. & Nagai, M. 2016. Identification of further diversity among posaviruses. *Arch Virol,* 161**,** 3541-3548.

Sasaki, M., Orba, Y., Ueno, K., Ishii, A., Moonga, L., Hang'ombe, B. M., Mweene, A. S., Ito, K. & Sawa, H. 2015. Metagenomic analysis of the shrew enteric virome reveals novel viruses related to human stool-associated viruses. *J Gen Virol,* 96**,** 440-52.

Sauvage, V., Gomez, J., Barray, A., Vandenbogaert, M., Boizeau, L., Tagny, C. T., Rakoto, O., Bizimana, P., Guitteye, H., Cire, B. B., Soumana, H., Tchomba, J. S., Caro, V. & Laperche, S. 2018. High prevalence of cyclovirus Vietnam (CyCV-VN) in plasma samples from Madagascan healthy blood donors. *Infect Genet Evol,* 66**,** 9-12.

Saylors, K., Ngo Tri, T., Tran Khanh, T., Bach Tuan, K., Wertheim, H. F., Baker, S., Ngo Thi, H. & Bryant, J. E. 2015. Mobilising community-based research on zoonotic infections: A case study of longitudinal cohorts in Vietnam. *Gateways: International Journal of Community Research and Engagement,* 8**,** 20.

Scheidegger, J. 2013. Michigan veterinarian believes canine mystery illness is zoonotic. *DVM360 Magazine.*

Schlauder, G. G., Dawson, G. J., Erker, J. C., Kwo, P. Y., Knigge, M. F., Smalley, D. L., Rosenblatt, J. E., Desai, S. M. & Mushahwar, I. K. 1998. The sequence and phylogenetic analysis of a novel hepatitis E virus isolated from a patient with acute hepatitis reported in the United States. *J Gen Virol,* 79 (Pt 3)**,** 447-56.

Schlottau, K., Forth, L., Angstwurm, K., Hoper, D., Zecher, D., Liesche, F., Hoffmann, B., Kegel, V., Seehofer, D., Platen, S., Salzberger, B., Liebert, U. G., Niller, H. H., Schmidt, B.,

Matiasek, K., Riemenschneider, M. J., Brochhausen, C., Banas, B., Renders, L., Moog, P., Wunderlich, S., Seifert, C. L., Barreiros, A., Rahmel, A., Weiss, J., Tappe, D., Herden, C., Schmidt-Chanasit, J., Schwemmle, M., Rubbenstroth, D., Schlegel, J., Pietsch, C., Hoffmann, D., Jantsch, J. & Beer, M. 2018. Fatal Encephalitic Borna Disease Virus 1 in Solid-Organ Transplant Recipients. *N Engl J Med,* 379**,** 1377-1379.

Schlottau, K., Hoffmann, B., Homeier-Bachmann, T., Fast, C., Ulrich, R. G., Beer, M. & Hoffmann, D. 2017a. Multiple detection of zoonotic variegated squirrel bornavirus 1 RNA in different squirrel species suggests a possible unknown origin for the virus. *Arch Virol,* 162**,** 2747-2754.

Schlottau, K., Jenckel, M., Van Den Brand, J., Fast, C., Herden, C., Hoper, D., Homeier-Bachmann, T., Thielebein, J., Mensing, N., Diender, B., Hoffmann, D., Ulrich, R. G., Mettenleiter, T. C., Koopmans, M., Tappe, D., Schmidt-Chanasit, J., Reusken, C. B., Beer, M. & Hoffmann, B. 2017b. Variegated Squirrel Bornavirus 1 in Squirrels, Germany and the Netherlands. *Emerg Infect Dis,* 23**,** 477-481.

Scholtissek, C. 1990. Pigs as the "mixing vessel" for the creation of new pandemic influenza A viruses. *Med Princ Pract,* 2**,** 65-71.

Scholtissek, C., Koennecke, I. & Rott, R. 1978. Host range recombinants of fowl plague (influenza A) virus. *Virology,* 91**,** 79-85.

Shan, T., Li, L., Simmonds, P., Wang, C., Moeser, A. & Delwart, E. 2011. The fecal virome of pigs on a high-density farm. *J Virol,* 85**,** 11697-708.

Shan, T., Wang, C., Cui, L., Yu, Y., Delwart, E., Zhao, W., Zhu, C., Lan, D., Dai, X. & Hua, X. 2010. Picornavirus salivirus/klassevirus in children with diarrhea, China. *Emerg Infect Dis,* 16**,** 1303-5.

Shang, P., Misra, S., Hause, B. & Fang, Y. 2017. A Naturally Occurring Recombinant Enterovirus Expresses a Torovirus Deubiquitinase. *J Virol,* 91.

Sharp, P. M. & Hahn, B. H. 2008. AIDS: prehistory of HIV-1. *Nature,* 455**,** 605-6.

Shi, M., Lin, X. D., Chen, X., Tian, J. H., Chen, L. J., Li, K., Wang, W., Eden, J. S., Shen, J. J., Liu, L., Holmes, E. C. & Zhang, Y. Z. 2018. The evolutionary history of vertebrate RNA viruses. *Nature,* 556**,** 197-202.

Shi, M., Lin, X. D., Tian, J. H., Chen, L. J., Chen, X., Li, C. X., Qin, X. C., Li, J., Cao, J. P., Eden, J. S., Buchmann, J., Wang, W., Xu, J., Holmes, E. C. & Zhang, Y. Z. 2016. Redefining the invertebrate RNA virosphere. *Nature*.

Shimizu, K., Hamaguchi, S., Ngo, C. C., Li, T. C., Ando, S., Yoshimatsu, K., Yasuda, S. P., Koma, T., Isozumi, R., Tsuda, Y., Fujita, H., Pham, T. T., Le, M. Q., Dang, A. D., Nguyen, T. Q., Yoshida, L. M., Ariyoshi, K. & Arikawa, J. 2016. Serological evidence of infection with rodent-borne hepatitis E virus HEV-C1 or antigenically related virus in humans. *J Vet Med Sci,* 78**,** 1677-1681.

Shiota, T., Li, T. C., Yoshizaki, S., Kato, T., Wakita, T. & Ishii, K. 2013. The hepatitis E virus capsid C-terminal region is essential for the viral life cycle: implication for viral genome encapsidation and particle stabilization. *J Virol,* 87**,** 6031-6.

Shukla, P., Nguyen, H. T., Faulk, K., Mather, K., Torian, U., Engle, R. E. & Emerson, S. U. 2012. Adaptation of a genotype 3 hepatitis E virus to efficient growth in cell culture depends on an inserted human gene segment acquired by recombination. *J Virol,* 86**,** 5697-707.

Siddharth, S., Cyril, C. Y. Y., Shusheng, W., Jianpiao, C., Anna Jin-Xia, Z., Kit-Hang, L., Tom, W. H. C., Jasper, F. W. C., Wan-Mui, C., Jade, L. L. T., Rex, K. H. a.-Y., Vincent, C. C. C., Honglin, C., Susanna, K. P. L., Patrick, C. Y. W., Ning-Shao, X., Chung-Mau, L. & Kwok-Yung, Y. 2018. Rat Hepatitis E Virus as Cause of Persistent Hepatitis after Liver Transplant. *Emerging Infectious Disease journal,* 24.

Sikorski, A., Massaro, M., Kraberger, S., Young, L. M., Smalley, D., Martin, D. P. & Varsani, A. 2013. Novel myco-like DNA viruses discovered in the faecal matter of various animals. *Virus Res,* 177**,** 209-16.

Simner, P. J., Miller, S. & Carroll, K. C. 2018. Understanding the Promises and Hurdles of Metagenomic Next-Generation Sequencing as a Diagnostic Tool for Infectious Diseases. *Clin Infect Dis,* 66**,** 778-788.

Simpson, D. I., Smith, C. E., Marshall, T. F., Platt, G. S., Way, H. J., Bowen, E. T., Bright, W. F., Day, J., Mcmahon, D. A., Hill, M. N., Bendell, P. J. & Heathcote, O. H. 1976. Arbovirus infections in Sarawak: the role of the domestic pig. *Trans R Soc Trop Med Hyg,* 70**,** 66-72.

Sing, T., Sander, O., Beerenwinkel, N. & Lengauer, T. 2005. ROCR: visualizing classifier performance in R. *Bioinformatics,* 21**,** 3940-1.

Sinha, R., Stanley, G., Gulati, G. S., Ezran, C., Travaglini, K. J., Wei, E., Chan, C. K. F., Nabhan, A. N., Su, T., Morganti, R. M., Conley, S. D., Chaib, H., Red-Horse, K., Longaker, M. T., Snyder, M. P., Krasnow, M. A. & Weissman, I. L. 2017. Index Switching Causes "Spreading-Of-Signal" Among Multiplexed Samples In Illumina HiSeq 4000 DNA Sequencing. *bioRxiv*.

Skewes-Cox, P., Sharpton, T. J., Pollard, K. S. & Derisi, J. L. 2014. Profile hidden Markov models for the detection of viruses within metagenomic sequence data. *PLoS One,* 9**,** e105067.

Smith, G. J., Vijaykrishna, D., Bahl, J., Lycett, S. J., Worobey, M., Pybus, O. G., Ma, S. K., Cheung, C. L., Raghwani, J., Bhatt, S., Peiris, J. S., Guan, Y. & Rambaut, A. 2009. Origins and evolutionary genomics of the 2009 swine-origin H1N1 influenza A epidemic. *Nature,* 459**,** 1122-5.

Smits, S. L., Bodewes, R., Ruiz-Gonzalez, A., Baumgartner, W., Koopmans, M. P., Osterhaus, A. D. & Schurch, A. C. 2014a. Assembly of viral genomes from metagenomes. *Front Microbiol,* 5**,** 714.

Smits, S. L., Bodewes, R., Ruiz-Gonzalez, A., Baumgartner, W., Koopmans, M. P., Osterhaus, A. D. & Schurch, A. C. 2015. Recovering full-length viral genomes from metagenomes. *Front Microbiol,* 6**,** 1069.

Smits, S. L., Poon, L. L., Van Leeuwen, M., Lau, P. N., Perera, H. K., Peiris, J. S., Simon, J. H. & Osterhaus, A. D. 2011. Genogroup I and II picobirnaviruses in respiratory tracts of pigs. *Emerg Infect Dis,* 17**,** 2328-30.

Smits, S. L., Schapendonk, C. M., Van Beek, J., Vennema, H., Schurch, A. C., Schipper, D., Bodewes, R., Haagmans, B. L., Osterhaus, A. D. & Koopmans, M. P. 2014b. New viruses in idiopathic human diarrhea cases, the Netherlands. *Emerg Infect Dis,* 20**,** 1218-22.

Smits, S. L., Van Leeuwen, M., Schapendonk, C. M., Schurch, A. C., Bodewes, R., Haagmans, B. L. & Osterhaus, A. D. 2012. Picobirnaviruses in the human respiratory tract. *Emerg Infect Dis,* 18**,** 1539-40.

Smuts, H., Kew, M., Khan, A. & Korsman, S. 2014. Novel hybrid parvovirus-like virus, NIH-CQV/PHV, contaminants in silica column-based nucleic acid extraction kits. *J Virol,* 88**,** 1398.

Song, L., Zhou, Y., He, J., Zhu, H., Huang, R., Mao, P. & Duan, Q. 2008. Comparative sequence analyses of a new mammalian reovirus genome and the mammalian reovirus S1 genes from six new serotype 2 human isolates. *Virus Genes,* 37**,** 392-9.

Spielman, A. 1994. The emergence of Lyme disease and human babesiosis in a changing environment. *Ann N Y Acad Sci,* 740**,** 146-56.

Spinner, M. L. & Di Giovanni, G. D. 2001. Detection and identification of mammalian reoviruses in surface water by combined cell culture and reverse transcription-PCR. *Appl Environ Microbiol,* 67**,** 3016-20.

Steere, A. C., Coburn, J. & Glickstein, L. 2004. The emergence of Lyme disease. *J Clin Invest,* 113**,** 1093-101.

Steyer, A., Gutierrez-Aguire, I., Kolenc, M., Koren, S., Kutnjak, D., Pokorn, M., Poljsak-Prijatelj, M., Racki, N., Ravnikar, M., Sagadin, M., Fratnik Steyer, A. & Toplak, N. 2013. High similarity of novel orthoreovirus detected in a child hospitalized with acute gastroenteritis to mammalian orthoreoviruses found in bats in Europe. *J Clin Microbiol,* 51**,** 3818-25.

Stowers, C. C., Haselton, F. R. & Boczko, E. M. 2010. An Analysis of Quantitative PCR Reliability Through Replicates Using the C Method. *J Biomed Sci Eng,* 3**,** 459-469.

Strong, M. J., Xu, G., Morici, L., Splinter Bon-Durant, S., Baddoo, M., Lin, Z., Fewell, C., Taylor, C. M. & Flemington, E. K. 2014. Microbial contamination in next generation sequencing: implications for sequence-based analysis of clinical samples. *PLoS Pathog,* 10**,** e1004437.

Sumilo, D., Asokliene, L., Bormane, A., Vasilenko, V., Golovljova, I. & Randolph, S. E. 2007. Climate change cannot explain the upsurge of tick-borne encephalitis in the Baltics. *PLoS One,* 2**,** e500.

Sumilo, D., Bormane, A., Asokliene, L., Vasilenko, V., Golovljova, I., Avsic-Zupanc, T., Hubalek, Z. & Randolph, S. E. 2008. Socio-economic factors in the differential upsurge of tick-borne encephalitis in Central and Eastern Europe. *Rev Med Virol,* 18**,** 81-95.

Sun, H., Gao, H., Chen, M., Lan, D., Hua, X., Wang, C., Yuan, C., Yang, Z. & Cui, L. 2015. New serotypes of porcine teschovirus identified in Shanghai, China. *Arch Virol,* 160**,** 831-5.

Sun, W., Zhang, H., Zheng, M., Cao, H., Lu, H., Zhao, G., Xie, C., Cao, L., Wei, X., Bi, J., Yi, C., Yin, G. & Jin, N. 2019. The detection of canine circovirus in Guangxi, China. *Virus Res,* 259**,** 85-89.

Sun, X. & Xu, W. C. 2014. Fast Implementation of DeLong's Algorithm for Comparing the Areas Under Correlated Receiver Operating Characteristic Curves. *Ieee Signal Processing Letters,* 21**,** 1389-1393.

Suzuki, T., Hasebe, A., Miyazaki, A. & Tsunemitsu, H. 2015. Analysis of genetic divergence among strains of porcine rotavirus C, with focus on VP4 and VP7 genotypes in Japan. *Virus Res,* 197**,** 26-34.

Tacharoenmuang, R., Komoto, S., Guntapong, R., Ide, T., Sinchai, P., Upachai, S., Yoshikawa, T., Tharmaphornpilas, P., Sangkitporn, S. & Taniguchi, K. 2016. Full Genome Characterization of Novel DS-1-Like G8P[8] Rotavirus Strains that Have Emerged in Thailand: Reassortment of Bovine and Human Rotavirus Gene Segments in Emerging DS-1-Like Intergenogroup Reassortant Strains. *PLoS One,* 11**,** e0165826.

Takahashi, K., Kitajima, N., Abe, N. & Mishiro, S. 2004. Complete or near-complete nucleotide sequences of hepatitis E virus genome recovered from a wild boar, a deer, and four patients who ate the deer. *Virology,* 330**,** 501-5.

Takahashi, M., Kobayashi, T., Tanggis, Jirintai, S., Mulyanto, Nagashima, S., Nishizawa, T., Kunita, S. & Okamoto, H. 2016. Production of monoclonal antibodies against the ORF3 protein of rat hepatitis E virus (HEV) and demonstration of the incorporation of the ORF3 protein into enveloped rat HEV particles. *Arch Virol,* 161**,** 3391-3404.

Takhampunya, R., Korkusol, A., Pongpichit, C., Yodin, K., Rungrojn, A., Chanarat, N., Promsathaporn, S., Monkanna, T., Thaloengsok, S., Tippayachai, B., Kumfao, N., Richards, A. L. & Davidson, S. A. 2019. Metagenomic Approach to Characterizing

Disease Epidemiology in a Disease-Endemic Environment in Northern Thailand. *Front Microbiol,* 10**,** 319.

Tan Le, V., Thai Le, H., Phu, N. H., Nghia, H. D., Chuong, L. V., Sinh, D. X., Phong, N. D., Mai, N. T., Man, D. N., Hien, V. M., Vinh, N. T., Day, J., Chau, N. V., Hien, T. T., Farrar, J., De Jong, M. D., Thwaites, G., Van Doorn, H. R. & Chau, T. T. 2014. Viral aetiology of central nervous system infections in adults admitted to a tertiary referral hospital in southern Vietnam over 12 years. *PLoS Negl Trop Dis,* 8**,** e3127.

Tan Le, V., Van Doorn, H. R., Nghia, H. D., Chau, T. T., Tu Le, T. P., De Vries, M., Canuti, M., Deijs, M., Jebbink, M. F., Baker, S., Bryant, J. E., Tham, N. T., Nt, B. K., Boni, M. F., Loi, T. Q., Phuong Le, T., Verhoeven, J. T., Crusat, M., Jeeninga, R. E., Schultsz, C., Chau, N. V., Hien, T. T., Van Der Hoek, L., Farrar, J. & De Jong, M. D. 2013. Identification of a new cyclovirus in cerebrospinal fluid of patients with acute central nervous system infections. *MBio,* 4**,** e00231-13.

Tan, Y. F., Teng, C. L., Chua, K. B. & Voon, K. 2017. Pteropine orthoreovirus: An important emerging virus causing infectious disease in the tropics? *J Infect Dev Ctries,* 11**,** 215-219.

Tappe, D., Schlottau, K., Cadar, D., Hoffmann, B., Balke, L., Bewig, B., Hoffmann, D., Eisermann, P., Fickenscher, H., Krumbholz, A., Laufs, H., Huhndorf, M., Rosenthal, M., Schulz-Schaeffer, W., Ismer, G., Hotop, S. K., Bronstrup, M., Ott, A., Schmidt-Chanasit, J. & Beer, M. 2018. Occupation-Associated Fatal Limbic Encephalitis Caused by Variegated Squirrel Bornavirus 1, Germany, 2013. *Emerg Infect Dis,* 24**,** 978-987.

Tappe, D., Schmidt-Chanasit, J., Rauch, J., Allartz, P. & Herden, C. 2019. Immunopathology of Fatal Human Variegated Squirrel Bornavirus 1 Encephalitis, Germany, 2011-2013. *Emerg Infect Dis,* 25**,** 1058-1065.

Taylor, L. H., Latham, S. M. & Woolhouse, M. E. 2001. Risk factors for human disease emergence. *Philos Trans R Soc Lond B Biol Sci,* 356**,** 983-9.

Tei, S., Kitajima, N., Takahashi, K. & Mishiro, S. 2003. Zoonotic transmission of hepatitis E virus from deer to human beings. *Lancet,* 362**,** 371-3.

Teo, E. C., Tan, B. H., Purdy, M. A., Wong, P. S., Ting, P. J., Chang, P. J., Oon, L. L., Sue, A., Teo, C. G. & Tan, C. K. 2017. Hepatitis E in Singapore: A Case-Series and Viral Phylodynamics Study. *Am J Trop Med Hyg,* 96**,** 922-928.

Thaiwong, T., Wise, A. G., Maes, R. K., Mullaney, T. & Kiupel, M. 2016. Canine Circovirus 1 (CaCV-1) and Canine Parvovirus 2 (CPV-2): Recurrent Dual Infections in a Papillon Breeding Colony. *Vet Pathol,* 53**,** 1204-1209.

The Lancet Infectious Diseases 2017. Cholera in Yemen: war, hunger, disease...and heroics. *Lancet Infect Dis,* 17**,** 781.

Theze, J., Li, T., Du Plessis, L., Bouquet, J., Kraemer, M. U. G., Somasekar, S., Yu, G., De Cesare, M., Balmaseda, A., Kuan, G., Harris, E., Wu, C. H., Ansari, M. A., Bowden, R., Faria, N. R., Yagi, S., Messenger, S., Brooks, T., Stone, M., Bloch, E. M., Busch, M., Munoz-Medina, J. E., Gonzalez-Bonilla, C. R., Wolinsky, S., Lopez, S., Arias, C. F., Bonsall, D., Chiu, C. Y. & Pybus, O. G. 2018. Genomic Epidemiology Reconstructs the Introduction and Spread of Zika Virus in Central America and Mexico. *Cell Host Microbe,* 23**,** 855-864 e7.

Thimmasandra Narayanappa, A., Sooryanarain, H., Deventhiran, J., Cao, D., Ammayappan Venkatachalam, B., Kambiranda, D., Leroith, T., Heffron, C. L., Lindstrom, N., Hall, K., Jobst, P., Sexton, C., Meng, X. J. & Elankumaran, S. 2015. A novel pathogenic Mammalian orthoreovirus from diarrheic pigs and Swine blood meal in the United States. *MBio,* 6**,** e00593-15.

Thompson, C. N., Phan, M. V., Hoang, N. V., Minh, P. V., Vinh, N. T., Thuy, C. T., Nga, T. T., Rabaa, M. A., Duy, P. T., Dung, T. T., Phat, V. V., Nga, T. V., Tu Le, T. P., Tuyen, H. T.,

Yoshihara, K., Jenkins, C., Duong, V. T., Phuc, H. L., Tuyet, P. T., Ngoc, N. M., Vinh, H., Chinh, N. T., Thuong, T. C., Tuan, H. M., Hien, T. T., Campbell, J. I., Chau, N. V., Thwaites, G. & Baker, S. 2015. A prospective multi-center observational study of children hospitalized with diarrhea in Ho Chi Minh City, Vietnam. *Am J Trop Med Hyg,* 92**,** 1045-52.

Thuy, N. T., Thu, N. T., Son, N. G., Ha Le, T. T., Hung, V. K., Nguyen, N. T. & Khoa Do, V. A. 2013. Genetic analysis of ORF5 porcine reproductive and respiratory syndrome virus isolated in Vietnam. *Microbiol Immunol,* 57**,** 518-26.

Tischler, G. & Leonard, S. 2014. biobambam: tools for read pair collation based algorithms on BAM files. *Source Code for Biology and Medicine,* 9**,** 13.

To, K. K. W., Chan, W. M., Li, K. S. M., Lam, C. S. F., Chen, Z., Tse, H., Lau, S. K. P., Woo, P. C. Y. & Yuen, K. Y. 2017. High prevalence of four novel astrovirus genotype species identified from rodents in China. *J Gen Virol,* 98**,** 1004-1015.

Tokita, H., Harada, H., Gotanda, Y., Takahashi, M., Nishizawa, T. & Okamoto, H. 2003. Molecular and serological characterization of sporadic acute hepatitis E in a Japanese patient infected with a genotype III hepatitis E virus in 1993. *J Gen Virol,* 84**,** 421-7.

Tran, H. T., Ushijima, H., Quang, V. X., Phuong, N., Li, T. C., Hayashi, S., Xuan Lien, T., Sata, T. & Abe, K. 2003. Prevalence of hepatitis virus types B through E and genotypic distribution of HBV and HCV in Ho Chi Minh City, Vietnam. *Hepatol Res,* 26**,** 275-280.

Trifonov, V. & Rabadan, R. 2010. Frequency analysis techniques for identification of viral genetic data. *mBio,* 1.

Trujillo, A. A., Mccaustland, K. A., Zheng, D. P., Hadley, L. A., Vaughn, G., Adams, S. M., Ando, T., Glass, R. I. & Monroe, S. S. 2006. Use of TaqMan real-time reverse transcription-PCR for rapid detection, quantification, and typing of norovirus. *Journal of Clinical Microbiology,* 44**,** 1405-1412.

Tyler, K. L., Barton, E. S., Ibach, M. L., Robinson, C., Campbell, J. A., O'donnell, S. M., Valyi-Nagy, T., Clarke, P., Wetzel, J. D. & Dermody, T. S. 2004. Isolation and molecular characterization of a novel type 3 reovirus from a child with meningitis. *J Infect Dis,* 189**,** 1664-75.

Van Der Hoek, L., Pyrc, K., Jebbink, M. F., Vermeulen-Oost, W., Berkhout, R. J., Wolthers, K. C., Wertheim-Van Dillen, P. M., Kaandorp, J., Spaargaren, J. & Berkhout, B. 2004. Identification of a new human coronavirus. *Nat Med,* 10**,** 368-73.

Van Dijk, M., Hilderink, H., Rooij, W., Rutten, M., Ashton, R., Kartikasari, K. & Lan, V. C. 2013. Land-use change, food security and climate change in Vietnam; A global-to-local modelling approach. The Hague: LEI-Wageningen UR.

Van Dung, N., Anh, P. H., Van Cuong, N., Hoa, N. T., Carrique-Mas, J., Hien, V. B., Sharp, C., Rabaa, M., Berto, A., Campbell, J., Baker, S., Farrar, J., Woolhouse, M. E., Bryant, J. E. & Simmonds, P. 2016. Large-scale screening and characterization of enteroviruses and kobuviruses infecting pigs in Vietnam. *J Gen Virol,* 97**,** 378-88.

Van Man, N., Luan Le, T., Trach, D. D., Thanh, N. T., Van Tu, P., Long, N. T., Anh, D. D., Fischer, T. K., Ivanoff, B., Gentsch, J. R., Glass, R. I. & Vietnam Rotavirus Surveillance, N. 2005. Epidemiological profile and burden of rotavirus diarrhea in Vietnam: 5 years of sentinel hospital surveillance, 1998-2003. *J Infect Dis,* 192 Suppl 1**,** S127-32.

Van Nguyen, D., Van Nguyen, C., Bonsall, D., Ngo, T., Carrique-Mas, J., Pham, A., Bryant, J., Thwaites, G., Baker, S., Woolhouse, M. & Simmonds, P. 2018. Detection and Characterization of Homologues of Human Hepatitis Viruses and Pegiviruses in Rodents and Bats in Vietnam. *Viruses,* 10**,** 102.

Varsani, A. & Krupovic, M. 2017. Sequence-based taxonomic framework for the classification of uncultured single-stranded DNA viruses of the family Genomoviridae. *Virus Evol,* 3**,** vew037.

Varsani, A. & Krupovic, M. 2018. Smacoviridae: a new family of animal-associated single-stranded DNA viruses. *Arch Virol,* 163**,** 2005-2015.

Vazquez-Castellanos, J. F., Garcia-Lopez, R., Perez-Brocal, V., Pignatelli, M. & Moya, A. 2014. Comparison of different assembly and annotation tools on analysis of simulated viral metagenomic communities in the gut. *BMC Genomics,* 15**,** 37.

Victoria, J. G., Kapoor, A., Li, L., Blinkova, O., Slikas, B., Wang, C., Naeem, A., Zaidi, S. & Delwart, E. 2009. Metagenomic analyses of viruses in stool samples from children with acute flaccid paralysis. *J Virol,* 83**,** 4642-51.

Wakuda, M., Pongsuwanna, Y. & Taniguchi, K. 2005. Complete nucleotide sequences of two RNA segments of human picobirnavirus. *J Virol Methods,* 126**,** 165-9.

Wang, L., Fu, S., Cao, L., Lei, W., Cao, Y., Song, J., Tang, Q., Zhang, H., Feng, Y., Yang, W. & Liang, G. 2015a. Isolation and identification of a natural reassortant mammalian orthoreovirus from least horseshoe bat in China. *PLoS One,* 10**,** e0118598.

Wang, L. F. & Eaton, B. T. 2007. Bats, civets and the emergence of SARS. *Curr Top Microbiol Immunol,* 315**,** 325-44.

Wang, W., Lin, X. D., Guo, W. P., Zhou, R. H., Wang, M. R., Wang, C. Q., Ge, S., Mei, S. H., Li, M. H., Shi, M., Holmes, E. C. & Zhang, Y. Z. 2015b. Discovery, diversity and evolution of novel coronaviruses sampled from rodents in China. *Virology,* 474**,** 19-27.

Wang, Y., Zhao, J., Zheng, M., Liu, Z., Li, W., Fu, X., Lin, Y., Yuan, J., Zhao, J., Shen, Q., Wang, X., Wang, H. & Yang, S. 2018. A novel cardiovirus in wild rats. *Virol J,* 15**,** 58.

Wang, Z. D., Wang, B., Wei, F., Han, S. Z., Zhang, L., Yang, Z. T., Yan, Y., Lv, X. L., Li, L., Wang, S. C., Song, M. X., Zhang, H. J., Huang, S. J., Chen, J., Huang, F. Q., Li, S., Liu, H. H., Hong, J., Jin, Y. L., Wang, W., Zhou, J. Y. & Liu, Q. 2019. A New Segmented Virus Associated with Human Febrile Illness in China. *N Engl J Med,* 380**,** 2116-2125.

Ward, P., Leblanc, D., Houde, A. & Brassard, J. 2015. Analysis of complete genome sequences and a V239A substitution in the helicase domain of swine hepatitis E virus strains isolated in Canada. *Arch Virol,* 160**,** 1767-73.

Warren, C. J. & Sawyer, S. L. 2019. How host genetics dictates successful viral zoonosis. *PLoS Biol,* 17**,** e3000217.

Weiss, R. A. & Mcmichael, A. J. 2004. Social and environmental risk factors in the emergence of infectious diseases. *Nat Med,* 10**,** S70-6.

Wertheim, H. F. L., Nguyen, H. N., Taylor, W., Lien, T. T. M., Ngo, H. T., Nguyen, T. Q., Nguyen, B. N. T., Nguyen, H. H., Nguyen, H. M., Nguyen, C. T., Dao, T. T., Nguyen, T. V., Fox, A., Farrar, J., Schultsz, C., Nguyen, H. D., Nguyen, K. V. & Horby, P. 2009. Streptococcus suis, an Important Cause of Adult Bacterial Meningitis in Northern Vietnam. *PLOS ONE,* 4**,** e5973.

Who Emro 2019. MERS Situation Update November 2019.

Wilburn, L., Yodmeeklin, A., Kochjan, P., Saikruang, W., Kumthip, K., Khamrin, P. & Maneekarn, N. 2017. Molecular detection and characterization of picobirnaviruses in piglets with diarrhea in Thailand. *Arch Virol,* 162**,** 1061-1066.

Wildlife Conservation Society 2008. Commercial wildlife farms in Vietnam: a problem or solution for conservation? Hanoi, Vietnam.

Wilke, A., Bischof, J., Gerlach, W., Glass, E., Harrison, T., Keegan, K. P., Paczian, T., Trimble, W. L., Bagchi, S., Grama, A., Chaterji, S. & Meyer, F. 2016. The MG-RAST metagenomics database and portal in 2015. *Nucleic Acids Res,* 44**,** D590-4.

Willner, D., Furlan, M., Haynes, M., Schmieder, R., Angly, F. E., Silva, J., Tammadoni, S., Nosrat, B., Conrad, D. & Rohwer, F. 2009a. Metagenomic analysis of respiratory tract DNA viral communities in cystic fibrosis and non-cystic fibrosis individuals. *PLoS One,* 4**,** e7370.

Willner, D., Thurber, R. V. & Rohwer, F. 2009b. Metagenomic signatures of 86 microbial and viral metagenomes. *Environ Microbiol,* 11**,** 1752-66.

Wilson, M. R., Sample, H. A., Zorn, K. C., Arevalo, S., Yu, G., Neuhaus, J., Federman, S., Stryke, D., Briggs, B., Langelier, C., Berger, A., Douglas, V., Josephson, S. A., Chow, F. C., Fulton, B. D., Derisi, J. L., Gelfand, J. M., Naccache, S. N., Bender, J., Dien Bard, J., Murkey, J., Carlson, M., Vespa, P. M., Vijayan, T., Allyn, P. R., Campeau, S., Humphries, R. M., Klausner, J. D., Ganzon, C. D., Memar, F., Ocampo, N. A., Zimmermann, L. L., Cohen, S. H., Polage, C. R., Debiasi, R. L., Haller, B., Dallas, R., Maron, G., Hayden, R., Messacar, K., Dominguez, S. R., Miller, S. & Chiu, C. Y. 2019. Clinical Metagenomic Sequencing for Diagnosis of Meningitis and Encephalitis. *N Engl J Med,* 380**,** 2327-2340.

Wolfe, N. D., Daszak, P., Kilpatrick, A. M. & Burke, D. S. 2005a. Bushmeat hunting, deforestation, and prediction of zoonoses emergence. *Emerg Infect Dis,* 11**,** 1822-7.

Wolfe, N. D., Heneine, W., Carr, J. K., Garcia, A. D., Shanmugam, V., Tamoufe, U., Torimiro, J. N., Prosser, A. T., Lebreton, M., Mpoudi-Ngole, E., Mccutchan, F. E., Birx, D. L., Folks, T. M., Burke, D. S. & Switzer, W. M. 2005b. Emergence of unique primate T-lymphotropic viruses among central African bushmeat hunters. *Proc Natl Acad Sci U S A,* 102**,** 7994-9.

Wolfe, N. D., Switzer, W. M., Carr, J. K., Bhullar, V. B., Shanmugam, V., Tamoufe, U., Prosser, A. T., Torimiro, J. N., Wright, A., Mpoudi-Ngole, E., Mccutchan, F. E., Birx, D. L., Folks, T. M., Burke, D. S. & Heneine, W. 2004a. Naturally acquired simian retrovirus infections in central African hunters. *Lancet,* 363**,** 932-7.

Wolfe, N. D., Switzer, W. M., Folks, T. M., Burkes, D. S. & Heneine, W. 2004b. Simian retroviral infections in human beings - Reply. *Lancet,* 364**,** 139-140.

Wommack, K. E., Bhavsar, J., Polson, S. W., Chen, J., Dumas, M., Srinivasiah, S., Furman, M., Jamindar, S. & Nasko, D. J. 2012. VIROME: a standard operating procedure for analysis of viral metagenome sequences. *Stand Genomic Sci,* 6**,** 427-39.

Wommack, K. E., Bhavsar, J. & Ravel, J. 2008. Metagenomics: read length matters. *Appl Environ Microbiol,* 74**,** 1453-63.

Wong, C. C., Thean, S. M., Ng, Y., Kang, J. S. L., Ng, T. Y., Chau, M. L., Koh, T. H. & Chan, K. P. 2019. Seroepidemiology and genotyping of hepatitis E virus in Singapore reveal rise in number of cases and similarity of human strains to those detected in pig livers. *Zoonoses Public Health*.

Woo, P. C., Lau, S. K., Teng, J. L., Tsang, A. K., Joseph, M., Wong, E. Y., Tang, Y., Sivakumar, S., Bai, R., Wernery, R., Wernery, U. & Yuen, K. Y. 2014. Metagenomic analysis of viromes of dromedary camel fecal samples reveals large number and high diversity of circoviruses and picobirnaviruses. *Virology,* 471-473C**,** 117-125.

Woo, P. C., Teng, J. L., Bai, R., Wong, A. Y., Martelli, P., Hui, S. W., Tsang, A. K., Lau, C. C., Ahmed, S. S., Yip, C. C., Choi, G. K., Li, K. S., Lam, C. S., Lau, S. K. & Yuen, K. Y. 2016. High Diversity of Genogroup I Picobirnaviruses in Mammals. *Front Microbiol,* 7**,** 1886.

Wood, D. E. & Salzberg, S. L. 2014. Kraken: ultrafast metagenomic sequence classification using exact alignments. *Genome Biol,* 15**,** R46.

Woolhouse, M., Ashworth, J., Bogaardt, C., Tue, N. T., Baker, S., Thwaites, G. & Phuc, T. M. 2019. Sample descriptors linked to metagenomic sequencing data from human and animal enteric samples from Vietnam. *Sci Data,* 6**,** 202.

Woolhouse, M. E. & Adair, K. 2013. Ecological and taxonomic variation among human RNA viruses. *J Clin Virol,* 58**,** 344-5.

Woolhouse, M. E. & Gowtage-Sequeria, S. 2005. Host range and emerging and reemerging pathogens. *Emerg Infect Dis,* 11**,** 1842-7.

Woolhouse, M. E. J., Adair, K. & Brierley, L. 2013. RNA Viruses: A Case Study of the Biology of Emerging Infectious Diseases. *Microbiol Spectr,* 1.

World Health Organization 2018a. 2018 Annual review of diseases prioritized under the Research and Development Blueprint.

World Health Organization. 2018b. *Ebola virus disease* [Online]. Available: https://www.who.int/en/news-room/fact-sheets/detail/ebola-virus-disease [Accessed 18/06/2019].

Worobey, M. 2008. Phylogenetic evidence against evolutionary stasis and natural abiotic reservoirs of influenza A virus. *J Virol,* 82**,** 3769-74.

Worobey, M., Gemmel, M., Teuwen, D. E., Haselkorn, T., Kunstman, K., Bunce, M., Muyembe, J. J., Kabongo, J. M., Kalengayi, R. M., Van Marck, E., Gilbert, M. T. & Wolinsky, S. M. 2008. Direct evidence of extensive diversity of HIV-1 in Kinshasa by 1960. *Nature,* 455**,** 661-4.

Wright, E. S. & Vetsigian, K. H. 2016a. Inhibitory interactions promote frequent bistability among competing bacteria. *Nature Communications,* 7.

Wright, E. S. & Vetsigian, K. H. 2016b. Quality filtering of Illumina index reads mitigates sample cross-talk. *BMC Genomics,* 17**,** 876.

Wu, Z., Lu, L., Du, J., Yang, L., Ren, X., Liu, B., Jiang, J., Yang, J., Dong, J., Sun, L., Zhu, Y., Li, Y., Zheng, D., Zhang, C., Su, H., Zheng, Y., Zhou, H., Zhu, G., Li, H., Chmura, A., Yang, F., Daszak, P., Wang, J., Liu, Q. & Jin, Q. 2018. Comparative analysis of rodent and small mammal viromes to better understand the wildlife origin of emerging infectious diseases. *Microbiome,* 6**,** 178.

Wu, Z., Yang, L., Ren, X., He, G., Zhang, J., Yang, J., Qian, Z., Dong, J., Sun, L., Zhu, Y., Du, J., Yang, F., Zhang, S. & Jin, Q. 2016. Deciphering the bat virome catalog to better understand the ecological diversity of bat viruses and the bat origin of emerging infectious diseases. *ISME J,* 10**,** 609-20.

Xia, L., Fan, Q., He, B., Xu, L., Zhang, F., Hu, T., Wang, Y., Li, N., Qiu, W., Zheng, Y., Matthijnssens, J. & Tu, C. 2014. The complete genome sequence of a G3P[10] Chinese bat rotavirus suggests multiple bat rotavirus inter-host species transmission events. *Infect Genet Evol,* 28**,** 1-4.

Xiao, P., Li, C., Zhang, Y., Han, J., Guo, X., Xie, L., Tian, M., Li, Y., Wang, M., Liu, H., Ren, J., Zhou, H., Lu, H. & Jin, N. 2018. Metagenomic Sequencing From Mosquitoes in China Reveals a Variety of Insect and Human Viruses. *Front Cell Infect Microbiol,* 8**,** 364.

Xu, B., Zhi, N., Hu, G., Wan, Z., Zheng, X., Liu, X., Wong, S., Kajigaya, S., Zhao, K., Mao, Q. & Young, N. S. 2013. Hybrid DNA virus in Chinese patients with seronegative hepatitis discovered by deep sequencing. *Proc Natl Acad Sci U S A,* 110**,** 10264-9.

Yamamoto, D., Ghosh, S., Kuzuya, M., Wang, Y. H., Zhou, X., Chawla-Sarkar, M., Paul, S. K., Ishino, M. & Kobayashi, N. 2011. Whole-genome characterization of human group C rotaviruses: identification of two lineages in the VP3 gene. *J Gen Virol,* 92**,** 361-9.

Yang, J., Yang, F., Ren, L., Xiong, Z., Wu, Z., Dong, J., Sun, L., Zhang, T., Hu, Y., Du, J., Wang, J. & Jin, Q. 2011. Unbiased parallel detection of viral pathogens in clinical samples by use of a metagenomic approach. *J Clin Microbiol,* 49**,** 3463-9.

Yang, X. L., Tan, B., Wang, B., Li, W., Wang, N., Luo, C. M., Wang, M. N., Zhang, W., Li, B., Peng, C., Ge, X. Y., Zhang, L. B. & Shi, Z. L. 2015. Isolation and identification of bat viruses closely related to human, porcine and mink orthoreoviruses. *J Gen Virol,* 96**,** 3525-31.

Ye, Y., Choi, J. H. & Tang, H. 2011. RAPSearch: a fast protein similarity search tool for short reads. *BMC Bioinformatics,* 12**,** 159.

Yinda, C. K., Ghogomu, S. M., Conceicao-Neto, N., Beller, L., Deboutte, W., Vanhulle, E., Maes, P., Van Ranst, M. & Matthijnssens, J. 2018. Cameroonian fruit bats harbor divergent

viruses, including rotavirus H, bastroviruses, and picobirnaviruses using an alternative genetic code. *Virus Evol,* 4**,** vey008.

Yinda, C. K., Vanhulle, E., Conceicao-Neto, N., Beller, L., Deboutte, W., Shi, C., Ghogomu, S. M., Maes, P., Van Ranst, M. & Matthijnssens, J. 2019. Gut Virome Analysis of Cameroonians Reveals High Diversity of Enteric Viruses, Including Potential Interspecies Transmitted Viruses. *mSphere,* 4.

Yu, J. M., Ao, Y. Y., Liu, N., Li, L. L. & Duan, Z. J. 2015. Salivirus in Children and Its Association with Childhood Acute Gastroenteritis: A Paired Case-Control Study. *PLoS One,* 10**,** e0130977.

Yu, J. M., Li, J. S., Ao, Y. Y. & Duan, Z. J. 2013a. Detection of novel viruses in porcine fecal samples from China. *Virol J,* 10**,** 39.

Yu, J. M., Li, X. Y., Ao, Y. Y., Li, L. L., Liu, N., Li, J. S. & Duan, Z. J. 2013b. Identification of a novel picornavirus in healthy piglets and seroepidemiological evidence of its presence in humans. *PLoS One,* 8**,** e70137.

Yu, X., Li, B., Fu, Y., Jiang, D., Ghabrial, S. A., Li, G., Peng, Y., Xie, J., Cheng, J., Huang, J. & Yi, X. 2010. A geminivirus-related DNA mycovirus that confers hypovirulence to a plant pathogenic fungus. *Proc Natl Acad Sci U S A,* 107**,** 8387-92.

Yu, X. J., Liang, M. F., Zhang, S. Y., Liu, Y., Li, J. D., Sun, Y. L., Zhang, L., Zhang, Q. F., Popov, V. L., Li, C., Qu, J., Li, Q., Zhang, Y. P., Hai, R., Wu, W., Wang, Q., Zhan, F. X., Wang, X. J., Kan, B., Wang, S. W., Wan, K. L., Jing, H. Q., Lu, J. X., Yin, W. W., Zhou, H., Guan, X. H., Liu, J. F., Bi, Z. Q., Liu, G. H., Ren, J., Wang, H., Zhao, Z., Song, J. D., He, J. R., Wan, T., Zhang, J. S., Fu, X. P., Sun, L. N., Dong, X. P., Feng, Z. J., Yang, W. Z., Hong, T., Zhang, Y., Walker, D. H., Wang, Y. & Li, D. X. 2011. Fever with thrombocytopenia associated with a novel bunyavirus in China. *N Engl J Med,* 364**,** 1523-32.

Zaccaria, G., Malatesta, D., Scipioni, G., Di Felice, E., Campolo, M., Casaccia, C., Savini, G., Di Sabatino, D. & Lorusso, A. 2016. Circovirus in domestic and wild carnivores: An important opportunistic agent? *Virology,* 490**,** 69-74.

Zaki, A. M., Van Boheemen, S., Bestebroer, T. M., Osterhaus, A. D. & Fouchier, R. A. 2012. Isolation of a novel coronavirus from a man with pneumonia in Saudi Arabia. *N Engl J Med,* 367**,** 1814-20.

Zanluca, C., Melo, V. C., Mosimann, A. L., Santos, G. I., Santos, C. N. & Luz, K. 2015. First report of autochthonous transmission of Zika virus in Brazil. *Mem Inst Oswaldo Cruz,* 110**,** 569-72.

Zell, R., Delwart, E., Gorbalenya, A. E., Hovi, T., King, A. M. Q., Knowles, N. J., Lindberg, A. M., Pallansch, M. A., Palmenberg, A. C., Reuter, G., Simmonds, P., Skern, T., Stanway, G., Yamashita, T. & Consortium, I. R. 2017. ICTV Virus Taxonomy Profile: Picornaviridae. *Journal of General Virology,* 98**,** 2421-2422.

Zhang, B., Tang, C., Yue, H., Ren, Y. & Song, Z. 2014. Viral metagenomics analysis demonstrates the diversity of viral flora in piglet diarrhoeic faeces in China. *J Gen Virol,* 95**,** 1603-11.

Zhang, C., Liu, L., Wang, P., Liu, S., Lin, W., Hu, F., Wu, W., Chen, W. & Cui, S. 2011. A potentially novel reovirus isolated from swine in northeastern China in 2007. *Virus Genes,* 43**,** 342-9.

Zhang, Y. Z., Shi, M. & Holmes, E. C. 2018. Using Metagenomics to Characterize an Expanding Virosphere. *Cell,* 172**,** 1168-1172.

Zhao, G., Krishnamurthy, S., Cai, Z., Popov, V. L., Travassos Da Rosa, A. P., Guzman, H., Cao, S., Virgin, H. W., Tesh, R. B. & Wang, D. 2013. Identification of novel viruses using VirusHunter--an automated data analysis pipeline. *PLoS One,* 8**,** e78470.

Zhao, Y., Tang, H. & Ye, Y. 2012. RAPSearch2: a fast and memory-efficient protein similarity search tool for next-generation sequencing data. *Bioinformatics,* 28**,** 125-6.

Appendix. Viral signals in study populations

Table A.1 Viral signals in human samples

OTU	N° signals	Size (rp)	Accession nrs.	Species	Host	Average hit ID (%)	Cat.
Rotavirus	374	18-31961192	56 records, incl. KU550279 (65)	*Rotavirus A*	Human	97.72-100	IIIb
Rotavirus	10	71991-8733958	7 records, incl. KX363412 (3)	*Rotavirus A*	Swine	97.26-98.84	IIIb
Alphatorquevirus	113	19-138300	61 records, incl. AF247137 (6)-includes other anellovirus records	Various torque teno viruses*	Human	79.62-99.34	I
Picobirnavirus	55	18-15381	21 records, incl. KR827418 (12)	Uninformative (GI, GII)	Human	78.36-99.35	IIIb
Picobirnavirus	14	23-7625	KP941111	Uninformative (GI)	Fox	88.02-96.42	IIIb
Picobirnavirus	10	22-12416	4 records, incl. KM254165 (7)	Uninformative (GI, GII)	Chicken	81.43-93.48	IIIb
Picobirnavirus	8	19-2606	KF861773 (4), KJ650571 (3), HM070240 (1)	Uninformative (GI)	Swine	78.43-94.31	IIIb
Picobirnavirus	6	38-42697	KF792838	Uninformative (GII)	Cat	88.85-93.74	IIIb
Picobirnavirus	5	18-169	5 records, incl. KT334936 (1)	Uninformative (GI)	Macaque	82.37-96.96	IIIb
Picobirnavirus	4	21-976	KM573806 (2), KM573802 (1), KM573807 (1)	Uninformative (GI)	Dromedary camel	78.54-81.16	IIIb
Picobirnavirus	1	219	KT934307	Uninformative (GI)	Wolf	78.07	IIIb
Picobirnavirus	1	27	KJ476132	Uninformative (GI)	Turkey	93.16	IIIb
Picobirnavirus	1	70	JQ776552	Uninformative (GI)	California sea lion	NA	IIIb
Enterovirus	27	56-997267	13 records, incl. KU355273 (7)	*Enterovirus B*	Human	91.78-97.28	I
Enterovirus	17	27-201391	10 records, incl. KP289363 (4)	*Enterovirus A*	Human	89.88-99.67	I
Enterovirus	11	25-33194	6 records, incl. KR399988 (4)	*Enterovirus C*	Human	89.86-99.61	I
Enterovirus	7	26-1350	6 records, incl. KX398052 (2)	*Rhinovirus A*	Human	88.18-97.20	I
Enterovirus	5	31-279	5 records, incl. EF077279 (1)	*Rhinovirus C*	Human	94.44-96.33	I
Enterovirus	2	384-1098	FJ445155 (1), JX074048 (1)	*Rhinovirus B*	Human	91.11-98.20	I
Norovirus	64	43-2840517	17 records, incl. AB972505 (26)	*Norwalk virus*	Human	90.75-99.43	I

OTU	N° signals	Size (rp)	Accession nrs.	Species	Host	Average hit ID (%)	Cat.
Mastadenovirus	19	88-10356906	KF303071 (16), L19443 (3)	*Human mastadenovirus F*	Human	99.19-99.77	I
Mastadenovirus	9	27-146010	8 records, incl. LC068720 (4)	*Human mastadenovirus C*	Human	98.71-99.92	I
Mastadenovirus	7	235-402166	KF268311 (4), KU361344 (2), EF564601 (1)	*Human mastadenovirus B*	Human	99.56-99.85	I
Mastadenovirus	7	18-436	6 records, incl. AB695622 (2)	*Human mastadenovirus D*	Human	97.72-100	I
Mastadenovirus	5	97-32128	X73487 (3), GU191019 (1), KF268119 (1)	*Human mastadenovirus A*	Human	98.34-99.16	I
Parechovirus	37	22-178604	7 records, incl. KT626011 (22)	*Parechovirus A*	Human	89.00-97.74	I
Sapovirus	32	23-2720401	11 records, incl. KP067444 (9)	*Sapporo virus*	Human	83.78-98.11	I
Betatorquevirus	26	20-1106	10 records, incl. AB041962 (7)	Various torque teno mini viruses*	Human	73.74-97.88	I
Cytomegalovirus	23	18-1591	7 records, incl. KP973642 (15)	*Human betaherpesvirus 5*	Human	97.90-99.74	I
Mamastrovirus	17	19-3488594	HQ398856 (12), KF157967 (5)	*Mamastrovirus 1*	Human	97.88-98.78	I
Mamastrovirus	3	107-5339	AB823732 (2), AB829252 (1)	*Mamastrovirus 6*	Human	98.10-98.54	I
Mamastrovirus	2	353-815	JX857868 (1), KJ920197 (1)	*Mamastrovirus 9*	Human	97.39-98.35	I
Gammatorquevirus	18	18-941	12 records, incl. JX157237 (5)	Various torque teno midi viruses*	Human	77.94-98.94	I
Polyomavirus	7	18-49	5 records, incl. KJ128381 (2)	*Human polyomavirus 5*	Human	99.46-99.84	I
Polyomavirus	5	19-215	4 records, incl. JX273163 (2)	*Human polyomavirus 2*	Human	99.71-99.93	I
Polyomavirus	2	560-1538	KX787894	*Human polyomavirus 4*	Human	99.49-99.65	I
Cyclovirus	6	24-999	KF031471	*Human associated cyclovirus 8*	Chicken	96.80-98.80	IIIb
Cyclovirus	6	33-264	KF031466 (3), KF031468 (3)	*Human associated cyclovirus 8*	Human	98.07-99.31	IIIb
Cyclovirus	1	22	KC771281	*Human associated cyclovirus 9*	Human	91.32	I
Salivirus	12	90-67222	KT182636 (10), GQ253930 (1), KM023140 (1)	*Salivirus A*	Human	92.58-98.45	I
Betapapillomavirus	6	31-390	4 records, incl. Y15176 (3)	*Betapapillomavirus 2*	Human	90.39-98.91	I

OTU	N° signals	Size (rp)	Accession nrs.	Species	Host	Average hit ID (%)	Cat.
Betapapillomavirus	4	25-50	U31785 (2), FM955837 (1), U31782 (1)	*Betapapillomavirus 1*	Human	97.33-99.50	I
Betapapillomavirus	1	31	AY382779	*Betapapillomavirus 5*	Human	99.75	I
Cosavirus	4	27-1270	KJ194505 (3), FJ438908 (1)	*Cosavirus D*	Human	88.97-91.20	I
Cosavirus	3	55-18507	KJ396940 (2), JN867756 (1)	*Cosavirus A*	Human	91.24-96.66	I
Cosavirus	1	45	KM516909	*Cosavirus B*	Human	89.91	I
Cosavirus	1	181	FJ555055	*Cosavirus E*	Human	89.69	I
Kobuvirus	8	22-344997	FJ890523 (7), GQ927711 (1)	*Aichivirus A*	Human	95.60-97.55	I
Orthohepadnavirus	8	24-5195	6 records, incl. KP341010 (3)	*Hepatitis B virus*	Human	98.48-99.78	I
Cardiovirus	7	23-2720	5 records, incl. FR682076 (2)	*Cardiovirus B*	Human	90.64-98.02	I
Bocaparvovirus	5	78-16043	KX373885	*Primate bocaparvovirus 1*	Human	99.67-99.79	I
Bocaparvovirus	1	28	KM624025	*Primate bocaparvovirus 2*	Human	99.85	I
Gemykrogvirus	5	21-425	KJ938717	*Caribou associated gemykrogvirus 1*	Caribou	86.88-89.90	IV
Porprismacovirus	3	20-22	KP233190	*Human associated porprismacovirus 2*	Chimpanzee	88.58-91.49	I
Gemykibivirus	2	22-29	KP133080	*Human associated gemykibivirus 2*	NA (sewage)	98.78-98.94	I
Hepacivirus	2	70-120	KT234509 (1), KU645407 (1)	*Hepacivirus C*	Human	97.43-97.90	I
Lymphocryptovirus	2	18-811	LC137018	*Human gammaherpesvirus 4*	Human	99.69-99.90	I
Morbillivirus	2	13960-253813	KT732232	*Measles morbillivirus*	Human	99.32-99.50	I
Rubulavirus	2	184-424	KF483663	*Human rubulavirus 4*	Human	96.41-96.53	I
Alphapapillomavirus	1	80	EF117891	*Alphapapillomavirus 4*	Human	99.62	I
Betacoronavirus	1	187	KX344031	*Betacoronavirus 1*	Human	99.57	I
Circovirus	1	46	KC241982	*Canine circovirus*	Dog	88.82	IV
Flavivirus	1	18	GU131898	*Dengue virus*	Human	99.48	I

OTU	N° signals	Size (rp)	Accession nrs.	Species	Host	Average hit ID (%)	Cat.
Molluscipoxvirus	1	18	U60315	*Molluscum contagiosum virus*	Human	99.67	I
Orthohepevirus	1	72	AB720034	*Orthohepevirus A*	Human	97.9	I
Orthopneumovirus	1	46	KX765977	*Human orthopneumovirus*	Human	99.6	I
Orthoreovirus	1	42	KT444545	*Mammalian orthoreovirus*	Bat	99	IIIb
Pegivirus	1	19	LT009493	*Pegivirus C*	Human	92.13	I

Table A.2 Viral signals in swine samples

OTU	Nº signals	Size (rp)	Accession nrs.	Species	Host	Average hit ID (%)	Cat.
Picobirnavirus	71	20-14990	13 records, incl. KU892530 (23)	Uninformative (GI, GII)	Human	79.02-99.20	IIIb
Picobirnavirus	65	31-16215	7 records, incl. KM573802 (25)	Uninformative (GI)	Dromedary camel	79.37-90.22	IIIb
Picobirnavirus	31	45-3492	4 records, incl. KM254164 (16)	Uninformative (GI)	Chicken	78.24-83.89	IIIb
Picobirnavirus	22	59-7393	13 records, incl. KC841459 (3)	Uninformative (GI, GII)	Swine	79.32-99.12	IIIb
Picobirnavirus	16	48-3157	4 records, incl. KF823810 (10)	Uninformative (GI)	Fox	80.24-87.62	IIIb
Picobirnavirus	13	32-3813	KR902505	Uninformative (GI)	Horse	80.08-87.05	IIIb
Picobirnavirus	8	77-3419	KJ476133 (6), KJ476129 (1), KJ476131 (1)	Uninformative (GI)	Turkey	79.99-86.42	IIIb
Picobirnavirus	7	98-2177	JQ776552	Uninformative (GI)	California sea lion	79.36-83.19	IIIb
Picobirnavirus	7	39-2187	KT335010 (5), KT334917 (1), KT334992 (1)	Uninformative (GI)	Macaque	83.67-91.27	IIIb
Picobirnavirus	7	29-6603	5 records, incl. KJ135810 (2), KJ135877 (2)	Uninformative (GI)	NA (wastewater/sewage)	94.22-97.89	IIIb
Picobirnavirus	5	35-5346	AB828072 (4), KP843617 (1)	Uninformative (GI)	Cow	94.45-95.52	IIIb
Picobirnavirus	4	60-291	JF755420	Uninformative (GI)	Meadow vole	78.06-79.51	IIIb
Picobirnavirus	1	494	JF755419	Uninformative (GI)	Mouse	79.36	IIIb
Picobirnavirus		561	KT934308	Uninformative (GI)	Wolf	79.17	IIIb
Picobirnavirus	1	105	KF823812	Uninformative (GI)	Genet	83.37	IIIb

OTU	N° signals	Size (rp)	Accession nrs.	Species	Host	Average hit ID (%)	Cat.
Mamastrovirus	146	62-212582	8 records, incl. JX556692 (78)	Unassigned (proposed MAstV26-27, PAstV4-like)	Swine	84.64-95.86	IIa
Mamastrovirus	87	81-140253	8 records, incl. KP982872 (27)	Unassigned (proposed MAstV32, PAstV2-like)	Swine	85.14-92.82	IIa
Mamastrovirus	8	109-8539	KJ571486	Unassigned (proposed MAstV32, PAstV2-like)	Porcupine	86.01-87.64	IIa
Mamastrovirus	4	64-49143	JX556693 (2), JF713711 (1), KP747574 (1)	Unassigned (proposed MAstV24, PAstV5-like)	Swine	90.47-94.77	IIa
Mamastrovirus	3	157-3291	KF787112 (2), GQ914773 (1)	*Mamastrovirus 3*	Swine	88.94-89.46	IIa
Mamastrovirus	3	81-757	JX556691	Unassigned (proposed MAstV22, PAstV3-like)	Swine	86.90-94.76	IIa
Bocaparvovirus	129	24-25612	35 records, incl. KF206168 (19)	*Ungulate bocaparvovirus 5*	Swine	88.44-99.30	IIa
Bocaparvovirus	17	155-929340	LC090198	*Ungulate bocaparvovirus 3*	Swine	99.43-99.75	IIa
Bocaparvovirus	13	27-1747	6 records, incl. KM402139 (6)	*Ungulate bocaparvovirus 2*	Swine	95.61-98.15	IIa
Bocaparvovirus	7	32-1854	KJ622366	*Ungulate bocaparvovirus 4*	Swine	98.59-98.82	IIa
Porprismacovirus	36	19-315	KP233192	*Gorilla associated porprismacovirus 1*	Gorilla	84.30-89.79	IIc
Porprismacovirus	35	21-2787	KJ577810	*Porcine associated porprismacovirus 4*	Swine	86.76-93.65	IIa
Porprismacovirus	24	19-315	KJ577811	*Porcine associated porprismacovirus 5*	Swine	81.57-91.27	IIa
Porprismacovirus	17	18-259	KJ577816	*Porcine associated porprismacovirus 9*	Swine	86.94-92.83	IIa
Porprismacovirus	14	37-456	KJ577818 (12), KC545226 (2)	*Porcine associated porprismacovirus 2*	Swine	85.75-92.16	IIa
Porprismacovirus	13	19-437	KC545230	*Porcine associated porprismacovirus 3*	Swine	86.62-91.60	IIa
Porprismacovirus	9	38-215	4 records, incl. KJ577813 (3)	*Porcine associated porprismacovirus 7*	Swine	83.98-88.74	IIa
Porprismacovirus	1	21	KT862226	*Porcine associated porprismacovirus 1*	Hare	91.85	IIa

OTU	N° signals	Size (rp)	Accession nrs.	Species	Host	Average hit ID (%)	Cat.
Enterovirus	105	20-70095	10 records, incl. JQ818253 (35)	*Enterovirus G*	Swine	81.19-96.22	IIa
Enterovirus	2	52-91	KJ641696	Unassigned (Bat picornavirus)	Bat	76.60-76.90	IIc
Teschovirus	91	19-2795	31 records, incl. AF296111 (10)	*Teschovirus A*	Swine	77.39-97.31	IIa
Po-Circo-like virus	62	20-3866	JF713717 (41), JF713716 (21)	Unassigned (Po-Circo-like virus 21 & 22)	Swine	84.25-94.76	IIa
Po-Circo-like virus	13	25-1902	JF713718	Unassigned (Po-Circo-like virus 41)	Swine	88.08-93.74	IIa
Po-Circo-like virus	13	27-2537	JF713719	Unassigned (Po-Circo-like virus 51)	Swine	91.03-95.49	IIa
Sapelovirus	82	19-57143	10 records, incl. KJ821020 (17)	*Sapelovirus A*	Swine	86.17-95.68	IIa
Pasivirus	69	18-5015	JX491648 (43), KM259923 (20), JQ316470 (6)	*Pasivirus A*	Swine	83.09-93.60	IIa
Huchismacovirus-like	61	18-623	7 records, incl. KU203352 (26)	Unassigned / unclear	Swine	80.66-97.62	IIa
Mastadenovirus	35	19-377	AB026117	*Porcine mastadenovirus A*	Swine	91.61-97.75	IIa
Mastadenovirus	16	53-82832	AF289262	*Porcine mastadenovirus C*	Swine	97.32-98.95	IIa
Posavirus 1	31	22-17030	6 records, incl. LC123280 (16)	Unclassified	Swine	91.22-96.34	IIa
Posavirus 3	29	30-4583	KR019688 (14), KT833079 (12), KT833078 (3)	Unclassified	Swine	83.90-89.25	IIa
Dependoparvovirus	19	18-693	4 records, incl. JX896667 (10)	Unassigned or taxonomy unclear	Swine	88.63-99.26	IIa
Dependoparvovirus	7	24-765	DQ335246	*Adeno-associated dependoparvovirus B*	Goat	86.46-88.62	IIa
Dependoparvovirus	1	19	FJ688147	*Adeno-associated dependoparvovirus B*	Swine	88.59	IIa
Sapovirus	23	19-3559	11 records, incl. AB221130 (7)	*Sapporo virus*	Swine	81.05-93.13	IIa
Copiparvovirus	13	22-688	KF999684 (10), KX384822 (2), KX384823 (1)	Unassigned (Porcine parvovirus 6)	Swine	97.44-99.06	IIa

OTU	N° signals	Size (rp)	Accession nrs.	Species	Host	Average hit ID (%)	Cat.
Copiparvovirus	9	22-810	KU745628 (5), JQ868716 (3), GU978966 (1)	*Ungulate copiparvovirus 2*	Swine	99.04-99.84	IIa
Rotavirus	9	56-2239	6 records, incl. KX362470 (2)	*Rotavirus C*	Swine	94.63-99.11	IIIa
Rotavirus	6	36-10734	5 records, incl. KX363352 (2)	*Rotavirus A*	Swine	97.68-98.44	IIIb
Rotavirus	3	19-1904	KX363301 (1), LC169985 (1), LC169990 (1)	*Rotavirus A*	Human	98.86-99.58	IIIb
Rotavirus	2	27-2083	KX362406	*Rotavirus B*	Swine	96.87-97.02	IIa
Rotavirus	2	19 -330	KX362517 (1), KX362530 (1)	*Rotavirus H*	Swine	97.99-99.07	IIa
Orthohepevirus	15	20-3058	4 records, incl. JQ679014 (8)	*Orthohepevirus A*	Human	89.26-94.08	IIIa
Orthohepevirus	5	40-1559	AB740232 (4), KJ507956 (1)	*Orthohepevirus A*	Swine	91.21-93.58	IIIa
Kobuvirus	18	20-735	12 records, incl. KM051987 (3)	*Aichivirus C*	Swine	87.43-97.85	IIa
Cyclovirus	12	21-314	KF031471	*Human associated cyclovirus 8*	Chicken	98.89-99.44	IIIb
Cyclovirus	3	19-23	GQ404890 (2), GQ404855 (1)	*Human associated cyclovirus 7*	Human	82.83-83.70	IV
Cyclovirus	2	21-212	KF031468 (1), KM392289 (1)	*Human associated cyclovirus 8*	Human	98.40-98.68	IIIb
Circovirus	16	21-572	KM042415 (9), KX828241 (4), KT336603 (3)	*Porcine circovirus 2*	Swine	98.86-99.89	IIa
Torovirus	12	19-1123	JQ860350 (10), KM403390 (2)	*Porcine torovirus*	Swine	74.92-93.43	IIa
Tetraparvovirus	6	25-147	KU167029 (4), KU167028 (2)	*Ungulate tetraparvovirus 2*	Swine	99.00-99.58	IIa
Tetraparvovirus	5	23-167	4 records, incl. KP245947 (2)	*Ungulate tetraparvovirus 3*	Swine	98.20-99.00	IIa
Protoparvovirus	9	20-76	KU867071 (5), KT965075 (4)	Unassigned (porcine bufavirus)	Swine	96.47-98.39	IIa
Protoparvovirus	1	27	JX568157	*Ungulate protoparvovirus 1*	Swine	98.55	IIa
St-Valerien swine virus	6	20-118	AB863586	Unassigned (St-Valerien swine virus)	Swine	89.40-91.78	IIa
Suipoxvirus	5	28-136	AF410153	*Swinepox virus*	Swine	97.05-98.60	IIa

OTU	N° signals	Size (rp)	Accession nrs.	Species	Host	Average hit ID (%)	Cat.
Orthoreovirus	4	24-1043	4 records, incl. JX486065 (1)	*Mammalian orthoreovirus*	Swine	91.99-96.87	IIIb
Pestivirus	3	22-39	KU194229 (2), KR011347 (1)	Unassigned (atypical porcine pestivirus)	Swine	85.57-89.47	IIa
Norovirus	2	23-122	AB126320 (1), HQ392821 (1)	*Norwalk virus*	Swine	85.49-89.23	IIa
Gemykibivirus	1	47	KF371634	*Black robin associated gemykibivirus 1*	Black robin	76.04	IIc
Arterivirus	1	839	KU950370	*Porcine reproductive and respiratory syndrome virus 2*	Swine	99.47	IIa
Kappatorquevirus	1	53	JQ406846	*Torque teno sus virus k2b*	Swine	95.78	IIa

Table A.3 Viral signals in rat samples

OTU	N^{o} signals	Size (rp)	Accession nrs.	Species	Host	Average hit ID (%)	Cat.
Rotavirus	78	20-42784	4 records, incl. U03556 (50)	*Rotavirus B*	Black rat (*Rattus rattus*)	81.63-88.99	IIb
Rotavirus	3	21-69	KF541282	*Rotavirus A*	Human	90.48-91.42	IIIb
Rotavirus	3	862-33384	KJ879450	*Rotavirus A*	Norway rat (*Rattus norvegicus*)	85.19-85.92	IIIb
Rotavirus	1	4008	KF649188	*Rotavirus A*	Bat	95.32	IIIb
Picobirnavirus	25	20-88620	4 records, incl. KF823810 (11)	Uninformative (GI)	Fox	77.70-84.85	IIIb
Picobirnavirus	12	19-8361	JF755419	Uninformative (GI)	Mouse	79.95-83.58	IIIb
Picobirnavirus	11	22-11460	4 records, incl. KM254161 (6)	Uninformative (GI)	Chicken	77.08-85.88	IIIb
Picobirnavirus	5	200-1715	HM070240 (3), KF861773 (2)	Uninformative (GI)	Swine	80.27-84.39	IIIb
Picobirnavirus	3	41-2572	KJ476125 (1), KJ476126 (1), KJ476133 (1)	Uninformative (GI)	Turkey	79.87-85.87	IIIb
Picobirnavirus	3	196-3365	JF755420	Uninformative (GI)	Meadow vole (*Microtus pennsylvanicus*)	81.07-82.07	IIIb
Picobirnavirus	3	57-720	KM573798 (1), KM573801 (1), KM573807 (1)	Uninformative (GI)	Dromedary camel	79.70-82.95	IIIb
Picobirnavirus	2	19-80	KT934307 (1), KT934308 (1)	Uninformative (GI)	Wolf	77.81-80.78	IIIb
Picobirnavirus	1	93	KJ135916	Uninformative (GI)	NA (wastewater/sewage)	83.27	IIIb
Picobirnavirus	1	466	KT335056	Uninformative (GI)	Macaque	85.98	IIIb
Mamastrovirus	39	18-218758	HM450382	Unassigned (proposed MAstV25)	Norway rat (*Rattus norvegicus*)	87.48-93.40	IIb
Mamastrovirus	16	31-99069	KT599569 (15), KT599570 (1)	Unassigned (unclear)	Macaque	81.21-88.01	IIc

OTU	N° signals	Size (rp)	Accession nrs.	Species	Host	Average hit ID (%)	Cat.
Mamastrovirus	3	86-580	KJ571449 (2), KJ571470 (1)	Unassigned (*E. cachinus* AstV Group 1)	Kachin red-backed vole (*Eothenomys cachinus*)	76.58-77.82	IIc
Mamastrovirus	1	82	KP404149	*Mamastrovirus 5*	Dog	76.97	IIc
Protoparvovirus	43	18-78200	KJ641666	*Rodent protoparvovirus 1*	Pomona roundleaf bat (*Hipposideros pomona*)	89.21-98.44	IIb
Protoparvovirus	10	27-179354	JX627317 (9), AF332883 (1)	*Rodent protoparvovirus 1*	Rat	89.08-98.99	IIb
Protoparvovirus	5	28-1477	V01115 (3), FJ445512 (2)	*Rodent protoparvovirus 1*	Mouse (*Mus musculus*)	86.98-91.41	IIb
Kobuvirus	42	19-43618	JQ898342	*Aichivirus A*	NA (sewage)	87.21-90.75	IIb
Hunnivirus	41	18-1136366	KT944213 (30), KT944212 (11)	Unassigned	Ricefield rat (*Rattus argentiventer*)	95.53-97.81	IIb
Mastadenovirus	27	18-544	HM049560	*Murine mastadenovirus B*	Mouse (*Mus musculus*)	83.34-85.91	IIb
Orthohepevirus	23	24-28857	5 records, incl. AB847307 (17)	*Orthohepevirus C*	Black rat (*Rattus rattus*)	84.65-88.60	IIIa
Cardiovirus	12	22-835	JX257003	*Cardiovirus A*	Wood mouse (*Apodemus sylvaticus*)	88.40-90.67	IIb
Cardiovirus	4	22-41023	JQ864242	*Cardiovirus C*	Norway rat (*Rattus norvegicus*)	86.19-89.15	IIb
Cardiovirus	3	21-51147	AB090161 (2), KJ950912 (1)	*Cardiovirus B*	Norway rat (*Rattus norvegicus*)	76.28-78.16	IIc
Cardiovirus	2	66-135	EU723237	*Cardiovirus B*	Human/rodent (controversial)	74.92-75.28	IIc
Cardiovirus	1	3462	AB747253	*Cardiovirus B*	Human	79.72	IIc
Betacoronavirus	15	52-155696	KM349744	Unassigned (ChRCoV HKU24)	Norway rat (*Rattus norvegicus*)	94.76-96.23	IIb
Betacoronavirus	4	21-1977	KF294371 (3), KF294372 (1)	*Murine coronavirus*	Rats (various species)	88.99-93.18	IIb

OTU	N° signals	Size (rp)	Accession nrs.	Species	Host	Average hit ID (%)	Cat.
Rosavirus	12	18-1810	KX783422 (5), KX783423 (5), KX783421 (2)	Unassigned (candidate sp. Rosavirus B)	Norway rat (*Rattus norvegicus*)	83.91-93.87	IIb
Norovirus	6	22-231	5 records, incl. JN975493 (2)	*Norwalk virus*	Mouse (*Mus musculus*)	80.03-86.47	IIb
Norovirus	6	23-430	JX486102	*Norwalk virus*	Norway rat (*Rattus norvegicus*)	74.28-77.70	IIb
Dependoparvovirus	10	104-6140	DQ100362	Unassigned or taxonomy unclear	Mouse (*Mus musculus*)	82.88-84.55	IIb
Dependoparvovirus	1	547	DQ100363	Unassigned or taxonomy unclear	Norway rat (*Rattus norvegicus*)	84.92	IIb
Enterovirus	4	176-1544	KJ641693	Unassigned (bat picornavirus)	Bat	77.61-79.09	IIc
Enterovirus	4	93-2234	KJ950883	*Rabovirus A*	Norway rat (*Rattus norvegicus*)	79.80-81.89	IIc
Enterovirus	1	544	JX627573	*Sapelovirus B*	Macaque	76.39	IIc
Rabovirus	8	23-1259	KP233897	*Rabovirus A*	Norway rat (*Rattus norvegicus*)	84.97-86.99	IIb
Sapovirus	4	27-355	KJ950882 (3), KJ950881 (1)	*Sapporo virus*	Norway rat (*Rattus norvegicus*)	79.30-80.31	IIb
Parechovirus	1	73615	HF677705	*Parechovirus C*	African wood mouse (*Hylomyscus* sp.)	77.02	IIb
Cyclovirus	1	27	KF031468	*Human associated cyclovirus 8*	Human	99.08	IIIb
Mosavirus	1	36	JF973687	*Mosavirus A*	Canyon mouse (*Peromyscus crinitus*)	77.24	IIb
Pipapillomavirus	1	24	GQ180114	*Pipapillomavirus 2*	Norway rat (*Rattus norvegicus*)	85.31	IIb