Understanding virus and microbial evolution in wildlife through meta-transcriptomics

Ayda Susana Ortiz Baez

2022

Faculty of Science, School of Life and Environmental Sciences

The University of Sydney

A thesis submitted in fulfilment of the requirements for the degree of Doctor of Philosophy

Statement of Originality

This is to certify that to the best of my knowledge the content of this thesis is my own work. This thesis has not been submitted for any degree or other purposes.

I certify that the intellectual content of this thesis is the product of my own work and that all the assistance received in preparing this thesis and sources have been acknowledged.

Ayda Susana Ortiz-Baez

Authorship attribution statement

Research chapters

Chapter 2 of this thesis is published as: Ortiz-Baez, A. S., Eden, J. S., Moritz, C., & Holmes, E. C. (2020). A divergent articulavirus in an Australian gecko identified using meta-transcriptomics and protein structure comparisons. *Viruses*, 12(6), 613. https://doi.org/10.3390/v12060613. *I was the lead author on the paper, analysed the data and wrote the drafts of the manuscript.*

Chapter 3 of this thesis is published as: Turnbull, O. M., Ortiz-Baez, A. S., Eden, J. S., Shi, M., Williamson, J. E., Gaston, T. F., Zhang, Y. Z., Holmes, E. C. & Geoghegan, J. L. (2020). Meta-transcriptomic identification of divergent Amnoonviridae in fish. *Viruses*, 12(11), 1254. https://doi.org/10.3390/v12111254. *I was the joint first author of this paper. I conducted part (50%) of the data analysis and co-wrote the drafts of the manuscript.*

Chapter 4 of this thesis is published as: Ortiz-Baez, A. S., Holmes, E. C., Charon, J., Pettersson, J. H., & Hesson, J. C. (2022). Meta-transcriptomics reveals potential virus transfer between Aedes communis mosquitoes and their parasitic water mites. *Virus evolution*, 8(2), veac090. https://doi.org/10.1093/ve/veac090. *I was the lead author on the paper, analysed the data and wrote the drafts of the manuscript.*

Chapter 5 of this thesis is published as: Ortiz-Baez, A. S., Shi, M., Hoffmann, A. A., & Holmes, E. C. (2021). RNA virome diversity and *Wolbachia* infection in individual *Drosophila simulans* flies.

The Journal of general virology, 102(10), 001639. https://doi.org/10.1099/jgv.0.001639. *I was the lead author on the paper, analysed the data and wrote the drafts of the manuscript.*

Chapter 6 of this thesis is published as: Ortiz-Baez, A. S., Jaenson, T.G.T., Holmes E.C., & Pettersson J.H.-O., Wilhelmsson, P. (2022). Substantial viral and bacterial diversity at the bat–tick interface. *Microbial Genomics* 9(13), 000942. *I was the lead author on the paper, analysed the data and wrote the drafts of the manuscript.*

Chapter 7 of this thesis is published as: Ortiz-Baez, A. S., Cousins, K., Eden, J. S., Chang, W. S., Harvey, E., Pettersson, J. H.-O., Carver S., Polkinghorne A., Šlapeta J., Rose K., & Holmes, E. C. (2020). Meta-transcriptomic identification of *Trypanosoma* spp. in native wildlife species from Australia. *Parasites & vectors*, 13(1), 1-10. https://doi.org/10.1186/s13071-020-04325-6. *I was the lead author on the paper, analysed the data and wrote the drafts of the manuscript.*

In addition to the statements above, in cases where I am not the corresponding author of a published item, permission to include the published material has been granted by the corresponding author.

Ayda Susana Ortiz-Baez

13th December 2022

As supervisor for the candidature upon which this thesis is based, I can confirm that the authorship attribution statements above are correct.

Prof. Edward C. Holmes

14 December 2022

Appendix

This appendix contains four additional peer-reviewed publications that I contributed to as a co-author during my PhD candidature at the University of Sydney.

1. Cunha, M. D. P., Duarte-Neto, A. N., Pour, S. Z., **Ortiz-Baez, A. S.**, Černý, J., Pereira, B. B. S., Braconi, C. T., Ho, Y. L., Perondi, B., Sztajnbok, J., Alves, V. A. F., Dolhnikoff, M., Holmes, E. C., Saldiva, P. H. N., & Zanotto, P. M. A. (2019). Origin of the São Paulo Yellow Fever epidemic of 2017-2018 revealed through molecular epidemiological analysis of fatal cases. *Scientific reports*, *9*(1), 20418. https://doi.org/10.1038/s41598-019-56650-1. *I contributed to the bioinformatic analysis and data curation.*

2. Campbell, S. J., Ashley, W., Gil-Fernandez, M., Newsome, T. M., Di Giallonardo, F., **Ortiz-Baez, A. S.**, ... & Geoghegan, J. L. (2020). Red fox viromes in urban and rural landscapes. *Virus Evolution*, *6*(2), veaa065. https://doi.org/10.1093/ve/veaa065. *I analysed part of the data, reviewed and edited the manuscript.*

3. Geoghegan, J. L., Di Giallonardo, F., Wille, M., **Ortiz-Baez, A. S.**, Costa, V. A., Ghaly, T., ... & Holmes, E. C. (2021). Virome composition in marine fish revealed by meta-transcriptomics. *Virus evolution*, *7*(1), veab005. https://doi.org/10.1093/ve/veab005. *I analysed part of the data, reviewed, and edited the manuscript.*

4. Van Brussel, K., Mahar, J. E., **Ortiz-Baez, A. S.**, Carrai, M., Spielman, D., Boardman, W. S., ... & Holmes, E. C. (2022). Faecal virome of the Australian grey-headed flying fox from urban/suburban environments contains novel coronaviruses, retroviruses and sapoviruses. *Virology*, *576*, 42-51. https://doi.org/10.1016/j.virol.2022.09.002. *I contributed with the formal analysis, writing – review & editing, and preparation of the manuscript.*

Acknowledgements

I begin by acknowledging and paying respect to the traditional owners of the land on which I undertake my doctoral studies: the Gadigal people of the Eora Nation. It is upon their ancestral lands that the University of Sydney is built. I also pay respect to the knowledge embedded forever within the Aboriginal Custodianship of Country.

I could not have undertaken this journey without my dear family. Their endless love, wisdom and support provide me all the inspiration and courage to pursue this dream. I am also especially grateful to Jhon Sarria, his love, insights and encouragement were essential throughout this important stage.

I would like to express my deepest gratitude to my advisor, Professor Edward (Eddie!) Holmes, for his guidance, support, and generosity. Since I was an undergrad student, it was a dream to work with him, and I am very grateful for that tremendous opportunity. I am very fortunate to undertake my doctoral studies under the supervision of such a great scientist and human being. His insights and expertise were essential to my growth as a researcher, and I am appreciative of all the opportunities given to me.

I am also thankful to the past and present members in the Holmes group (The Holmies) for all their kind help and support, as well as for the cherished time we spent together throughout my PhD studies. A special thanks to Wei-Shan Chang (Vv), Callum Le Lay, Kate Van Brussel, Jackie Mahar, Sabrina Sadiq, Ashleigh Porter, and Erin Harvey, for all the advice, silliness, and friendship.

I would like to extend my sincere thanks to John Pettersson, Jackie Mahar, Justine Charon, Jan P. Buchmann, Mang Shi, John-Sebastian Eden, Cixiu Li, Wei-Shan Chang, Erin Harvey, Michelle Wille, Cadhla Firth, and Jemma L. Geoghegan for sharing with me their valuable feedback and expertise so many times, and for giving me amazing collaboration opportunities. Their

suggestions and advice helped me to refine and improve my work and their support was important to the success of my projects.

I am grateful to Ambrosio Torres, Christian Julián Arenas-Villabona, Victor Pimentel, Daniel Miranda, Cinthy Jimenez, Marielton dos Passos, Juliette Gualdron, Maria Fernanda Carreño, Ines Duran, Oscar Sanabria, and Wilson Acuña for their good cheering, advise, and friendship. Even from different corners of the world, they were always present, close, and warm. I am grateful to have such supportive and caring friends.

Lastly, I will be forever thankful to the University of Sydney and the Australian government for my scholarships and the unvaluable opportunity to study in this great Country, Australia.

List of figures

List of tables

Abstract

Wildlife harbors a substantial and largely undocumented diversity of RNA viruses and microbial life forms. RNA viruses and microbes are also arguably the most diverse and dynamic entities on Earth, infecting a wide range of hosts throughout the tree of life and thriving in multiple environments. Despite their evident importance, there are major limitations in our knowledge of the diversity, ecology, and evolution of RNA viruses and microbial communities. These gaps stem from a variety of factors, including biased sampling and the difficulty in accurately identifying highly divergent sequences through sequence similarity-based analyses alone. The implementation of meta-transcriptomic sequencing has greatly contributed to narrowing this gap. In particular, the rapid increase in the number of newly described RNA viruses over the last decade provides a glimpse of the remarkable diversity within the RNA virosphere. The central goal in this thesis was to determine the diversity of RNA viruses associated with wildlife, particularly in an Australian context. To this end I exploited cutting-edge meta-transcriptomic and bioinformatic approaches to reveal the RNA virus diversity within diverse animal taxa, tissues, and environments, with a special focus on the highly divergent "dark matter" of the virome that has largely been refractory to sequence analysis. Similarly, I used these approaches to detect targeted common microbes circulating in vertebrate and invertebrate fauna. Another important goal was to assess the diversity of RNA viruses and microbes as a cornerstone within a new eco-evolutionary framework. By doing so, this thesis encompasses multiple disciplines including virus discovery, viral host-range distributions, microbial-virus and host–parasite interactions, phylogenetic analysis, and pathogen surveillance. In sum, the research presented in this thesis expands the known RNA virosphere as well as the detection and surveillance of targeted microbes in wildlife, providing new insights into the diversity, evolution, and ecology of these agents in nature.

Chapter 1

General introduction

CHAPTER 1 General Introduction

1.1 The RNA virosphere

Viruses are likely the most abundant, diverse, and ubiquitous biological entities on Earth. The more we learn about viruses, the more challenging it becomes to fully explain their origin and diversity, as well as their impact on the biology and evolution of other life forms (Berliner et al., 2018). Estimates based on bacteriophages, which may represent the majority of existing viruses, suggest that there are $>10^{31}$ total viruses that infect bacteria alone (Breitbart & Rohwer, 2005; Cobián Güemes et al., 2016). Similarly, existing projections for other groups have estimated approximately 87 million existing eukaryotic viruses (Geoghegan & Holmes, 2017). In terms of diversity, the astronomical number of viruses within the global virosphere is represented by approximately 10^7–10^9 viruses that are sufficiently different to be considered distinct species (Koonin et al., 2022). Notably, the International Committee on Taxonomy of Viruses (ICTV) currently recognizes 10,434 virus species, which means that only a miniscule proportion of the total virosphere have been discovered and classified to date (Geoghegan & Holmes, 2017). Although these figures could easily be underestimated given the massive number of viruses found in recent studies (Edgar et al., 2022; Neri et al., 2022; Shi, Lin, et al., 2018; Shi et al., 2016), they provide a sense of the vastness of the virosphere, and the modest steps taken to expand the known viral diversity (**Figure 1.1**). This idea is central to this thesis.

It has been proposed that viruses evolved on multiple times independently (i.e. they have polyphyletic origin), with perhaps each of the six currently classified virus realms having an independent origin: the *Adnaviria, Duplodnaviria, Monodnaviria, Riboviria, Varidnaviria* and *Ribozyviria* (Koonin et al., 2022). These realms were established based on the phylogenetic analysis of virus hallmark genes (VHG) that are relatively conserved across multiple taxa. Depending on the realm, these genes encode structural or non-structural viral proteins (**Table 1**) (Koonin et al., 2020, 2022). Of particular relevance to this thesis, the realm *Riboviria* encompasses all RNA viruses, including those that utilize either an RNA-dependent RNA

polymerase (RdRp) or a reverse transcriptase (RT) for genome replication. Although the RdRp is key feature for *Riboviria*, it is still too poorly a conserved gene to accurately elucidate the deep evolutionary relationships within this realm. Conversely, the three-dimensional structure of the RdRp exhibits greater evolutionary conservation (Bruenn, 2003; Ferrer-Orta et al., 2006; te Velthuis, 2014), although there have been few attempts to reconstruct phylogenies based on structure alone.

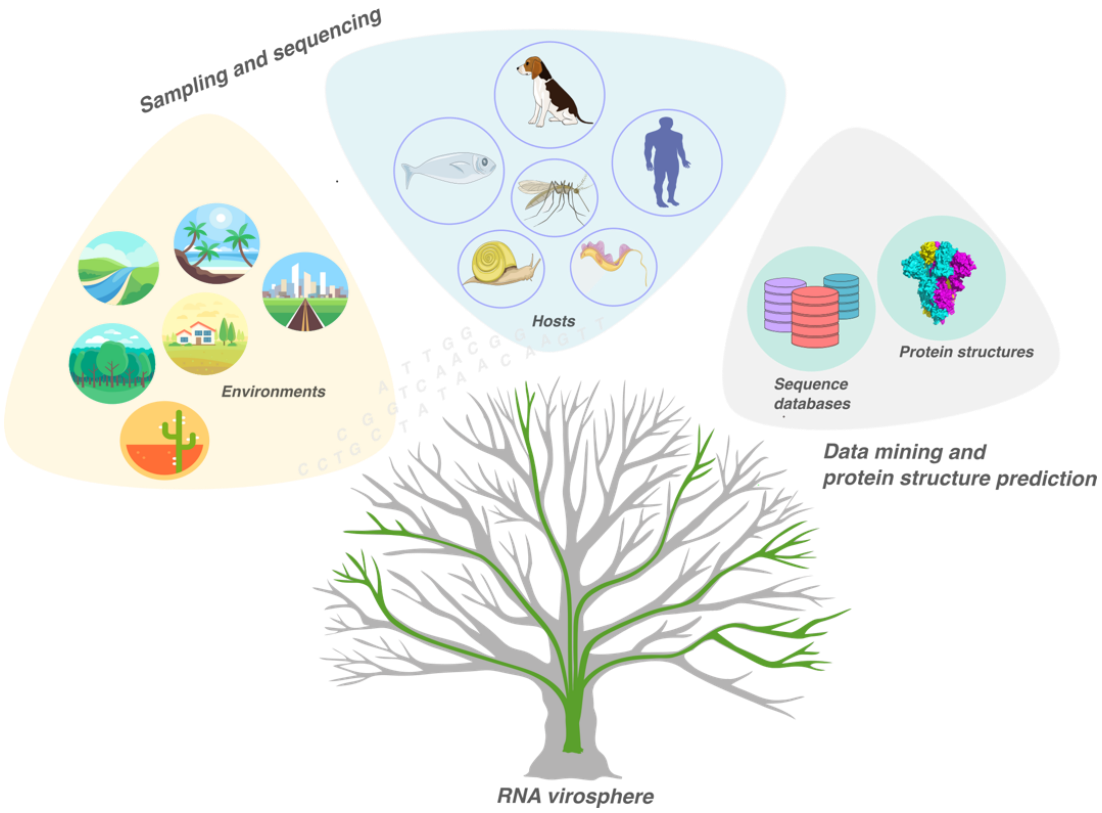


Figure 1.1 Hypothetical representation of the diversity of the RNA virosphere. The tree shown in grey represents the entire RNA virosphere, while those branches in green indicate those viruses described to date. Data sources contributing to the expansion of the known virosphere are shown on the top of tree. The tree was adapted from Zhang et al. 2019. Animal icons by Servier https://smart.servier.com/ is licensed under CC-BY 3.0 Unported https://creativecommons.org/licenses/by/3.0/

Table 1.1 List of conserved genes that characterize the current virus realms. Source: Koonin et al. 2022

Realm	Virus	Host range	Hallmark genes
Riboviria	(+) ssRNA viruses (-) ssRNA dsRNA viruses	Eukaryota, Bacteria	RNA-dependent RNA polymerase, Reverse transcriptase
Monodnaviria	ssDNA viruses dsDNA viruses	Eukaryota, Bacteria, Archaea	Endonuclease
Duplodnaviria	dsDNA tailed bacterial viruses Archaeal viruses Herpesviruses Mirusviruses	Eukaryota, Bacteria, Archaea	Major capsid protein, terminase, portal protein, capsid maturation protease
Varidnaviria	Bacterial viruses Archaeal viruses Eukaryote viruses	Eukaryota, Bacteria, Archaea	Vertical jelly-roll major capsid protein
Adnaviria	Archaeal viruses	Archaea	Major capsid protein
Ribozyviria	Hepatitis delta virus Viroid-like circular RNAs	Eukaryota	Nucleocapsid protein

The RdRp domain resembles a closed right-hand shape owing to the presence of three subdomains: the fingers, palm, and thumb. Seven conserved catalytic motifs (denoted A–G) are distributed between the fingers (motifs F–G) and palm (motifs A–E) subdomains, which are directly involved in the RNA synthesis (Jia & Gong, 2019; te Velthuis, 2014) (**Figure 1.2**). Some viruses harbour an additional motif H that is present in the thumb subdomain (Ramaswamy et al., 2022). Motifs A, B and C in the palm subdomain are generally well-conserved between viruses even at higher taxonomic levels, allowing the recent development of domain-based tools for virus discovery and classification (**Figure 1.2**) (Babaian & Edgar, 2022; Charon et al., 2022b; te Velthuis, 2014). For instance, motif C typically consists of Gly-Asp-Asp (GDD) residues, resulting in a major metal binding site in the RdRp of most RNA viruses. However, reported exceptions to this canonical composition include negative-sense single-stranded RNA viruses exhibiting GDN and SDD, and birnaviruses displaying an ADN tripeptide (Charon et al., 2022b; Gorbalenya et al., 2002).

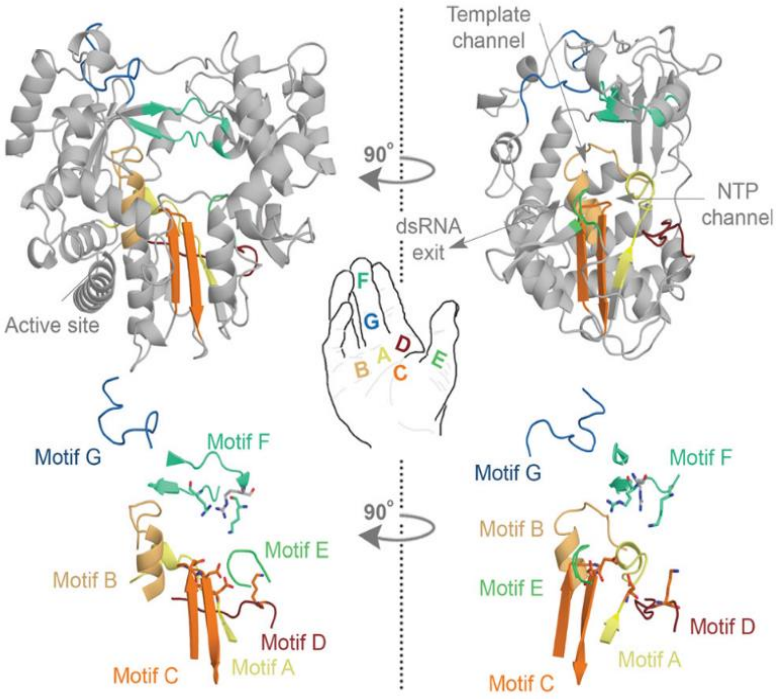


Figure 1.2 Hand-like structure of the RNA virus RNA polymerase (RdRp). Motifs (A–E) are showed in colours. Adapted from te Velthuis, 2014.

1.1.2 The expansion of known global RNA virome

RNA viruses exhibit a remarkable diversity and evolve at great rapidity due to very high mutation rates. In turn, these high rates of mutation are in large part explained by the reduced or absent proofreading activity of the RdRp, resulting in error rates that range from 10^{-6} to 10^{-4} substitutions/nucleotide/cell infection (Sanjuán et al., 2010). The interplay of ongoing mutation and natural selection has shaped this genetic variation into the current staggering diversity of RNA viruses. Indeed, RNA viruses are adapted to virtually all environments and life forms (**Figure 1.1**), although it is striking that no *bona fide* RNA viruses from Archaea have yet been

identified. While earlier research was mainly focused on human and veterinary pathogenic viruses, innovations in sequencing techniques and computational methods have allowed contemporary studies to venture into a global virome approach to better understand the diversity, evolution, and ecology of all RNA viruses in nature. For instance, in recent years ambitious studies have covered a diverse range of animal species and environments by generating and analyzing enormous amounts of sequencing data or mining massive data that is publicly available in sequence databases (Chen et al., 2021; Edgar et al., 2022; Shi et al., 2018; Shi et al., 2016). Despite these efforts, a sampling bias towards some biological groups such as mammals (Chordata) and dipterans (Arthropoda) within the Animalia remains (**Figure 1.3**) (Harvey & Holmes, 2022).

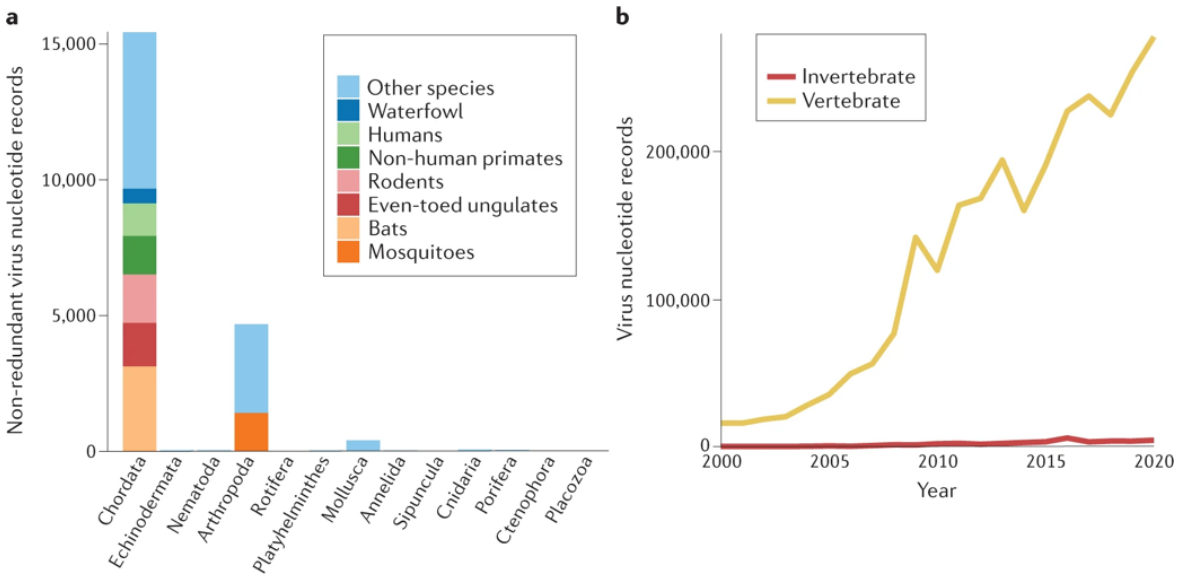


Figure 1.3 Representation of RNA viruses associated with vertebrate and invertebrate hosts in the NCBI GenBank databases. A) Distribution of recorded viruses by animal host taxa. B) Temporal trends in the record of vertebrate and invertebrate-associated viruses over the past two decades. Taken from Harvey and Holmes, 2022.

The substantial discovery of RNA viruses has also shaken their taxonomic classification and the way we think about how viruses are related to each other. Modern taxonomy stretches back to the 18th Century with Carl Linnaeus, who provided a hierarchical system and the

principles for classifying and naming the biological diversity. The current taxonomic classification of viruses reflects the traditional taxonomic ranks used for cellular organisms (i.e. kingdom, phylum, class, order, family, genus, species), in addition to realm, the highest rank in the taxonomic pyramid (Gorbalenya et al., 2020). Thus, new viruses are assigned to existing categories within each rank or lead to the creation of new taxa, reflecting the dynamism of virus taxonomy. The most recent report of the ICTV (https://ictv.global/) approved 174 taxonomic proposals, including the creation of new species, families and order categories, as well as the renaming of viruses to binomial species names (i.e. genus name followed by species epithet) (Walker et al., 2022). As part of the demarcation criteria, the taxonomic classification of viruses considers different characters including genetic similarity, genome composition, serological distances, protein structure, and hallmark genes and proteins. However, unlike the taxonomic grouping of cellular forms, virus taxonomy does not necessarily mirror the evolutionary history of a particular RNA viral group, since characters may have evolved independently or resulted from horizontal gene transfer (HGT) (Breitbart & Rohwer, 2005; Gorbalenya et al., 2020; Koonin et al., 2022), and classifications are often based on the analysis of a single gene (the RdRP) only.

The assessment of phylogenetic relationships is central to comprehend the evolutionary history of RNA viruses (Bamford et al., 2005a). In the context of virus surveillance and discovery, phylogenetic analysis provides a key tool for inferring key aspects of virus evolution and ecology. The branching pattern of a tree represents a hypothesis of how virus taxa are related, the amount of genetic change through time, and their shared ancestors (**Figure 1.4**). Indeed, the position of a particular virus in the tree can reveal information on likely host-associations or even potential sources of contamination (e.g. a virus from a mammalian sample that groups with aquatic viruses) (Cobbin et al., 2021; Shi et al., 2018). Hence, the metadata annotation of viral taxa provides a more complete phylogenetic context for newly discovered viruses, making it easier to identify evolutionary relationships as well as patterns of trait distribution (e.g. geographic location, clinical phenotype, host, sampling environment) across phylogenetic history.

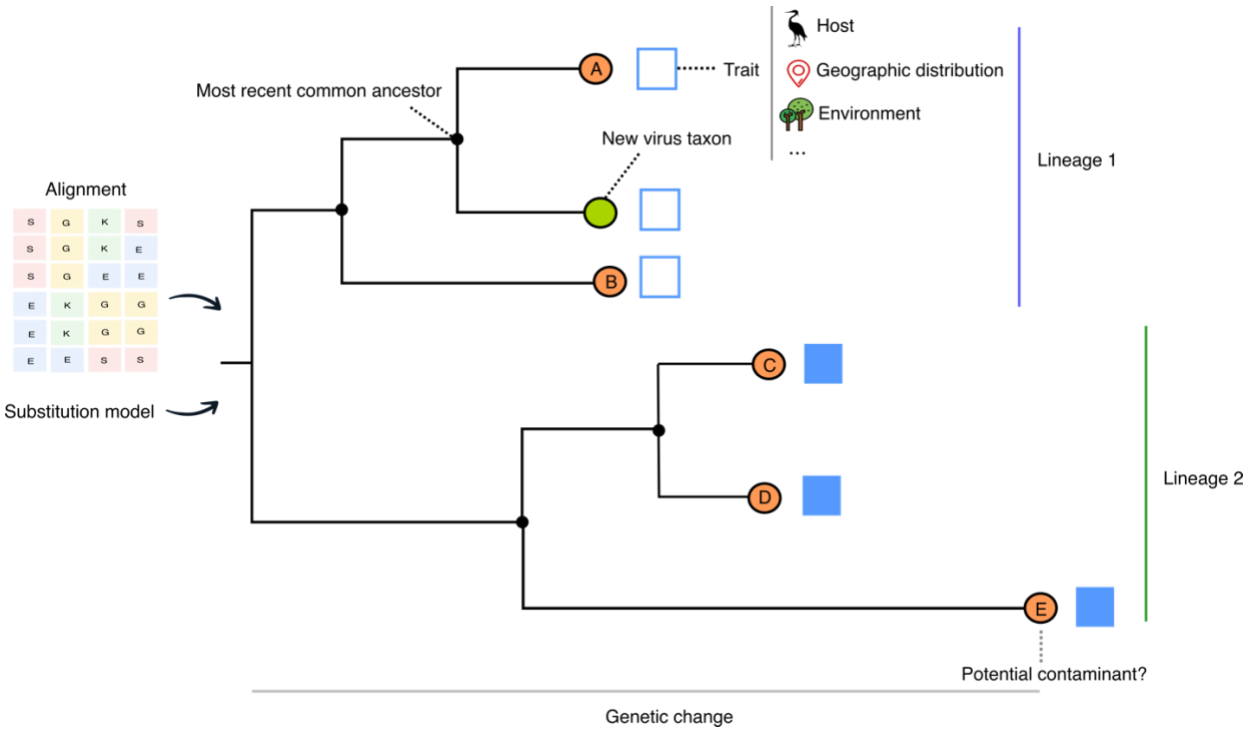


Figure 1.4 Schematic representation of a phylogenetic tree with hypothetical character traits. The phylogenetic position of the new virus taxon shows a close relationship to terminal taxa A and B within lineage 1. A hypothetical trait is annotated for terminal taxa A–F, whereas the shade of the blue square represents the distribution of the trait state throughout the tree. Long branches suggest missannotation or potential cross-contamination. Amino acid alignments are used to assess the phylogenetic relationships of divergent viruses since they are more conserved than nucleotide sequences.

In addition, evolutionary and epidemiological processes, such as the mode of transmission of a virus, can leave their stamp on the shape of virus phylogenies (Poon et al., 2013). For instance, the ladder-like phylogeny of the haemagglutinin (HA) gene of Influenza A virus (IAV) is typically associated with acute viral infections and antigenic drift (i.e., immune escape), whereas viruses associated with persistent viral infections such as the human immunodeficiency virus (HIV) result in “star-like” trees that are characterized by short-internal branches and long branches at the tips (Colijn & Plazzotta, 2018). On the other hand, phylogenies can provide important information on the occurrence of virus-host co-divergence

and host-switching by assessing tree congruence (i.e. the extent of phylogenetic mismatch). For instance, comparisons of vertebrate and virus phylogenies suggests multiple cross-species transmission of influenza-like viruses among this group of organisms (Shi et al, 2018). Phylogenetic analysis also assists in the taxonomic classification of novel viruses. As a case in point, due to its close phylogenetic relationships and genomic similarity to other bat betacoronaviruses (Andersen et al., 2020; Holmes et al., 2021), the recently discovered SARS-CoV-2 is classified within the subgenus *Sarbecovirus*, genus *Betacoronavirus* in the family *Coronaviridae*. The current nomenclature of SARS-CoV-2 variants is similarly based on phylogenetic analyses and lineage annotation (Rambaut et al., 2020).

1.1.3 Virus-host associations

The study of virus-host associations is often a challenging task in virus discovery projects. Indeed, it is estimated that over 40% of records in the NCBI lack host assignation (Cobbin et al., 2021). This, in part, is due to the remarkable ability of some RNA viruses to infect diverse host species, even across kingdoms (**Table 1**). Arthropod-borne viruses (i.e., arboviruses) represent a typical example of RNA viruses with a broad host range. These viruses can replicate in both arthropod and vertebrate hosts, displaying adaptability to the differing conditions of these cell environments (Hanley & Weaver, 2008; Shope & Meegan, 1997). Notably, this strategy has arisen independently in multiple families within *Riboviria*. Providence virus (*Tetraviridae*) provides another example of cross-species transmission. This virus has been isolated from Lepidoptera and is able to replicate in plants and mammalian cells (Jiwaji et al., 2019). Similarly, virus host ranges have been shown to be intricately dynamic over the evolutionary time scale. For instance, it has been hypothesized that plant-associated viruses from the order *Bunyavirales* have arthropod origins (German et al., 2020; Junglen, 2016). Just as it is exciting to understand the evolution of virus-host ranges, so is expanding the host range of known groups of viruses, and virus discovery projects have made great contributions in this regard (Li et al., 2015; Shi et al., 2018).

Another difficulty in assessing virus-host associations lies in the fact that viruses can infect a particular host or symbionts or dietary-associated viruses within that host. This is a common error that sometimes leads to incorrect host assignment (Cobbin et al., 2021). It is similarly difficult to assign the natural host of viruses in samples containing diverse microbial communities such as stool, water, or soil. Despite the challenges of virus-host inference based on *in silico* data, helpful approaches, including comparisons of virus composition and relative abundance within and across samples, compositional analyses (e.g. dinucleotide composition and codon usage) as well as small RNA profiling, are available and assist in revealing the natural hosts of newly discovered viruses (di Giallonardo et al., 2017; Kapoor et al., 2010; Mlotshwa et al., 2008; Webster et al., 2016a).

1.2 Diversity of viruses and microbes in wildlife

1.2.1 Surveys of microbial diversity

The discovery and early detection of microbes in wildlife can assist in the surveillance of microbial agents as well as in the monitoring and prevention of zoonotic spillovers (Artois et al., 2009; Lipkin, 2013; Woods et al., 2019). As such, the combination of wildlife surveys and genome sequencing can provide insights into the introduction of a particular microorganism in a host population, its distribution, host range and diversity (Lipkin, 2013). Despite this, we currently lack knowledge on the composition of the microbial communities associated with native wildlife as well as their impact on the ecology and evolution of their hosts. This becomes especially important when these microbial agents threaten populations of native fauna. For example, the spread the fungus *Batrachochytrium dendrobatidis* has been associated with disease outbreaks and increased mortality of amphibian populations (~ 700 species) globally (Lips, 2016; van Rooij et al., 2015). It is also hypothesized that the Maclear's rat (*Rattus macleari*), endemic to Christmas Island, was decimated to extinction by trypanosome parasites likely introduced by black rats (*Rattus rattus*) (Wyatt et al., 2008). Likewise, wildlife fauna can serve as reservoir hosts of pathogenic agents for other animals, including livestock and humans.

As a case in point, trypanosomiasis in wildlife is suspected to facilitate the infection of numerous mammalian species with *Trypanosoma cruzi* and *T. evansi* (Kasozi et al., 2021). Strikingly, while some pathogenic trypanosome species are well characterized, we do not know the host range distribution and impact of many trypanosome species on wildlife (Smith et al., 2008).

Apart from the analysis of samples derived from vertebrate animals, the survey of vector species offers a practical and non-invasive way to assess the composition of targeted microbial agents in wildlife. Accordingly, the survey of arthropods with vectorial or parasitic roles such as mosquitoes, mites, flies, fleas, lice, ticks, and midges can reveal the circulation of common and unusual infectious agents in animal fauna (**Figure 1.5**) (Cohen et al., 2017). Examples of this include the detection of tick-borne pathogenic bacteria such as *Borrelia burgdorferi* and *Rickettsia* spp., the etiological agents of Lyme disease and Spotted fever, respectively (Johnson et al., 1984; Stewart et al., 2017) (although it is important to note that *Borrelia burgdorferi* has not been detected in Australia). Although vector surveys mainly focus on the detection of microbes and viruses of public health and veterinary importance, vector surveys facilitate investigations of other components of the symbiont microbiota (Bonnet & Pollet, 2021). A well-known example is the detection of the endosymbiotic bacterium *Wolbachia* sp. in natural populations of *Drosophila simulans*, sometimes revealing beneficial effects on the fitness (i.e. survival and reproduction) of infected flies over time (Qiu et al., 2014a; Weeks et al., 2007). Similarly, the presence of *Coxiella* and *Francisella* bacteria in ticks have been associated with B vitamin and cofactor synthesis (Duron et al., 2018; Greay et al., 2018; Wu-Chuang et al., 2021). Hence, vector surveys can shed light on the microbial communities associated with both vertebrate and invertebrate hosts.

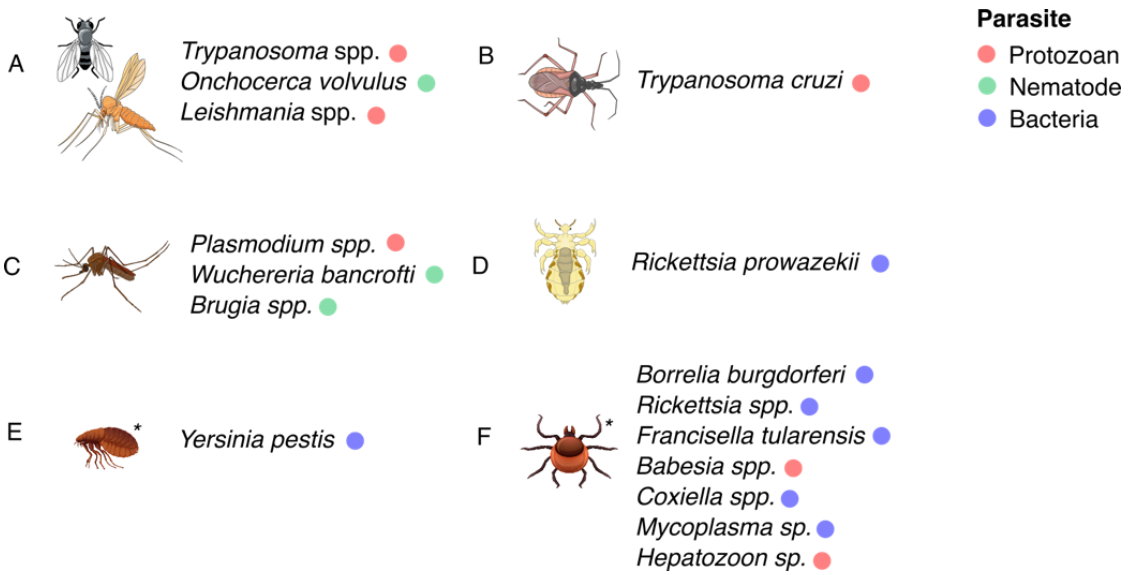


Figure 1.5 Common arthropod vectors of microbial parasites present in vertebrate animals. (A) flies, (B), triatomine bugs (C), mosquitoes (D), lice (E) fleas, (F) ticks. Examples of parasites are indicated next to each arthropod. Icons by Servier https://smart.servier.com/ is licensed under CC-BY 3.0 Unported https://creativecommons.org/licenses/by/3.0/ and Vecteezy (*) www.vecteezy.com.

1.2.2 The invertebrate-associated virome

Within the Animalia, the invertebrate phylum Arthropoda is the most diverse and abundant group on Earth. Current estimates of arthropod species richness range from 5 to 10 million (Ødegaard, 2000; Stork, 2018), thereby accounting for over 83% of all animal species. The evolutionary history of arthropods traces back to the Cambrian, more than 500 million years ago (Gould, 1994). Over this extended time period arthropods have evolved complex life cycles that involve different developmental stages, as well as adaptations to inhabit virtually all environments. Likewise, arthropods have thrived as free-living forms or by stablishing a variety of host-symbiont interactions (e.g. parasitism and mutualism) with other species, playing a central role in terrestrial and aquatic ecosystems. To the same extent that arthropods continue to defy our understanding of animal biodiversity, a growing number of studies of the arthropod RNA virome have revealed astonishing levels of diversity, leading to the discovery of novel

viruses and even higher taxa, including new viral genera and families (Käfer et al., 2019a; Li et al., 2015). For instance, recent meta-transcriptomic research on arthropods has revealed divergent virus lineages such as the *Chuviridae* family and the quenyaviruses. The *Chuviridae* (*Jingchuvirales*) was first identified in Arthropoda (Li et al., 2015), although they have also been documented in other invertebrates, including nematodes and arachnids. These negative-sense single-stranded RNA (-)ssRNA viruses exhibit diverse genome organizations, including segmented, non-segmented and circular genomes, and can also be found as endogenous viral elements in their invertebrate hosts (Dezordi et al., 2020; Wallau, 2022). Similarly, the quenyaviruses are a recently proposed family of segmented ssRNA viruses originally identified in *Drosophila* and which appear to be divergent to other RNA viruses (Obbard et al., 2020).

Although most studies of invertebrate-associated viruses have focused on species of scientific, ecological and/or socio-economic importance, such as *Drosophila melanogaster, Ixodes sp*. *Apis mellifera, and Aedes aegypti*, the expanded research on invertebrates virome has showed that RNA viruses that were once thought to be restricted to vertebrates have relatives with invertebrate hosts, including those within the *Hantaviridae*, *Orthomyxoviridae*, *Paramyxoviridae* and *Rhabdoviridae* (**Figure 1.6**) (Harvey & Holmes, 2022; Käfer et al., 2019; Shi, et al., 2018). Hence, viruses that were previously thought to be exclusively associated with mammals in fact have evolutionary histories that date back to the time of the invertebrates. In addition, numerous families that infect plants also share ancestry with arthropod viruses (Chen et al., 2019; Herath et al., 2020; Kormelink et al., 2011). Although this could point an even older evolutionary ancestry, in most cases it likely reflects the intimate interactions between plants and arthropods that enables cross-species virus transmission. Moreover, the survey of invertebrates has also led to the discovery of viruses associated with their parasites and microbiota, such as RNA viruses belonging to the *Partitiviridae*, *Totiviridae* and *Narnaviridae* viral families that infect plants, fungi, and protists (Charon et al., 2019; Urayama et al., 2022; Webster et al., 2015). From a broad perspective, the megadiversity of invertebrates gives us a notion of the vast diversity of viruses yet to be discovered and helps highlight major taxonomic gaps in sampling (**Figure 1.2**).

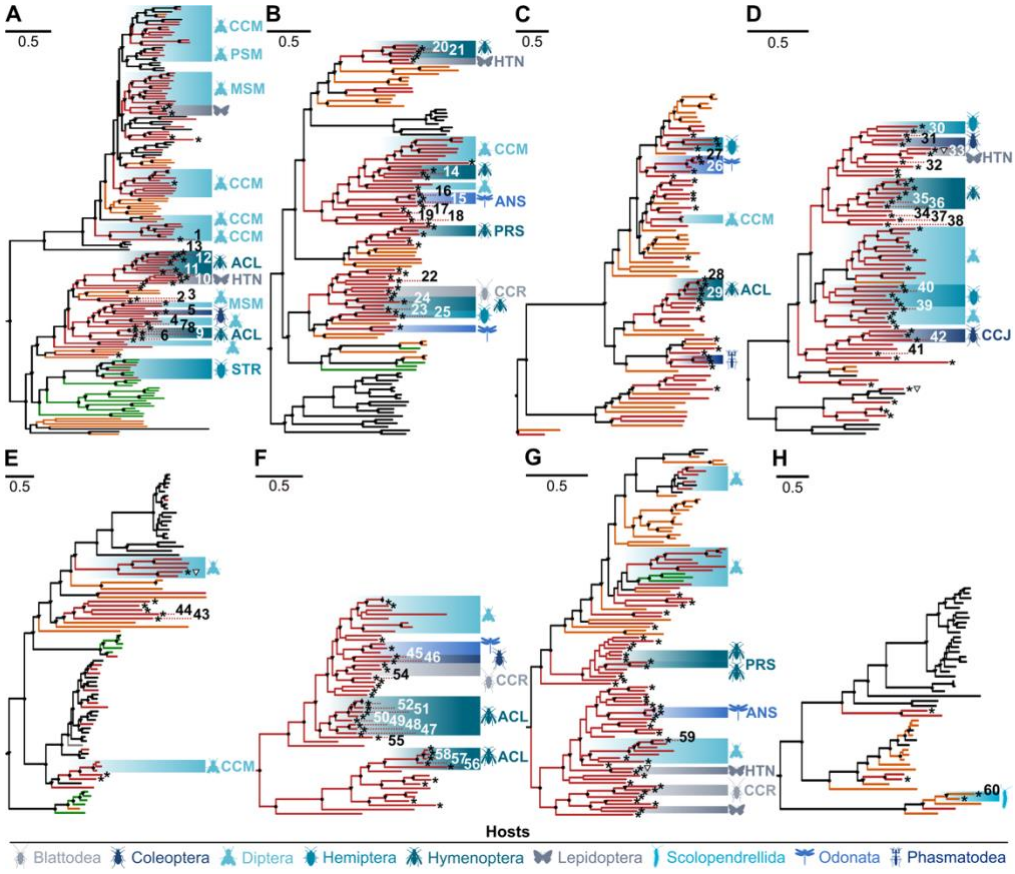


Figure 1.6 Phylogenetic trees showing virus-invertebrate associations across different RNA virus families. Branches are coloured according to the host: vertebrates (black), insects (red), other invertebrates (orange), plants (green). RNA virus phylogenies correspond to the families (A) *Rhabdoviridae*; (B) *Xinmoviridae*, *Nyamiviridae, Bornaviridae*, *Artoviridae, Lispiviridae, Paramyxoviridae, Sunviridae, Filoviridae*, and *Pneumoniviridae*; (C) *Chuviridae, Qinviridae*, and *Yueviridae*; (D) *Orthomyxoviridae*; (E) *Hantaviridae, Cruliviridae, Peribunyaviridae,* and *Fimoviridae*; (F) *Phasmaviridae*; (G) *Phenuiviridae*; (H) *Arenaviridae, Mypoviridae, Nairoviridae*, and *Wupedeviridae*. Hosts infraorders are indicated as Blattodea: Cockroaches (CCR); Coleoptera: Cucujiformia (CCJ); Diptera: Culicomorpha (CCM), Muscomorpha (MSM), Psychodomorpha (PSM); Hemiptera: Stenorrhyncha (STR); Hymenoptera:

Aculeata (ACL), Parasitica (PRS); Lepidoptera: Heteroneura (HTN); Odonata: Anisoptera (ANS). Taken from Käfer et al. 2019.

1.2.3 The vertebrate-associated virome

Vertebrates (phylum Chordata) have evolved into complex life forms with diverse morphologies, tissues, organ systems, immune mechanisms, and physiologies that have allowed their adaptation to terrestrial, aquatic, and aerial environments. Although major sampling biases for vertebrate groups such as mammals and birds has influenced our perception of the known virus diversity (**Figure 1.3**), recent large-scale surveys of other animal groups have led to the reinterpretation of the evolutionary history and ecology of vertebrate RNA viruses (Harvey & Holmes, 2022; Shi et al., 2018). Many of the newly discovered RNA viruses have seemingly accompanied the evolution of vertebrates, indicative of a complex pattern of virus-host co-divergence (i.e. parallel evolution) and cross-species virus transmission (i.e. viral host jumps) that have shaped virus phylogenetic history (Shi et al., 2018). For example, the phylogeny of the virus family *Arteriviridae* broadly reflects the evolutionary relationships of their vertebrate hosts, whereas the viruses within the *Picornaviridae* and *Coronaviridae* show more frequent viral host jumps across vertebrates (**Figure 1.7**) (Shi et al., 2018).

The identification of novel viruses in different vertebrate groups has also expanded the known host range for a variety of virus taxa. A remarkable example of this is found in the fish virome. Recent research has shown that fish harbour enormous viral diversity, including RNA viruses within the *Astroviridae*, *Coronaviridae, Filoviridae*, *Hantaviridae* and *Arenaviridae* that were previously exclusively associated with mammals, birds, and reptiles (Geoghegan et al., 2018; Shi et al., 2018). Hence, the likely age of these families can now be extended to at least the age of the Osteichthyes. Similarly, the discovery of dimarhabdoviruses and flaviviruses in fish suggest an early association with vertebrates (Geoghegan et al., 2018; Shi et al., 2018). Moreover, studies of the fish virome have revealed highly divergent viruses such as tilapia lake virus – TiLV (*Tilapinevirus tilapiae)* – which was detected in tilapia fish, giving rise to the novel family *Amnoonviridae* within the order *Articulavirares* (Bacharach et al., 2016; Eyngor et al., 2014).

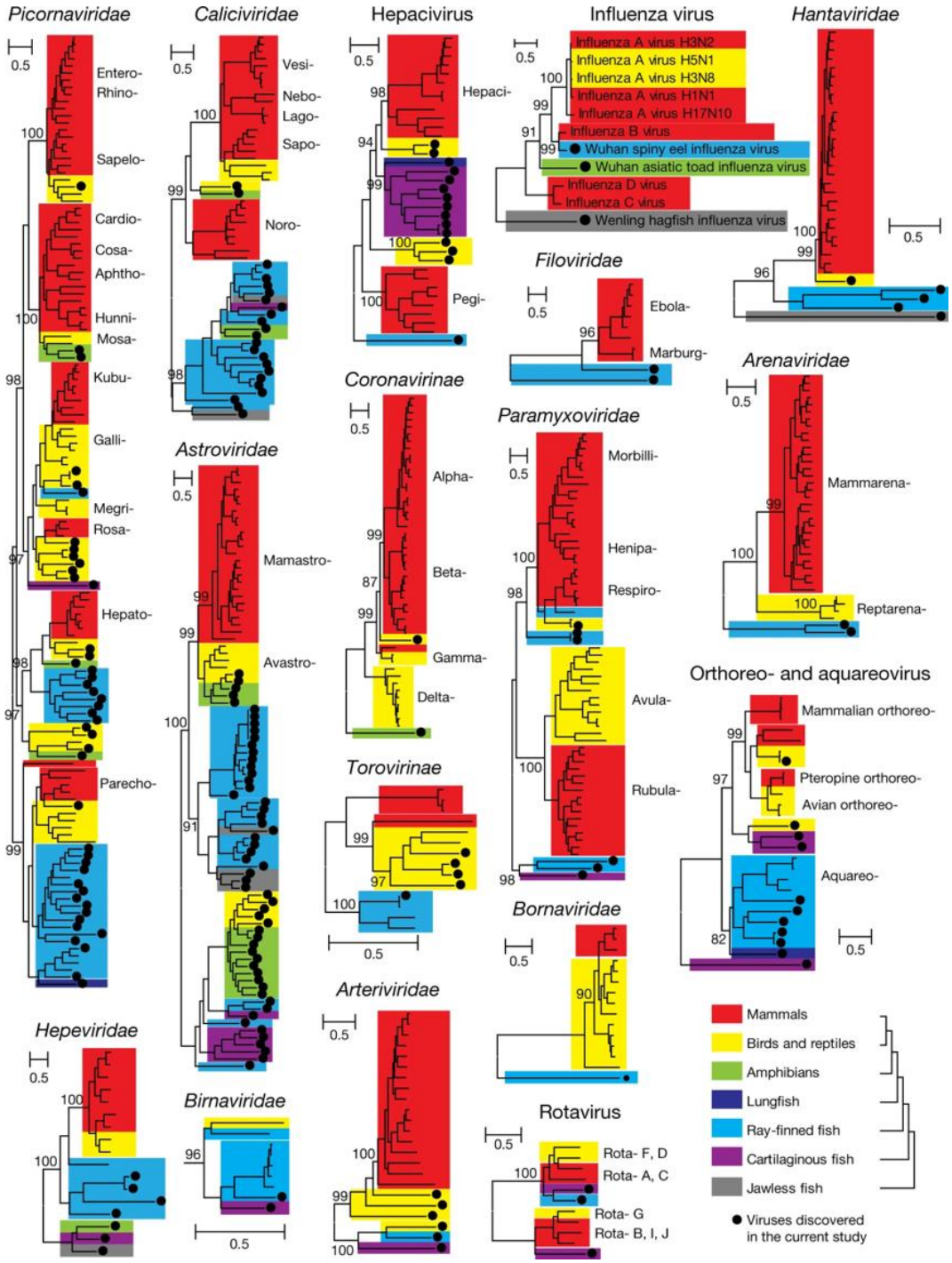

Picornaviridae
Entero-
Rhino-
Sapelo-
Cardio-
Cosa-
Aphtho-
Hunni-
Mosa-
Kubu-
Galli-
Megri-
Rosa-
Hepato-
Parecho-
Caliciviridae
Vesi-
Nebo-
Lago-
Sapo-
Noro-
Astroviridae
Mamastro-
Avastro-
Hepeviridae
Birnaviridae
Hepacivirus
Hepaci-
Pegi-
Coronavirinae
Alpha-
Beta-
Gamma-
Delta-
Torovirinae
Arteriviridae
Influenza virus
Influenza A virus H3N2
Influenza A virus H5N1
Influenza A virus H3N8
Influenza A virus H1N1
Influenza A virus H17N10
Influenza B virus
Wuhan spiny eel influenza virus
Wuhan asiatic toad influenza virus
Influenza D virus
Influenza C virus
Wenling hagfish influenza virus
Filoviridae
Ebola-
Marburg-
Paramyxoviridae
Morbilli-
Henipa-
Respiro-
Avula-
Rubula-
Bornaviridae
Rotavirus
Rota- F, D
Rota- A, C
Rota- G
Rota- B, I, J
Hantaviridae
Arenaviridae
Mammarena-
Reptarena-
Orthoreo- and aquareovirus
Mammalian orthoreo-
Pteropine orthoreo-
Avian orthoreo-
Aquareo-
Mammals
Birds and reptiles
Amphibians
Lungfish
Ray-finned fish
Cartilaginous fish
Jawless fish
Viruses discovered in the current study

Figure 1.7 Maximum likelihood phylogenetic trees showing virus-vertebrate host associations in 17 families of RNA viruses. Hosts are highlighted according to the color-codes used in the vertebrate phylogeny (right-bottom). Novel viruses within each host group are indicated with solid black circles. Virus taxonomic categories are shown above each tree. Taken from Shi et al. 2018.

Given the multicellular organization of vertebrates into tissues and organs, RNA viruses can also exhibit a broad cell tropism involving several tissues (i.e. systemic infection) or a restricted affinity to a particular tissue. This has important implications for virus discovery and the detection of RNA viruses in vertebrate hosts, as virus sampling tends to be directed to specific tissues. For example, flaviviruses such as hepatitis C virus, dengue virus, and yellow fever virus show a strong hepatotropism, whereas west Nile, japanese encephalitis virus and zika virus have a high tropism for cells of the nervous system (Bailey & Diamond, 2022; Best, 2016). In contrast, enteroviruses (*Picornaviridae*) and noroviruses (*Caliciviridae*) are commonly found in the gut (Muehlenbachs et al., 2015; Shi 2018; Wobus, 2018).

Wildlife vertebrates can also serve as important reservoirs for RNA viruses. Non-human primates are suspected to be the primarily reservoirs of arboviruses such as yellow fever virus and dengue virus in zoonotic cycles (Kuno et al., 2017). Common vertebrate reservoirs often include rodents, birds, and bats (Calisher et al., 2006; Causey & Edwards, 2008; Luis et al., 2013). For example, aquatic wild birds serve as natural reservoirs of Influenza A virus, whereas rabies virus is associated with bats as reservoir hosts (Calisher et al., 2006; Webby & Webster, 2001). Nonetheless, the reservoirs for most RNA viruses are unknown. This is the case for SARS-Cov-2, which is suspected to have a mammalian reservoir based on its close relatives found in bats (particularly horseshoe bats) and pangolins (Andersen et al., 2020; Zhou et al., 2021). Clearly, vertebrate hosts play a key role in the maintenance, emergence, and diversification of RNA viruses, although the full host for most viruses is clearly unknown (as is also the case for SARS-CoV-2). Exploring the vertebrate virome therefore remains a major task and an active area of research.

1.3 Virus and microbial discovery in the next-generation sequencing era

1.3.1 An overview of next-generation sequencing

Over the last two decades the development of sequencing technologies and associated computational tools have revolutionized virus and microbial discovery. Precursor technologies include Sanger sequencing, a first generation technology that initiated the field of nucleic acid sequencing and became the benchmark for research and clinical diagnostics (Grada & Weinbrecht, 2013; Sanger et al., 1977). Despite its historical importance and great utility, Sanger sequencing has limitations, including low throughput, that it is not cost effective, and has a reduced sensitivity to detect low frequent variants (**Table 1.2**). Ambitious initiatives such as the Human Genome Project and the 1000 Genome project highlighted the limitations of conventional sequencing and triggered important technological advances that promoted the development of next-generation sequencing (NGS) technologies (Morey et al., 2013).

Table 1.2 Comparison of Sanger sequencing and massively parallel sequencing (Next generation sequencing) technologies.

Feature	Sanger sequencing	Next generation sequencing
Method	Chain-termination	Variable. e.g., sequencing-by-synthesis, sequencing-by-ligation, long read sequencing[1]
Throughput	Low	High
Cost	Cost-time effective for a low number of samples	Cost-time effective for a large number of samples
Sensitivity	Low sensitivity to detect low frequency variants.	High sensitivity to detect low frequency variants
Per-base accuracy	High (99 – 99.99%)	High (99 – 99.999%)
Scalability	Low	High
Sample enrichment	Cloning vectors, enrichment PCR	Optional (e.g., hybridization-based, and PCR-based target/Amplicon Sequencing methods, circularisation, etc.)
Read length	Long reads up to ~1,000 bp	Variable (150 bp – 2.3 Mb)
Workflow	Simple data analysis	Complex pipelines and require bioinformatic training.

[1]Source: Goodwin et al., 2016

First introduced in 2004, NGS is now classified into second and third generation technologies depending on the length of the sequences generated. Second generation technologies are represented by platforms such as Illumina, 454 Life Sciences, DNBSEQ and SOLiD which focus on short-read sequencing (< 1000 bp), whereas third generation technologies are dominated by platforms Pacific Biosciences (PacBio) and Oxford Nanopore Technologies that produce far longer read lengths (up to > 800 Kb) (Hu et al., 2021; Pollard et al., 2018). The preference of one technology over another is determined by the intrinsic advantages and limitations of the sequencing approach associated to each NGS platform. For instance, long-read NGS performs are superior to short-read NGS in terms of spanning longer regions and enabling the high-resolution sequencing of transposable elements, repetitive regions, and complex structural variants (Metzker, 2009; Pollard et al., 2018; Shahid & Slotkin, 2020). Moreover, recent innovations in this field have facilitated the development of portable devices such as the Oxford Nanopore MinION sequencer that have played an important role in pathogen surveillance in clinical and fieldwork settings (de Vries et al., 2022; Greninger et al., 2015; Quick et al., 2016). Conversely, short-read sequencing provides a powerful approach to characterize whole-genomes with high depth of coverage, throughput, and accuracy (Goodwin et al., 2016). Because short-read NGS has a wider applicability in *de novo* sequence assembly and virus/microbial discovery in metagenomics, and was used extensively in this thesis, I will describe short-read NGS in more detail throughout this section.

In general, short-read NGS relies upon three main steps: (*i*) Sampling processing, (*ii*) library preparation and (*iii*) sequencing (**Figure 1.8**). Samples might include whole organisms, blood, *in vitro* cell culture, water, urine, stool, as well as animal and plant tissues. Sample quality is critical for efficient sequencing. As a consequence, samples should be as fresh as possible and stored at very low temperatures (e.g. -80 ºC) to prevent degradation of RNA by RNA ribonucleases before use. During sampling processing, chemical or mechanical methods are used to disrupt the integrity of tissues. The genetic material is then extracted and separated from potential chemical contaminants that might interfere with downstream stages. This step is achieved using a variety of methods and reagents as well as protocols and kits tailored to

process specific biological and environmental samples. Once the RNA is isolated, it is reverse transcribed into cDNA which is the input for library preparation. This latter involves the fragmentation of the genetic material and the ligation of adapter sequences at each end of molecules (Goodwin et al., 2016). Finally, during sequencing the molecules are clonally amplified (i.e., template enrichment) in a flowcell or bead-based system, and the order of nucleotides in the DNA templates is decoded. Depending on the platform, it is possible to sequence template fragments in one (single-end sequencing) or both directions (paired-end sequencing) by adding indexes to the ends (Hu et al., 2021; Metzker, 2009). The incorporation of nucleotides in the new template is signaled using methods such as fluorescence imaging (Goodwin et al., 2016; Metzker, 2009). As a result, multiple samples can be analyzed, and vast amounts of raw sequence data are generated in parallel. Notably, the constant improvements to library preparation kits contributes to optimize the sequencing process by reducing costs, the number of steps and time involved in laboratory procedures.

Figure 1.8 Overview of common steps in meta-transcriptomics analyses for virus and microbial discovery. Icons by Servier https://smart.servier.com/ is licensed under CC-BY 3.0 Unported https://creativecommons.org/licenses/by/3.0/; Illumina_miseq icon by DBCLS https://togotv.dbcls.jp/en/pics.html is licensed under CC-BY 4.0 Unported https://creativecommons.org/licenses/by/4.0/.

The capabilities of short-read sequencing have also been expanded to the field of RNA research (Stark et al., 2019). RNA sequencing (RNA-seq) or meta-transcriptomics (when total RNA is sequenced) enables the sequencing of coding and non-coding RNA. It therefore provides a convenient means to characterize and quantify the whole repertoire of RNA molecules in a sample, including the host transcriptome, but also the RNA from microbial parasites as well as RNA viruses and DNA viruses that are expressing in the targeted host (Shi et al., 2018b). This technique has opened a window to explore the virosphere and achieve a better sense of its composition and scale in nature (Obbard, 2018). In this context, features and technological developments of RNA-seq such as RNA ribosomal (rRNA) depletion during library preparation, RNA target-enrichment, small RNAs profiling as well as detection of both sense and antisense transcripts (Ozsolak & Milos, 2010) have facilitated the capture of key information on the quantitative assessment and composition of RNA viruses, genome polarity, and even virus-host associations via antiviral response (Batson et al., 2021; Obbard et al., 2020; Shi et al., 2018). However, some methodological challenges remain regarding its implementation, relatively high-cost (particularly for low-income countries), biases introduced by cDNA synthesis and template switching, detection and coverage of low-abundance variants (Han et al., 2015; Ozsolak & Milos, 2010). For microbial and virus discovery, further challenges are also present in the form of computational analyses. These will be discussed in the next section (Cobbin et al., 2021; Harvey & Holmes, 2022).

Overall, the growing access and affordability of NGS has enabled its implementation into various research and industry fields. Indeed, the pressure imposed by the COVID-19 pandemic on coronavirus surveillance since 2020 has greatly boosted the use and improvement of NGS technologies, and it is likely only matter of time before NGS becomes an accessible tool in

public health and veterinary institutions globally. Aside from virus and microbial discovery, some applications of NGS include transcriptome profiling, whole-genome sequencing (WGS), pathogen detection and population/evolutionary genetics. These have been reviewed in detail elsewhere (Kulski, 2016; Metzker, 2009; Morey et al., 2013; Reuter et al., 2015). Thus, NGS will continue to revolutionize the way we approach the study of life, and the questions that can be addressed through genetically screening diverse living forms and their environments (**Figure 1.1**).

1.3.2 Meta-transcriptomic data analysis

The development of NGS in molecular biology has triggered important findings accompanied by the generation of millions of sequences stored at public databases (Reuter et al., 2015). Similarly, the emergence of NGS has impacted virology research and led to the implementation of viral bioinformatics analysis as an alternative, but compatible, approach to traditional *in vitro* experimentation. Indeed, the relevance and importance of computational tools for analyzing and interpreting the large volume of virus sequencing data is becoming increasingly evident (**Figure 1.9**). In this respect, the implementation of meta-transcriptomic methods has led to the characterization of known, unknown and uncultured viral communities, in turn accelerating the pace of virus discovery and the exploration of the RNA virosphere (Obbard, 2018; Shi et al., 2018b).

There is a plethora of bioinformatic pipelines available for the analysis of sequencing data derived from meta-transcriptomics (Ho & Tzanetakis, 2014; Kalantar et al., 2020; Neri et al., 2022; Zhao et al., 2017). Since the analysis of metagenomic samples can be challenging, pipelines can be tailored for specific applications depending on the research question, virus type and sample characteristics. For example, virus discovery from clinical and environmental samples might require different steps to filter out host-derived sequences or database selection. Despite variations in methodological approaches, the use of defined pipelines provides a logical order of steps to guide the process of virus discovery from metagenomic

datasets to minimize artifacts and increases reproducibility of results (**Figure 1.8**). Generally, the standard workflow of meta-transcriptomic analyses starts with the quality control of raw read data (*e.g.* N and GC content, adaptor removal, quality-trimming) prior to the assembly of the remaining reads into contigs. Poor quality reads might act as confounding factors during the assembly process, in which an inadequate pre-processing of reads might strongly bias downstream steps (Andrews, 2010; Cantalupo & Pipas, 2019).

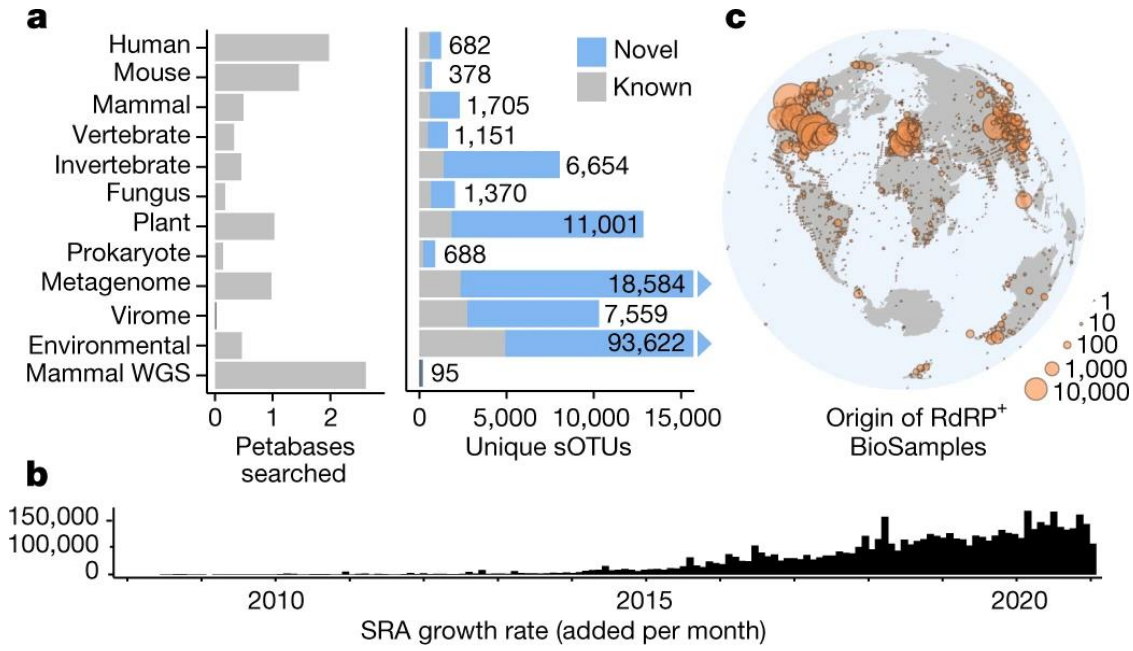


Figure 1.9 Distribution of sequencing data in the Sequence Read Archive (SRA) database. The number of available SRAs is displayed by (A) sample type, (B) release date, and (C) geography. Taken from Edgar et al. 2022 (https://serratus.io).

Because host ribosomal RNA (rRNA) sequences are often detected in metagenomics data, even when ribosomal depletion methods are implemented during library preparation, additional steps might be required to filter out host-derived sequences (*i.e. Bacteria*, *Eukaryota* and *Archaea* hosts) using such tools as the SILVA and RFAM databases (Griffiths-Jones et al., 2003; Quast et al., 2013). The major advantage of this strategy is that it increases the relative signal of viruses compared to background noise, which impact estimates of virus abundance,

and reduces computational load. However, the accidental removal of viral sequences could also impact sequence assembly and virus detection.

Sequence assembly is another critical step in metagenomic studies, especially as genetic material is commonly derived from multiple hosts in virological studies. Metagenome assemblers achieve this challenging task through two approaches: reference-based and *de novo* assembly. During reference-based assembly, reads are mapped to a guide sequence based on their similarity to a specific target. Although useful, this represents a limitation for the detection of unknown or unrelated sequences. In contrast, *de novo* assembly is a reference-free approach in which contigs are assembled using graph strategies such as the de Bruijn graph and Overlap-Layout-Consensus algorithms (Behizadi et al., 2022; Cantalupo & Pipas, 2019; Hölzer & Marz, 2019). Although *de novo* assembly enhances the discovery of uncharacterized sequences, including novel viruses, it is a computationally demanding process and assemblers often perform poorly on repetitive regions, high redundancy data and low expressed transcripts (*i.e.* rare variants) that might induce chimera formation or partial assemblies (Freedman et al., 2021; Liao et al., 2019). However, it is always possible to combine both approaches as well as assembly tools to increase accuracy and extend the length of assemblies (Cantalupo & Pipas, 2019; Hölzer & Marz, 2019).

Virus discovery studies by meta-transcriptomics largely rely on the comparison of assembly contigs against available sequences in public databases (**Table 1.3**). Sequence similarity searches are performed by programs (*e.g.* blastp, blastn, blastx, MegaBLAST, DIAMOND) that align query contigs to known sequences stored in various databases such as the NCBI nucleotide and non-redundant protein databases (**Table 1.3**) (Buchfink et al., 2015). The expected output is a collection of hits ranked by their similarity, bit score or expectation value (e-value); the latter number confers statistical significance to matches (*i.e.* pairwise alignments) and is influenced by the size of the database and the length of the query. An e-value less than 1 indicates that a particular hit is less likely to occur by chance alone given the size of the database. Thus, the lower the e-value the more likely to attribute “significance” to a hit

(Wheeler & Bhagwat, 2007). Although significant similarity might be related to shared homology, it is misleading to infer that similar sequences are always homologous, since similarity might result from convergent evolution. The opposite scenario is also possible, and two homologous sequences might display limited similarity due to the accumulation of changes since their time of divergence (Pearson, 2013).

Table 1.3 List of common databases used in meta-transcriptomics analyses for microbial and virus discovery.

Database	Data type	Search database tools	Link
NCBI/nt	Nucleotide sequences	GenBank, Basic local alignment search tool, blastn, megablast	https://www.ncbi.nlm.nih.gov/genbank/
NCBI/nr	Protein sequences	GenBank, Basic local alignment search tool, blastp, blastx	https://www.ncbi.nlm.nih.gov/genbank/
SRA	Raw reads	SRA Toolkit, Magic-BLAST	https://www.ncbi.nlm.nih.gov/sra
CDD	Protein domains	*rpsblast, rpstblastn,* InterProScan (Batch) CD-Search	https://www.ncbi.nlm.nih.gov/Structure/cdd/cdd.shtml
Pfam	Protein domains and protein family models	InterProScan, HMMER, HHblits	https://www.ebi.ac.uk/interpro/
RCSB PDB	Protein and nucleic acids 3D structures	PDB web portal, HMMER, HHpred	https://www.rcsb.org/
AlphaFoldDB	Protein Structure predictions	PDB web portal, AlphaFold web portal, HMMER	https://alphafold.ebi.ac.uk/
RefSeq	Reference nucleotide and protein sequences	GenBank, Basic local alignment search tool, blastp, blastx, blastn,	https://www.ncbi.nlm.nih.gov/refseq/

		Entrez	
TSA	Transcriptome data	tblastn	https://www.ncbi.nlm.nih.gov/genbank/
UniProt	Protein sequences and functional information	UniProt web portal, InterProScan, HMMER, HHpred, HHblits	https://www.uniprot.org/
SILVA	Ribosomal RNA (rRNA) sequences: small (16S/18S, SSU) and large subunit (23S/28S, LSU)	SortMeRNA	https://www.arb-silva.de
NCBI Taxonomy	Taxonomic data	Taxonomy browser, Entrez	https://www.ncbi.nlm.nih.gov/Taxonomy/Browser/wwwtax.cgi
UniVec	non-redundant vector database	Blastn, VecScreen program	https://www.ncbi.nlm.nih.gov/tools/vecscree/ ftp://ftp.ncbi.nlm.nih.gov/pub/UniVec/

Traditionally, comparisons between virus queries and databases are performed at the nucleotide or protein sequence level. Different search algorithms scan databases in a variety of ways depending on the input, reference database, and the desired sensitivity (**Table 1.3)**. For example, nucleotide-to-nucleotide BLAST searches (blastn) are particularly useful for identifying viruses that are moderately/closely related (>40% id) to known viral sequences in nucleotide-based databases (Pearson, 2013). In contrast, translated nucleotide searches (blastx) compare all six reading frames of a query sequence against a protein database, enabling a more sensitive approach for the detection of distantly related viruses (>30% id). Default e-values during similarity BLAST searches are commonly set to 10. Setting more relaxed or stringent e-values can also be used to produce different levels of sensitivity, although this might also lead to a greater number of false positives.

To deal with the difficulty of detecting viral sequences that share marginal similarity with known viruses (< 20% identity), alternative approaches have been adopted including structure prediction and hidden Markov model-based profiles (HMM-profiles). These are discussed in the next section (Bigot et al., 2019; Charon et al., 2022; Kelley et al., 2015; Söding, 2005). Similarly, within some biological groups, other methods such as virus-derived small RNA profiling and CRISPR spacers searching have gained popularity for predicting viral sequences and virus-host associations (Neri et al., 2022; Singh et al., 2012; Snyder et al., 2010; Webster et al., 2015).

Prediction of open reading frames (ORFs) from putative viral sequences are used for protein annotation by comparing these against databases such as Pfam and CCD (**Table 1.3**) (Marchler-Bauer et al., 2011; Mistry et al., 2021). These resources classify protein domain families, providing additional evidence on the genome assembly completeness, functionality, virus-host cell interactions, and taxonomic identity of viral contigs (Chen et al., 2012; Sobhy, 2016). For instance, the identification of conserved virus motifs and domains in RNA-dependent RNA polymerases (RdRp) and capsids can assist the classification of RNA virus orthologous detected through meta-transcriptomics (Bramley et al., 2020; Simmonds, 2015; Wolf et al., 2020).

Contamination is a recurrent peril in metagenomic analyses, and in most gene sequencing applications. The sources of viral contamination are multiple and diverse. They can come from reagents and controls as well as be introduced in all steps associated with sample handling, including sample collection and processing, library preparation and sequencing (Cobbin et al., 2021; Holmes, 2019; Porter et al., 2021). Circular Rep-Encoding Single-Stranded (CRESS) DNA viruses, circoviruses, and other single-stranded DNA (ssDNA) viruses are well-documented examples of viral contaminants in reagents (Holmes, 2019; Porter et al., 2021). Similarly, cross-contamination between samples can occur due to misassignment of reads in pooled libraries during sequencing. This phenomenon is referred to as index-hopping, in which free adapters from multiplexed libraries are swapped during library preparation, resulting in

false positive detection of sequencing reads among samples (Guenay-Greunke et al., 2021; van der Valk et al., 2020). To mitigate the occurrence of index-hopping, it is advisable the use of unique dual indexing for pooled libraries, storage of libraries and pools at low temperatures (T ≤ –20ºC), and pooling of samples with similar expression profiles (Costello et al., 2018; Guenay-Greunke et al., 2021; Illumina, 2017). Index-hopped strands can also be detected computationally by comparing the relative abundance of assemblies with the (average) expected rates of index-hopping associated with each sequencing platform. For example, abundance values below 0.1% of the highest count for a particular viral assembly among libraries are often assumed as index-hopping artifacts in virus discovery studies (Illumina, 2017, 2022; Le Lay et al., 2020; Wenqiang et al., 2022).

Finally, aside from the profiling of the viral composition present in metagenomic data sets, abundance quantification is an essential component in determining virus diversity. There are several expression units to assess the RNA abundance (Bedre, 2022; Corchete et al., 2020; Tarazona et al., 2011). Such metrics might consider such aspects as the number of mapped reads, sequencing depth, length of the target sequence, and library size (Bedre, 2022; Tarazona et al., 2011). However, abundance assessment across samples is expected to perform poorly when libraries are generated under different experimental conditions and sequencing protocols (Zhao et al., 2020). Further, variation in RNA concentration and rRNA levels between samples impact the accuracy of abundance estimates (Zhao et al., 2020). To avoid misinterpretation regarding fluctuations in abundance levels across libraries, virus abundance estimates are compared with those of stably expressed host genes such as the cytochrome c oxidase subunit I (*COX1*), 12S rRNA (*12S*) and 16S rRNA (*16S*), thereby providing a reference for determining expression levels within and between libraries. Hence, abundance quantification not only sheds light on the relative contribution of a virus of interest to the RNA repertoire within a library, but also provides insights into transcript expression levels, contaminant viral sequences (*e.g.* index-hopping), and potential host-associations (Cobbin et al., 2021; Geoghegan et al., 2021; Pettersson et al., 2019; Shi et al., 2017; Wille et al., 2018).

Overall, NGS data analysis workflows are dynamic and continuing to evolve in response to the challenges imposed by research questions and sequencing data. Clearly, NGS has accelerated the pace of virus discovery and previous gold standard approaches such as Sanger sequencing and PCR now play an assisting role in verifying metagenomic findings.

1.3.4 NGS discovery analyses and host associations

Since metagenomic sequencing enables the characterization of the entire collection of nucleic acids present in one sample, this opens a window of opportunity for parallel viral and microbial discovery. For instance, gut metagenomic samples might comprise sequences from the animal host, diet components, as well as host protozoans, fungi, bacteria, and viruses (**Figure 1.10**). This makes the prediction of virus-host associations a challenging process, particularly for viruses that infect a wide range of hosts, including those from different taxonomic domains. As a specific case in point, viruses within the family *Totiviridae* have been associated with both fungal and protozoan hosts (Koonin et al., 2015). Similarly, viruses belonging to the families *Flaviviridae* and *Nyamiviridae* can infect invertebrate and vertebrate hosts (Dietzgen et al., 2021; Shi et al., 2016a). Horizontal virus transfer and gene module shuffling are plausible explanations for the host range diversification observed in some virus families (Dolja & Koonin, 2018; Koonin et al., 2015).

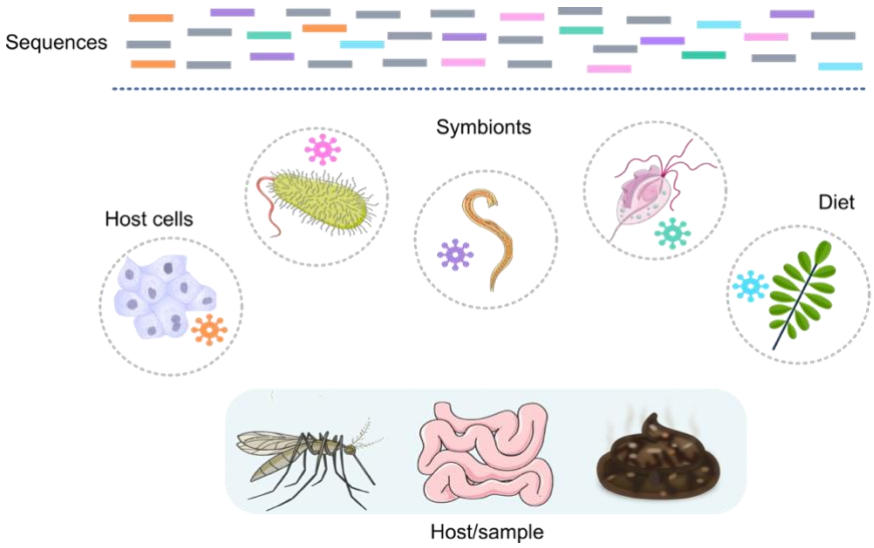


Figure 1.10 Characterization of common components present in meta-transcriptomics libraries. Libraries might include sequences from the host, co-infecting bacteria, protozoans and nematodes, diet and viruses. Viruses and viral sequences are indicated with different colors. Icons by Servier https://smart.servier.com/ is licensed under CC-BY 3.0 Unported https://creativecommons.org/licenses/by/3.0/.

Because of the ubiquity of viruses and their dependance on life forms, sample profiling is central to pinpoint potential hosts. Taxonomic sequence classifiers make use of public sequence databases to determine the composition of metagenomic samples, providing an overview of the spectrum of possible hosts for a target virus (Kim et al., 2016; Marcelino et al., 2020; Wood & Salzberg, 2014). Although sample profiling narrows the search for potential virus-host associations, it is still a vague and limited means to establish definitive associations (Dolja & Koonin, 2018). Supportive evidence can come from examining the relative abundance or the host-associations of close relatives in a phylogenetic tree (**Figure 1.4**). However, more accurate methods include virus isolation, and the use of small RNAs in invertebrates, as well as CRISPR spacers in bacteria and archaea (Freije & Sabeti, 2021; Mull et al., 2022; Obbard et al., 2020; Shmakov et al., 2020; Webster et al., 2015).

Along with the massive detection of RNA viruses, NGS has also enabled the characterization and discovery of microbial communities present in samples from multicellular hosts, although to a lesser extent (Edgar et al., 2022; Fraser et al., 2000; Shi et al., 2016) (**Figure 1.9**). Unlike most RNA viruses, microorganisms exhibit larger and more complex genomes, which are characterized by numerous genes and regulatory regions involved in metabolic and ecological processes. The rapid advancement in comparative and functional genomics has transformed the understanding of microbial genomes and genotype-phenotype associations (Kobras et al., 2021). However, the characterization of microbial diversity is primality based on targeting molecular markers in both mitochondrial and nuclear compartments (Burki et al., 2021; Mitreva, 2017; Obiol et al., 2020; Roux et al., 2011). Specifically, hallmark genes, such as 16S rRNA in prokaryotes and the small subunit 18S rRNA (SSU) in eukaryotes, are widely used

for taxonomic identification and phylogenetic analysis. For instance, sequence analysis of 16S RNA have shown to be effective for assessing the composition of bacterial communities in microbiome studies (Johnson et al., 2019; Mitreva, 2017; Tran et al., 2017), and diagnosing known pathogens in clinical samples (Aggarwal et al., 2020; Woo et al., 2008).

Conveniently, the detection of marker genes during bulk RNA sequencing also enables the surveillance and discovery of potential pathogenic organisms in wildlife (Ko et al., 2022). Therefore, through a meta-transcriptomic pipeline it is possible to investigate the presence of targeted microorganisms across different host species and environments (Doyle et al., 2017; Ko et al., 2022). Although molecular surveillance of pathogens in wildlife has predominantly focused on characterizing bacteria and viruses (Chang et al., 2021; Khoo et al., 2016; Lv et al., 2018; Qiu et al., 2014; Shi et al., 2017), neglected groups such as fungi and protists are increasingly gaining terrain (Burki et al., 2021; Caron et al., 2016). For example, the application of NGS methods combined with data mining to study kinetoplastid protists led to the discovery of free-living forms in the Prokinetoplastina subsclass, providing preliminary insights into the evolution of endosymbiosis, parasitism, and associated characters in Kinetoplastea (Tikhonenkov et al., 2021). Furthermore, meta-transcriptomic data have also revealed the presence of trypanosomatids (e.g. *Trypanosoma*, *Blechomonas* and *Leptomonas*) in invertebrate hosts across virome studies in Australia (Gofton et al., 2022; Harvey et al., 2019; Harvey et al., 2019a; Shi et al., 2017). However, there are gaps in the current knowledge about the distribution, ecology, and evolution of trypanosomatids in wildlife vertebrate populations. This research topic is addressed in this dissertation in Chapter 6.

1.4 Profiles and protein structure prediction to detect divergent viral sequences

As noted above, primary sequence similarity-based searches rely on the comparison of query sequences with known sequences available in public databases. This approach exploits the shared similarity between sequences, which is detected through the interplay between algorithms and global and local alignments. However, this approach cannot detect sequences

that share limited similarity with those already present in sequence databases. The development of sequence profile methods has brought new opportunities for the detection of distantly related sequences and, in turn, the discovery of divergent viruses. A profile can be defined as the quantitative description of a protein alignment (**Figure 1.11**) (Gribskov et al., 1987). In practice, this involves the quantification of the relative frequency for each amino acid residue in a given position along the alignment, which is summarized in a position-specific scoring matrix (Sander & Schneider, 1991; Thompson et al., 2008). Likewise, the use of probabilistic inference models such as profile hidden Markov models (profile HMMs) has enabled the capture of information on the occurrence of gaps, insertions, and substitutions, providing a more precise and detailed description of the variability of each residue position in a multiple sequence alignment (**Figure 1.11**) (Böer, 2016; Eddy, 1995, 1998).

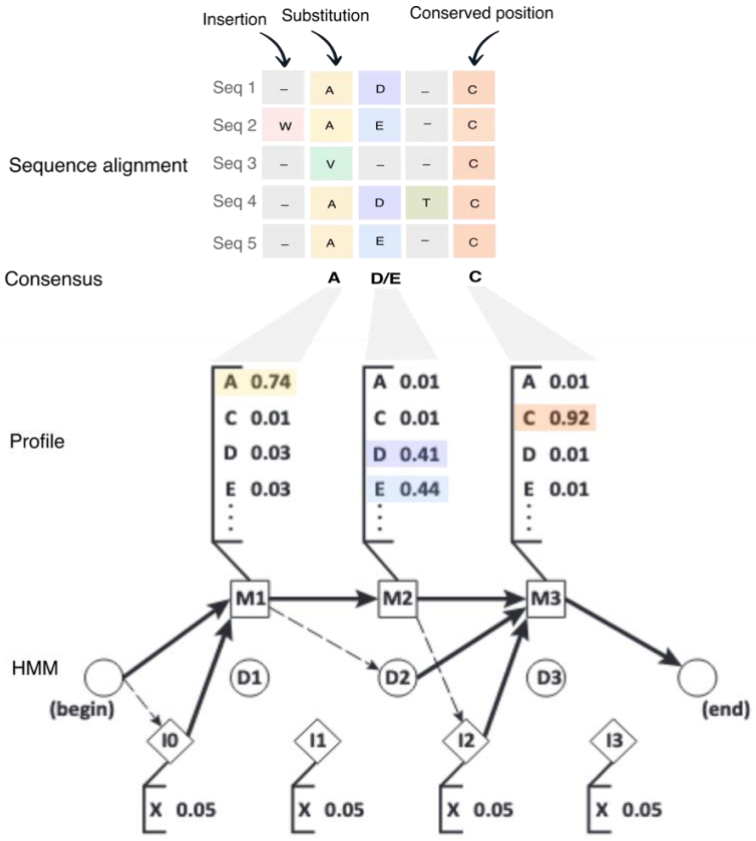


Figure 1.11 Overview of a typical HMM profile. The occurrence of insertions, deletions and substitutions is shown across the positions of a protein sequence alignment. The consensus sequence is indicated below the alignment. To build a profile, the relative frequency (probability distribution) of the 20 amino

acids in each position of the multiple sequence alignment is computed based on an HMM probabilistic model. In the HMM model, match (M), insert (I) and delete (D) states are indicated with squares, diamonds, and circles shapes, respectively. High state transition probabilities are displayed with bold arrows. Adapted from Boer 2016 and Eddy 1995.

Profile HMM search algorithms (e.g. *phmmer, hmmsearch, jackhammer, hhblits*) enable protein-protein, protein–profile, and profile–profile comparisons by interrogating profile HMM libraries and sequence databases (**Table 3**) (Finn et al., 2011; Potter et al., 2018). These algorithms are implemented in software suits such as HMMER (http://hmmer.org) and HHpred (http://toolkit.genzentrum.lmu.de/hhblits/) that perform similarity searches at remarkable speed against a specific database or even multiple databases, maximizing the use of available sequence resources (Finn et al., 2011; Potter et al., 2018; Remmert et al., 2011). In a profile HMM search the queries correspond to either a single protein sequence, a protein alignment or a profile, while the target databases include sequence collections such as the NCBI/nr, PDB and Pfam. Given that profile HMMs searches provide a more sensitive approach to detect remote homology, it is feasible to scrutinize unclassified sequences from meta-transcriptomic data sets, offering an alternative avenue to gain further insights into what has been termed viral "dark matter" (i.e. sequences with < 30% amino acid identity) (Charon et al., 2022; Roux et al., 2015). For example, the combination of profile HMMs searches, and homology modeling has proven to be a successful strategy for the detection of sequences sharing residual similarity (< 10% amino acid identity) with known viral sequences based on core components of the viral RdRp (Charon et al., 2022). Protein structure prediction is another promising approach in the case of the residual levels (< 10% aa id) of genetic similarity often observed among RNA viruses, helping us to expand the limits of the known RNA virosphere. Indeed, the process resembles a feedback loop, in which the newly detected and highly divergent viruses are effectively used as "baits" to identify hidden relatives in public databases. Since homologous proteins are likely to be conserved throughout evolution, this supports the prediction of protein structures from related proteins with known 3D structures (Centeno et al., 2005; Dunbrack, 2006; Illergård et al., 2009).

The broad idea behind the detection of remote homology using homology modeling (template-based modeling) relies on generate query-template comparisons to guide a 3D model prediction. Protein structure prediction applications such as the Phyre2 server enable the analysis of hundreds of sequences by mining structure databases (Kelley et al., 2015). The general process involves several steps, including *(i)* alignment of a target sequence against distantly related sequences (≤ 20% id), *(ii)* secondary structure prediction, *(iii)* Profile HMM building, (*iv*) profile-profile HMM search, *(v)* loop modeling and *(vi)* side-chain placement. Likewise, the assessment of the output is made based on parameters such as the confidence score (>90%), coverage and sequence identity (**Figure 1.12**) (Kelley et al., 2015).

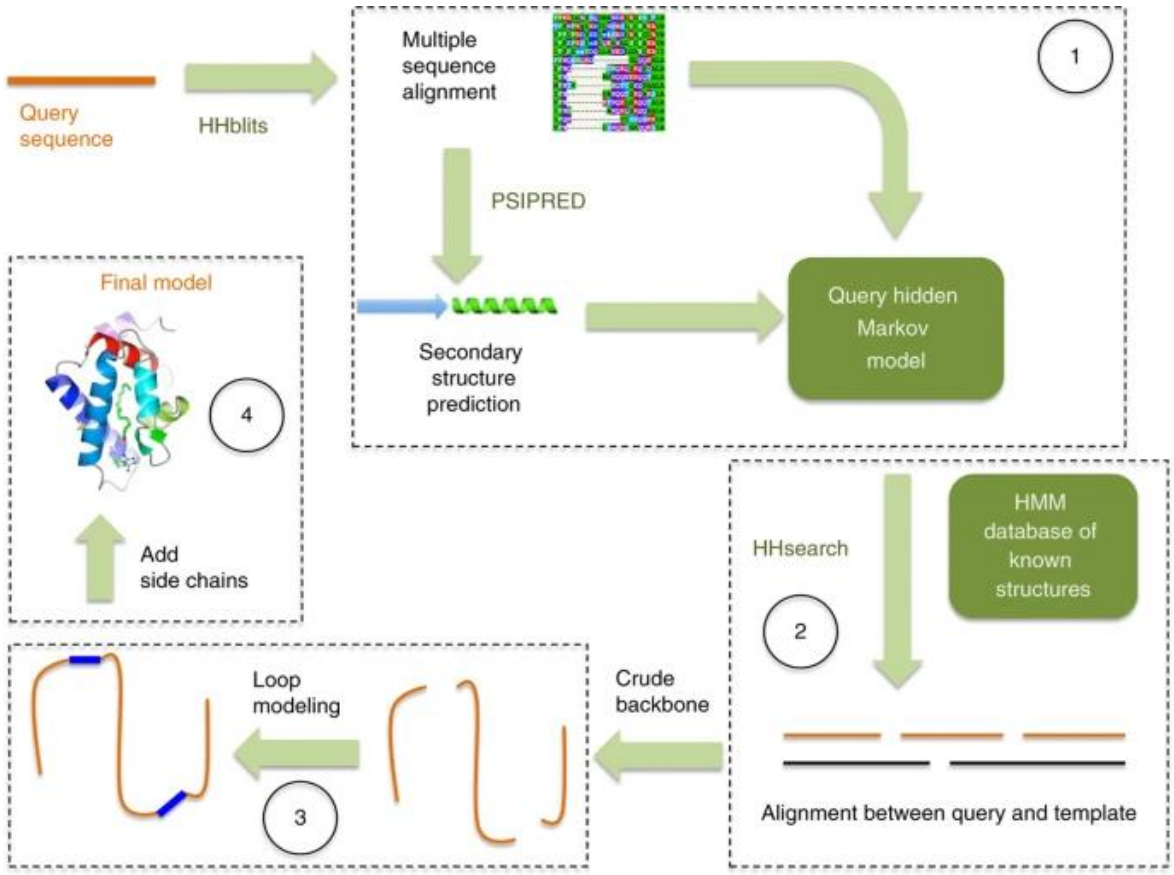


Figure 1.12 Protein structure prediction process using the "normal" mode of Phyre2. Key steps are numbered in a clockwise order. The query sequence is compared against a database of sequences with ≤ 20% sequence identity. The alignment as well as derived secondary structure prediction data are combined reconstruct a HMM profile. (2) The HMM profile is compared with a HMM database of known

structures. (3) Loop modelling based on template proteins. (4) Addition of side chains to 3D predicted structure protein. Algorithms are indicated in green. Taken from Kelley et al. 2015.

Although homology modelling takes advantage of existing protein structures as templates, the main drawback of this approach is precisely the dependency on available structures deposited in public databases such as the RCSB Protein Data Bank (RCSB PDB). In particular, the viral RdRp structures represents only a tiny fraction (< 1%) of the database size (**Figure 1.13**), which biases protein comparisons and increases the likelihood of false positives. However, this might progressively be alleviated by the implementation of new approaches such as artificial intelligence and machine learning methods. In this respect, the recent development of AlphaFold has vastly expanded the boundaries of RCSB PDB by predicting a remarkable number of computing structure models of proteins (~ 200 million protein structures) (Jumper et al., 2021).

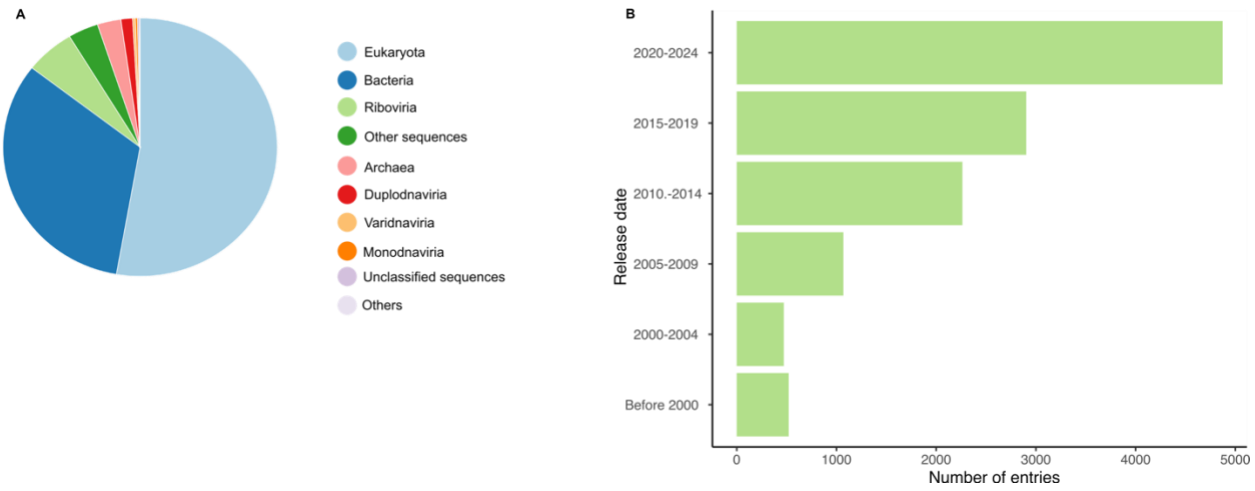


Figure 1.13 Proportion of structure entries in the RCSB PDB. (A) Entries classified by taxonomic group. (B) Distribution of Riboviria virus entries over time. Source https://www.rcsb.org/.

Five decades since the creation of the RCSB PDB, it continues to ensure the storage, access, visualization, and analysis of structural data, exhibiting an annual growth rate of 10% (**Figure 1.13**). To date, the number of available protein structures deposited in the RCSB Protein Data

Bank (RCSB PDB, https://www.rcsb.org) corresponds to ~198,000 experimental structures and ~1,000,000 computed structure models (CSM) from AlphaFold and RoseTTA fold (Baek et al., 2021; Jumper et al., 2021). As a repository system of 3D structures, the RCSB PDB represents a comprehensive collection of structural data and the primary search source for template-based modelling and protein prediction.

1.5 Thesis rationale

Fundamental research on RNA viral and microbial diversity in wildlife is paramount to enhance our understanding of key aspects of their ecology, evolution, host range and distribution. The development of meta-transcriptomics was therefore central to recent advances in the discovery and description of these agents in nature, enabling us to explore a broad spectrum of research questions ranging from the origin of the known RNA virosphere to the emergence of infectious pathogens. In this thesis, I aimed to determine the diversity, abundance and distribution of RNA viruses and common microbial life forms circulating in wildlife and reveal how this diversity contributes to shape aspects of their ecology and evolution in nature. By doing so, I assessed a diverse array of hosts, tissues and environments, largely within an Australian context.

This thesis encompasses six research-based chapters that contribute to both the discovery and detection of infectious agents using meta-transcriptomics by addressing a range of topics as follows:

Chapter 2 expands the known RNA virosphere by identifying a highly divergent virus (*Lauta virus*) within the order *Articulavirales* (negative-sense RNA viruses) in an Australian reptile. I also determined whether protein structure conservation in the viral RdRp enables the detection of highly divergent viruses within the viral dark matter. Hence, this study exploited the potential of protein structure prediction, together with meta-transcriptomics, to uncover the hidden diversity of articulaviruses in wildlife.

Chapter 3 continues to explore the diversity of articulaviruses in vertebrate hosts. Specifically, this study asked whether relatives of the novel *Lauta virus* were associated with host species placed deeper in the phylogeny of vertebrates. I therefore used meta-transcriptomics and data mining of published transcriptomes to detect divergent viruses within the family *Amnoonviridae* in fish from both from marine and freshwater environments. In addition, the composition, and biological and/or ecological drivers of the RNA virome of 19 marine fish species were investigated in a collaborative publication included in the supplementary material.

Chapter 4 Compares the RNA virome of *Aedes communis* mosquitoes and their parasitic mites to identify the occurrence of potential virus transfer through the host–parasite interaction. This study characterized the RNA virome of mite-free and mite-detached mosquitoes, as well as their parasitic mites, revealing a substantial RNA virus diversity that is shared between mosquitoes and mites.

Chapter 5 assesses whether the RNA virome diversity of *Drosophila simulans* flies varies with the presence or absence of the endosymbiotic *Wolbachia* bacteria. In this study, individual *Drosophila simulans* flies sampled from Western Australia were sequenced to establish whether the *w*Au strain of *Wolbachia* confers antiviral protection against the natural RNA virome of *D. simulans*.

Chapter 6 investigates the RNA virome diversity of *Carios vespertilionis* ticks parasitizing Soprano pipistrelle bats from Sweden to determine whether these ectoparasites carry tick-borne viruses along with bat-associated viruses of public health importance. This study revealed a substantial diversity of novel RNA viruses and provided the first report of Issyk-Kul virus circulating in Sweden. This chapter also assesses the suitability of tick surveys for detecting common pathogenic tick-borne bacteria.

Chapter 7 provides the first meta-transcriptomic detection of *Trypanosoma* spp. in wildlife. This chapter implemented a meta-transcriptomic approach to assesses the diversity, tissue tropism, and distribution of trypanosomes in endemic Australian fauna, expanding the known genetic diversity and host range for these important parasites and providing new insights into their evolutionary history.

Finally, I discuss these findings considering the current body of knowledge in the area, as well as the challenges and limitations identified, the impact of the work performed and the general research area, as well as future research directions. In addition, the last section of this thesis includes published collaborative co-authored research that further characterise the ecology and evolution of the RNA virosphere. These studies addressed a variety of questions, including: Is the RNA virome of the red fox structured by the rural/urban landscape? What is the virome composition in marine fish, and what are the determinants that shape that diversity? What is the origin of largest outbreak of yellow fever of the 21st century in the Americas, and can we trace the circulation of this virus in Brazil? Furthermore, is there any evidence of zoonotic spillover? How diverse is the faecal RNA virome of the Australian grey-headed flying foxes in urban and suburban settings? Can the faecal RNA virome shed light on the presence bat-associated viruses?

Overall, I expect that the research work presented in this thesis to contribute to a more comprehensive understanding of the ecology and evolution of the RNA virus world, as well as the formulation of specific hypotheses that will further advance the field.

1.6 References

Aggarwal, D., Kanitkar, T., Narouz, M., Azadian, B. S., Moore, L. S. P., & Mughal, N. (2020). Clinical utility and cost-effectiveness of bacterial 16S rRNA and targeted PCR based diagnostic testing in a UK microbiology laboratory network. *Scientific Reports 2020 10:1*, *10*(1), 1–9. https://doi.org/10.1038/s41598-020-64739-1

Andersen, K. G., Rambaut, A., Lipkin, W. I., Holmes, E. C., & Garry, R. F. (2020a). The proximal origin of SARS-CoV-2. *Nature Medicine 2020 26:4*, *26*(4), 450–452. https://doi.org/10.1038/s41591-020-0820-9

Andrews, S. (2010). *FastQC: a quality control tool for high throughput sequence data.*

Artois, M., Bengis, R., Delahay, R. J., Duchêne, M. J., Duff, J. P., Ferroglio, E., Gortazar, C., Hutchings, M. R., Kock, R. A., Leighton, F. A., Mörner, T., & Smith, G. C. (2009). Wildlife disease surveillance and monitoring. *Management of Disease in Wild Mammals*, 187–213. https://doi.org/10.1007/978-4-431-77134-0_10

Babaian, A., & Edgar, R. (2022). Ribovirus classification by a polymerase barcode sequence. *PeerJ*, *10*, e14055. https://doi.org/10.7717/peerj.14055

Bacharach, E., Mishra, N., Briese, T., Zody, M. C., Tsofack, J. E. K., Zamostiano, R., Berkowitz, A., Ng, J., Nitido, A., Corvelo, A., Toussaint, N. C., Abel Nielsen, S. C., Hornig, M., del Pozo, J., Bloom, T., Ferguson, H., Eldar, A., & Lipkin, W. I. (2016). Characterization of a novel orthomyxo-like virus causing mass die-offs of Tilapia. *MBio*, *7*(2). https://doi.org/10.1128/MBIO.00431-16

Baek, M., DiMaio, F., Anishchenko, I., Dauparas, J., Ovchinnikov, S., Lee, G. R., Wang, J., Cong, Q., Kinch, L. N., Dustin Schaeffer, R., Millán, C., Park, H., Adams, C., Glassman, C. R., DeGiovanni, A., Pereira, J. H., Rodrigues, A. v., van Dijk, A. A., Ebrecht, A. C., … Baker, D. (2021). Accurate prediction of protein structures and interactions using a three-track neural network. *Science*, *373*(6557), 871–876. https://doi.org/10.1126/SCIENCE.ABJ8754

Bailey, A. L., & Diamond, M. S. (2022). Hepatopathology of flaviviruses. *Journal of Hepatology*, *0*(0). https://doi.org/10.1016/j.jhep.2022.05.024

Bamford, D. H., Grimes, J. M., & Stuart, D. I. (2005). What does structure tell us about virus evolution? *Current Opinion in Structural Biology*, *15*, 655–663.

Batson, J., Dudas, G., Haas-Stapleton, E., Kistler, A. L., Li, L. M., Logan, P., Ratnasiri, K., & Retallack, H. (2021). Single mosquito metatranscriptomics identifies vectors, emerging pathogens and reservoirs in one assay. *ELife*, *10*. https://doi.org/10.7554/eLife.68353

Bedre, R. (2022). *Gene expression units explained: RPM, RPKM, FPKM, TPM, DESeq, TMM, SCnorm, GeTMM, and ComBat-Seq*. https://www.reneshbedre.com/blog/expression_units.html

Behizadi, K., Jafarzadeh, N., & Iranmanesh, A. (2022). Graph Theoretical Strategies in De Novo Assembly. *IEEE Access*, *10*, 9328–9339. https://doi.org/10.1109/access.2022.3144113

Berliner, A. J., Mochizuki, T., & Stedman, K. M. (2018). Astrovirology: Viruses at Large in the Universe. https://doi.org/10.1089/ast.2017.1649

Best, S. M. (2016). Current Biology. *Current Biology*, *26*, R1258–R1260. https://doi.org/10.1146/annurev

Bigot, T., Temmam, S., Pérot, P., & Eloit, M. (2019). RVDB-prot, a reference viral protein database and its HMM profiles. *f1000research*, *8*. https://doi.org/10.12688/f1000research.18776.2

Böer, J. (2016). Multiple alignment using hidden Markov models Seminar Hot Topics in Bioinformatics. *Proteins*, *4*, 1–14.

Bonnet, S. I., & Pollet, T. (2021). Update on the intricate tango between tick microbiomes and tick-borne pathogens. *Parasite Immunology*, *43*(5), e12813. https://doi.org/10.1111/PIM.12813

Bramley, J. C., Yenkin, A. L., Zaydman, M. A., DiAntonio, A., Milbrandt, J. D., & Buchser, W. J. (2020). Domain-centric database to uncover structure of minimally characterized viral genomes. *Scientific Data 2020 7:1*, *7*(1), 1–11. https://doi.org/10.1038/s41597-020-0536-1

Breitbart, M., & Rohwer, F. (2005). Here a virus, there a virus, everywhere the same virus? *Trends in Microbiology*, *13*(6), 278–284. https://doi.org/10.1016/j.tim.2005.04.003

Bruenn, J. A. (2003). A structural and primary sequence comparison of the viral RNA-dependent RNA polymerases. *Nucleic Acids Research*, *31*(7), 1821–1829. https://doi.org/10.1093/nar/gkg277

Buchfink, B., Xie, C., & Huson, D. H. (2015). Fast and sensitive protein alignment using DIAMOND. *Nature Methods*, *12*(1), 59–60. https://doi.org/10.1038/nmeth.3176

Burki, F., Sandin, M. M., & Jamy, M. (2021). Diversity and ecology of protists revealed by metabarcoding. *Current Biology*, *31*(19), R1267–R1280. https://doi.org/10.1016/J.CUB.2021.07.066

Calisher, C. H., Childs, J. E., Field, H. E., Holmes, K. v., & Schountz, T. (2006). Bats: Important Reservoir Hosts of Emerging Viruses. *Clinical Microbiology Reviews*, *19*(3), 531–545. https://doi.org/10.1128/CMR.00017-06

Cantalupo, P. G., & Pipas, J. M. (2019). Detecting viral sequences in NGS data. *Current Opinion in Virology*, *39*, 41–48. https://doi.org/10.1016/j.coviro.2019.07.010

Caron, D. A., Alexander, H., Allen, A. E., Archibald, J. M., Armbrust, E. V., Bachy, C., Bell, C. J., Bharti, A., Dyhrman, S. T., Guida, S. M., Heidelberg, K. B., Kaye, J. Z., Metzner, J., Smith, S. R., & Worden, A. Z. (2016). Probing the evolution, ecology and physiology of marine protists using transcriptomics. *Nature Reviews Microbiology 2016 15:1*, *15*(1), 6–20. https://doi.org/10.1038/nrmicro.2016.160

Causey, D., & Edwards, S. v. (2008). Ecology of Avian Influenza Virus in Birds. *The Journal of Infectious Diseases*, *197*(Supplement_1), S29–S33. https://doi.org/10.1086/524991

Centeno, N. B., Planas-Iglesias, J., & Oliva, B. (2005). Comparative modelling of protein structure and its impact on microbial cell factories. *Microbial Cell Factories*, *4*(1), 1–11. https://doi.org/10.1186/1475-2859-4-20

Chang, W. S., Rose, K., & Holmes, E. C. (2021). Meta-transcriptomic analysis of the virome and microbiome of the invasive Indian myna (Acridotheres tristis) in Australia. *One Health*, *13*, 100360. https://doi.org/10.1016/j.onehlt.2021.100360

Charon, J., Buchmann, J. P., Sadiq, S., & Holmes, E. C. (2022). RdRp-scan: A bioinformatic resource to identify and annotate divergent RNA viruses in metagenomic sequence data. *Virus Evolution*, *8*(2). https://doi.org/10.1093/ve/veac082

Charon, J., Grigg, M. J., Eden, J. S., Piera, K. A., Rana, H., William, T., Rose, K., Davenport, M. P., Anstey, N. M., & Holmes, E. C. (2019). Novel RNA viruses associated with Plasmodium vivax in human malaria and Leucocytozoon parasites in avian disease. *PLOS Pathogens*, *15*(12), e1008216. https://doi.org/10.1371/journal.ppat.1008216

Chen, T. W., Gan, R. R. C., Wu, T. H., Lin, W. C., & Tang, P. (2012). VIP DB — A viral protein domain usage and distribution database. *Genomics*, *100*(3), 149–156. https://doi.org/10.1016/j.ygeno.2012.06.008

Chen, Y., Dessau, M., Rotenberg, D., Rasmussen, D. A., & Whitfield, A. E. (2019). Entry of bunyaviruses into plants and vectors. *Advances in Virus Research*, *104*, 65–96. https://doi.org/10.1016/bs.aivir.2019.07.001

Chen, Y.-M., Sadiq, S., Tian, J.-H., Chen, X., Lin, X.-D., Shen, J.-J., Chen, H., Hao, Z.-Y., Yang, W.-D., Zhou, Z.-C., Wu, J., Li, F., Wang, H.-W., Xu, Q.-Y., Wang, W., Gao, W.-H., Holmes, E. C., & Zhang,

Y.-Z. (2021). RNA virome composition Is shaped by sampling ecotype. *SSRN Electronic Journal*. https://doi.org/10.2139/ssrn.3934022

Cobbin, J. C., Charon, J., Harvey, E., Holmes, E. C., & Mahar, J. E. (2021). Current challenges to virus discovery by meta-transcriptomics. *Current Opinion in Virology*, *51*, 48–55.

Cobián Güemes, A. G., Youle, M., Cantú, V. A., Felts, B., Nulton, J., & Rohwer, F. (2016). Viruses as Winners in the Game of Life. *Https://Doi.Org/10.1146/Annurev-Virology-100114-054952*, *3*, 197–214. https://doi.org/10.1146/annurev-virology-100114-054952

Cohen, J., Powderly, W. G., & Opal, S. M. (2017). Infectious Diseases, 2-Volume Set. *Infectious Diseases, 2-Volume Set*, 1-1779.e1. https://doi.org/10.1016/c2013-1-00044-3

Colijn, C., & Plazzotta, G. (2018). A Metric on Phylogenetic Tree Shapes. *Systematic Biology*, *67*(1), 113. https://doi.org/10.1093/sysbio/syx046

Corchete, L. A., Rojas, E. A., Alonso-López, D., de Las Rivas, J., Gutiérrez, N. C., & Burguillo, F. J. (2020). Systematic comparison and assessment of RNA-seq procedures for gene expression quantitative analysis. *Scientific Reports 2020 10:1*, *10*(1), 1–15. https://doi.org/10.1038/s41598-020-76881-x

Costello, M., Fleharty, M., Abreu, J., Farjoun, Y., Ferriera, S., Holmes, L., Granger, B., Green, L., Howd, T., Mason, T., Vicente, G., Dasilva, M., Brodeur, W., DeSmet, T., Dodge, S., Lennon, N. J., & Gabriel, S. (2018). Characterization and remediation of sample index swaps by non-redundant dual indexing on massively parallel sequencing platforms. *BMC Genomics*, *19*(1), 1–10. https://doi.org/10.1186/s12864-018-4703-0

de Vries, E. M., Cogan, N. O. I., Gubala, A. J., Mee, P. T., O'riley, K. J., Rodoni, B. C., & Lynch, S. E. (2022). Rapid, in-field deployable, avian influenza virus haemagglutinin characterisation tool using MinION technology. *Scientific Reports |*, *12*, 11886. https://doi.org/10.1038/s41598-022-16048-y

Dezordi, F. Z., Vasconcelos, C. R. dos S., Rezende, A. M., & Wallau, G. L. (2020). In and Outs of Chuviridae Endogenous Viral Elements: Origin of a Potentially New Retrovirus and Signature of Ancient and Ongoing Arms Race in Mosquito Genomes. *Frontiers in Genetics*, *11*, 1291. https://doi.org/10.3389/fgene.2020.542437

di Giallonardo, F., Schlub, T. E., Shi, M., & Holmes, E. C. (2017). Dinucleotide Composition in Animal RNA Viruses Is Shaped More by Virus Family than by Host Species. *Journal of Virology*, *91*(8). https://doi.org/10.1128/jvi.02381-16

Dietzgen, R. G., Firth, A. E., Jiāng, D., Junglen, S., Kondo, H., Kuhn, J. H., Paraskevopoulou, S., & Vasilakis, N. (2021). ICTV virus taxonomy profile: Nyamiviridae 2021. *Journal of General Virology*, *102*(11), 001681. https://doi.org/10.1099/jgv.0.001681

Dolja, V. v., & Koonin, E. v. (2018). Metagenomics reshapes the concepts of RNA virus evolution by revealing extensive horizontal virus transfer. *Virus Research*, *244*, 36–52.

Doyle, C. J., O'Toole, P. W., & Cotter, P. D. (2017). Metagenome-based surveillance and diagnostic approaches to studying the microbial ecology of food production and processing environments. *Environmental Microbiology*, *19*(11), 4382–4391. https://doi.org/10.1111/1462-2920.13859

Dunbrack, R. L. (2006). Sequence comparison and protein structure prediction. *Current Opinion in Structural Biology*, *16*(3), 374–384. https://doi.org/10.1016/j.sbi.2006.05.006

Duron, O., Morel, O., Noël, V., Buysse, M., Binetruy, F., Lancelot, R., Loire, E., Ménard, C., Bouchez, O., Vavre, F., & Vial, L. (2018). Tick-Bacteria Mutualism Depends on B Vitamin Synthesis Pathways. *Current Biology*, *28*(12), 1896-1902.e5. https://doi.org/10.1016/J.CUB.2018.04.038

Eddy, S. R. (1995). *Multiple alignment using hidden Markov models*. www.aaai.org

Eddy, S. R. (1998). Profile hidden Markov models. *Bioinformatics (Oxford, England)*, *14*(9), 755–763. https://doi.org/10.1093/bioinformatics/14.9.755

Edgar, R. C., Taylor, J., Lin, V., Altman, T., Barbera, P., Meleshko, D., Lohr, D., Novakovsky, G., Buchfink, B., Al-Shayeb, B., Banfield, J. F., de la Peña, M., Korobeynikov, A., Chikhi, R., & Babaian, A. (2022). Petabase-scale sequence alignment catalyses viral discovery. *Nature 2022 602:7895*, *602*(7895), 142–147. https://doi.org/10.1038/s41586-021-04332-2

Eyngor, M., Zamostiano, R., Tsofack, J. E. K., Berkowitz, A., Bercovier, H., Tinman, S., Lev, M., Hurvitz, A., Galeotti, M., Bacharach, E., & Eldar, A. (2014). Identification of a Novel RNA Virus Lethal to Tilapia. *Journal of Clinical Microbiology*, *52*(12), 4137. https://doi.org/10.1128/JCM.00827-14

Ferrer-Orta, C., Arias, A., Escarmís, C., & Verdaguer, N. (2006). A comparison of viral RNA-dependent RNA polymerases. *Current Opinion in Structural Biology*, *16*(1), 27–34. https://doi.org/10.1016/j.sbi.2005.12.002

Finn, R. D., Clements, J., & Eddy, S. R. (2011). HMMER web server: interactive sequence similarity searching. *Nucleic Acids Research*, *39*(Web Server issue), W29. /pmc/articles/PMC3125773/

Fraser, C. M., Eisen, J. A., & Salzberg, S. L. (2000). Microbial genome sequencing. *Nature 2000 406:6797*, *406*(6797), 799–803. https://doi.org/10.1038/35021244

Freedman, A. H., Clamp, M., & Sackton, T. B. (2021). Error, noise and bias in de novo transcriptome assemblies. *Molecular Ecology Resources*, *21*(1), 18–29. https://doi.org/10.1111/1755-0998.13156

Freije, C. A., & Sabeti, P. C. (2021). Detect and destroy: CRISPR-based technologies for the response against viruses. *Cell Host and Microbe*, *29*(5), 689–703. https://doi.org/10.1016/j.chom.2021.04.003

Geoghegan, J. L., di Giallonardo, F., Wille, M., Ortiz-Baez, A. S., Costa, V. A., Ghaly, T., Mifsud, J. C. O., Turnbull, O. M. H., Bellwood, D. R., Williamson, J. E., & Holmes, E. C. (2021). Virome composition in marine fish revealed by meta-transcriptomics. *Virus Evolution*, *7*(1), 5. https://doi.org/10.1093/ve/veab005

Geoghegan, J. L., Giallonardo, F. di, Cousins, K., Shi, M., Williamson, J. E., Holmes, E. C., & Bashir, M. (2018). Hidden diversity and evolution of viruses in market fish. *Virus Evolution*, *4*(2). https://doi.org/10.1093/VE/VEY031

Geoghegan, J. L., & Holmes, E. C. (2017). Predicting virus emergence amid evolutionary noise. *Open Biology*, *7*(10). https://doi.org/10.1098/rsob.170189

German, T. L., Lorenzen, M. D., Grubbs, N., & Whitfield, A. E. (2020). New technologies for studying negative-strand RNA viruses in plant and arthropod hosts. *Molecular Plant-Microbe Interactions*, *33*(3), 382–393. https://doi.org/10.1094/mpmi-10-19-0281-fi

Gofton, A. W., Blasdell, K. R., Taylor, C., Banks, P. B., Michie, M., Roy-Dufresne, E., Poldy, J., Wang, J., Dunn, M., Tachedjian, M., & Smith, I. (2022). Metatranscriptomic profiling reveals diverse tick-borne bacteria, protozoans and viruses in ticks and wildlife from Australia. *Transboundary and Emerging Diseases*, *69*(5), e2389–e2407. https://doi.org/10.1111/tbed.14581

Goodwin, S., McPherson, J. D., & McCombie, W. R. (2016). Coming of age: ten years of next-generation sequencing technologies. *Nature Reviews Genetics*, *17*(6), 333–351. https://doi.org/10.1038/nrg.2016.49

Gorbalenya, A. E., Krupovic, M., Mushegian, A., Kropinski, A. M., Siddell, S. G., Varsani, A., Adams, M. J., Davison, A. J., Dutilh, B. E., Harrach, B., Harrison, R. L., Junglen, S., King, A. M. Q., Knowles, N. J., Lefkowitz, E. J., Nibert, M. L., Rubino, L., Sabanadzovic, S., Sanfaçon, H., … Kuhn, J. H. (2020). The new scope of virus taxonomy: partitioning the virosphere into 15 hierarchical ranks. *Nature Microbiology 2020 5:5*, *5*(5), 668–674. https://doi.org/10.1038/s41564-020-0709-x

Gorbalenya, A. E., Pringle, F. M., Zeddam, J. L., Luke, B. T., Cameron, C. E., Kalmakoff, J., Hanzlik, T. N., Gordon, K. H. J., & Ward, V. K. (2002). The Palm Subdomain-based Active Site is Internally Permuted in Viral RNA-dependent RNA Polymerases of an Ancient Lineage. *Journal of Molecular Biology*, *324*(1), 47–62. https://doi.org/10.1016/S0022-2836(02)01033-1

Gould, S. J. (1994). The evolution of life on the earth. *Scientific American*, *271*(4), 84–91. https://doi.org/10.1038/scientificamerican1094-84

Grada, A., & Weinbrecht, K. (2013). Next-generation sequencing: methodology and application. *The Journal of Investigative Dermatology*, *133*(8). https://doi.org/10.1038/JID.2013.248

Greay, T. L., Gofton, A. W., Paparini, A., Ryan, U. M., Oskam, C. L., & Irwin, P. J. (2018). Recent insights into the tick microbiome gained through next-generation sequencing. *Parasites & Vectors 2018 11:1*, *11*(1), 1–14. https://doi.org/10.1186/S13071-017-2550-5

Greninger, A. L., Naccache, S. N., Federman, S., Yu, G., Mbala, P., Bres, V., Stryke, D., Bouquet, J., Somasekar, S., Linnen, J. M., Dodd, R., Mulembakani, P., Schneider, B. S., Muyembe-Tamfum, J. J., Stramer, S. L., & Chiu, C. Y. (2015). Rapid metagenomic identification of viral pathogens in clinical samples by real-time nanopore sequencing analysis. *Genome Medicine*, *7*(1), 1–13. https://doi.org/10.1186/s13073-015-0220-9

Gribskov, M., McLachlan, A. D., & Eisenberg, D. (1987). Profile analysis: detection of distantly related proteins. *Proceedings of the National Academy of Sciences*, *84*(13), 4355–4358. https://doi.org/10.1073/pnas.84.13.4355

Griffiths-Jones, S., Bateman, A., Marshall, M., Khanna, A., & Eddy, S. R. (2003). Rfam: an RNA family database. *Nucleic Acids Research*, *31*(1), 439. https://doi.org/10.1093/NAR/GKG006

Guenay-Greunke, Y., Bohan, D. A., Traugott, M., & Wallinger, C. (2021). Handling of targeted amplicon sequencing data focusing on index hopping and demultiplexing using a nested

metabarcoding approach in ecology. *Scientific Reports*, *11*(1), 19510. https://doi.org/10.1038/s41598-021-98018-4

Han, Y., Gao, S., Muegge, K., Zhang, W., & Zhou, B. (2015). Advanced applications of RNA sequencing and challenges. *Bioinformatics and Biology Insights*, *9*, 29–46. https://doi.org/10.4137/BBI.S28991

Hanley, K. A., & Weaver, S. C. (2008). Arbovirus Evolution. *Origin and Evolution of Viruses*, 351–391. https://doi.org/10.1016/B978-0-12-374153-0.00016-3

Harvey, E., & Holmes, E. C. (2022). Diversity and evolution of the animal virome. *Nature Reviews Microbiology*, *0123456789*. https://doi.org/10.1038/s41579-021-00665-x

Harvey, E., Rose, K., Eden, J. S., Lawrence, A., Doggett, S. L., & Holmes, E. C. (2019). Identification of diverse arthropod associated viruses in native Australian fleas. *Virology*, *535*, 189–199. https://doi.org/10.1016/J.VIROL.2019.07.010

Harvey, E., Rose, K., Eden, J.-S., Lo, N., Abeyasuriya, T., Shi, M., Doggett, S. L., & Holmes, E. C. (2019a). Extensive diversity of RNA viruses in Australian ticks. *Journal of Virology*, *93*(3). https://doi.org/10.1128/JVI.01358-18

Herath, V., Romay, G., Urrutia, C. D., & Verchot, J. (2020). Family Level Phylogenies Reveal Relationships of Plant Viruses within the Order Bunyavirales. *Viruses 2020, Vol. 12, Page 1010*, *12*(9), 1010. https://doi.org/10.3390/V12091010

Ho, T., & Tzanetakis, I. E. (2014). Development of a virus detection and discovery pipeline using next generation sequencing. *Virology*, *471*, 54–60. https://doi.org/10.1016/j.virol.2014.09.019

Holmes, E. C. (2019). Reagent contamination in viromics: all that glitters is not gold. *Clinical Microbiology and Infection*, *25*(10), 1167–1168. https://doi.org/10.1016/J.CMI.2019.06.019

Holmes, E. C., Goldstein, S. A., Rasmussen, A. L., Robertson, D. L., Crits-Christoph, A., Wertheim, J. O., Anthony, S. J., Barclay, W. S., Boni, M. F., Doherty, P. C., Farrar, J., Geoghegan, J. L., Jiang, X., Leibowitz, J. L., Neil, S. J. D., Skern, T., Weiss, S. R., Worobey, M., Andersen, K. G., … Rambaut, A. (2021). The origins of SARS-CoV-2: A critical review. *Cell*, *184*(19), 4848–4856. https://doi.org/10.1016/j.cell.2021.08.017

Hölzer, M., & Marz, M. (2019). De novo transcriptome assembly: A comprehensive cross-species comparison of short-read RNA-Seq assemblers. *GigaScience*, *8*(5). https://doi.org/10.1093/gigascience/giz039

Hu, T., Chitnis, N., Monos, D., & Dinh, A. (2021). Next-generation sequencing technologies: An overview. *Human Immunology*, *82*(11), 801–811. https://doi.org/10.1016/j.humimm.2021.02.012

Illergård, K., Ardell, D. H., & Elofsson, A. (2009). Structure is three to ten times more conserved than sequence—A study of structural response in protein cores. *Proteins: Structure, Function, and Bioinformatics*, *77*(3), 499–508. https://doi.org/10.1002/PROT.22458

Illumina. (2017). *Effects of Index Misassignment on Multiplexing and Downstream Analysis*.

Illumina. (2022). *Minimizing Index Hopping*. https://www.illumina.com/techniques/sequencing/ngs-library-prep/multiplexing/index-hopping.html

Jia, H., & Gong, P. (2019). A structure-function diversity survey of the rna-dependent rna polymerases from the positive-strand rna viruses. *Frontiers in Microbiology*, *10*(AUG), 1945. https://doi.org/10.3389/fmicb.2019.01945

Jiwaji, M., Matcher, G. F., de Bruyn, M. M., Awando, J. A., Moodley, H., Waterworth, D., Jarvie, R. A., & Ann Dorrington, R. (2019). Providence virus: An animal virus that replicates in plants or a plant virus that infects and replicates in animal cells? *PLOS ONE*, *14*(6), e0217494. https://doi.org/10.1371/journal.pone.0217494

Johnson, J. S., Spakowicz, D. J., Hong, B. Y., Petersen, L. M., Demkowicz, P., Chen, L., Leopold, S. R., Hanson, B. M., Agresta, H. O., Gerstein, M., Sodergren, E., & Weinstock, G. M. (2019). Evaluation of 16S rRNA gene sequencing for species and strain-level microbiome analysis. *Nature Communications 2019 10:1*, *10*(1), 1–11. https://doi.org/10.1038/s41467-019-13036-1

Johnson, R. C., Schmid, G. P., & Hyde, F. W. (1984). Borrelia burgdorferi sp. nov.: Etiologic agent of Lyme disease. *International Journal of Systematic Bacteriology*, *34*(4), 496–497. https://doi.org/10.1099/00207713-34-4-496

Jumper, J., Evans, R., Pritzel, A., Green, T., Figurnov, M., Ronneberger, O., Tunyasuvunakool, K., Bates, R., Žídek, A., Potapenko, A., Bridgland, A., Meyer, C., Kohl, S. A. A., Ballard, A. J., Cowie, A., Romera-Paredes, B., Nikolov, S., Jain, R., Adler, J., … Hassabis, D. (2021). Highly accurate

protein structure prediction with AlphaFold. *Nature 2021 596:7873*, *596*(7873), 583–589. https://doi.org/10.1038/s41586-021-03819-2

Junglen, S. (2016). Evolutionary origin of pathogenic arthropod-borne viruses — a case study in the family Bunyaviridae. *Current Opinion in Insect Science*, *16*, 81–86. https://doi.org/10.1016/J.COIS.2016.05.017

Käfer, S., Paraskevopoulou, S., Zirkel, F., Wieseke, N., Donath, A., Petersen, M., Jones, T. C., Liu, S., Zhou, X., Middendorf, M., Junglen, S., Misof, B., & Drosten, C. (2019a). Re-assessing the diversity of negative strand RNA viruses in insects. *PLOS Pathogens*, *15*(12), e1008224. https://doi.org/10.1371/JOURNAL.PPAT.1008224

Kalantar, K. L., Carvalho, T., de Bourcy, C. F. A., Dimitrov, B., Dingle, G., Egger, R., Han, J., Holmes, O. B., Juan, Y. F., King, R., Kislyuk, A., Lin, M. F., Mariano, M., Morse, T., Reynoso, L. v., Cruz, D. R., Sheu, J., Tang, J., Wang, J., … DeRisi, J. L. (2020). IDseq-An open source cloud-based pipeline and analysis service for metagenomic pathogen detection and monitoring. *GigaScience*, *9*(10), 1–14. https://doi.org/10.1093/GIGASCIENCE/GIAA111

Kapoor, A., Simmonds, P., Lipkin, W. I., Zaidi, S., & Delwart, E. (2010). Use of Nucleotide Composition Analysis To Infer Hosts for Three Novel Picorna-Like Viruses. *Journal of Virology*, *84*(19), 10322–10328. https://doi.org/10.1128/JVI.00601-10

Kasozi, K. I., Zirintunda, G., Ssempijja, F., Buyinza, B., Alzahrani, K. J., Matama, K., Nakimbugwe, H. N., Alkazmi, L., Onanyang, D., Bogere, P., Ochieng, J. J., Islam, S., Matovu, W., Nalumenya, D. P., Batiha, G. E. S., Osuwat, L. O., Abdelhamid, M., Shen, T., Omadang, L., & Welburn, S. C. (2021). Epidemiology of Trypanosomiasis in Wildlife—Implications for Humans at the Wildlife Interface in Africa. *Frontiers in Veterinary Science*, *8*, 565. https://doi.org/10.3389/FVETS.2021.621699

Kelley, L. A., Mezulis, S., Yates, C. M., Wass, M. N., & Sternberg, M. J. E. (2015). The Phyre2 web portal for protein modeling, prediction and analysis. *Nature Protocols 2015 10:6*, *10*(6), 845–858. https://doi.org/10.1038/nprot.2015.053

Khoo, J. J., Chen, F., Kho, K. L., Ahmad Shanizza, A. I., Lim, F. S., Tan, K. K., Chang, L. Y., & AbuBakar, S. (2016). Bacterial community in *Haemaphysalis* ticks of domesticated animals from the Orang Asli communities in Malaysia. *Ticks and Tick-Borne Diseases*, *7*(5), 929–937. https://doi.org/10.1016/J.TTBDIS.2016.04.013

Kim, D., Song, L., Breitwieser, F. P., & Salzberg, S. L. (2016). Centrifuge: Rapid and sensitive classification of metagenomic sequences. *Genome Research*, *26*(12), 1721–1729. https://doi.org/10.1101/GR.210641.116/-/DC1

Ko, K. K. K., Chng, K. R., & Nagarajan, N. (2022). Metagenomics-enabled microbial surveillance. *Nature Microbiology 2022 7:4*, *7*(4), 486–496. https://doi.org/10.1038/s41564-022-01089-w

Kobras, C. M., Fenton, A. K., & Sheppard, S. K. (2021). Next-generation microbiology: from comparative genomics to gene function. *Genome Biology 2021 22:1*, *22*(1), 1–16. https://doi.org/10.1186/S13059-021-02344-9

Koonin, E. v., Dolja, V. v., & Krupovic, M. (2015). Origins and evolution of viruses of eukaryotes: The ultimate modularity. In *Virology* (Vols. 479–480, pp. 2–25). Academic Press Inc. https://doi.org/10.1016/j.virol.2015.02.039

Koonin, E. v., Dolja, V. v., Krupovic, M., Varsani, A., Wolf, Y. I., Yutin, N., Zerbini, F. M., & Kuhn, J. H. (2020). Global Organization and Proposed Megataxonomy of the Virus World. *Microbiology and Molecular Biology Reviews*, *84*(2). https://doi.org/10.1128/MMBR.00061-19

Koonin, E. v., Krupovic, M., & Dolja, V. v. (2022). The global virome: How much diversity and how many independent origins? *Environmental Microbiology*. https://doi.org/10.1111/1462-2920.16207

Kormelink, R., Garcia, M. L., Goodin, M., Sasaya, T., & Haenni, A. L. (2011). Negative-strand RNA viruses: The plant-infecting counterparts. *Virus Research*, *162*(1–2), 184–202. https://doi.org/10.1016/J.VIRUSRES.2011.09.028

Kulski, J. K. (2016). Next-Generation Sequencing — An Overview of the History, Tools, and "Omic" Applications. *Next Generation Sequencing - Advances, Applications and Challenges*. https://doi.org/10.5772/61964

Kuno, G., Mackenzie, J. S., Junglen, S., Hubálek, Z., Plyusnin, A., & Gubler, D. J. (2017). Vertebrate Reservoirs of Arboviruses: Myth, Synonym of Amplifier, or Reality? *Viruses 2017, Vol. 9, Page 185*, *9*(7), 185. https://doi.org/10.3390/V9070185

le Lay, C., Shi, M., Buček, A., Bourguignon, T., Lo, N., & Holmes, E. C. (2020). Unmapped RNA Virus Diversity in Termites and their Symbionts. *Viruses*, *12*(10). https://doi.org/10.3390/V12101145

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. (2015). Unprecedented genomic diversity of RNA viruses in arthropods reveals the ancestry of negative-sense RNA viruses. *ELife*, *4*(4). https://doi.org/10.7554/ELIFE.05378

Liao, X., Li, M., Zou, Y., Wu, F. X., Yi-Pan, & Wang, J. (2019). Current challenges and solutions of de novo assembly. *Quantitative Biology 2019 7:2*, *7*(2), 90–109. https://doi.org/10.1007/S40484-019-0166-9

Lipkin, W. I. (2013). The changing face of pathogen discovery and surveillance. *Nature Reviews Microbiology 2013 11:2*, *11*(2), 133–141. https://doi.org/10.1038/nrmicro2949

Lips, K. R. (2016). Overview of chytrid emergence and impacts on amphibians. *Philosophical Transactions of the Royal Society B: Biological Sciences*, *371*(1709). https://doi.org/10.1098/RSTB.2015.0465

Luis, A. D., Hayman, D. T. S., O'Shea, T. J., Cryan, P. M., Gilbert, A. T., Pulliam, J. R. C., Mills, J. N., Timonin, M. E., Willis, C. K. R., Cunningham, A. A., Fooks, A. R., Rupprecht, C. E., Wood, J. L. N., & Webb, C. T. (2013). A comparison of bats and rodents as reservoirs of zoonotic viruses: are bats special? *Proceedings of the Royal Society B: Biological Sciences*, *280*(1756). https://doi.org/10.1098/RSPB.2012.2753

Lv, J., Fernández de Marco, M. del M., Goharriz, H., Phipps, L. P., McElhinney, L. M., Hernández-Triana, L. M., Wu, S., Lin, X., Fooks, A. R., & Johnson, N. (2018). Detection of tick-borne bacteria and babesia with zoonotic potential in *Argas (Carios) vespertilionis* (Latreille, 1802) ticks from British bats. *Scientific Reports*, *8*(1), 1865. https://doi.org/10.1038/s41598-018-20138-1

Marcelino, V. R., Clausen, P. T. L. C., Buchmann, J. P., Wille, M., Iredell, J. R., Meyer, W., Lund, O., Sorrell, T. C., & Holmes, E. C. (2020). CCMetagen: Comprehensive and accurate identification of eukaryotes and prokaryotes in metagenomic data. *Genome Biology*, *21*(1), 1–15. https://genomebiology.biomedcentral.com/articles/10.1186/s13059-020-02014-2

Marchler-Bauer, A., Lu, S., Anderson, J. B., Chitsaz, F., Derbyshire, M. K., DeWeese-Scott, C., Fong, J. H., Geer, L. Y., Geer, R. C., Gonzales, N. R., Gwadz, M., Hurwitz, D. I., Jackson, J. D., Ke, Z., Lanczycki, C. J., Lu, F., Marchler, G. H., Mullokandov, M., Omelchenko, M. v., … Bryant, S. H. (2011). CDD: a Conserved Domain Database for the functional annotation of proteins. *Nucleic Acids Research*, *39*(suppl_1), D225–D229. https://doi.org/10.1093/NAR/GKQ1189

Metzker, M. L. (2009). Sequencing technologies — the next generation. *Nature Reviews Genetics 2010 11:1*, *11*(1), 31–46. https://doi.org/10.1038/nrg2626

Mistry, J., Chuguransky, S., Williams, L., Qureshi, M., Salazar, G. A., Sonnhammer, E. L. L., Tosatto, S. C. E., Paladin, L., Raj, S., Richardson, L. J., Finn, R. D., & Bateman, A. (2021). Pfam: The protein families database in 2021. *Nucleic Acids Research*, *49*(D1), D412–D419. https://doi.org/10.1093/NAR/GKAA913

Mitreva, M. (2017). The Microbiome in Infectious Diseases. In *Infectious Diseases* (Vol. 1, pp. 68-74.e2). Elsevier. https://doi.org/10.1016/B978-0-7020-6285-8.00008-3

Mlotshwa, S., Pruss, G. J., & Vance, V. (2008). Small RNAs in viral infection and host defense. *Trends in Plant Science*, *13*(7), 375–382. https://doi.org/10.1016/J.TPLANTS.2008.04.009

Morey, M., Fernández-Marmiesse, A., Castiñeiras, D., Fraga, J. M., Couce, M. L., & Cocho, J. A. (2013). A glimpse into past, present, and future DNA sequencing. *Molecular Genetics and Metabolism*, *110*(1–2), 3–24. https://doi.org/10.1016/J.YMGME.2013.04.024

Muehlenbachs, A., Bhatnagar, J., & Zaki, S. R. (2015). Tissue tropism, pathology and pathogenesis of enterovirus infection. *The Journal of Pathology*, *235*(2), 217–228. https://doi.org/10.1002/PATH.4438

Mull, N., Carlson, C. J., Forbes, K. M., & Becker, D. J. (2022). Virus isolation data improve host predictions for New World rodent orthohantaviruses. *Journal of Animal Ecology*, *91*(6), 1290–1302. https://doi.org/10.1111/1365-2656.13694

Neri, U., Wolf, Y. I., Roux, S., Camargo, A. P., Lee, B., Kazlauskas, D., Chen, I. M., Ivanova, N., Zeigler Allen, L., Paez-Espino, D., Bryant, D. A., Bhaya, D., Krupovic, M., Dolja, V. v., Kyrpides, N. C., Koonin, E. v., Gophna, U., Narrowe, A. B., Probst, A. J., … Wang, Z. (2022). Expansion of the global RNA virome reveals diverse clades of bacteriophages. *Cell*. https://doi.org/10.1016/J.CELL.2022.08.023

Obbard, D. J. (2018). Expansion of the metazoan virosphere: progress, pitfalls, and prospects. *Current Opinion in Virology*, *31*, 17–23. https://doi.org/10.1016/J.COVIRO.2018.08.008

Obbard, D. J., Shi, M., Roberts, K. E., Longdon, B., & Dennis, A. B. (2020). A new lineage of segmented RNA viruses infecting animals. *Virus Evolution*, *6*(1), 61. https://academic.oup.com/ve/article/6/1/vez061/5708932

Obiol, A., Giner, C. R., Sánchez, P., Duarte, C. M., Acinas, S. G., & Massana, R. (2020). A metagenomic assessment of microbial eukaryotic diversity in the global ocean. *Molecular Ecology Resources*, *20*(3), 718–731. https://doi.org/10.1111/1755-0998.13147

Ødegaard, F. (2000). How many species of arthropods? Erwin's estimate revised. *Biological Journal of the Linnean Society*, *71*(4), 583–597. https://doi.org/10.1111/J.1095-8312.2000.TB01279.X

Ozsolak, F., & Milos, P. M. (2010). RNA sequencing: advances, challenges and opportunities. *Nature Reviews Genetics 2010 12:2*, *12*(2), 87–98. https://doi.org/10.1038/nrg2934

Pearson, W. R. (2013). An Introduction to Sequence Similarity ("Homology") Searching. *Current Protocols in Bioinformatics*, *42*(1), 3.1.1-3.1.8. https://doi.org/10.1002/0471250953.BI0301S42

Pettersson, J. H. O., Shi, M., Eden, J. S., Holmes, E. C., & Hesson, J. C. (2019). Meta-transcriptomic comparison of the RNA viromes of the mosquito vectors *Culex pipiens* and *Culex torrentium* in Northern Europe. *Viruses 2019, Vol. 11, Page 1033*, *11*(11), 1033. https://doi.org/10.3390/V11111033

Pollard, M. O., Gurdasani, D., Mentzer, A. J., Porter, T., & Sandhu, M. S. (2018). Long reads: their purpose and place. *Human Molecular Genetics*, *27*(R2), R234–R241. https://doi.org/10.1093/HMG/DDY177

Poon, A. F. Y., Walker, L. W., Murray, H., McCloskey, R. M., Harrigan, P. R., & Liang, R. H. (2013). Mapping the Shapes of Phylogenetic Trees from Human and Zoonotic RNA Viruses. *PLoS ONE*, *8*(11), 78122. https://doi.org/10.1371/JOURNAL.PONE.0078122

Porter, A. F., Cobbin, J., Li, C.-X., Eden, J.-S., & Holmes, E. C. (2021). Metagenomic identification of viral sequences in laboratory reagents. *Viruses*, *13*(11), 2122. https://doi.org/10.3390/v13112122

Potter, S. C., Luciani, A., Eddy, S. R., Park, Y., Lopez, R., & Finn, R. D. (2018). HMMER web server: 2018 update. *Nucleic Acids Research*, *46*(W1), W200–W204. https://doi.org/10.1093/NAR/GKY448

Qiu, Y., Nakao, R., Ohnuma, A., Kawamori, F., & Sugimoto, C. (2014a). Microbial Population Analysis of the Salivary Glands of Ticks; A Possible Strategy for the Surveillance of Bacterial Pathogens. *PLOS ONE*, *9*(8), e103961. https://doi.org/10.1371/JOURNAL.PONE.0103961

Quast, C., Pruesse, E., Yilmaz, P., Gerken, J., Schweer, T., Yarza, P., Peplies, J., & Glöckner, F. O. (2013). The SILVA ribosomal RNA gene database project: improved data processing and web-based tools. *Nucleic Acids Research*, *41*(Database issue), D590. https://doi.org/10.1093/NAR/GKS1219

Quick, J., Loman, N. J., Duraffour, S., Simpson, J. T., Severi, E., Cowley, L., Bore, J. A., Koundouno, R., Dudas, G., Mikhail, A., Ouédraogo, N., Afrough, B., Bah, A., Baum, J. H. J., Becker-Ziaja, B., Boettcher, J. P., Cabeza-Cabrerizo, M., Camino-Sánchez, Á., Carter, L. L., … Carroll, M. W. (2016). Real-time, portable genome sequencing for Ebola surveillance. *Nature 2016 530:7589*, *530*(7589), 228–232. https://doi.org/10.1038/nature16996

Ramaswamy, K., Rashid, M., Ramasamy, S., Jayavelu, T., & Venkataraman, S. (2022). Revisiting Viral RNA-Dependent RNA Polymerases: Insights from Recent Structural Studies. *Viruses 2022, Vol. 14, Page 2200*, *14*(10), 2200. https://doi.org/10.3390/V14102200

Rambaut, A., Holmes, E. C., O'Toole, Á., Hill, V., McCrone, J. T., Ruis, C., du Plessis, L., & Pybus, O. G. (2020). A dynamic nomenclature proposal for SARS-CoV-2 lineages to assist genomic epidemiology. *Nature Microbiology 2020 5:11*, *5*(11), 1403–1407. https://doi.org/10.1038/s41564-020-0770-5

Remmert, M., Biegert, A., Hauser, A., & Söding, J. (2011). HHblits: lightning-fast iterative protein sequence searching by HMM-HMM alignment. *Nature Methods 2011 9:2*, *9*(2), 173–175. https://doi.org/10.1038/nmeth.1818

Reuter, J. A., Spacek, D. v., & Snyder, M. P. (2015). High-throughput sequencing technologies. *Molecular Cell*, *58*(4), 586–597. https://doi.org/10.1016/J.MOLCEL.2015.05.004

Roux, S., Enault, F., le Bronner, G., & Debroas, D. (2011). Comparison of 16S rRNA and protein-coding genes as molecular markers for assessing microbial diversity (Bacteria and Archaea) in ecosystems. *FEMS Microbiology Ecology*, *78*(3), 617–628. https://doi.org/10.1111/J.1574-6941.2011.01190.X

Roux, S., Hallam, S. J., Woyke, T., & Sullivan, M. B. (2015). Viral dark matter and virus–host interactions resolved from publicly available microbial genomes. *ELife*, *4*(JULY2015). https://doi.org/10.7554/ELIFE.08490.001

Sander, C., & Schneider, R. (1991). Database of homology-derived protein structures and the structural meaning of sequence alignment. *Proteins: Structure, Function, and Bioinformatics*, *9*(1), 56–68. https://doi.org/10.1002/PROT.340090107

Sanger, F., Nicklen, S., & Coulson, A. R. (1977). DNA sequencing with chain-terminating inhibitors. *Proceedings of the National Academy of Sciences of the United States of America*, *74*(12), 5463. https://doi.org/10.1073/PNAS.74.12.5463

Sanjuán, R., Nebot, M. R., Chirico, N., Mansky, L. M., & Belshaw, R. (2010). Viral Mutation Rates. *Journal of Virology*, *84*(19), 9733–9748. https://doi.org/10.1128/JVI.00694-10

Shahid, S., & Slotkin, R. K. (2020). The current revolution in transposable element biology enabled by long reads. *Current Opinion in Plant Biology*, *54*, 49–56. https://doi.org/10.1016/J.PBI.2019.12.012

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 2018 556:7700*, *556*(7700), 197–202. https://doi.org/10.1038/s41586-018-0012-7

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*, *540*(7634), 539–543. https://doi.org/10.1038/nature20167

Shi, M., Lin, X.-D., Vasilakis, N., Tian, J.-H., Li, C.-X., Chen, L.-J., Eastwood, G., Diao, X.-N., Chen, M.-H., Chen, X., Qin, X.-C., Widen, S. G., Wood, T. G., Tesh, R. B., Xu, J., Holmes, E. C., & Zhang, Y.-Z. (2016a). Divergent Viruses Discovered in Arthropods and Vertebrates Revise the Evolutionary History of the Flaviviridae and Related Viruses. *Journal of Virology*, *90*(2), 659–669. https://doi.org/10.1128/jvi.02036-15

Shi, M., Neville, P., Nicholson, J., Eden, J.-S., Imrie, A., & Holmes, E. C. (2017). High-Resolution Metatranscriptomics Reveals the Ecological Dynamics of Mosquito-Associated RNA Viruses in Western Australia. *Journal of Virology*, *91*(17), 680–697. https://doi.org/10.1128/JVI.00680-17

Shi, M., Zhang, Y. Z., & Holmes, E. C. (2018). Meta-transcriptomics and the evolutionary biology of RNA viruses. *Virus Research*, *243*, 83–90. https://doi.org/10.1016/j.virusres.2017.10.016

Shi, W., Shi, M., Que, T. C., Cui, X. M., Ye, R. Z., Xia, L. Y., Hou, X., Zheng, J. J., Jia, N., Xie, X., Wu, W. C., He, M. H., Wang, H. F., Wei, Y. J., Wu, A. Q., Zhang, S. F., Pan, Y. S., Chen, P. Y., Wang, Q., …

Cao, W. C. (2022). Trafficked Malayan pangolins contain viral pathogens of humans. *Nature Microbiology 2022 7:8*, *7*(8), 1259–1269. https://doi.org/10.1038/s41564-022-01181-1

Shmakov, S. A., Wolf, Y. I., Savitskaya, E., Severinov, K. v., & Koonin, E. v. (2020). Mapping CRISPR spaceromes reveals vast host-specific viromes of prokaryotes. *Communications Biology*, *3*(1), 1–9. https://doi.org/10.1038/s42003-020-1014-1

Shope, R. E., & Meegan, J. M. (1997). Arboviruses. *Viral Infections of Humans*, 151–183. https://doi.org/10.1007/978-1-4899-0036-4_6

Simmonds, P. (2015). Methods for virus classification and the challenge of incorporating metagenomic sequence data. *Journal of General Virology*, *96*(6), 1193–1206. https://doi.org/10.1099/JGV.0.000016

Singh, K., Kaur, R., & Qiu, W. (2012). New virus discovery by deep sequencing of small RNAs. In *Methods in Molecular Biology* (Vol. 883, pp. 177–191). Humana Press Inc. https://doi.org/10.1007/978-1-61779-839-9_14

Smith, A., Clark, P., Averis, S., Lymbery, A. J., Wayne, A. F., Morris, K. D., & Thompson, R. A. (2008). Trypanosomes in a declining species of threatened Australian marsupial, the brush-tailed bettong Bettongia penicillata (Marsupialia: Potoroidae). *Parasitology*, *135*(11), 1329–1335. https://doi.org/10.1017/S0031182008004824

Snyder, J. C., Bateson, M. M., Lavin, M., & Young, M. J. (2010). Use of cellular CRISPR (clusters of regularly interspaced short palindromic repeats) spacer-based microarrays for detection of viruses in environmental samples. *Applied and Environmental Microbiology*, *76*(21), 7251–7258. https://doi.org/10.1128/AEM.01109-10

Sobhy, H. (2016). A Review of Functional Motifs Utilized by Viruses. *Proteomes*, *4*(1). https://doi.org/10.3390/PROTEOMES4010003

Söding, J. (2005). Protein homology detection by HMM–HMM comparison. *Bioinformatics*, *21*(7), 951–960. https://doi.org/10.1093/BIOINFORMATICS/BTI125

Stark, R., Grzelak, M., & Hadfield, J. (2019). RNA sequencing: the teenage years. *Nature Reviews Genetics*, *20*(11), 631–656. https://doi.org/10.1038/s41576-019-0150-2

Stewart, A., Armstrong, M., Graves, S., & Hajkowicz, K. (2017). Rickettsia australis and Queensland Tick Typhus: A Rickettsial Spotted Fever Group Infection in Australia. *The American Journal of Tropical Medicine and Hygiene*, *97*(1), 24. https://doi.org/10.4269/AJTMH.16-0915

Stork, N. E. (2018). How many species of insects and other terrestrial arthropods are there on Earth? *Https://Doi.Org/10.1146/Annurev-Ento-020117-043348*, *63*, 31–45. https://doi.org/10.1146/ANNUREV-ENTO-020117-043348

Tarazona, S., García-Alcalde, F., Dopazo, J., Ferrer, A., & Conesa, A. (2011). Differential expression in RNA-seq: a matter of depth. *Genome Research*, *21*(12), 2213–2223. https://doi.org/10.1101/GR.124321.111

te Velthuis, A. J. W. (2014). Common and unique features of viral RNA-dependent polymerases. *Cellular and Molecular Life Sciences*, *71*(22), 4403–4420. https://doi.org/10.1007/S00018-014-1695-Z

Thompson, J. D., Ueffing, M., & Schaeffer-Reiss, C. (Eds.). (2008). *Functional Proteomics* (Vol. 484). Humana Press. https://doi.org/10.1007/978-1-59745-398-1

Tikhonenkov, D. v., Gawryluk, R. M. R., Mylnikov, A. P., & Keeling, P. J. (2021). First finding of free-living representatives of Prokinetoplastina and their nuclear and mitochondrial genomes. *Scientific Reports 2021 11:1*, *11*(1), 1–21. https://doi.org/10.1038/s41598-021-82369-z

Tran, Q., Pham, D. T., & Phan, V. (2017). Using 16S rRNA gene as marker to detect unknown bacteria in microbial communities. *BMC Bioinformatics*, *18*(14), 155–161. https://doi.org/10.1186/S12859-017-1901-8

Urayama, S. I., Takaki, Y., Chiba, Y., Zhao, Y., Kuroki, M., Hagiwara, D., & Nunoura, T. (2022). Eukaryotic Microbial RNA Viruses—Acute or Persistent? Insights into Their Function in the Aquatic Ecosystem. *Microbes and Environments*, *37*(5), ME22034. https://doi.org/10.1264/JSME2.ME22034

van der Valk, T., Vezzi, F., Ormestad, M., Dalén, L., & Guschanski, K. (2020). Index hopping on the Illumina HiseqX platform and its consequences for ancient DNA studies. *Molecular Ecology Resources*, *20*(5), 1171–1181. https://doi.org/10.1111/1755-0998.13009

van Rooij, P., Martel, A., Haesebrouck, F., & Pasmans, F. (2015). Amphibian chytridiomycosis: a review with focus on fungus-host interactions. *Veterinary Research 2015 46:1*, *46*(1), 1–22. https://doi.org/10.1186/S13567-015-0266-0

Walker, P. J., Siddell, S. G., Lefkowitz, E. J., Mushegian, A. R., Adriaenssens, E. M., Alfenas-Zerbini, P., Dempsey, D. M., Dutilh, B. E., García, M. L., Curtis Hendrickson, R., Junglen, S., Krupovic, M., Kuhn, J. H., Lambert, A. J., Łobocka, M., Oksanen, H. M., Orton, R. J., Robertson, D. L., Rubino, L., … Zerbini, F. M. (2022). Recent changes to virus taxonomy ratified by the International Committee on Taxonomy of Viruses (2022). *Archives of Virology*, *167*(11), 2429–2440. https://doi.org/10.1007/S00705-022-05516-5

Wallau, G. L. (2022). RNA virus EVEs in insect genomes. *Current Opinion in Insect Science*, *49*, 42–47. https://doi.org/10.1016/J.COIS.2021.11.005

Webby, R. J., & Webster, R. G. (2001). Emergence of influenza A viruses. *Philosophical Transactions of the Royal Society of London. Series B: Biological Sciences*, *356*(1416), 1817–1828. https://doi.org/10.1098/RSTB.2001.0997

Webster, C. L., Longdon, B., Lewis, S. H., & Obbard, D. J. (2016). Twenty-five new viruses associated with the drosophilidae (Diptera). *Evolutionary Bioinformatics*, *12*, 13–25. https://doi.org/10.4137/EBO.S39454

Webster, C. L., Waldron, F. M., Robertson, S., Crowson, D., Ferrari, G., Quintana, J. F., Brouqui, J.-M., Bayne, E. H., Longdon, B., Buck, A. H., Lazzaro, B. P., Akorli, J., Haddrill, P. R., & Obbard, D. J. (2015). The discovery, distribution, and evolution of viruses associated with *Drosophila melanogaster*. *PLoS Biology*, *13*(7), e1002210. https://doi.org/10.1371/journal.pbio.1002210

Weeks, A. R., Turelli, M., Harcombe, W. R., Reynolds, K. T., & Hoffmann, A. A. (2007). From Parasite to Mutualist: Rapid Evolution of Wolbachia in Natural Populations of Drosophila. *PLOS Biology*, *5*(5), e114. https://doi.org/10.1371/journal.pbio.0050114

Wheeler, D., & Bhagwat, M. (2007). BLAST QuickStart. *Methods in Molecular Biology*, *395*, 149–176. https://doi.org/10.1385/1-59745-514-8:149

Wille, M., Eden, J. S., Shi, M., Klaassen, M., Hurt, A. C., & Holmes, E. C. (2018). Virus–virus interactions and host ecology are associated with RNA virome structure in wild birds. *Molecular Ecology*, *27*(24), 5263–5278. https://doi.org/10.1111/MEC.14918

Wobus, C. E. (2018). The Dual Tropism of Noroviruses. *Journal of Virology*, *92*(16). https://doi.org/10.1128/JVI.01010-17

Wolf, Y. I., Silas, S., Wang, Y., Wu, S., Bocek, M., Kazlauskas, D., Krupovic, M., Fire, A., Dolja, V. v., & Koonin, E. v. (2020). Doubling of the known set of RNA viruses by metagenomic analysis of an aquatic virome. *Nature Microbiology 2020 5:10*, *5*(10), 1262–1270. https://doi.org/10.1038/s41564-020-0755-4

Woo, P. C. Y., Lau, S. K. P., Teng, J. L. L., Tse, H., & Yuen, K. Y. (2008). Then and now: use of 16S rDNA gene sequencing for bacterial identification and discovery of novel bacteria in clinical microbiology laboratories. *Clinical Microbiology and Infection*, *14*(10), 908–934. https://doi.org/10.1111/J.1469-0691.2008.02070.X

Wood, D. E., & Salzberg, S. L. (2014). Kraken: Ultrafast metagenomic sequence classification using exact alignments. *Genome Biology*, *15*(3), 1–12. https://doi.org/10.1186/GB-2014-15-3-R46

Woods, R., Reiss, A., Cox-Witton, K., Grillo, T., & Peters, A. (2019). The Importance of Wildlife Disease Monitoring as Part of Global Surveillance for Zoonotic Diseases: The Role of Australia. *Tropical Medicine and Infectious Disease 2019, Vol. 4, Page 29*, *4*(1), 29. https://doi.org/10.3390/tropicalmed4010029

Wu-Chuang, A., Hodžić, A., Mateos-Hernández, L., Estrada-Peña, A., Obregon, D., & Cabezas-Cruz, A. (2021). Current debates and advances in tick microbiome research. *Current Research in Parasitology & Vector-Borne Diseases*, *1*, 100036. https://doi.org/10.1016/J.CRPVBD.2021.100036

Wyatt, K. B., Campos, P. F., Gilbert, M. T. P., Kolokotronis, S. O., Hynes, W. H., DeSalle, R., Daszak, P., MacPhee, R. D. E., & Greenwood, A. D. (2008). Historical Mammal Extinction on Christmas Island (Indian Ocean) Correlates with Introduced Infectious Disease. *PLOS ONE*, *3*(11), e3602. https://doi.org/10.1371/journal.pone.0003602

Zhang, Y.-Z., Shi, M., & Holmes, E. C. (2018). Using metagenomics to characterize an expanding virosphere. *Cell*, *172*(6), 1168–1172. https://doi.org/10.1016/j.cell.2018.02.043

Zhao, G., Wu, G., Lim, E. S., Droit, L., Krishnamurthy, S., Barouch, D. H., Virgin, H. W., & Wang, D. (2017). VirusSeeker, a computational pipeline for virus discovery and virome composition analysis. *Virology*, *503*, 21–30. https://doi.org/10.1016/J.VIROL.2017.01.005

Zhao, S., Ye, Z., & Stanton, R. (2020). Misuse of RPKM or TPM normalization when comparing across samples and sequencing protocols. *RNA*, *26*(8), 903–909. https://doi.org/10.1261/rna.074922.120

Zhou, H., Ji, J., Chen, X., Bi, Y., Li, J., Wang, Q., Hu, T., Song, H., Zhao, R., Chen, Y., Cui, M., Zhang, Y., Hughes, A. C., Holmes, E. C., & Shi, W. (2021). Identification of novel bat coronaviruses sheds light on the evolutionary origins of SARS-CoV-2 and related viruses. *Cell*, *184*(17), 4380-4391.e14. https://doi.org/10.1016/j.cell.2021.06.008

Chapter 2

A Divergent Articulavirus in an Australian Gecko Identified Using Meta-Transcriptomics and Protein Structure Comparisons

---

Ayda Susana Ortiz-Baez, John-Sebastian Eden, Craig Moritz and Edward C. Holmes

viruses

Article

CHAPTER 2 Adivergent *Articulavirus* in an Australian gecko identified using meta-transcriptomics and protein structure comparisons

Ayda Susana Ortiz-Baez[1], John-Sebastian Eden[1,2], Craig Moritz[3] and Edward C. Holmes[1]*

[1] Marie Bashir Institute for Infectious Diseases and Biosecurity, School of Life and Environmental Sciences and School of Medical Sciences, The University of Sydney, Sydney, New South Wales 2006, Australia.

[2] Centre for Virus Research, Westmead Institute for Medical Research, Westmead, NSW 2145, Australia.

[3] Research School of Biology & Centre for Biodiversity Analysis, The Australian National University, Acton, ACT 6201, Australia.

* Correspondence: edward.holmes@sydney.edu.au

Received: 2 May 2020; Accepted: 3 June 2020; Published: 4 June 2020;
https://doi.org/10.3390/v12060613

2.1 Abstract

The discovery of highly divergent RNA viruses is compromised by their limited sequence similarity to known viruses. Evolutionary information obtained from protein structural modelling offers a powerful approach to detect distantly related viruses based on the conservation of tertiary structures in key proteins such as the RNA-dependent RNA polymerase (RdRp). We utilised a template-based approach for protein structure prediction from amino acid sequences to identify distant evolutionary relationships among viruses detected in meta-transcriptomic sequencing data from Australian wildlife. The best predicted protein structural model was compared with the results of similarity searches against protein databases. Using this combination of meta-transcriptomics and protein structure prediction

we identified the RdRp (PB1) gene segment of a divergent negative-sense RNA virus, denoted *Lauta virus* (LTAV), in a native Australian gecko (*Gehyra lauta*). The presence of this virus was confirmed by PCR and Sanger sequencing. Phylogenetic analysis revealed that *Lauta virus* is a newly described genus within the family *Amnoonviridae,* order *Articulavirales*, that is most closely related to the fish virus *Tilapia tilapinevirus* (TiLV). These findings provide important insights into the evolution of negative-sense RNA viruses and structural conservation of the viral replicase among members of the order *Articulavirales*.

Keywords: virus discovery; protein structure; meta-transcriptomics; *Tilapia tilapinevirus*; *Articulavirales*; *Amnoonviridae*; RNA virus; *Lauta virus*; gecko

---

2.2. Introduction

The development of next-generation sequencing technologies (NGS), including total RNA sequencing (meta-transcriptomics), has revolutionized studies of virome diversity and evolution (Shi et al., 2018; Thermes, 2014; Zhang et al., 2019). Despite this, the discovery of highly divergent viruses remains challenging because of the often limited (or no) primary sequence similarity between putative novel viruses and those for which genome sequences are already available (R. Rose et al., 2016; Shi et al., 2016a; Zhang et al., 2018). For example, it is possible that the small number of families of RNA viruses found in bacteria, as well as their effective absence in archaeabacteria, in reality reflects the difficulties in detecting highly divergent sequences rather than their true absence from these taxa (Zhang et al., 2019).

The conservation of protein structures in evolution and the limited number of proteins folds (fold space) in nature form the basis of template-based protein structure prediction (Deng et al., 2018a), providing a powerful way to reveal the origins and evolutionary history of viruses (Bamford et al., 2005b; Holmes, 2011). Indeed, the utility of protein structural similarity in revealing key aspects of virus evolution is well known (Bamford et al., 2005c; Benson et al., 2004). For instance, double-strand (ds) DNA viruses including the thermophilic archaeal virus

STIV, enterobacteria phage PRD1, and human adenovirus exhibit conserved viral capsids, suggesting a deep common ancestry (G. Rice et al., 2004). Thus, protein structure prediction utilising comparisons to solved protein structures can assist in the identification of novel viruses (Baker & Sali, 2001; Deng et al., 2018b). Herein, we use this method as an alternative approach to virus discovery.

There is a growing availability of three-dimensional structural data in curated databases such as the Protein Data Bank (PDB), with approximately 11,000 viral protein solved structures that can be used in comparative studies. Importantly, these include a limited number (around 115) structures of the RNA-dependent RNA polymerase (RdRp) from a variety of viral groups. Viral RdRp proteins are catalytic proteins (∼460 to ∼1930 residues) implicated in the low-fidelity replication of the genetic material of RNA viruses (Jia & Gong, 2019). The tertiary structure of the RdRp contains an active site (core component), and three subdomains: palm, fingers and thumb that resemble a right-hand shape. These subdomains include seven catalytic motifs (G, F1–3, A, B, C, D and E) that are central to polymerase function (Černý et al., 2014; te Velthuis, 2014). The palm subdomain comprises several key conserved motifs (denoted A–E), including the aspartate residues (xDD) in motif C, that constitute a highly conserved element in the RdRp that is central to catalytic activity (te Velthuis, 2014). The RdRp exhibits the highest level of sequence similarity (although still limited) among RNA viruses, and hence is expected to contain relatively well conserved protein structures. Exploiting such structural features in combination with metagenomic data will undoubtedly improve our ability to detect divergent viruses in nature, particularly in combination with wildlife surveillance (Shi 2018; Shi et al., 2016; Zhang et al., 2018).

The International Committee on Taxonomy of Viruses (ICTV) recently introduced the *Amnoonviridae* as a newly recognized family of negative-strand RNA viruses present in fish (ICTV Master Species List 2018b.v2). Together with the *Orthomyxoviridae*, the *Amnoonviridae* are classified in the order *Articulavirales*, describing a set of negative-sense RNA viruses with segmented genomes. While the *Orthomyxoviridae* includes seven genera, four of these

comprise influenza viruses (FLUV), and to date the family *Amnoonviridae* comprises a single genus – *Tilapinevirus* – which in turn includes only a single species - *Tilapia tilapinevirus* or Tilapia Lake virus (TiLV).

TiLV was originally identified in farmed tilapine populations (*Oreochromis niloticus*) in Israel and Ecuador (Bacharach et al., 2016). The virus has now been described in wild and hybrid tilapia in several countries in the Americas, Africa, Asia, and Southeast Asia (Ahasan et al., 2020; Jansen et al., 2019b; Pulido et al., 2019). TiLV has been associated with high morbidity and mortality in infected animals. Pathological manifestations include syncytial hepatitis, skin erosion and encephalitis (Jansen et al., 2019c; Subramaniam et al., 2019a). TiLV was initially classified as a putative orthomyxo-like virus based on weak sequence resemblance (~17% amino acid identity) in the PB1 segment that contains the RdRp, as well as the presence of conserved 5′ and 3′ termini (Bacharach et al., 2016). While both the *Orthomyxoviridae* and *Amnoonviridae* have negative-sense, segmented genomes, the genomic organization of the *Amnoonviridae* comprises 10 instead of 7-8 segments (Al-Hussinee et al., 2018; Bacharach et al., 2016; Subramaniam et al., 2019), and their genomes are shorter (~10 kb) than those of the *Orthomyxoviridae* (~12-15 kb). To date, however, only the RdRp (encoded by a 1641 bp PB1 sequence) has been reliably defined and most segments carry proteins of unknown function. Importantly, comparisons of TiLV RdRp with sequences from members of the *Orthomyxoviridae* revealed the presence of four conserved amino acid motifs (I-IV) of size 4-9 residues each (Bacharach et al., 2016) that effectively comprise a "molecular fingerprint" for the order.

Unlike other members of the *Articulavirales* (Payne, 2017), TiLV appears to have a limited host range and has been only documented in tilapia (*O. niloticus*, *O*. sp.) and hybrid tilapia (*O. niloticus* x *O. aureus*). Herein, we report the discovery of a divergent virus from an Australian gecko (*Gehyra lauta*) using a combination of meta-transcriptomic and structure-based approaches, and employ a phylogenetic analysis to reveal its relationship to TiLV. Our work suggests that this Gecko virus likely represents a novel genus within the *Amnoonviridae*.

2.3 Materials and Methods

2.3.1. Sample collection

A total of seven individuals corresponding to the reptile species *Carlia amax, Carlia gracilis, Carlia munda, Gehyra lauta, Gehyra nana, Heteronotia binoei,* and *Heteronotia planiceps* were collected alive in 2013 from Queensland, Australia. Specimens were identified by mtDNA typing and/or morphological data. Livers were harvested and stored in RNAlater at -80°C before downstream processing. All sampling was conducted in accordance with animal ethics approval (#A2012/14) from the Australian National University and collection permits from the Parks and Wildlife Commission of the Northern Territory (#45090), the Australian Government (#AU-COM2013-192), and the Department of Environment and Conservation (#SF009270).

2.3.2. Sampling processing and sequencing

RNA extraction was performed using the RNeasy Plus minikit (Qiagen, Hilden, Germany) following manufacturer's instructions. Each of the seven livers were extracted individually and then pooled in equal amounts. For RNA sequencing, ribosomal RNA (rRNA) was depleted using the RiboZero (epidemiology) depletion kit and libraries were prepared with the TruSeq stranded RNA library prep kit before sequencing on an Illumina HiSeq 2500 platform (100 bp paired end reads). Library preparation and sequencing was performed by the Australian Genome Research Facility (AGRF), generating a total of 22,394,787 paired end reads for the pooled liver RNA library.

2.3.3. De novo assembly and sequence annotation

Raw Illumina reads were trimmed of sequencing adapters and low-quality bases with Trimmomatic v0.38 (Bolger et al., 2014a). The trimmed reads were then *de novo* assembled into contigs (transcripts) using Trinity v2.6.6 (Grabherr et al., 2011a) with default parameter settings. Contig abundance was estimated with RSEM (B. Li & Dewey, 2011) and shown as the

numbers of transcripts per million (TPM). For sequence annotation, contigs were compared against the NCBI nucleotide (nt) and non-redundant (nr) protein databases (nr) using BLASTn v.2.8 (released on 2018-03-28) (Altschul et al., 1990) and DIAMOND v.0.921 (Buchfink et al., 2015), respectively.

2.3.4. Protein structure prediction for virus detection

To further screen the meta-transcriptomic data, all the assembled sequences below the assigned threshold (e-value $\geq 10^{-5}$) were assigned as "orphan" contigs (n= 293,586). These were then analysed using a protein structure-informed approach. Specifically, orphan contigs were translated into all six open reading frames (ORFs) using the getorf program (P. Rice et al., 2000a) to identify continuous ORFs of at least 1000nt in length (n=57). To detect distant sequence homologies and predict viral protein structures, this subset of translated ORFs were then analysed using a template-based modelling approach as implemented in Phyre2 (http://www.sbg.bio.ic.ac.uk/phyre2) (Kelley et al., 2015). In brief, target proteins were compared against proteins of known structure via homology modelling and fold recognition, followed by loop modelling and sidechain fitting (Kelley et al., 2015). In total, 6 of 14 confident (i.e. confidence values >90%) matches to known viral structures were identified. These included a single match to the RdRp of a vertebrate-associated virus, and the queried contig was selected for downstream analyses. Annotations from the predicted model were used as preliminary data for tentative taxonomic assignment and protein classification. The structural alignment between the PDB of the predicted model and the PDB of the template was performed using TM-align v.20190822 (Zhang & Skolnick, 2005) with default settings, and visualized using PyMOLv.2.3.5 (Schrödinger & DeLano, 2015).

2.3.5. Annotation of the newly discovered virus

To corroborate the viral origin of the predicted protein structure and gain insights into its taxonomic classification, we conducted parallel comparisons using DIAMOND (Buchfink et al., 2015) against the GenBank non-redundant (nr) database (https://www.ncbi.nlm.nih.gov/) and

the HMMER web server v2.41.1 (http://www.ebi.ac.uk/Tools/hmmer) against the following profile databases using default e-value cut-offs to assign significance: (i) reference proteomes v.2019_09, downloaded on 2019-10-03 (https://proteininformationresource.org/rps/), (ii) Uniprot v.2019_09, downloaded on 2019-10-03 (https://www.uniprot.org/), and (iii) Swiss-Prot v.2019_09, downloaded on 2019-10-03. Protein families were identified using Pfam v.32.0 (https://pfam.xfam.org/). In addition, conserved domains were annotated using the Conserved Domain Database (CDD) v.3.17 and the CD-search tool (http://www.ncbi.nlm.nih.gov/Structure/cdd/cdd.shtml). To detect additional contigs and better characterize the genome of the novel virus, we aligned the cDNA contigs against custom databases using DIAMOND v.0.9.32 (Buchfink et al., 2015), including (i) reference sequences corresponding to all the segments of TiLV (**Table S2.1**), and (ii) reference RdRp sequences from the order *Articulavirales* (**Table S2.2**). Given the divergent nature of these viruses, we considered all hits with e-value >10^{-4} in the analyses using DIAMOND.

2.3.6. Phylogenetic analysis

The predicted contig encoding the RdRp of the newly discovered virus was aligned with reference protein sequences of the order *Articulavirales* (**Table S2.3**). A multiple amino acid sequence alignment was performed using the E-INS-i algorithm as implemented in the MAFFT v7.450 program (Katoh & Standley, 2013a). Selection of the best-fit model of amino acid substitution was carried out using the Akaike Information criterion (AIC) and the Bayesian Information Criterion (BIC) with the standard model selection option (-m TEST) in IQ-TREE (Nguyen et al., 2015a). Phylogenetic analysis of these data was then performed using the Maximum Likelihood (ML) method available in IQ-TREE, with node support estimated with the ultra-fast bootstrap (UFBoot) approximation (1000 replicates) and the Shimodaira-Hasegawa approximate Likelihood ratio test (SH-aLRT). Sequencing reads are available at the NCBI Sequence Read Archive (SRA) under the Bioproject PRJNA626677 (BioSample: SAMN14647831; Sample name: VERT7; SRA: SRS6507258). The assembled sequence for the newly determined *Lauta virus* was deposited in GenBank under the accession number MT386081.

2.3.7.PCR validation

To validate the presence of the novel gecko amnoonvirus, and to identify the putative host species, we screened the individual liver RNA using RT-PCR. Briefly, cDNA was prepared using Superscript IV VILO master mix and RT-PCR was performed with the Platinum SuperFi Green PCR master mix and two primers sets targeting the gecko RdRp contig – F2V7 and F3V7 (**Table S2.4**). The resultant RT-PCR products were analysed by agarose gel electrophoresis and validated by Sanger sequencing.

2.4 Results

2.4.1. Virus discovery using meta-transcriptomics and protein structural features

We employed a meta-transcriptomic approach to screen a single pooled library containing liver RNA of seven Australian native reptile species (*Gehyra lauta, Carlia amax, Heteronotia binoei, Gehyra nana, Carlia gracilis, Carlia munda*, and *Heteronotia planiceps*). We focused on the *de novo* assembled contigs that had no significant hits using initial searches against the NCBI nucleotide and non-redundant databases. Accordingly, of 293,586 orphan contigs, 57 contained translatable ORFs of more than 1000 nt in length, and because we hypothesized that some may correspond to undetected virus sequences, we interrogated them using a protein structure prediction approach with template-based modelling (TBM) in Phyre2 (Kelley et al., 2015). From the 57 queried contigs, we obtained a 3D model of a 407 amino acid (1227 bp) contig with a high confidence hit (98.3%) to the RdRp catalytic subunit of a bat influenza A virus (family *Orthomyxoviridae*) (**Table 2.1, Figure 2.1a-b**). This level of confidence is indicative of a high probability of modelling success. Predicted secondary structures for the modeled protein corresponded to α-helix (50%) and β-strand (9%) conformations. In addition, the alignment coverage between our query and the viral template (PDB identifier: 4WSB) corresponded to 52% (213 residues) of the query sequence, while the proportion of identical amino acids (i.e. sequence identity) was 19% (**Table 2.1**). Despite this low sequence similarity, we observed

common folding patterns in the palm domain of the RdRp between the aligned protein structures (**Figure 2.1a**).

Table 2.1 Summary of analyses and parameters used for the detection of *Lauta virus*.

Analysis/database	Parameter (unit)	Value / Hit (e-value)
Trinity *de novo* assembly	Length (nt)	1227
	Predicted ORF length (aa)	407
	Coverage (# of reads)	35
	Abundance (TPM [1])	1.10
Phyre2/PDB	PDB molecule	RdRp catalytic subunit
	PDB title	Bat influenza a polymerase with bound vRNA promoter
	PDB identifier	4WSB
	Resolution	2.65
	Confidence (%)	98.3
	Coverage (%)	52
	Identity (%)	19
DIAMOND/nr	Match	[QES69295.1] Hypothetical protein (Tilapia lake virus), segment 1
	Similarity (%)	29
	e-value	1.30e-07
DIAMOND/custom db RdRp subunit PB1	Match	[YP_009246481] Hypothetical protein (Tilapia lake virus), segment 1
	Similarity (%)	29
	e-value	2.4e-14
HMMER/references proteomes	Taxonomy	Tilapia lake virus (3.9e-11)
	Domain architecture	Flu_PB1
HMMER/UniProt	Taxonomy	Tilapia lake virus (1.4e-10)
	Domain architecture	Flu_PB1
HMMER/SwissProt	Taxonomy	Infectious salmon anaemia virus RDRP_ISAV8, segment 2 (5.2e-3)
	Domain architecture	Flu_PB1
Pfam	Family	Flu_PB1 (1.8e-2)
	Description	Influenza RNA-dependent RNA polymerase subunit PB1

CDD	Domain hit	Flu_PB1 super family (6.43e-05)

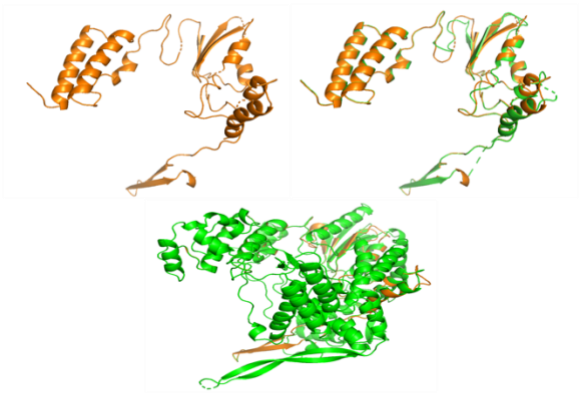
[1] TPM: transcripts per million.

To corroborate these findings, the structural results were compared with those obtained from other analyses based on primary sequence similarity searches against public databases (**Table 2.1**). This revealed matches to the RdRp subunit (PB1 gene segment) of different members of the order *Articulavirales*, including influenza A virus (FLUAV), TiLV, and Infectious salmon anaemia virus (ISAV). Comparisons of the assembled contigs against a custom database containing only members of the *Articulavirales* were then performed to improve sequence alignments. Accordingly, the best hit matches were obtained to TiLV (e-values $<10^{-15}$) (**Table 2.1**). To identify additional viral segments, the assembled contigs were aligned to the ten segments of TiLV using DIAMOND. A total of 87 contigs were scored across the genome, although we did not recover any significant hit for segments 2-10 likely because they are so divergent in sequence (**Table S2.1**).

2.4.2. Sequence alignment and phylogenetic relationships

We tentatively name the new virus identified here as *Lauta virus* (reflecting the species name of the gecko in which it was identified), abbreviated as LTAV. Multiple sequence alignment of the RdRp between *Lauta virus* and other members the order *Articulavirales* identified a number of well conserved amino acid motifs (I-IV) ranging in length from 5-11 amino acids in length (**Figure 2.2**). Phylogenetic analysis of the aligned RdRp region revealed that LTAV falls within the order *Articulavirales* and, along with TiLV (family *Amnoonviridae*), comprises a distinct monophyletic group. The close relationship between LTAV and TiLV was supported by high UFBoot/SH-aLRT values (99%/99%) (**Figure 2.1c**). Likewise, estimates of the amino acid identity in the RdRp showed a closer (but still distant) sequence similarity (15.35%) with TiLV than other members of the order *Articulavirales* (**Table 2.2**).

a



b


1
200
201
400
401
Confidence Key
High(9)
Low (0)

c

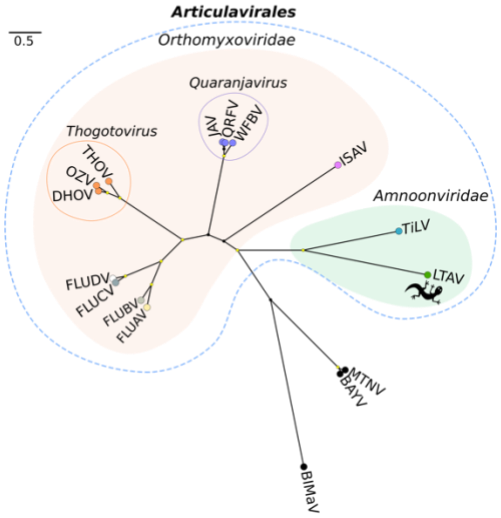

Articulavirales
0.5
Orthomyxoviridae
Quaranjavirus
JAV
ORFV
WFBV
Thogotovirus
THOV
OZV
DHOV
ISAV
Amnoonviridae
TiLV
LTAV
FLUDV
FLUCV
FLUBV
FLUAV
MTNV
BAYV
BlMaV

Figure 2.1 Protein structure prediction and phylogenetic relationships of *Lauta virus* (LTAV). (a) 3D model prediction of the RdRp subunit PB1 of LTAV (top left). Protein structure superposition in the aligned region between the predicted model for LTAV and the RdRp (PB1 gene) of influenza A virus (FLUAV) (top right). Protein structure superposition of the predicted model for LTAV and the entire RdRp subunit of FLUAV (bottom). The protein structure predicted for LTAV is displayed in orange and that of FLUAV in green. (b) Confidence summary of residues modelled. (c) Maximum likelihood tree depicting the phylogenetic relationships between LTAV and TiLV within the family *Amnoonviridae,* order *Articulavirales*. Families are indicated with colored filled bubbles. Tip labels are colored according to genus. Genera comprising multiple species are indicated with unfilled bubbles. Support values >= 95% UFBoot and 80% SH-aLRT are displayed with yellow-circle shapes at nodes. *Alphainfluenzavirus* (FLUBA); *Betainfluenzavirus* (FLUBV); *Deltainfluenzavirus* (FLUDV); *Gammainfluenzavirus* (FLUCV); *Dhori thogotovirus* (DHOV); Oz virus (OZV); *Thogoto thogotovirus* (THOV); *Quaranfil quaranjavirus* (QRFV); *Wellfleet Bay virus* (WFBV); *Johnston Atoll quaranjavirus* (JAV); *Salmon isavirus* (ISAV); *Tilapia tilapinevirus* (TiLV); *Lauta virus* (LTAV; gecko symbol); *Blueberry mosaic associated virus* (BIMaV); *Montano orthohantavirus* (MTNV); *Bayou orthohantavirus* (BAYV).

Table 2.2 Percentage of identical residues among members of the order *Articulavirales* and *Lauta virus*.

Virus classification			**Percentage of amino acid Identity [1]**		
Family	**Genus**	**Species**	**FLUAV**	**TiLV**	**LTAV**
Orthomyxoviridae	*Alphainfluenzavirus*	FLUAV	–	13.90	11.75
	Betainfluenzavirus	FLUBV	60.37	13.33	12.01
	Deltainfluenzavirus	FLUDV	39.03	14.62	11.53
	Gammainfluenzavirus	FLUCV	38.63	14.50	12.66
	Isavirus	ISAV	18.40	11.84	11.41
	Quaranjavirus	QRFV	22.94	13.68	11.46
	Thogotovirus	THOV	24.90	14.61	13.08
Amnoonviridae	*Tilapinevirus*	TiLV	13.90	–	15.35

[1] Percentage of identical bases/residues

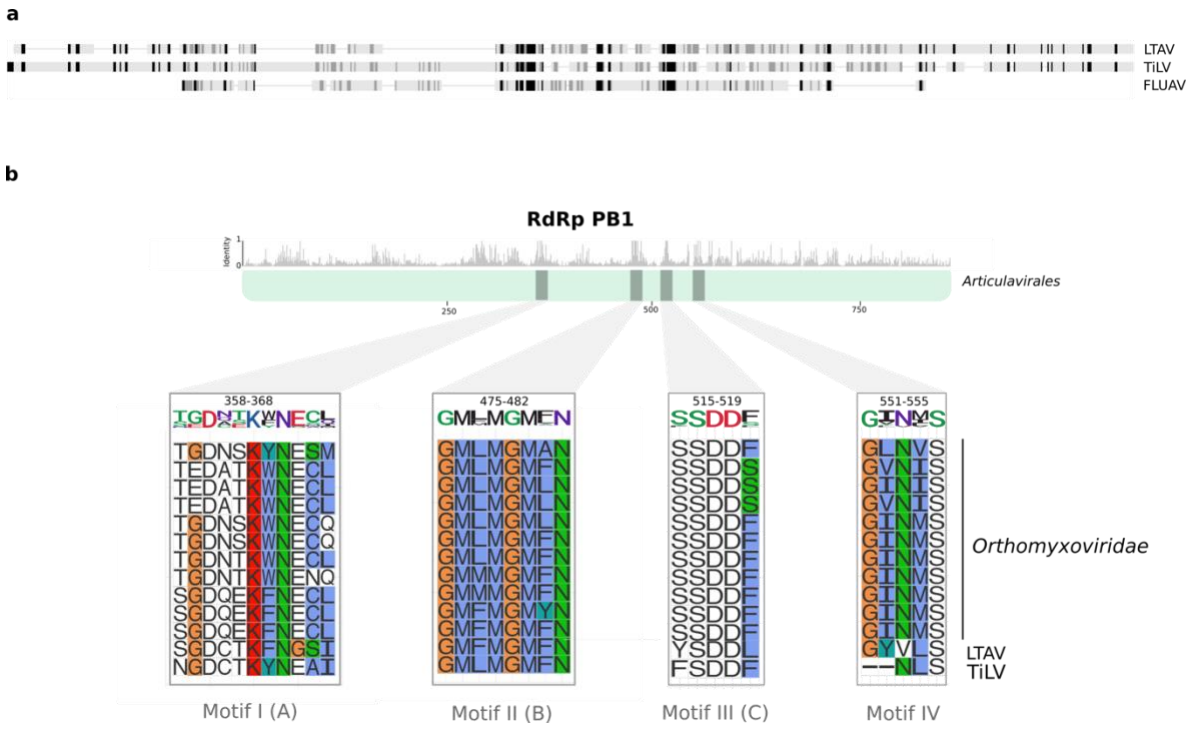


Figure 2.2 Conserved motifs in the RdRp subunit PB1 from the order *Articulavirales*. (a) Comparison of the *Lauta virus* RdRp sequence with the full-length PB1 sequence of TiLV and FLUAV. The gradient from black to light grey indicates the level of sequence similarity in the alignment. Highly conserved positions are shown in black. (b) Top panel shows the mean pairwise identity over all pairs in the column across the multiple sequence alignment among members of the order *Articulavirales*. Sequence motifs are shown with grey bars. The bottom panel depicts a magnified view of individual motifs. Letters in parenthesis denote the A–G RdRp motif nomenclature. The original amino acid residue position and standard logos are displayed in the top of each motif; the size of each character represents the level of sequence conservation. Amino acid residues in the alignment are coloured according to the Clustal colouring scheme.

2.4.3. Host association and in vitro validation

Lauta virus was initially identified in the pooled sequencing library comprising a mix of several Australian reptile species. To identify the exact host species, we screened each individual species sample separately using RT-PCR and Sanger sequencing. As a result, we

detected the presence of the novel *Lauta virus* RdRp sequence in liver tissue of *G. lauta* (paratype QM J96622) (**Figure S2.1**), a gecko species native to north-western Queensland and the north-eastern Northern territory in Australia (Oliver et al., 2020).

2.5 Discussion

Advances in protein modelling and sequence analysis based on structural comparisons with well-characterized protein templates constitute an attractive approach for the identification of highly divergent RNA viruses (Kelley et al., 2015). The RdRp is ubiquitous in RNA viruses with different genomic architectures and replication strategies, showing a conserved core with sequence motifs that adopt specific folds. The protein is critically required for RNA synthesis and replication in RNA viruses (i.e. template recognition, initiation, elongation and regulation) (te Velthuis, 2014). As proteins such as the RdRp play such a central role in the life-cycle of RNA viruses it is expected that structures and key motifs for catalytic functionality will be relatively well conserved through evolutionary history (Ng et al., 2008; Wolf et al., 2018). Based on this premise, it is expected that template-based protein structure modelling could be a powerful tool in the identification of highly divergent viruses (Deng et al., 2018b; Fiser, 2010; Kelley et al., 2015). Accordingly, we used protein structural similarity in combination with sequence and a profile similarity to identify a novel and divergent RNA virus in an Australian gecko (*G. lauta*).

We obtained a confident predicted 3D model for the RdRp of *Lauta virus* based on its structural similarity with the RdRp subunit PB1 of influenza virus (family *Orthomyxoviridae*) (**Figure 2.1a-b; Table 2.1**). Although the structural data suggested that *Lauta virus* belonged to the family *Orthomyxoviridae* (order *Articulavirales*) (Kelley et al., 2015), additional sequence analysis revealed a closer relationship to members of the *Amnoonviridae* (**Figure 2.1c**). In this context it is important to recall that biases in taxonomic assignment can occur because of the limited number of available proteins with known structures in the PDB. Although this is clearly a limitation, template-based approaches offer a tractable starting point for virus discovery and its taxonomic classification.

Although compromised by the large evolutionary distances involved, phylogenetic analysis among members of the order *Articulavirales* revealed that *Lauta virus* was most closely related to TiLV, in turn suggesting that it is a novel and divergent genus within the *Amnoonviridae*. To date, members of the *Amnoonviridae* have only been detected in fish (Bacharach et al., 2016), such that the discovery of *Lauta virus* expands the host range of this family. Indeed, given the huge genetic distance between TiLV and LTAV, we expect that further uncharacterised phylogenetic diversity exists in the *Amnoonviridae* especially in fish and reptiles, and that more studies using the form of genomic surveillance performed here will capture a far greater diversity of negative-sense RNA viruses (C.-X. Li et al., 2015; Shi et al., 2016).

Comparisons of the RdRp subunit PB1 from different articulaviruses revealed the presence of four well conserved motifs in *Lauta virus*, broadly consistent with observations made for TiLV (Bacharach et al., 2016). As suggested by several studies, motifs I-IV are critically implicated in the catalytic activity of PB1 (Biswas & Nayak, 1994; Chu et al., 2012). Despite minor variations, we identified the SDD (serine-aspartic acid-aspartic acid) sequence in motif III that is presumed to be essential for protein functionality in FLUV (Biswas & Nayak, 1994; Chu et al., 2012). Hence, the presence of well conserved motifs I-IV across the order *Articulavirales* may constitute effective molecular fingerprints for these viruses. Unfortunately, the marked lack of sequence similarity meant we did not recover any conclusive evidence regarding presence of other genome segments in *Lauta virus*. Further studies that include sequencing, microscopy, and cell culture techniques, are therefore required to fully characterize the genome of this novel virus.

The identification of a novel virus in an Australian gecko (*G. lauta*) highlights the importance of virus surveillance in native species. Although *Lauta virus* was detected in liver tissue, we currently cannot draw any conclusions regarding its pathogenic potential and impact on the health of *G. lauta*, particularly since a limited number of individuals were collected and all were apparently healthy. Additional research is therefore needed to establish the type of

biological interaction between *Lauta virus* and *G. lauta*. While a previous study reported the isolation of the arbovirus Charleville virus (family *Rhabdoviridae*) in *G. australis* (possibly *G. dubia* based on its distribution) collected in Queensland [36,37], this is the first report of a divergent articulavirus in reptiles. Taken together, these findings hint at a hidden diversity of RNA viruses in reptiles that remains to be characterized.

2.6 Supplementary Material

The following are available online at http://www.mdpi.com/1999-4915/12/6/613/s1.

Figure S2.1: PCR detection and host association of *Lauta virus*. (a–b) Agarose gels electrophoresis showing PCR products from two sets of primers that target a region in the PB1 gene segment (*RdRp*). Samples correspond to (c) liver tissue from seven different reptile species. A 355 bp PCR product was only amplified in *G. lauta*.

Table S2.1: Summary of the contig alignment to genomic segments of TiLV using DIAMOND. The relative abundance of each transcript was also calculated (see Methods).

Table S2.2: Summary of hits recovered after alignment of the untranslated contigs with reference protein sequences of the RdRp subunit PB1. The custom database included virus reference sequences from the order *Articulavirales*.

Table S2.3: List of virus sequences used in the phylogenetic analysis. All sequences correspond to the PB1 protein.

Table S2.4: Set of primers used for PCR and Sanger sequencing reactions.

Author Contributions: Conceptualization, E.C.H.; methodology, A.S.O.-B., E.C.H., and J.-S.E.; formal analysis, A.S.O.-B.; investigation, A.S.O.-B., E.C.H., and J.-S.E.; resources, C.M., J.-S.E and E.C.H.; writing—original draft preparation A.S.O.-B.; writing—review and editing E.C.H., J.-S.E. and C.M.; visualization, A.S.O.-B.; supervision, E.C.H. All authors have read and agreed to the published version of the manuscript.

2.6 References

Ahasan, M. S., Keleher, W., Giray, C., Perry, B., Surachetpong, W., Nicholson, P., Al-Hussinee, L., Subramaniam, K., & Waltzek, T. B. (2020). Genomic Characterization of Tilapia Lake Virus Isolates Recovered from Moribund Nile Tilapia (Oreochromis niloticus) on a Farm in the United States . *Microbiology Resource Announcements*, *9*(4). https://doi.org/10.1128/mra.01368-19

Al-Hussinee, L., Subramaniam, K., Ahasan, M. S., Keleher, B., & Waltzek, T. B. (2018). Complete genome sequence of a tilapia lake virus isolate obtained from Nile tilapia (Oreochromis niloticus). *Genome Announcements*, *6*(26). https://doi.org/10.1128/genomeA.00580-18

Altschul, S. F., Gish, W., Miller, W., Myers, E. W., & Lipman, D. J. (1990). Basic local alignment search tool. *Journal of Molecular Biology*, *215*(3), 403–410. https://doi.org/10.1016/S0022-2836(05)80360-2

Bacharach, E., Mishra, N., Briese, T., Zody, M. C., Kembou Tsofack, J. E., Zamostiano, R., Berkowitz, A., Ng, J., Nitido, A., Corvelo, A., Toussaint, N. C., Abel Nielsen, S. C., Hornig, M., Del Pozo, J., Bloom, T., Ferguson, H., Eldar, A., & Lipkin, W. I. (2016). Characterization of a Novel Orthomyxo-like Virus Causing Mass Die-Offs of Tilapia. *MBio*, *7*(2), e00431-16. https://doi.org/10.1128/mBio.00431-16

Baker, D., & Sali, A. (2001). Protein structure prediction and structural genomics. In *Science* (Vol. 294, Issue 5540, pp. 93–96). American Association for the Advancement of Science. https://doi.org/10.1126/science.1065659

Bamford, D. H., Grimes, J. M., & Stuart, D. I. (2005a). What does structure tell us about virus evolution? In *Current Opinion in Structural Biology* (Vol. 15, Issue 6, pp. 655–663). https://doi.org/10.1016/j.sbi.2005.10.012

Benson, S. D., Bamford, J. K. H., Bamford, D. H., & Burnett, R. M. (2004). Does Common Architecture Reveal a Viral Lineage Spanning All Three Domains of Life? *Molecular Cell*, *16*, 673–685.

Biswas, S. K., & Nayak, D. P. (1994a). Mutational analysis of the conserved motifs of influenza A virus polymerase basic protein 1. *Journal of Virology*, *68*(3), 1819–1826. https://doi.org/10.1128/jvi.68.3.1819-1826.1994

Bolger, A. M., Lohse, M., & Usadel, B. (2014). Trimmomatic: a flexible trimmer for Illumina sequence data. *Bioinformatics (Oxford, England)*, *30*(15), 2114–2120. https://doi.org/10.1093/bioinformatics/btu170

Buchfink, B., Xie, C., & Huson, D. H. (2015). Fast and sensitive protein alignment using DIAMOND. *Nature Methods*, *12*(1), 59–60. https://doi.org/10.1038/nmeth.3176

Černý, J., Černá Bolfíková, B., Valdés, J. J., Grubhoffer, L., & Růžek, D. (2014). Evolution of Tertiary Structure of Viral RNA Dependent Polymerases. *PLOS ONE*, *9*(5), 1–14. https://doi.org/10.1371/journal.pone.0096070

Chu, C., Fan, S., Li, C., Macken, C., Kim, J. H., Hatta, M., Neumann, G., & Kawaoka, Y. (2012a). Functional Analysis of Conserved Motifs in Influenza Virus PB1 Protein. *PLoS ONE*, *7*(5), e36113. https://doi.org/10.1371/journal.pone.0036113

Deng, H., Jia, Y., & Zhang, Y. (2018). Protein structure prediction. *International Journal of Modern Physics B*, *32*(18). https://doi.org/10.1142/S021797921840009X

Fiser, A. (2010). Template-based protein structure modeling. In *Methods in molecular biology (Clifton, N.J.)* (Vol. 673, pp. 73–94). Humana Press, Totowa, NJ. https://doi.org/10.1007/978-1-60761-842-3_6

Grabherr, M. G., Haas, B. J., Yassour, M., Levin, J. Z., Thompson, D. A., Amit, I., Adiconis, X., Fan, L., Raychowdhury, R., Zeng, Q., Chen, Z., Mauceli, E., Hacohen, N., Gnirke, A., Rhind, N., di Palma, F., Birren, B. W., Nusbaum, C., Lindblad-Toh, K., … Regev, A. (2011). Full-length transcriptome assembly from RNA-Seq data without a reference genome. *Nature Biotechnology*, *29*(7), 644–652. https://doi.org/10.1038/nbt.1883

Holmes, E. C. (2011). What Does Virus Evolution Tell Us about Virus Origins? *Journal of Virology*, *85*(11), 5247–5251. https://doi.org/10.1128/jvi.02203-10

Jansen, M. D., Dong, H. T., & Mohan, C. V. (2019a). Tilapia lake virus: a threat to the global tilapia industry? *Reviews in Aquaculture*, *11*(3), 725–739. https://doi.org/10.1111/raq.12254

Jia, H., & Gong, P. (2019). A structure-function diversity survey of the rna-dependent rna polymerases from the positive-strand rna viruses. *Frontiers in Microbiology*, *10*(AUG), 1945. https://doi.org/10.3389/FMICB.2019.01945/BIBTEX

Katoh, K., & Standley, D. M. (2013). MAFFT Multiple Sequence Alignment Software Version 7: Improvements in Performance and Usability. *Molecular Biology and Evolution*, *30*(4), 772–780. https://doi.org/10.1093/molbev/mst010

Kelley, L. A., Mezulis, S., Yates, C. M., Wass, M. N., & Sternberg, M. J. E. (2015). The Phyre2 web portal for protein modeling, prediction and analysis. *Nature Protocols 2015 10:6*, *10*(6), 845–858. https://doi.org/10.1038/nprot.2015.053

Li, B., & Dewey, C. N. (2011). RSEM: accurate transcript quantification from RNA-Seq data with or without a reference genome. *BMC Bioinformatics*, *12*(1), 323. https://doi.org/10.1186/1471-2105-12-323

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. (2015). Unprecedented genomic diversity of RNA viruses in arthropods reveals the ancestry of negative-sense RNA viruses. *ELife*, *4*. https://doi.org/10.7554/eLife.05378

Ng, K. K. S., Arnold, J. J., & Cameron, C. E. (2008). Structure-function relationships among RNA-dependent RNA polymerases. In *Current Topics in Microbiology and Immunology* (Vol. 320, pp. 137–156). Springer, Berlin, Heidelberg. https://doi.org/10.1007/978-3-540-75157-1_7

Nguyen, L.-T., Schmidt, H. A., von Haeseler, A., & Minh, B. Q. (2015). IQ-TREE: a fast and effective stochastic algorithm for estimating maximum-likelihood phylogenies. *Molecular Biology and Evolution*, *32*(1), 268–274. https://doi.org/10.1093/molbev/msu300

Oliver, P. M., Prasetya, A. M., Tedeschi, L. G., Fenker, J., Ellis, R. J., Doughty, P., & Moritz, C. (2020). Crypsis and convergence: Integrative taxonomic revision of the Gehyra australis group (Squamata: Gekkonidae) from northern Australia. *PeerJ*, *2020*(1), e7971. https://doi.org/10.7717/peerj.7971

Payne, S. (2017). Family Orthomyxoviridae. In *Viruses* (pp. 197–208). Elsevier. https://doi.org/10.1016/b978-0-12-803109-4.00023-4

Pulido, L. L. H., Mora, C. M., Hung, A. L., Dong, H. T., & Senapin, S. (2019). Tilapia lake virus (TiLV) from Peru is genetically close to the Israeli isolates. *Aquaculture*, *510*, 61–65. https://doi.org/10.1016/j.aquaculture.2019.04.058

Rice, G., Tang, L., Stedman, K., Roberto, F., Spuhler, J., Gillitzer, E., Johnson, J. E., Douglas, T., & Young, M. (2004). The structure of a thermophilic archaeal virus shows a double-stranded DNA

viral capsid type that spans all domains of life. *Proceedings of the National Academy of Sciences of the United States of America*, *101*(20), 7716–7720. https://doi.org/10.1073/pnas.0401773101

Rice, P., Longden, L., & Bleasby, A. (2000). EMBOSS: the European Molecular Biology Open Software Suite. *Trends in Genetics : TIG*, *16*(6), 276–277. https://doi.org/10.1016/S0168-9525(00)02024-2

Rose, R., Constantinides, B., Tapinos, A., Robertson, D. L., & Prosperi, M. (2016). Challenges in the analysis of viral metagenomes. *Virus Evolution*, *2*(2). https://doi.org/10.1093/VE/VEW022

Schrödinger, L., & DeLano, W. (2015). *The PyMOL Molecular Graphics System* (2.3.5).

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*, *540*(7634), 539–543. https://doi.org/10.1038/nature20167

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*(7700), 197–202. https://doi.org/10.1038/s41586-018-0012-7

Shi, M., Lin, X.-D., Vasilakis, N., Tian, J.-H., Li, C.-X., Chen, L.-J., Eastwood, G., Diao, X.-N., Chen, M.-H., Chen, X., Qin, X.-C., Widen, S. G., Wood, T. G., Tesh, R. B., Xu, J., Holmes, E. C., & Zhang, Y.-Z. (2016a). Divergent Viruses Discovered in Arthropods and Vertebrates Revise the Evolutionary History of the Flaviviridae and Related Viruses. *Journal of Virology*, *90*(2), 659–669. https://doi.org/10.1128/jvi.02036-15

Subramaniam, K., Ferguson, H. W., Kabuusu, R., & Waltzek, T. B. (2019a). Genome Sequence of Tilapia Lake Virus Associated with Syncytial Hepatitis of Tilapia in an Ecuadorian Aquaculture Facility. *Microbiology Resource Announcements*, *8*(18). https://doi.org/10.1128/mra.00084-19

te Velthuis, A. J. W. (2014). Common and unique features of viral RNA-dependent polymerases. *Cellular and Molecular Life Sciences*, *71*(22), 4403–4420. https://doi.org/10.1007/S00018-014-1695-Z

Thermes, C. (2014). Ten years of next-generation sequencing technology. *Trends in Genetics : TIG*, *30*(9), 418–426. https://doi.org/10.1016/j.tig.2014.07.001

Wolf, Y. I., Kazlauskas, D., Iranzo, J., Lucía-Sanz, A., Kuhn, J. H., Krupovic, M., Dolja, V. V., & Koonin, E. V. (2018). Origins and Evolution of the Global RNA Virome. *BioRxiv*, 451740. https://doi.org/10.1101/451740

Zhang, Y., & Skolnick, J. (2005). TM-align: a protein structure alignment algorithm based on the TM-score. *Nucleic Acids Research*.

Zhang, Y.-Z., Chen, Y.-M., Wang, W., Qin, X.-C., & Holmes, E. C. (2019a). Expanding the RNA Virosphere by Unbiased Metagenomics. *Annual Review of Virology*, *6*(1), annurev-virology-092818-015851. https://doi.org/10.1146/annurev-virology-092818-015851

Zhang, Y.-Z., Shi, M., & Holmes, E. C. (2018). Using metagenomics to characterize an expanding virosphere. *Cell*, *172*(6), 1168–1172. https://doi.org/10.1016/j.cell.2018.02.043

Chapter 3

Meta-transcriptomic identification of divergent *Amnoonviridae* in fish

---

Olivia M. H. Turnbull, Ayda Susana Ortiz-Baez, John-Sebastian Eden, Mang Shi, Jane E. Williamson, Troy F. Gaston , Yong-Zhen Zhang , Edward C. Holmes and Jemma L. Geoghegan

viruses

Communication

CHAPTER 3 Meta-transcriptomic identification of divergent *Amnoonviridae* in fish

Olivia M. H. Turnbull[1#], Ayda Susana Ortiz-Baez[2#], John-Sebastian Eden[2,3], Mang Shi[2,4], Jane E. Williamson[1], Troy F. Gaston[5], Yong-Zhen Zhang[6], Edward C. Holmes[2*], Jemma L. Geoghegan[7,8*]

[1.] Department of Biological Sciences, Macquarie University, Sydney, NSW 2109, Australia.

[2] Marie Bashir Institute for Infectious Diseases and Biosecurity, School of Life and Environmental Sciences and School of Medical Sciences, University of Sydney, NSW 2006, Australia.

[3] Centre for Virus Research, Westmead Institute for Medical Research, Westmead, NSW, Australia.

[4] School of Medicine, Sun Yat-sen University, Guangzhou, China.

[5] Troy Gaston: School of Life and Environmental Sciences, University of Newcastle, Ourimbah, NSW 2258, Australia.

[6] Shanghai Public Health Clinic Center & Institutes of Biomedical Sciences, Fudan University, Shanghai, China.

[7] Department of Microbiology and Immunology, University of Otago, Dunedin 9016, New Zealand.

[8] Institute of Environmental Science and Research, Wellington 5018, New Zealand.

[#]authors contributed equally

[*] Correspondence: edward.holmes@sydney.edu.au (E.C.H.); jemma.geoghegan@otago.ac.nz (J.L.G.)

Received: 7 October 2020; Accepted: 3 November 2020; Published: 4 November 2020

3.1 Abstract

Tilapia lake virus (TiLV) has caused mass mortalities in farmed and wild tilapia with serious economic and ecological consequences. Until recently, this virus was the sole member of the *Amnoonviridae*, a family within the order *Articulavirales* comprising

segmented negative-sense RNA viruses. We sought to identify additional viruses within the *Amnoonviridae* through total RNA sequencing (meta-transcriptomics) and data mining of published transcriptomes. Accordingly, we sampled marine fish species from both Australia and China and discovered several segments of two new viruses within the *Amnoonviridae*, tentatively called flavolineata virus and piscibus virus, respectively. In addition, by mining vertebrate transcriptome data, we identified nine additional virus transcripts matching to multiple genomic segments of TiLV in both marine and freshwater fish. These new viruses retained sequence conservation with the distantly related *Orthomyxoviridae* in the RdRp subunit PB1, but formed a distinct and diverse phylogenetic group. These data suggest that the *Amnoonviridae* have a broad host range within fish and that greater animal sampling will identify additional divergent members of the *Articulavirales*.

Keywords: meta-transcriptomics; virus discovery; *Amnoonviridae*; *Articulavirales*; fish; tilapia lake virus; evolution; phylogeny

---

3.2. Introduction

The *Amnoonviridae* are a recently described family of segmented and enveloped negative-sense RNA viruses associated with disease in fish. Until recently, the *Amnoonviridae* comprised only a single species, *tilapinevirus* or tilapia lake virus (TiLV) (Bacharach et al., 2016; Eyngor et al., 2014), which is associated with high rates of morbidity and mortality in both farmed and wild tilapia (*Oreochromis niloticus* and *Oreochromis niloticus* x *O. aureus* hybrid). As the second most farmed fish globally (Barange, 2018) and an important subsistence organism for farmers and high value markets (Fitzsimmons, 2015), tilapia contribute USD $7.5 billion annually to the aquaculture industry. Outbreaks of TiLV have resulted in significant economic and ecological loss. The virus causes gross lesions of the eyes and skin, while also impacting brain, liver and kidney tissue (Eyngor et al., 2014), with associated mortality rates up to 90% (Behera et al., 2018; Surachetpong et al., 2017). While ongoing surveillance has detected the virus across numerous countries in Asia, Africa and South America (Jansen et al., 2019a), whether related viruses infect other fish hosts remains unclear.

The *Amnoonviridae* are members of the order *Articulavirales* that also includes the *Orthomyxoviridae* (ICTV, 2018) that are particularly well-known because they contain the mammalian and avian influenza viruses. Unlike the *Amnoonviridae*, the *Orthomyxoviridae*, and closely related but unclassified orthomyxo-like viruses, infect a broad range of host species comprising both invertebrates and vertebrates. Notably, a divergent member of the *Amnoonviridae*, *Lauta virus*, was recently identified in an Australian gecko (Ortiz-Baez et al., 2020), strongly suggesting that members of this family are present in a wider range of vertebrate hosts. In addition, the large phylogenetic distance between *Lauta virus* and TiLV suggests that the former may even constitute a new genus within the *Amnoonviridae*, with the long branches throughout the *Articulavirales* phylogeny likely indicative of very limited sampling.

To help address whether the *Amnoonviridae* might be present in a wider range of vertebrate taxa, we screened for their presence using a meta-transcriptomic analysis of marine fish sampled in Australia and China, combined with data mining of published transcriptomes.

3.3 Materials and Methods

3.3.1 Fish collection in Australia

Fish samples were collected from the Bass Strait (40°15′ S–42°20′ S, 147°05′ E–148°35′ E), Australia, in November 2018. The fish species collected included *Rhombosolea tapirina*, *Platycephalus bassensis*, *Platycephalus speculator*, *Trachurus declivis*, *Trachurus novaezelandiae, Scorpaena papillosa*, *Pristiophorus nudipinnis*, *Pentaceropsis recurvirostris* and *Meuschenia flavolineata*. Fish were caught via repeated research trawls on the fisheries' training vessel, Bluefin, following the methodology outlined in (J. M. Park et al., 2017). Ten individuals from each species were caught and stored separately. Gill tissues were dissected and snap frozen at -20°C on the vessel, and then stored in a -80°C freezer

at Macquarie University, Sydney. Sampling was conducted under the approval of the University of Tasmania Animal Ethics Committee, approval number A0015366.

3.3.2 Fish collection in China

As well as Australia, we analysed the transcriptome data derived from a previous study of the viromes of fish sampled from the South China Sea (Shi et al., 2018). For that study, the fish species sampled and subsequently pooled for shotgun RNA sequencing included *Proscyllium habereri*, *Urolophus aurantiacus*, *Rajidae sp.*, *Eptatretus burgeri*, *Heterodontus zebra*, *Dasyatis bennetti, Acanthopagrus latus, Epinephelus awoara, Conger japonicus, Siganus canaliculatus, Glossogobius circumspectus, Halichoeres nigrescens,* and *Boleophthalmus pectinirostris*. Liver samples from each species were pooled and stored in a -80°C freezer. The procedures for sampling and sample processing were approved by the ethics committee of the National Institute for Communicable Disease Control and Prevention of the China CDC.

3.3.3 RNA sequencing

For RNA extraction, frozen tissue was partially thawed and submerged in lysis buffer containing 1% ß-mercaptoethanol and 0.5% Reagent DX before tissues were homogenized together with TissueRupture (Qiagen). The homogenate was centrifuged to remove any potential tissue residues, and RNA from the clear supernatant was extracted using the Qiagen RNeasy Plus Mini Kit. RNA was quantified using NanoDrop (ThermoFisher). RNA isolated from the Australian samples was pooled for each host species, whereas RNA isolated from the Chinese samples was pooled from all species (Shi 2018), resulting in 3μg per pool (250ng per individual). Libraries were constructed using the TruSeq Total RNA Library Preparation Protocol (Illumina) and host ribosomal RNA (rRNA) was depleted using the Ribo-Zero-Gold Kit (Illumina) to facilitate virus discovery. Fish caught in Australia were

subject to paired-end (100 bp) sequencing performed on the NovaSeq 500 platform (Illumina) carried out by the Australian Genome Research Facility (AGRF). RNA sequencing of the pooled fish sampled from China were sequenced on the HiSeq 2500 platform (Illumina) at BGI Tech (Shenzhen).

3.3.4 Transcript sequence similarity searching for novel amnoonviruses

Sequencing reads were first quality trimmed then assembled *de novo* using Trinity RNA-Seq (v.2.11.0) (Haas et al., 2013). The assembled contigs were annotated based on similarity searches against the National Center for Biotechnology Information (NCBI) nucleotide (nt) and non-redundant protein (nr) databases using BLASTn and Diamond BLASTX (v.2.0.2) (Buchfink et al., 2014). To infer the evolutionary relationships of the amnoonviruses newly discovered the translated viral contigs were combined with representative protein sequences from TiLV and *Lauta virus* obtained from NCBI GenBank. The sequences retrieved were then aligned with those generated here using MAFFT (v7.4) employing the E-INS-i algorithm. Ambiguously aligned regions were removed using trimAl (v.1.2) (Capella-Gutiérrez et al., 2009). To estimate phylogenetic trees, we utilized the maximum likelihood approach available in IQ-TREE (v 1.6.8) (Nguyen et al., 2015a), selecting the best-fit model of amino acid substitution with ModelFinder (Kalyaanamoorthy, Minh, Wong, von Haeseler, et al., 2017), and using 1000 bootstrap replicates to assess nodal support. Phylogenetic trees were annotated with FigTree (v.1.4.2).

3.3.5 PCR confirmation

To further confirm the presence of flavolineata virus in the yellow-striped leatherjacket collection, 10µl of extracted RNA was transcribed into cDNA using SuperScript® VILO™ reverse transcriptase (Invitrogen, CA USA). PCR amplification was

performed using Platinum™ II Hot-Start PCR Master Mix (2X) (Invitrogen, CA, USA) and 3 sets of primers (**Table S3.1**) designed to cover different regions of the virus sequence. PCR products were visualized on 2% agarose gel stained with SYBR® Safe (Invitrogen, CA USA).

3.3.6 TSA mining

To identify additional novel vertebrate viruses within the *Amnoonviridae* we screened *de novo* transcriptome assemblies available at the NCBI Transcriptome Shotgun Assembly (TSA) database (https://www.ncbi.nlm.nih.gov/genbank/tsa/). Amino acid sequences of flavolineata virus, piscibus virus and TiLV were queried against the assemblies using the translated Basic Local Alignment Search Tool (tBLASTn) algorithm. We restricted the search to transcriptomes within the Vertebrata (taxonomic identifier: 7742). Putative virus contigs were subsequently queried using BLASTx against the non-redundant virus database.

3.3.7 Virus naming

New viruses identified in this study are tentatively named by drawing from the names of their host species.

3.4 Results

3.4.1 Identification of a novel Amnoonviridae in yellow-striped leatherjacket

As part of a large virological survey on nine species of marine fish our meta-transcriptomic analysis identified several segments of a novel member of the *Amnoonviridae*, tentatively named flavolineata virus, in a sequencing library of 10 pooled individuals of yellow-striped leatherjacket (*Meuschenia flavolineata*) sampled from the Bass Strait off the coast of

Tasmania, Australia. No amnoonviruses were identified in the remaining eight fish species. We identified a complete, highly divergent protein in which a Diamond BLASTx analysis revealed 37% amino acid identity to TiLV segment 1, characterized as the PB1 subunit (Genbank accession: QJD15207.1, e-value: $2.0x10^{-80}$, query coverage 95%), with a GC composition of 48.2% and a standardised abundance of 0.00004% of the total non-rRNA library. The presence of flavolineata virus was further confirmed in the unpooled samples using RT-PCR (**Figure S3.1**).

3.4.2 Identification of a novel Amnoonviridae in pooled marine fish from the South China Sea

An additional novel member of the *Amnoonviridae*, in which we have provisionally termed piscibus virus, was identified in a pool of various marine species (including sharks, eels, stingrays, jawless fish and perch-like fish) sampled in the South China Sea as described previously (Shi et al., 2018). Specifically, we identified a short contig (270 nucleotides) that shared highest amino acid sequence similarity (48.8%, e-value: $1.5x10^{-15}$) to flavolineata virus using a custom database including the known members of the family *Amnoonviridae*. In addition, a comparison to the NCBI nr database showed that piscibus virus had 48.5% amino acid similarity (e-value: $2.0x10^{-07}$) to the PB1 subunit of the TiLV RdRp. The GC composition of the assembled sequence was 49.2% and it had a standardized abundance of 0.0001% of the total non-rRNA library. Despite the limited contig length for piscibus virus, such that its status as a *bona fide* novel virus will need to be confirmed with additional sequencing, we did identify conserved motifs within the PB1 subunit (see below).

3.4.3 Identification of novel Amnoonviridae in published transcriptomes

To identify additional novel vertebrate viruses within the *Amnoonviridae* we screened *de novo* transcriptome assemblies available at NCBI's TSA database. In doing so we

identified nine further potentially novel viruses in fish matching segments 1-4 of TiLV (**Table 3.1**).

Relatives of the *Amnoonviridae* were identified in ray-finned fish species (Actinopterygii) from marine (*Lepidonotothen nudifrons* and *Chionodraco hamatus*) and freshwater ecosystems (*Gymnocypris przewalskii*, *Gymnocypris namensis, Micropterus dolomieu, Oxygymnocypris stewartia*, *Schizothorax plagiostomus* and *Silurus asotus*) (**Table 3.1**). All viral sequences corresponded to segments 1-4 and ranged from 209-1784 nucleotides in length. The putative segments shared 26-51% sequence identity with TiLV. Most of the identified viral sequences corresponded to segment 1, containing the RdRp and covered motifs II and III (**Figure 3.1**). Notably, no other vertebrate class within the TSA were identified as potential hosts of these viruses.

3.4.4 Evolutionary relationships of novel Amnoonviridae

We next performed phylogenetic analysis of the RdRp subunit (segment 1) across the order *Articulavirales* (**Figure 3.2**). This revealed two distinct clades of fish viruses within the *Amnoonviridae* (with 83% bootstrap support). The original member of this virus family, TiLV, grouped with flavolineata virus, piscibus virus, dolomieu virus, namensis virus and hamatus virus in one clade. The second fish virus clade comprised the newly identified stewartii virus, plagiostomus virus, przewalskii virus, asotus virus 1, and asotus virus *2*. *Lauta* virus, identified in a native Australian gecko, appears to form a distinct lineage, suggestive of a separate genus. This phylogenetic analysis clearly illustrates the diversity of these viruses within both marine and freshwater fish, with no apparent host taxonomic structure (**Figure 3.2)**

Table 3.1 Novel viruses identified in this study.

Virus name	Host	Geographic location of sample collection	Detection method (NCBI accession of TSA data)	Contig length matching segment 1 or 4 of TiLV (nt)	Closest amino acid match to segment 1 or 4 (GenBank accession)
flavolineata virus	*Meuschenia flavolineata*	Australia	Fish sampling + meta-transcriptomics	1536	37% TiLV (QJD15207.1)
piscibus virus	Pooled marine fish (see methods)	China	Fish sampling + meta-transcriptomics	270	49% TiLV (QJD15207.1)
dolomieu virus	*Micropterus dolomieu*		TSA search (GDQU01066121.1, GDQU01106321.1, GDQU01283605.1, GDQU01532168.1)	1440	34% TiLV (QJD15204.1)
namensis virus	*Gymnocypris namensis*		TSA search (GHYH01080462.1, GHYH01005036.1, GHYH01084204.1)	1503	35% TiLV (AOE22913.1)
hamatus virus	*Chionodraco hamatus*		TSA search (GFMN01088333.1)	321	51% TiLV (QJD15205.1)
stewartii virus	*Oxygymnocypris stewartii*		TSA search (GIBO01031171.1, GIBO01013027.1)	1743	28% TiLV (QJD15204.1)
plagiostomus virus	*Schizothorax plagiostomus*		TSA search (GHXZ01024367.1, GHXZ01079240.1)	366	39% TiLV (AOE22912.1)
przewalskii virus	*Gymnocypris przewalskii*		TSA search (GHYJ01002273.1, GHYJ01008047.1, GHYJ01010906.1)	1761	26% TiLV (QJD15208.1)
asotus virus 1	*Silurus asotus*		TSA search (GHGF01026383.1, GHGF01034639.1, GHGF01033499.1, GHGF01028660.1, GHGF01037407.1)	1710	32% TiLV (QMT29723.1)
asotus virus 2	*Silurus asotus*		TSA search (GHGF01016319.1, GHGF01027066.1, GHGF01047620.1)	1719	29% TiLV (QJD15204.1)
nudifrons virus	*Lindbergichthys nudifrons*		TSA search (HACN01008153.1)	1032 (segment 4)	44% f*lavolineata virus* (segment 4)

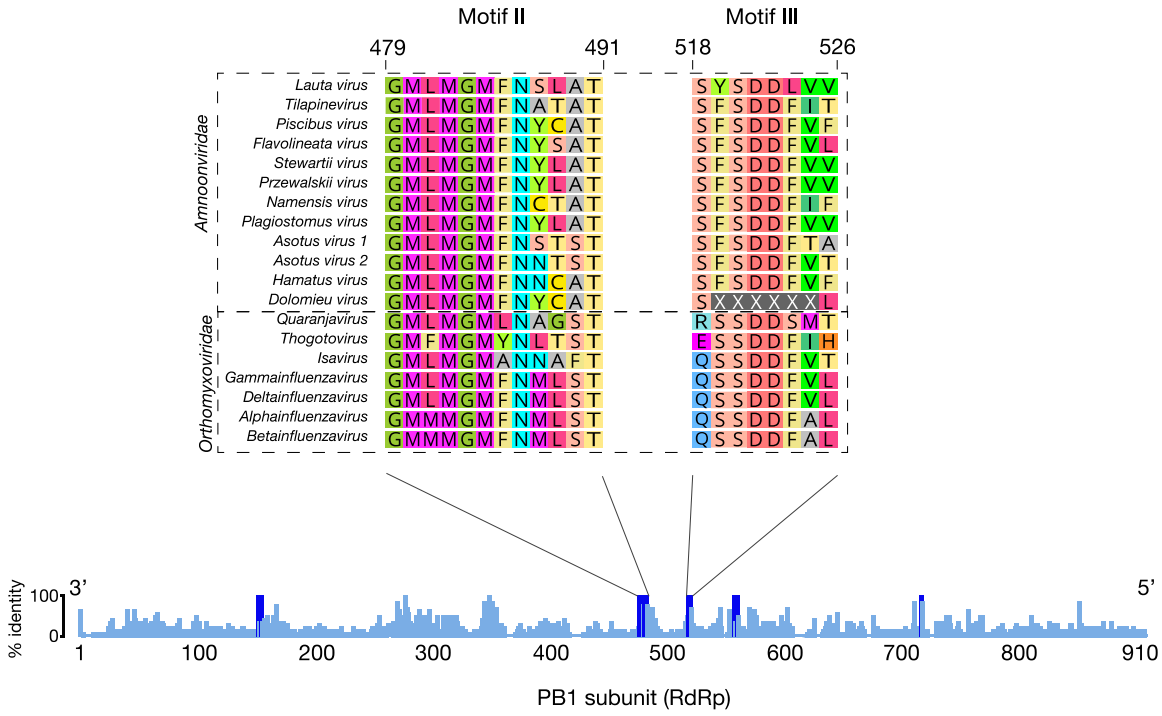


Figure 3.1 Alignment of viruses within the order *Articulavirales* in the RdRp subunit PB1. Blue bars illustrate the mean pairwise identity over all the pairs in each column where royal blue highlights conserved motifs with 100% identity. The sequences of two motifs (II and III) are shown where all sequences overlapped.

3.4.5 Genome composition of the novel Amnoonviridae

Ten of the novel viruses identified included segment 1, corresponding to the RdRp subunit PB1, and sharing clear sequence homology with different members of the *Articulavirales* including the *Orthomyxoviridae* (**Figure 3.1**). These viruses had a closest genetic match to TiLV, ranging from 28-51% sequence similarity at the amino acid level to segment 1 (**Table 3.1**). Segment 4 was the only segment found from the tentatively named nudifrons virus (**Table 3.1, Table S3.2**).

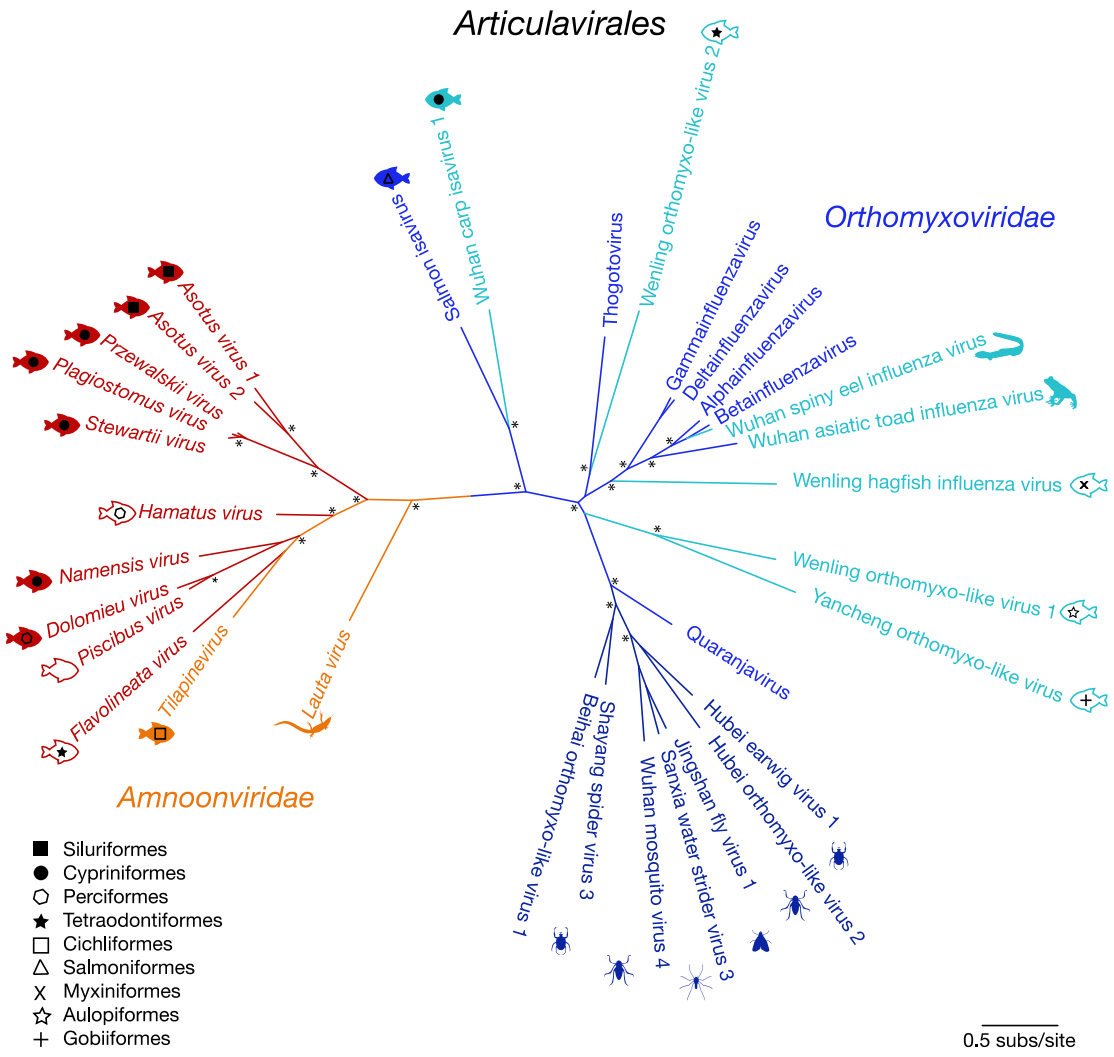


Figure 3.2 Unrooted maximum likelihood phylogenetic tree of the PB1 subunit showing the topological position of 10 of the 11 newly discovered viruses (red) that shared sequence similarity to segment 1 within the order Articulavirales (*Amnoonviridae*: orange; *Orthomyxoviridae*: shades of blue). Tilapinevirus (TiLV) and the recently discovered *Lauta virus* were the only viruses previously identified in this family.Fish viruses are annotated with fish symbols (filled: freshwater; outline: marine), and fish order corresponds to shapes illustrated by the key. All branches are scaled according to the number of amino acid substitutions per site. An asterisk (*) illustrates nodes with bootstrap support > 70%.

Despite the lack of genomic characterization of TiLV, a sequence comparison across the *Articulavirales* revealed several conserved PB1 motifs, which included those described previously (Bacharach et al., 2016). Sequence similarities with other viral RNA-dependent RNA polymerases suggest that motif III plays a key functional role at the core of the transcriptase-replicase activity (Biswas & Nayak, 1994; Chu et al., 2012). Defined by the consensus serine-aspartic acid-aspartic acid (SDD) sequence in the *Articulavirales*, this motif is highly conserved and is critical for protein stability and function.

In contrast to the six to eight genomic coding segments that comprise viruses within the *Orthomyxoviridae*, TiLV contains 10 segments with open reading frames, none of which have been functionally characterized to date (Taengphu et al., 2020). While we were able to distinguish virus transcripts with sequence similarity to segments 1 – 4 of TiLV (**Figure 3.3**), it is possible that the other segments are present but too divergent in sequence to be detected, and this will need to be addressed in future studies. Indeed, the remaining segments of TiLV exhibit no sequence similarity to any other known viruses (Bacharach et al., 2016; Eyngor et al., 2014) or eukaryotic genes.

It is also of note that we found some evidence for phylogenetic incongruence between the topologies of the different gene segments, although this analysis is complicated by the differing numbers of viruses available for each segment, the short sequence alignments, and the highly divergent nature of the sequences being analysed. For example, flavolineata virus and TiLV appear as sister taxa in segment 1 yet are seemingly more divergent in segment 4 (**Figure 3.3**). Hence, this phylogenetic pattern tentatively suggests that amnoonviruses may have undergone reassortment in similar manner to influenza A viruses in the *Orthomyxoviridae*, although this will need to be confirmed with the addition of longer sequences and more taxa. Reassortment has previously been observed within circulating TiLV strains, which has added complexity to inferring its evolutionary history (Chaput et al., 2020).

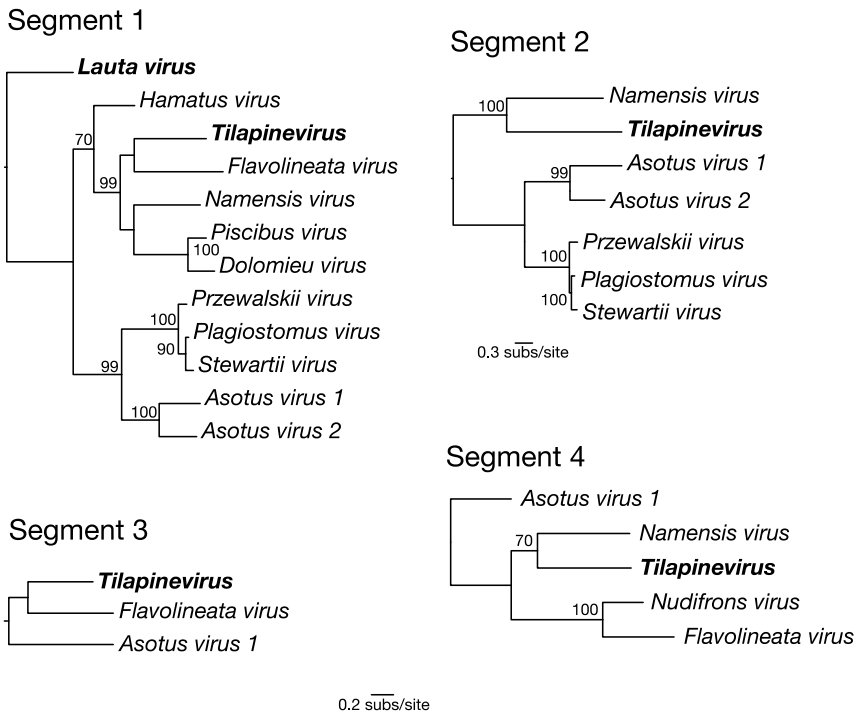


Figure 3.3 Maximum likelihood phylogenetic trees of genomic segments 1-4 for the new virus transcripts identified in this study within the *Amnoonviridae*. Viruses previously identified in this family are in bold. Bootstrap values >70% are shown. The segment 1 phylogeny was rooted using *Lauta virus* as the outgroup (as suggested by the tree in Figure 3.2). The remaining three segment phylogenies were then rooted to match the segment 1 tree. A branch scale bar represents 0.2 substitutions per site. See Table S3.2 for virus sequence details.

3.5 Discussion

Through both sampling marine fish and mining publicly available sequence data, we discovered 11 new viruses, all of which are the closest genetic relatives of TiLV. These viruses fall within the *Amnoonviridae*, which currently comprises only two viruses: TiLV and *Lauta virus*. The discovery of these new viruses expands our understanding of the host range of the *Amnoonviridae* to include host species across multiple taxonomic orders of freshwater and marine fish, including Cypriniformes, Siluriformes, Perciformes and Tetraodontiformes, and includes animals sampled in a range of geographic localities (Australia, China, North America,

Antarctica and Japan). Not only does the identification of these new viruses greatly increase the phylogenetic diversity in this newly identified group of viruses, but it may also provide insight into the potential origins and host range of TiLV, a virus that has major economic and ecological impacts on fisheries and aquaculture.

The viruses discovered here were highly divergent in sequence, likely limiting our ability to detect all genome segments present in the data. Nevertheless, sequence conservation within segment 1 across the entire taxonomic order strongly supports the inclusion of these new viruses within the *Amnoonviridae*. While we only found new viruses in fish and no other vertebrate classes, it is important to note that fish comprise 44% of currently available vertebrate transcriptomes (as of September 2020). With the expansion of these databases, it is likely we will identify additional highly divergent viruses within the *Amnoonviridae* and hence of the *Articulavirales* as a whole. The discovery of these 11 viruses invites further research into the true diversity and evolutionary origins of the *Amnoonviridae*.

3.6 Supplementary Material

The following are available online at http://www.mdpi.com/1999-4915/12/11/1254/s1

Table S3.1: List of primer sets used for the RT-PCR confirmation of flavolineata virus in specimens of *Meuschenia flavolineata*

Table S3.2: All virus transcripts identified in this study that fell across genomic segments within the *Amnoonviridae*

Figure S3.1: Agarose gels electrophoresis showing PCR products from three sets of primers that target a region in the PB1 gene segment (RdRp) for ten individuals of *Meuschenia flavolineata*.

Sequencing reads are available at the NCBI Sequence Read Archive (SRA). For piscibus virus see Bioproject PRJNA418053 (BioSample: SAMN08013970; Library name: BHFishG) and for flavolineata virus see Bioproject: PRJNA667570. Alignments with new virus transcripts are available at https://github.com/jemmageoghegan/Amnoonviridae-in-fish.

Author Contributions

Conceptualization, E.C.H. and J.L.G.; formal analysis, O.M.H.T, A.S.O.B. J.S.E, J.L.G; resources, J.E.W., T.F.G., M.S., Y.Z.Z, E.C.H and J.L.G.; writing—original draft preparation, O.M.H.T, A.S.O.B., J.S.E., E.C.H. and J.L.G.; writing—review and editing, all authors; funding acquisition, E.C.H. and J.L.G. All authors have read and agreed to the published version of the manuscript.

3.6 References

Bacharach, E., Mishra, N., Briese, T., Zody, M. C., Kembou Tsofack, J. E., Zamostiano, R., Berkowitz, A., Ng, J., Nitido, A., Corvelo, A., Toussaint, N. C., Abel Nielsen, S. C., Hornig, M., del Pozo, J., Bloom, T., Ferguson, H., Eldar, A., & Lipkin, W. I. (2016). Characterization of a Novel Orthomyxo-like Virus Causing Mass Die-Offs of Tilapia. *MBio*, *7*(2), e00431-16. https://doi.org/10.1128/mBio.00431-16

Barange, M. (2018). Fishery and aquaculture statistics. *FAO Yearbook. Fishery and Aquaculture Statistics= FAO Annuaire. Statistiques Des Peches et de l'Aquaculture= FAO Anuario. Estadisticas de Pesca y Acuicultura*, I–82.

Behera, B. K., Pradhan, P. K., Swaminathan, T. R., Sood, N., Paria, P., Das, A., Verma, D. K., Kumar, R., Yadav, M. K., Dev, A. K., Parida, P. K., Das, B. K., Lal, K. K., & Jena, J. K. (2018). Emergence of Tilapia Lake Virus associated with mortalities of farmed Nile Tilapia Oreochromis niloticus (Linnaeus 1758) in India. *Aquaculture*, *484*, 168–174. https://doi.org/10.1016/j.aquaculture.2017.11.025

Biswas, S. K., & Nayak, D. P. (1994). Mutational analysis of the conserved motifs of influenza A virus polymerase basic protein 1. *Journal of Virology*, *68*(3), 1819–1826. https://doi.org/10.1128/JVI.68.3.1819-1826.1994

Buchfink, B., Xie, C., & Huson, D. H. (2014). Fast and sensitive protein alignment using DIAMOND. *Nature Methods*, *12*(1), 59–60. https://doi.org/10.1038/NMETH.3176

Capella-Gutiérrez, S., Silla-Martínez, J. M., & Gabaldón, T. (2009). trimAl: A tool for automated alignment trimming in large-scale phylogenetic analyses. *Bioinformatics*, *25*(15), 1972–1973. https://doi.org/10.1093/BIOINFORMATICS/BTP348

Chaput, D. L., Bass, D., Alam, M. M., Hasan, N. al, Stentiford, G. D., van Aerle, R., Moore, K., Bignell, J. P., Haque, M. M., & Tyler, C. R. (2020). The segment matters: Probable reassortment of tilapia

lake virus (TiLV) complicates phylogenetic analysis and inference of geographical origin of new isolate from Bangladesh. *Viruses*, *12*(3). https://doi.org/10.3390/V12030258

Chu, C., Fan, S., Li, C., Macken, C., Kim, J. H., Hatta, M., Neumann, G., & Kawaoka, Y. (2012). Functional analysis of conserved motifs in influenza virus PB1 protein. *PloS One*, *7*(5). https://doi.org/10.1371/journal.pone.0036113

Eyngor, M., Zamostiano, R., Tsofack, J. E. K., Berkowitz, A., Bercovier, H., Tinman, S., Lev, M., Hurvitz, A., Galeotti, M., Bacharach, E., & Eldar, A. (2014). Identification of a novel RNA virus lethal to tilapia. *Journal of Clinical Microbiology*, *52*(12), 4137–4146. https://doi.org/10.1128/JCM.00827-14

Fitzsimmons, K. (2015). *Global Tilapia Market update 2015*. https://doi.org/10.13140/RG.2.1.2848.9448

Haas, B. J., Papanicolaou, A., Yassour, M., Grabherr, M., Blood, P. D., Bowden, J., Couger, M. B., Eccles, D., Li, B., Lieber, M., Macmanes, M. D., Ott, M., Orvis, J., Pochet, N., Strozzi, F., Weeks, N., Westerman, R., William, T., Dewey, C. N., … Regev, A. (2013). De novo transcript sequence reconstruction from RNA-seq using the Trinity platform for reference generation and analysis. *Nature Protocols*, *8*(8), 1494–1512. https://doi.org/10.1038/NPROT.2013.084

ICTV. (2018). *Virus Taxonomy: 2018a Release https://ictv.global/taxonomy*.

Jansen, M. D., Dong, H. T., & Mohan, C. V. (2019). Tilapia lake virus: a threat to the global tilapia industry? *Reviews in Aquaculture*, *11*(3), 725–739. https://doi.org/10.1111/RAQ.12254

Kalyaanamoorthy, S., Minh, B. Q., Wong, T. K. F., von Haeseler, A., & Jermiin, L. S. (2017). ModelFinder: Fast model selection for accurate phylogenetic estimates. *Nature Methods*, *14*(6), 587–589. https://doi.org/10.1038/NMETH.4285

Nguyen, L.-T., Schmidt, H. A., von Haeseler, A., & Minh, B. Q. (2015). IQ-TREE: a fast and effective stochastic algorithm for estimating maximum-likelihood phylogenies. *Molecular Biology and Evolution*, *32*(1), 268–274. https://doi.org/10.1093/molbev/msu300

Ortiz-Baez, A. S., Eden, J. S., Moritz, C., & Holmes, E. C. (2020). A divergent articulavirus in an australian gecko identified using meta-transcriptomics and protein structure comparisons. *Viruses*, *12*(6). https://doi.org/10.3390/v12060613

Park, J. M., Coburn, E., Platell, M. E., Gaston, T. F., Taylor, M. D., & Williamson, J. E. (2017). Diets and resource partitioning among three sympatric gurnards in Northeastern Tasmanian Waters, Australia. *Marine and Coastal Fisheries*, *9*(1), 305–319. https://doi.org/10.1080/19425120.2017.1320342

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*(7700), 197–202. https://doi.org/10.1038/S41586-018-0012-7

Surachetpong, W., Janetanakit, T., Nonthabenjawan, N., Tattiyapong, P., Sirikanchana, K., & Amonsin, A. (2017). Outbreaks of tilapia lake virus infection, Thailand, 2015–2016. *Emerging Infectious Diseases*, *23*(6), 1031–1033. https://doi.org/10.3201/EID2306.161278

Taengphu, S., Sangsuriya, P., Phiwsaiya, K., Debnath, P. P., Delamare-Deboutteville, J., Mohan, C. V., Dong, H. T., & Senapin, S. (2020). Genetic diversity of tilapia lake virus genome segment 1 from 2011 to 2019 and a newly validated semi-nested RT-PCR method. *Aquaculture*, *526*. https://doi.org/10.1016/j.aquaculture.2020.735423

Chapter 4

Meta-transcriptomics reveals potential virus transfer between *Aedes communis* mosquitoes and their parasitic water mites

---

Ayda Susana Ortiz-Baez, Edward C. Holmes, Justine Charon,

John H-O Pettersson and Jenny C. Hesson

VIRUS EVOLUTION

Article

CHAPTER 4 Meta-transcriptomics reveals potential virus transfer between Aedes communis mosquitoes and their parasitic water mites

Ayda Susana Ortiz-Baez,[1,†] Edward C. Holmes,[1] Justine Charon,[1] John H.-O. Pettersson,[1,2,3,*,†] and Jenny C. Hesson[3,4,*,†]

[1] Sydney Institute for Infectious Diseases, School of Medical Sciences, The University of Sydney, Sydney, NSW 2006, Australia

[2] Clinical Microbiology and Hospital Hygiene, Uppsala University Hospital, Dag Hammarskjölds väg 38, Uppsala SE-751 85, Sweden

[3] Zoonosis Science Center, Department of Medical Biochemistry and Microbiology, University of Uppsala, Husargatan 3, C8:3, Uppsala SE-751 23, Sweden

[4] Biologisk Myggkontroll, Nedre Dal˙alven Utvecklings AB, Vårdsätravägen 5, Uppsala SE 75646, Sweden

† Joint senior author

* Correspondence: john.pettersson@imbim.uu.se; jenny.hesson@imbim.uu.se

Received: 01 June 2022; Accepted: 23 September 2022; Published: 26 October 2020;
https://doi.org/10.1093/ve/veac090

4.1 Abstract

Arthropods harbor a largely undocumented diversity of RNA viruses. Some arthropods, like mosquitoes, can transmit viruses to vertebrates but are themselves parasitized by other arthropod species, such as mites. Very little is known about the viruses of these ectoparasites and how they move through the host-parasite relationship.

To address this, we determined the virome of both mosquitoes and the mites that feed on them. The mosquito *Aedes communis* is an abundant and widely distributed species in Sweden, in northern Europe. These dipterans are commonly parasitized by water mite larvae (Trombidiformes: Mideopsidae) that are hypothesized to impose negative selection pressures on the mosquito by reducing fitness. In turn, viruses are dual-host agents in the mosquito-mite interaction. We determined the RNA virus diversity of mite-free and mite-detached mosquitoes, as well as their parasitic mites, using meta-transcriptomic sequencing. Our results revealed an extensive RNA virus diversity in both mites and mosquitoes, including thirty-seven putative novel RNA viruses that cover a wide taxonomic range. Notably, a high proportion of viruses (20/37) were shared between mites and mosquitoes, while a limited number of viruses were present in a single host. Comparisons of virus composition and abundance suggest potential virus transfer between mosquitoes and mites during their symbiotic interaction. These findings shed light on virome diversity and ecology in the context of arthropod host-parasite-virus relationships.

Keywords: meta-transcriptomics, mosquito-borne viruses, arthropod-borne viruses, virus evolution, virome

---

4.2. Introduction

Arthropods can interact in various ways to establish symbiotic relationships in nature (Kaplan & Eubanks, 2005; Peng et al., 2013; Werblow et al., 2015a). Among these, parasitic associations have profound effects on host populations and community ecology (Vasquez et al., 2020). This symbiotic strategy allows a parasitic arthropod to exploit the resources of an arthropod host to survive and reproduce. In freshwater ecosystems, parasitic associations can be observed between water mites and other arthropods such as crustaceans and insects (Pozojević et al., 2019; Vasquez et al., 2020). However, such biotic

associations are not only confined to hosts and parasites. Relatedly, viruses are ubiquitous actors capable of permeating through arthropod symbiotic systems and interacting with either the parasite and/or the 'base host', resulting in a dynamic tripartite setting (i.e. host-parasite-virus system) (Di Prisco et al., 2016; Parratt & Laine, 2016a).

Parasitic mites can act as vectors or activators of viral diseases. For instance, RNA viruses such as Kashmir bee virus (KBV), sacbrood virus (SBV) and deformed wing virus (DWV) are often detected in honeybee colonies infested with *Varroa* mites (Dainat et al., 2009; Shen et al., 2005). Importantly, there are major gaps in our current knowledge of the diversity and biology of viruses associated with natural mite populations that parasitize mosquitoes and their vectorial capacity. Indeed, most research on viruses infecting mites are related to pathogens of mammals and plants (Poinar & Poinar, 1998; Valiente Moro et al., 2005; X. -j. Yu & Tesh, 2014a). In the same way, the relationship between parasitism (e.g. multiparasitism) and virus ecology at the mosquito–mite–virus interface remains to be determined (Auld et al., 2017).

Water mite larvae (Acari: Parasitengona: Hydrachnid) are obligate ectoparasites of culicid mosquitoes (Werblow et al., 2015b). Although the exact nature of the host-parasite relationship between mosquitoes and mites is uncertain, water mites exhibit predatory and parasitic behaviors on larval and adult stage mosquitoes, respectively (Atwa et al., 2017; Vasquez et al., 2020; Werblow et al., 2015c). During the biotic interaction, water mite larvae often attach to pre-imaginal stages or adult mosquitoes that provide the larvae with nutrients and transport to complete their life cycle (Werblow et al., 2015c). Once the larval stage is complete, water mites detach from the mosquito for post-larval (nymphal stages) and adult development, feeding on insect larvae, including mosquito eggs and larvae present in aquatic habitats (Atwa et al., 2017; Vasquez et al., 2020). Conversely, parasitism of mosquitoes by mites is usually associated with adverse effects on mosquito fitness (i.e. reduced reproductive ability and survival) (dos Santos et al.,

2016). Among these, mite infestation might impact flight, sexual maturity, and egg production in mosquitoes.

The snow-pool mosquito species, *Aedes communis* (De Geer 1976), is a monocyclic species with a Holarctic distribution, occurring in Eurasia and North America (Becker et al., 2010). It is commonly found not only in forested areas such as coniferous and temperate forests but also on the tundra (Medvedev et al., 2011). *Aedes communis* females commonly blood feed during twilight on a variety of vertebrates including, humans, rabbits, birds, rodents, and cattle. In Sweden, *A. communis* is abundant and widespread in spring and early summer (Lundström et al., 2013). The virome of *A. communis* is largely unknown but has been shown to include insect-specific viruses (ISVs) from the families *Phasmaviridae*, *Rhabdoviridae* and *Solemoviridae* of RNA viruses (Öhlund et al., 2019). Sporadic detections of different arboviruses have also been reported from this species (Campbell et al., 1991; Lvov et al., 2015), although it is not considered a vector species for any arbovirus (Campbell et al., 1991).

The use of metagenomic sequencing to characterize virus diversity has revolutionized our understanding of the evolutionary history, ecology, and distribution of RNA viruses in nature (Shi et al., 2018a; Zhang et al., 2018), transforming our ability to detect viruses in terms of scalability, speed, and accuracy. In particular, these studies have revealed an enormous number and diversity of viruses in invertebrates, including both ISVs and arboviruses, some of which fall into highly divergent lineages or RNA virus families (Li et al., 2015; Liu et al., 2015; Shi et al., 2016). Herein, we used meta-transcriptomics to reveal the virome diversity of *A. communis* and their parasitic mites and investigate whether mosquito-mite interactions can facilitate virus transfer among them. For this purpose, we compared the diversity and abundance of RNA viruses in mosquitoes parasitized by mites, mite-free mosquitoes, and parasitic mites to assess the viral community composition in the mosquito-mite interaction.

4.3 Materials and Methods

4.3.1 Sample collection

Aedes communis mosquitoes were collected within a mosquito control program (https://mygg.se/) across the river Dalälven floodplains in central Sweden (60.2888° N, 16.8970° E) between weeks 25 and 35 in 2014, 2019 and 2020, using the Centers for Disease Control and Prevention miniature light traps (CDC-traps) baited with dry ice. Morphological identification of mosquitoes was conducted using a stereomicroscope and the key provided by Becker et al. (2010) (Becker et al., 2010) on a chilled table. The mosquitoes collected were examined under a stereoscopic microscope for the presence of mites (*Mideopsis* sp.). Detected mites were removed and, together with the mosquitoes, were separated into groups of mites (K), mite-free mosquitoes (M) and mite-detached mosquitoes (MK) (infection load = 1–20 mites per mosquito) (**Supplementary Table S4.1**). The collected specimens were kept at -80°C until molecular processing.

4.3.2 Sample processing and sequencing

Samples were processed in three groups corresponding to M (n = 80), MK (n = 80), and K (n=160). In total, twenty-four sequencing libraries were prepared, eight libraries for each group. Samples were homogenized in pools of ten mosquitoes or twenty mites, using ZR BashingBead 0.1mm (Zymo Research, Irvine, CA, USA) for 180s using a TissueLyzer II (Qiagen). Total RNA was extracted from the homogenates using the ZymoBIOMICS DNA/RNA Miniprep Kit (Zymo Research, Irvine, California, USA) according to the manufacturer's instructions. Ribosomal RNA (rRNA) depleted RNA was extracted from each sample using the Ribo-Zero Gold kit (Illumina). Whole-transcriptome libraries were constructed using DNA nanoball technology (paired-end sequencing) on a DNBseq platform. Library preparation and sequencing were performed by the Beijing Genomics Institute, Hong Kong. For taxonomic identification of the most likely genus of mites

detached from the mosquitoes we compared our contigs against a custom database including Cox-1 amino acid sequences from mites (Trombidiformes; taxid: 83136). We also assessed the variation in *Cox-1* gene abundance across libraries using different reference sequences of *Mideopsis sp.* (**Figure S4.2**).

4.3.3 Sequence data processing

Sequence read quality assessment was performed with FastQC v0.11.8 (Andrews, 2010) and summarized using the MultiQC tool (Ewels et al., 2016a). Ribosomal reads of Archaea, Bacteria and Eukarya were filtered from the meta-transcriptomic data with the SortMeRNA v2.1b software (Kopylova et al., 2012). Reads were assembled into contigs using the metagenomic assembler MEGAHIT v1.2.9 with default settings (D. Li et al., 2015). Meta-transcriptome assembly evaluation was conducted using QUAST v4.3 (Gurevich et al., 2013). To reduce false-positives in the detection of viruses due to index-hopping, putative viruses were considered as present in a library if the total read count was $\geq$ 0.1% of the highest count for that virus across the libraries with at least two reads per sample. Taxonomic profiling of metagenomic data was conducted using CCMetagen v.1.2.4 (Marcelino et al., 2020) (Summary data available at doi: 10.6084/m9.figshare.20499726).

The sequencing reads and viral sequences identified in this study have been deposited in the SRA (Bioproject: PRJNA838788; Biosamples: SAMN28502431–SAMN28502454; SRA accession codes: SRR19268734–SRR19268757) and GenBank (ON860444–ON860480, OP555115-OP555127) databases, respectively.

4.3.4 Virus abundance and host association inference

Abundance was quantified as the number of reads per million mapped reads (RPM). Reads were mapped to the viral assemblies and the *Cox-1* gene as host marker

(JX040509.1 and MN362385.1) using the BBMap tool v.37.98 (sourceforge.net/projects/bbmap/). Contig assemblies were compared against the National Center for Biotechnology Information (NCBI) nucleotide (NCBI-nt) and non-redundant protein database (NCBI-nr) using DIAMOND v.2.0.9 with e-value cutoffs ≥ 1E-10 and ≥ 1E-4, respectively. To infer likely virus-host associations, we considered available data on *(i)* the virus prevalence within and between arthropod groups, *(ii)* abundance estimates, *(iii)* the closest hits in the BLAST/nr search, and *(iv)* phylogenetic relationships. To establish a likely host association, at least three of the four criteria had to be compatible.

4.3.5 Taxonomic assignment and protein annotation

Taxonomic information was collected from the NCBI Entrez taxonomy database using the NCBI-taxonomist tool v1.2.1 (https://pypi.org/project/ncbi-taxonomist/) (Buchmann & Holmes, 2020). Open reading frame (ORF) detection and sequence translation were performed on contigs >1000 nt with the program getORF v.6.6.0 (*-minsize* 600 *-find* 0), EMBOSS (P. Rice et al., 2000a). Classification of proteins and domain detection on predicted ORFs were performed using the InterProScan v5.51-85.0 software (Jones et al., 2014) with default search parameters, and the HMMER v3.3 program (hmmscan search) against the Pfam and PROSITE databases (Finn et al., 2011). To identify and annotate highly divergent viruses that were missed in the DIAMOND BLASTX search or that had similarities to taxonomically unassigned viruses, orphan contigs and unclassified viruses were run through the RdRp-scan resource with e-value 1E - 6 in the hmmscan search (Charon et al., 2022a). The completeness and quality of viral sequences were assessed by visual inspection and execution of the CheckV pipeline (Nayfach et al., 2020) and Prodigal v.2.6.3 (Hyatt et al., 2012).

4.3.6 Statistical analyses

To determine whether virus abundance levels differed significantly among the K, M and MK groups, we assessed the normality of the data corresponding to RPM values (raw and log10 transformed) by visual inspection and using the Shapiro–Wilk test. Since the data did not follow a normal distribution, comparisons were made using the Kruskal – Wallis chi-squared test and Pairwise Wilcoxon Rank Sum Test. All analyses were performed using the packages rstatix (Kassambara, 2021) and ggpubr (Kassambara, 2020) in R (R Core Team, 2021) (available at https://www.R-project.org/).

4.3.7 Phylogenetic analysis

Viral protein sequences for the RNA-dependent RNA polymerase (RdRp) identified in this study were aligned to a set of representative sequences publicly available at NCBI/GenBank according to the virus family, using Clustal Omega v.1.2.4 with default settings. The reference Quenyavirus sequences were obtained from Obbard et al. 2020 (Obbard et al., 2020). We assigned provisional names to novel viruses based on geographic locations from where they were collected. Selection of the best-fit model of sequence evolution and phylogenetic relationships within the virus families were assessed using the maximum likelihood (ML) method available in IQ-TREE v1.6.12 (-m TEST -alrt 1000 -bb 1000 -nt 4 -bnni) (Hoang et al., 2018a; Nguyen et al., 2015a). Nodal support was estimated with SH-aLRT and the ultrafast bootstrap (UFBoot). A total of 1000 replicates were run for both approaches and we used the option *bnni* to avoid overestimating branch supports with UFBoot. Tree visualization was conducted using the R software packages ggtree and ggplot2 (Wickham, 2016a; G. Yu et al., 2017).

4.4 Results

4.4.1 Extenstive RNA virome diversity in A. communis and their parasitic mites

A total of 160 mosquitoes and their parasitic mites were collected and pooled into twenty-four separate libraries, representing three different groups, to characterize the virome of each host and assess the virus prevalence across mites, and in both mite-free and mite-detached mosquitoes. Overall, we generated between 48.5 and 74 million pair reads per library, of which ~76 per cent corresponded to rRNA content. Meta-transcriptomic reads were *de novo* assembled into partial viral genomes from which we identified thirty-seven novel RNA viruses based on the presence of a viral RdRp: these represented eighteen families and fifteen orders of positive-stranded RNA (n = 8), negative-stranded RNA (n = 18) and double stranded RNA (n = 6) (**Table 4.1**). Three viruses were only classified to the level of phylum or class. One additional virus was taxonomically unclassified. The newly discovered viruses shared between 25.2 and 80.7% amino acid sequence similarity to the RdRp of the closest viral hit in the NCBI-nr (**Table 4.1**).

Notably, the mite-specific viruses were highly divergent. BLASTX similarity searches revealed that the majority of the virus contigs were related to arthropod-associated viruses (29/37), although we identified three viruses associated with those previously identified in nematodes (Wuchang romanomermis nematode virus 2; similarity = 33.5%), protozoans (Leptomonas moramango leishbunyavirus; similarity = 80.7%) and algae (Diatom colony associated dsRNA virus 11; similarity = 34.2%). Likewise, we identified three viruses in the families *Narnaviridae* and *Tymoviridae* that are most often associated with fungi or plants (**Table 4.1**).

Table 4.1 List of putative viruses discovered in this study and present in mite/mosquito hosts. Each viral sequence was compared with the NCBI non-redundant (nr) database using DIAMOND BLASTX. Hosts are represented with letters corresponding to mites (K), mite-free mosquitoes (M) and mite-detached mosquitoes (MK).

Virus name	Contig name	Length	Provisional classification	Best hit in the NCBI/nr	Similarity	E-value	Host
Smedsang bunya-like virus	k119_16122	2250	*Bunyavirales*	BBQ05095.1 RNA-dependent RNA polymerase [Culex pseudovishnui bunya-like virus]	34.5	1.34E-100	K, M, MK
Avesta bunya-like virus	k119_3430	1925	*Bunyavirales*	QNS17451.1 RNA-dependent RNA polymerase, partial [Serbia bunya-like virus 1]	57.4	2.84E-249	K, M, MK
Heby virus	k119_4879	1754	*Bunyavirales*	QGA70945.1 RNA-dependent RNA polymerase [Salari virus]	69	8.28E-286	M, MK
Buska virus	k119_17401	7663	*Bunyavirales*	AJG39275.1 RNA-dependent RNA polymerase [Zhee Mosquito virus]	39.4	0	K, M, MK
Gaddsjo leishbunyavirus	k119_11873	2442	*Bunyavirales Leishbuviridae*	ANJ59510.1 putative RNA dependent RNA polymerase [Leptomonas moramango leishbunyavirus]	80.7	0	M, MK
Sater virus	k119_979	13942	*Bunyavirales* Nairo-like	YP_009300680.1 RNA-dependent RNA polymerase [Shayang Spider Virus 1]	25.4	1.42E-278	K, MK
Fallet virus	k119_5606	7027	*Jingchuvirales Chuviridae*	API61887.1 RNA-directed RNA polymerase [Chuvirus Mos8Chu0]	63	0	K, M, MK
Hede virus	k119_2521	1378	*Amarillovirales Flaviviridae*	YP_009179222.1 polyprotein [Xinzhou spider virus 2]	31	2.07E-53	K
Broddbo narna-like virus	k119_1307	3122	*Wolframvirales Narnaviridae*	APG77272.1 RNA-dependent RNA polymerase, partial [Wenling narna-like virus 6]	35.9	6.23E-179	K
Hytton narna-like virus	k119_17837	3140	*Wolframvirales Narnaviridae*	AGW51768.2 putative RNA-dependent RNA polymerase-like protein [Ochlerotatus-associated narna-like	73.4	0	K, M, MK

				virus 2]			
Hedemora virus	k119_6373	16014	*Mononegavirales* Rhabdo-like	YP_009304476.1 RNA-dependent RNA polymerase [Tacheng Tick Virus 7]	29.2	5.70E-211	K
Sonnboviken virus	k119_1814	1016	*Mononegavirales* Lispi-like	QMP82230.1 RNA-dependent RNA polymerase, partial [Megalopteran arli-related virus OKIAV106]	25.2	2.02E-13	K
Fors virus	k119_3330	3921	*Mononegavirales* Lispi-like	YP_009342285.1 RNA-dependent RNA polymerase [Wuchang romanomermis nematode virus 2]	33.5	7.53E-171	K
Osterbannback virus	k119_16137	13129	*Mononegavirales*	QRW42735.1 RNA-dependent RNA polymerase [Gordis virus]	37.6	0	K, M, MK
Bro virus	k119_22347	6397	*Mononegavirales Xinmoviridae*	BBQ04817.1 RNA-dependent RNA polymerase [Culex tritaeniorhynchus Anphevirus]	41.2	0	M
Malby virus	k119_10539	6320	*Mononegavirales Xinmoviridae*	BBQ04817.1 RNA-dependent RNA polymerase [Culex tritaeniorhynchus Anphevirus]	40	0	K, M, MK
Pelarsalen rhabdo-like virus	k119_4181	6711	*Mononegavirales* Rhabdo-like	QHA33680.1 RdRp [Atrato Rhabdo-like virus 3]	48.6	0	K, M, MK
Tierp virus	k119_3941	2428	*Articulavirales Orthomyxoviridae*	QRW42655.1 polymerase PB1 [Usinis virus]	57.6	4.35e-311	K, MK
Husby virus	k119_19965	2448	*Articulavirales Orthomyxoviridae*	QGA70921.1 RNA-dependent RNA polymerase [Wuhan Mosquito Virus 4]	55.9	3.01E-305	K, M, MK
Kagbo partiti-like virus	k119_14506	1722	*Durnavirales Partitiviridae*	APG78217.1 RdRp [Hubei partiti-like virus 22]	60	2.19E-248	MK
Ormpussen virus	k119_12042	1630	*Durnavirales Partitiviridae*	AWY11085.1 orf1 [Galbut virus]	36	8.71E-90	K, M, MK
Hebron partiti-like virus	k119_1779	1699	*Durnavirales Partitiviridae*	APG78260.1 RdRp [Hubei partiti-like virus 19]	50.6	7.23E-194	K, M, MK
Hundmyran chaq-like virus	k119_19664	1476	*Durnavirales Partitiviridae*	AKH40308.1 orf1 [Chaq virus*]	52.9	3.8E-103	K, M, MK
Nor picorna-like virus	k119_7745	9260	*Picornavirales Iflaviridae*	AWC26954.1 polyprotein [Culex picorna-like virus 1]	53.9	0	M
Dalkarlsbo virus	k119_10044	2238	*Quenyaviridae*	QIQ61196.1 putative RNA dependent RNA polymerase [Nete virus]	39.7	5.40E-158	K, M, MK
Morgongava	k119_3685	5653	*Muvirales*	QGA70948.1 RNA-dependent RNA	33.9	3.85E-226	K, M,

virus			*Qinviridae*	polymerase [Vittskovle virus]			MK
Hallarsbo virus	k119_7784	4121	*Muvirales Qinviridae*	QLJ83493.1 RNA-dependent RNA polymerase [Fitzroy Crossing qinvirus 1]	33	1.09E-222	K, M, MK
Berg reo-like virus	k119_12531	4277	*Reovirales Sedoreoviridae*	QHA33824.1 putative RdRp [Atrato Reo-like virus]	63.8	0	M
Koversta virus	k119_2089	3950	*Ghabrivirales Totiviridae*	QHA33712.1 RdRp [Embera virus]	54.2	0	M
Disbo virus	k119_15173	2401	*Ghabrivirales Totiviridae*	YP_009552795.1 RNA dependent RNA polymerase [Diatom colony associated dsRNA virus 11]	34.2	2.53E-131	M
Karbo virus	k119_13033	1727	*Tymovirales Tymoviridae*	YP_009551972.1 polyprotein [Alfalfa virus F]	44.5	5.20E-139	M, MK
Ginka virga-like virus	k119_5852	2312	*Martellivirales Virgaviridae*	QHA33742.1 polyprotein [Atrato Virga-like virus 3]	57.1	4.76E-285	M, MK
Baggbo virus	k119_4924	4407	Unclassified Sobelivirales$^{\varphi}$	QIS87998.1 RNA-dependent RNA polymerase [Khabarov virus]	42.7	1.87E-75	K
Sala virus	k119_14050	5220	Unclassified Ellioviricetes$^{\varphi}$	AGW51765.1 RNA-dependent RNA polymerase-like protein [uncultured virus]	39.3	0	K, M, MK
Nedre virus	k119_4180	7799	Unclassified Ellioviricetes$^{\varphi}$	AGW51765.1 RNA-dependent RNA polymerase-like protein [uncultured virus]	39.2	0	K, MK
Kvarnon virus	k119_3338	1763	*Ortervirales Metaviridae*	QPF16710.1 putative RNA-dependent RNA polymerase [Aedes aegypti To virus 2]	51.3	7.60E-194	K, M, MK
Fullsta virus	k119_9880	2263	Unclassified Pisuviricota$^{\varphi}$	QFR59041.1 putative RNA dependent RNA polymerase, partial [Hanyang virus]	41.5	2.57E-157	K, M, MK

*Chaq virus is often considered either a satellite virus or a segment or galbut virus. $^{\varphi}$ Previously unclassified viruses annotated using RdRp-scan (Charon et al., 2022a).

Based on the phylogenetic analysis, we identified several putative novel viruses in mosquitoes and mites that shared close relationships to known RNA viruses within the families *Chuviridae*, *Flaviviridae*, *Metaviridae*, *Narnaviridae*, *Orthomyxoviridae*, *Partitiviridae*, *Iflaviridae*, *Qinviridae*, *Quenyaviridae*, *Sedoreoviridae, Totiviridae* and *Tymoviridae* (**Figure 4.1–4.5**). Of particular note was a novel pestivirus, tentatively named Hede virus, that exhibited ~31% amino acid sequence similarity to Xinzhou spider virus 2 previously discovered in spiders (Araneae), and the novel Kvarnon virus that shared ~51% similarity to the errantivirus Aedes aegypti To virus 2 (*Metaviridae*) (**Table 4.1, Figure 4.3B, Figure 4.4C**) (Shi et al., 2016a). Similarly, we identified viruses related to members of the *Leishbunyaviridae* and *Nairoviridae* within the order *Bunyavirales*, as well as members of the *Rhadboviridae*, *Lispiviridae* and *Xinmoviridae* in the order *Mononegavirales* of single-strand negative-sense RNA viruses (**Figure 4.1**). A small number of the novel viruses identified here grouped with unclassified RNA virus sequences in the *Bunyavirales* and *Mononegavirales*. Due to the limited similarity shared between the novel and known viruses, we only recovered partial genome/replicase sequences encoding conserved domains such as the RdRp and MTase, as well as segments encoding uncharacterized proteins (**Table S4.3, Figure 4.5**).

Although the newly discovered Baggbo virus, Fullsta virus, Sala virus and Nedre virus shared limited similarity with unclassified viruses (similarity = 39.2–42.7 per cent) (**Table 4.1**), we provided a broad taxonomic assignment for these viruses within the *Sobelivirales*, *Ellioviricetes* and *Pisuviricota* (**Table S4.2**). Also of note was that the newly identified viruses for which the taxonomic status could be assigned fell into distinct clades within several families, helping to fill the gaps in the phylogenies of these groups. In other cases, the putative viruses identified here are grouped together as sister taxa to each other. For example, the Hallarsbo virus fell as a sister taxon to Morgongava virus (*Qinviviridae*) as part of a clade of mosquito-associated viruses (**Figure 4.2D**), as did Malby virus and Bro virus (*Xinmoviridae*) (**Figure 4.1B**). Finally, some of the newly identified viral sequences occupied basal phylogenetic positions, such as Heby virus (*Bunyavirales*), which was a

sister to a clade comprising the newly discovered viruses Smedsang bunya-like virus, Avesta bunya-like virus, Buska virus and their closest known relatives in mosquitoes (**Figure 4.1A**). Similarly, Sater virus (a Nairo-like virus) shared common ancestry with other tick nairoviruses (**Figure 4.1A**).

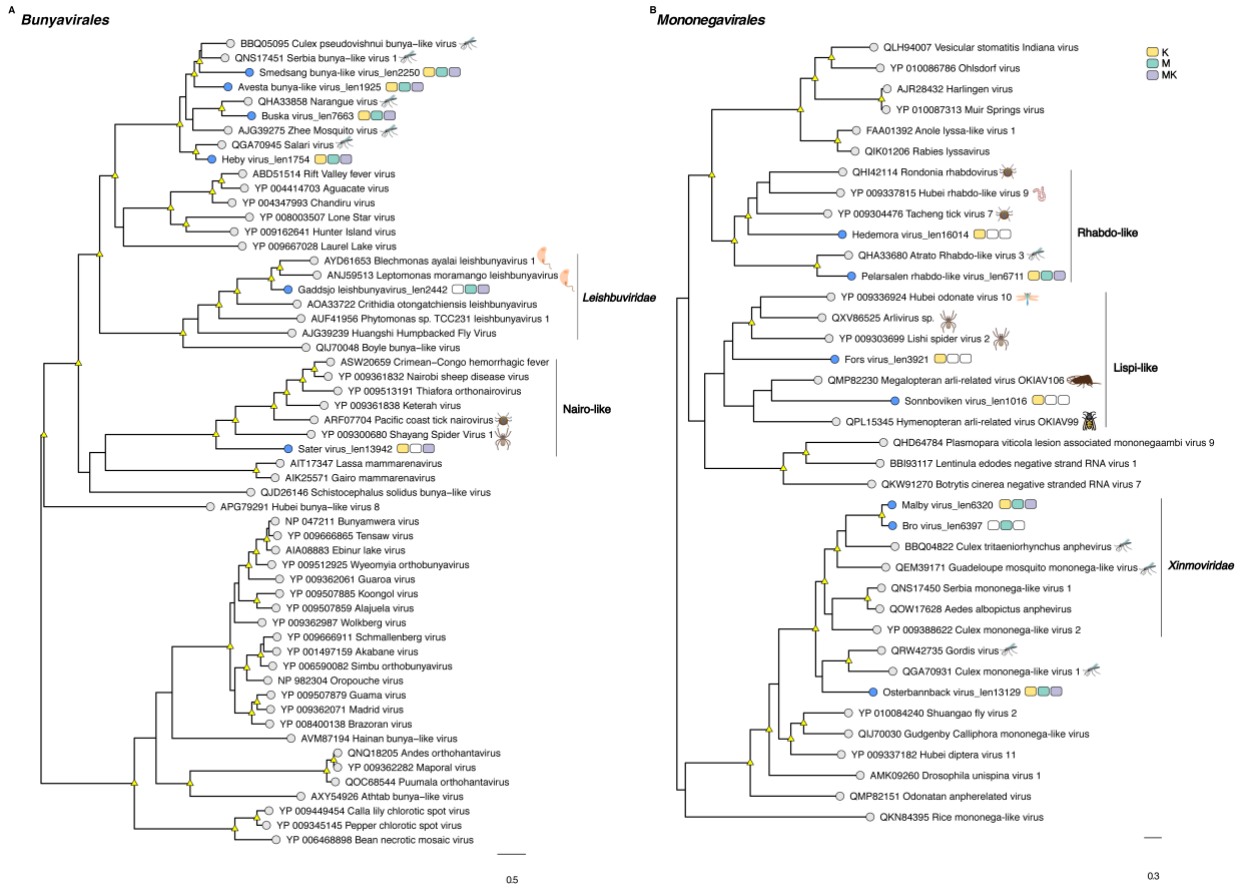


Figure 4.1 Phylogenetic relationships among the viruses found in this study and reference sequences within the RNA virus orders (A) *Bunyavirales* and (B) *Mononegavirales*. Phylogenetic trees were estimated using the Q.pfam+F+I+Γ_4 substitution model. Novel viruses are indicated with blue tip points and hosts are represented with three-pack bars corresponding to mite (K; yellow), mite-free mosquito (M; green), and mite-detached mosquito samples (MK; purple). Trees are based on the amino acid sequences of the putative RdRp. Nodal support values ≥ 80% SH-aLRT and ≥ 95% UFboot are denoted with yellow triangles at nodes. Scale bars indicate the number of amino acid substitutions per site and the trees are mid-point rooted for clarity only. Host species information (animal icons) is shown for the closest relatives of the novel viruses.

4.4.2 Composition and distribution of RNA viruses reveal host connectivity

To assess the differences in the virome composition between groups we determined virus prevalence across all the sequencing libraries generated here. This revealed a similar number of viruses present in M (n = 29), MK (n = 25) and K (n = 26) (**Figure 4.6**). Notably, six viruses were specific to the mite (K) libraries, eighteen viruses were shared between mosquitoes (M/MK) and mites (K), whereas five viruses were exclusively present in the mite-free (M) libraries (**Table 4.1**, **Figure 4.6**). The shared viruses were classified within the *Chuviridae*, *Narnaviridae*, *Orthomyxoviridae, Partitiviridae* and *Quenyaviridae*, as well as those assigned to the orders *Bunyavirales* and *Mononegavirales* (**Table 4.1, Figure 4.6**). Cross-reference between some MK and K pools was concordant with these results (**Figure S1)**. Within the *Partitiviridae,* Ormpussen galbut-like virus and Hundmyran chaq-like virus exhibited limited sequence similarity to galbut (similarity = 36%) and chaq virus (similarity = ~ 53%), respectively (**Table 4.1**). Remarkably, these viruses also co-occurred in most libraries (7/10) (**Figure 4.7, Figure S1**). In general, we found at least four viruses per library, with the exception of the mite-free mosquito library M3, which harboured seventeen novel viruses (**Figure 4.7**).

Although we expected that all viruses present in MK libraries (i.e. mosquitoes previously infested with mites) would also be found in either M or K libraries, Kagbo partiti-like virus was only recovered from a single MK library (**Figure 4.6–4.7**). In contrast, we observed common viruses among host-specific libraries, such as Hallarsbo virus (*Qinviridae*) and Kvarnon virus (*Metaviridae*) in mosquitoes and Baggbo virus (*Sobelivirales*) and Hedemora virus (*Mononegavirales*) in mites (**Table 4.1, Figure 4.7**). Only the putative Buska virus (*Bunyavirales*) was broadly distributed among most of the libraries (17/24).

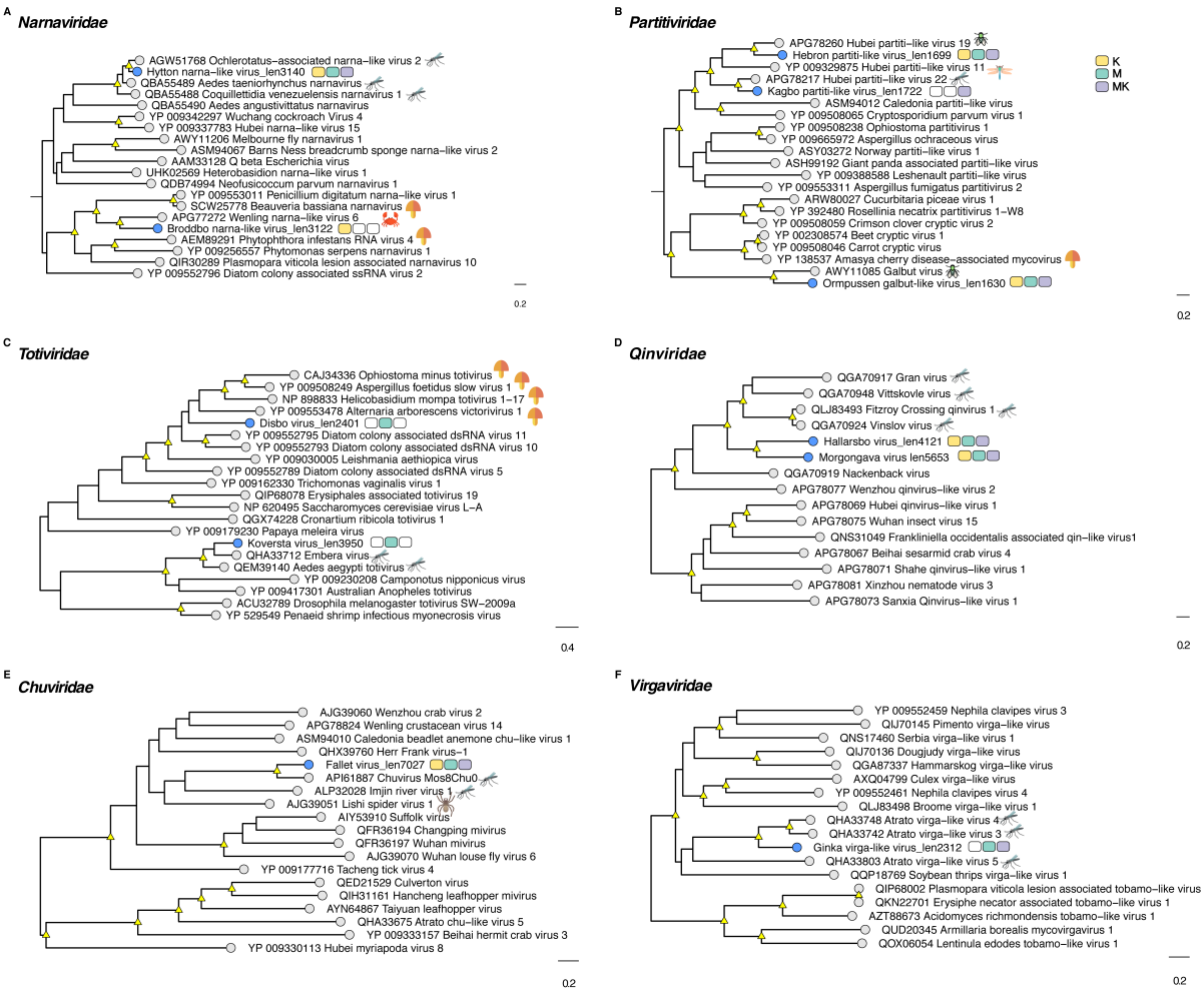


Figure 4.2 Phylogenetic relationships among the viruses found in this study and reference sequences within the RNA virus families (A) *Narnaviridae*, (B) *Partitiviridae*, (C) *Totiviridae*, (D) *Qinviridae*, (E) *Chuviridae* and (F) *Virgaviridae*. Phylogenetic trees were estimated using the

VT+F+I+G4 (*Narnaviridae*, *Partitiviridae,Virgaviridae*) and Q.pfam+F+I+Γ_4 (*Totiviridae, Qinviridae, Chuviridae*) substitution models. Novel viruses are indicated with blue tip points and hosts are represented with three-pack bars corresponding to mite (K; yellow), mite-free mosquito (M; green), and mite-detached mosquito samples (MK; purple). Trees are based on the amino acid sequences of the putative RdRp. Nodal support values ≥ 80% SH-aLRT and ≥ 95% UFboot are denoted with yellow triangles at nodes. Scale bars indicate the number of amino acid substitutions per site and the trees are mid-point rooted for clarity only. Host species information (animal icons) is shown for the closest relatives of the novel viruses.

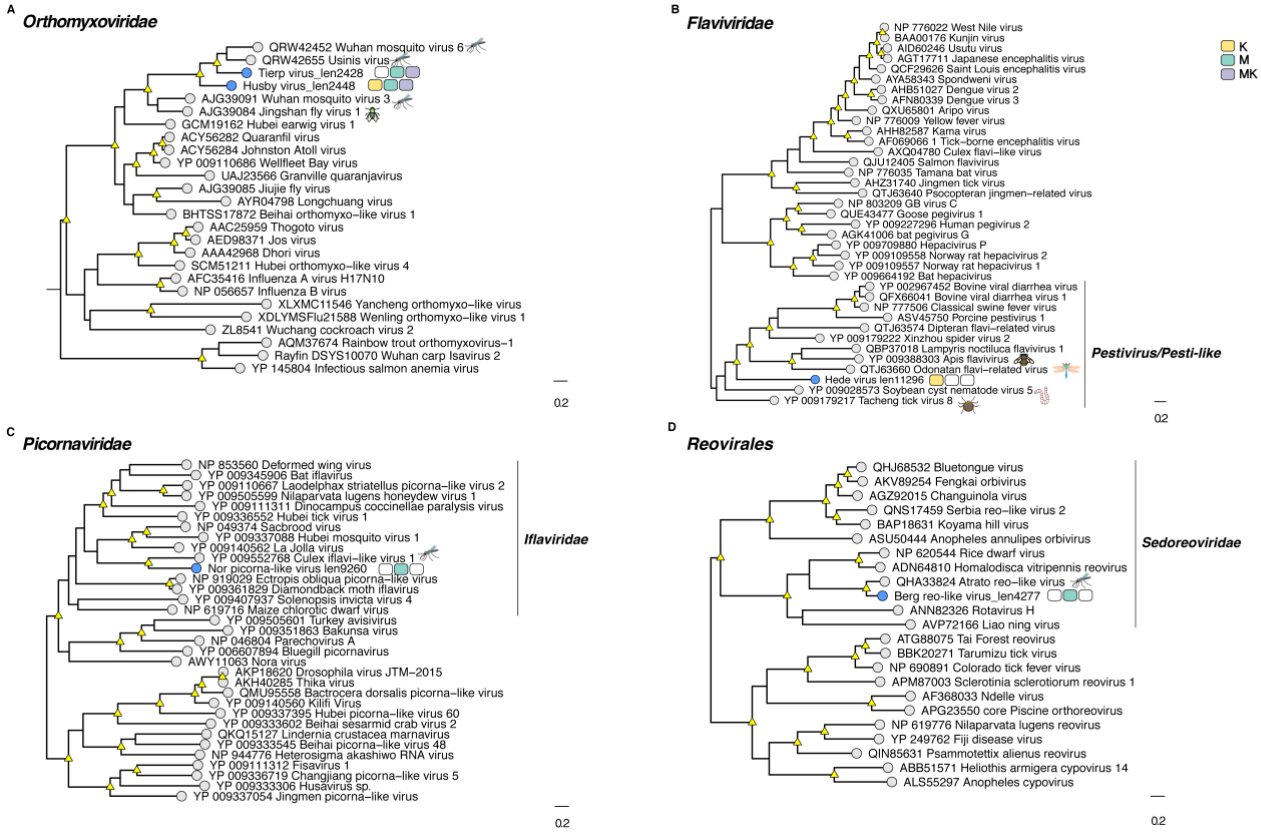


Figure 4.3 Phylogenetic relationships among the viruses found in this study and reference sequences within the RNA virus families (A) *Orthomyxoviridae*, (B) *Flaviviridae*, (C) *Picornaviridae* and (D) *Reovirales*. Phylogenetic trees were estimated using the VT+F+I+G4 (*Flaviviridae*, *Picornavirales* and *Sedoreoviridae*) and Q.pfam+F+I+ Γ_4 (*Orthomyxoviridae*) substitution models. Novel viruses are indicated with blue tip points and hosts are represented with three-pack bars corresponding to mite (K; yellow), mite-free mosquito (M; green), and mite-detached mosquito

samples (MK; purple). Trees are based on the amino acid sequences of the putative RdRp. Nodal support values ≥ 80% SH-aLRT and ≥ 95% UFboot are denoted with yellow triangles at nodes. Scale bars indicate the number of amino acid substitutions per site and the trees are mid-point rooted for clarity only. Host species information (animal icons) is shown for the closest relatives of the novel viruses.

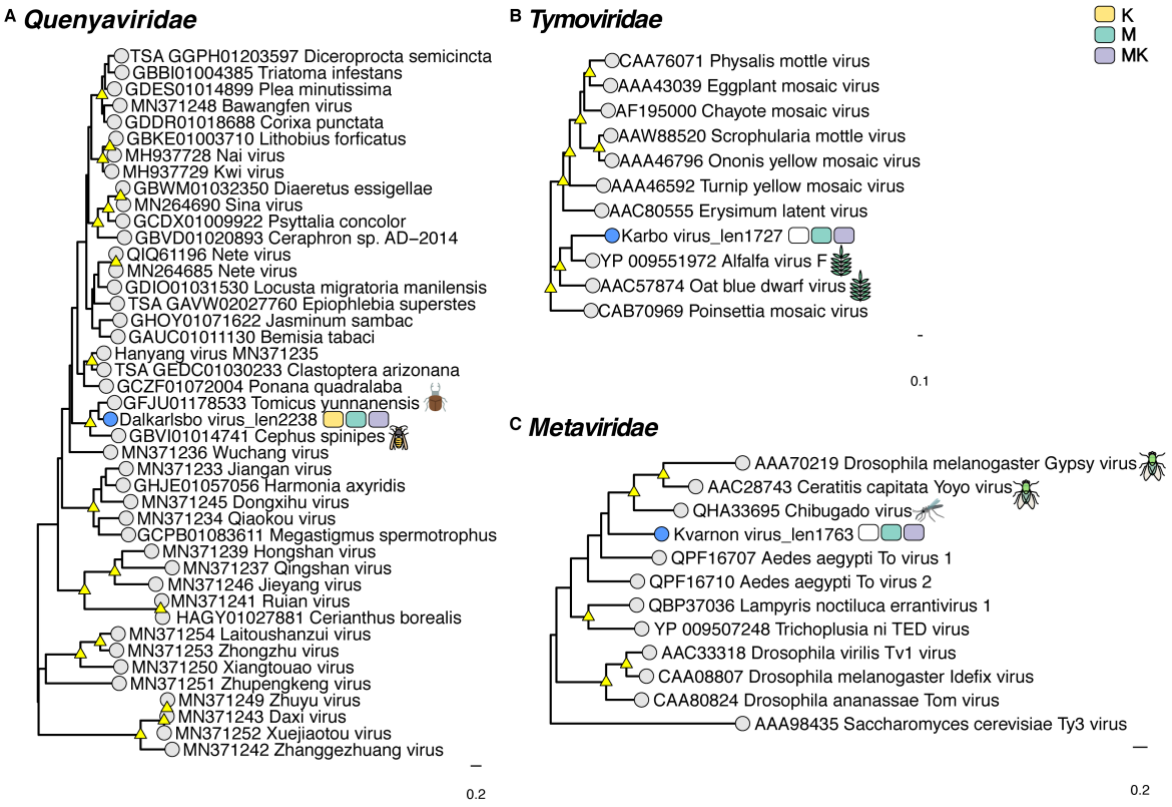


Figure 4.4 Phylogenetic relationships among the viruses found in this study and reference sequences within the RNA virus families (A) *Quenyaviridae* (B) *Tymoviridae*, and (C) *Metaviridae*. Phylogenetic trees were estimated using the rtREV+F+I+ Γ_4 substitution model. Novel viruses are indicated with blue tip points and hosts are represented with three-pack bars corresponding to mite (K; yellow), mite-free mosquito (M; green), and mite-detached mosquito samples (MK; purple). Trees are based on the amino acid sequences of the putative RdRp or RT. Nodal support values ≥ 80% SH-aLRT and ≥ 95% UFboot are denoted with yellow triangles at nodes. Scale bars indicate the number of amino acid substitutions per site and the trees are mid-point rooted for

clarity only. Host species information (animal icons) is shown for the closest relatives of the novel viruses.

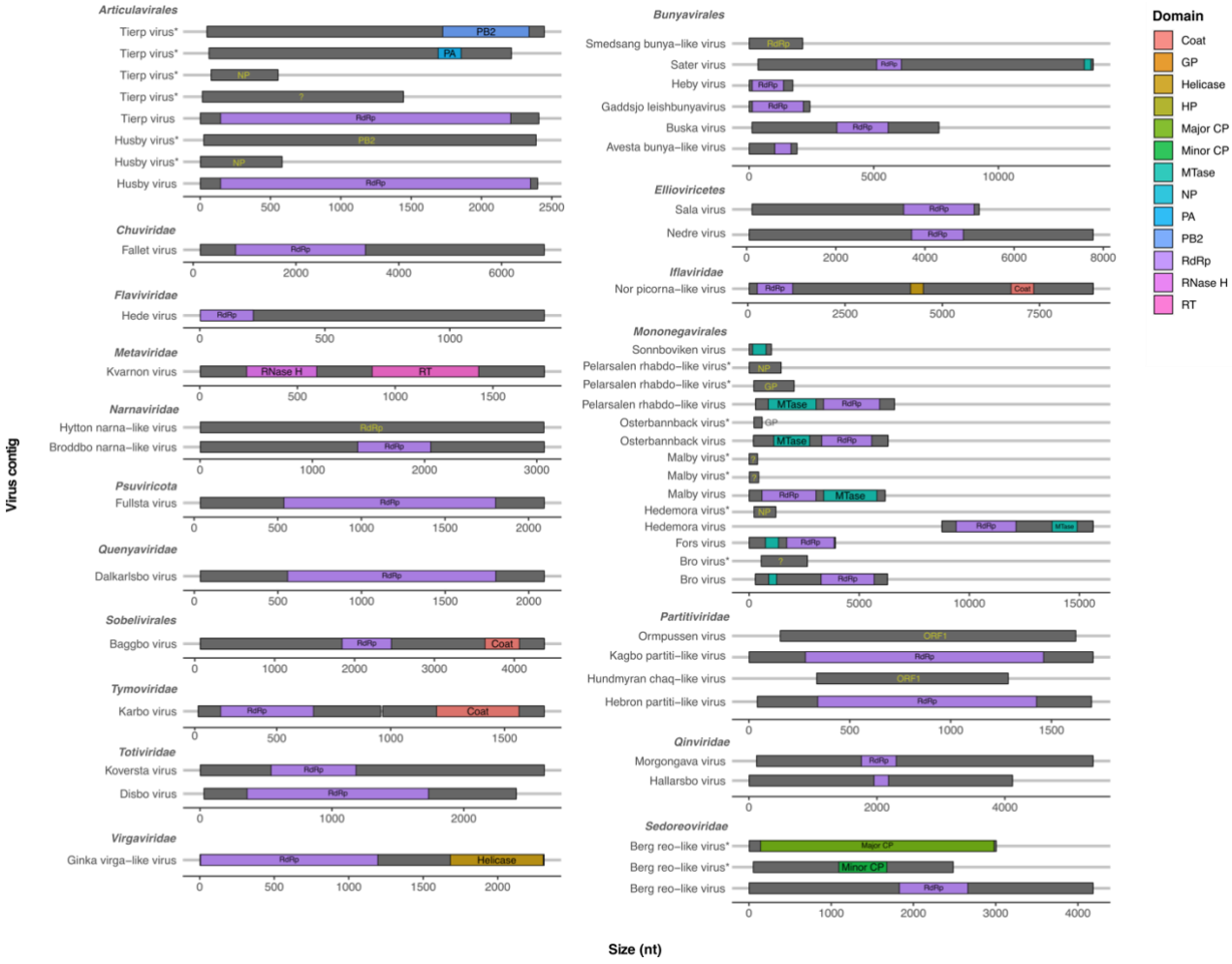


Figure 4.5 Schematic representation of protein domains found in the viral sequences identified in this study. Putative novel viruses are grouped by virus family. Diagrams represent predicted ORFs (grey), while domains are displayed as colored boxes (see legend). ORFs lacking conserved domains are annotated (text labels) based on the closest hit in the BLASTX search. The question marks (?) represent ORFs encoding hypothetical/unknown proteins. The ORFs size is shown as number of nucleotides and each ORF is located along the contig according to the predicted coordinates. Multiple segments of a virus are indicated by asterisks (*). GP: glycoprotein; HP: hypothetical protein; CP: core protein; MTase: metyltransferase; NP: nucleoprotein; PA:

polymerase acidic protein; PB2: polymerase basic protein 2; RdRp: RNA-dependent RNA polymerase; RNase H: ribonuclease H; RT: reverse transcriptase.

4.4.3 Virus abundance levels suggest host-associations and virome connectivity between mosquitoes and mites

We next assessed virus abundance for each newly discovered virus compared to host gene markers (**Figure 4.7**). Accordingly, virus abundance varied from 1 to 47,863 RPM (**Figure 4.7**). In contrast, the abundance of the reference mitochondrial gene marker *Cox-1* was more stable across host-specific libraries, ranging between 40 and 307 RPM for *Mideopsis* sp. and 43–348 RPM for *A. communis*. Importantly, reads from the mosquito host were detected in all the libraries (**Figure 4.7**). Comparisons of virus abundances between groups K, M and MK revealed no significant differences (KW = 2.68 p-value = 0.2617). We considered viruses with values > 1000 RPM (> 0.1% of ribosomal-depleted RNA) to be highly abundant. For example, Hytton narna−like virus (*Narnaviridae*, RPM = 20–39,637) exhibited the highest abundance and was present across mite and mosquito libraries (**Figure 4.7**). Nearly all viruses exclusively detected in mites showed very high abundance levels, such as Baggbo virus (*Sobelivirales*, RPM = 84–10,251), which was highly prevalent among K libraries (**Figure 4.5 and 4.7**). We were also able to identify some viruses that were abundant in mosquito libraries but were still present in mites at negligible levels, including Buska virus (*Bunyavirales*, RPM = 1–1,885), Osterbannback virus (*Partitiviridae,* RPM = 1–3771) and Fallet virus (*Chuviridae*, RPM = 6–11,594), supporting the idea that these viruses are likely more associated with mosquitoes (**Figure 4.7**). Likewise, although both Kvarnon virus (*Metaviridae,* RPM = 6–19) and Hallarsbo virus (*Quinviridae,* RPM = 2–93) were at low abundance, they were stably expressed and highly prevalent in M libraries while scarce in K libraries. It is also notable that viruses restricted to single libraries, such as Kagbo partiti−like virus (*Partitiviridae*, RPM = 125) and Disbo virus (*Totiviridae*, RPM = 135) were at low abundance (**Figure 4.7**). Overall, these results revealed differences in virus composition and abundance that might help demonstrate

virus-host associations as well as connectivity (i.e. potential virus transfer) through the host-parasite system.

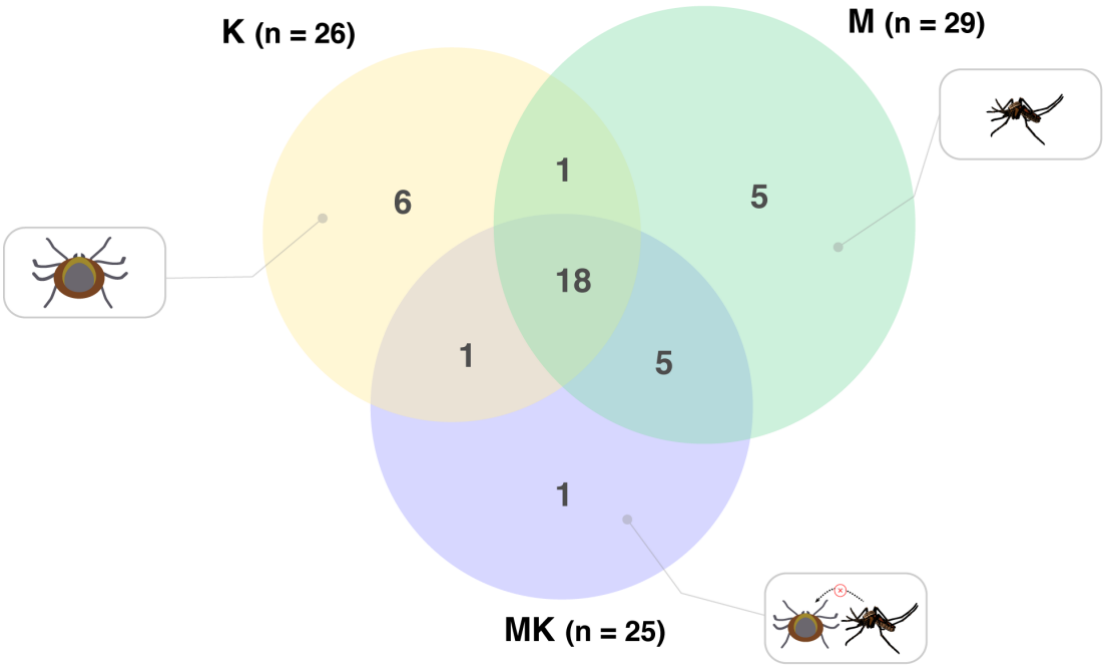


Figure 4.6 Comparison of the number of viruses shared between mites (K), mite-free mosquitoes (M), and mite-detached mosquito libraries (MK). The total number of viruses is indicated for each group. The size of the circle is proportional to the number of viruses identified in each host group. Color coding is the same as in Figure 4.1–4.4.

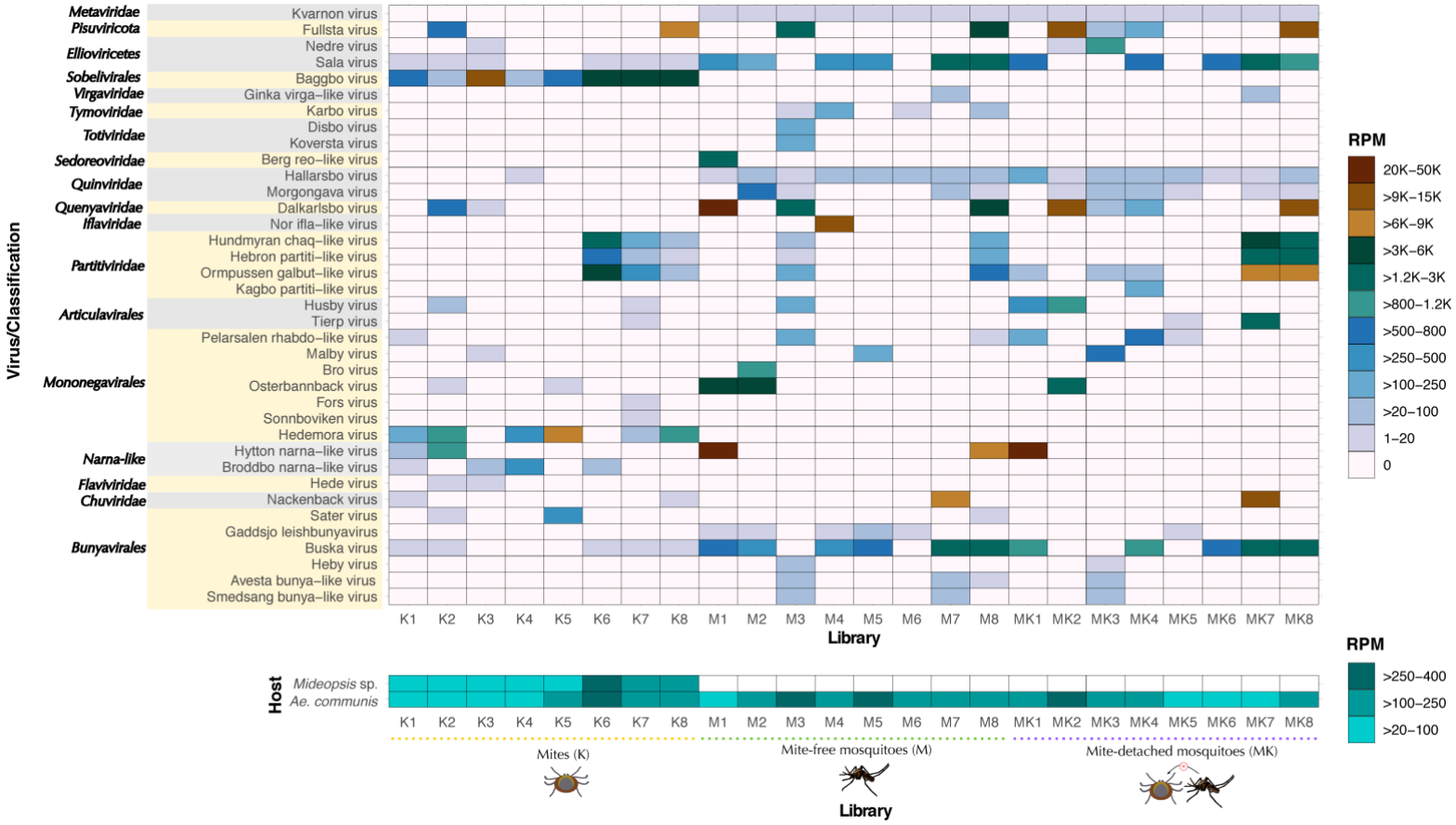


Figure 4.7 Overview of virus abundance quantified as reads per million (RPM). The newly discovered viruses are indicated on the y-axis. Abundance levels are color-coded as a heat map as specified in the legend. The queried libraries corresponding to mite (K), mite-free mosquito (M) and mite-detached mosquito (MK) are shown on the x-axis. Host abundance levels based on the *Cox-1* gene marker are represented in the bottom panel.

4.5 Discussion

Host–parasite relationships between mosquitoes and mites have impact on both arthropod and ecosystem ecology. However, aspects such as host and parasite virome diversity and composition, have largely been neglected within the mite-mosquito interaction. Here, we provide an overview of the diversity of RNA viruses in these arthropods, comparing their virome profile to investigate possible transfer events between both hosts. In line with previous studies on arthropod viromes (C.-X. Li et al., 2015; Obbard et al., 2020; Pettersson et al., 2017; Shi et al., 2016), we observed a high abundance of many diverse viruses, suggesting that many more arthropod viruses remain to be discovered (Harvey & Holmes, 2022; Junglen & Drosten, 2013; Shi et al., 2016).

Although mites are known to be vectors of some pathogens of medical and veterinary importance for vertebrate hosts (Gubler, 1988; Hubálek et al., 2014; Mullen & O'Connor, 2019; Weaver & Reisen, 2010; Yu & Tesh, 2014), these arthropods can also act as vectors of viral agents to other arthropods. For example, Varroa mites (*Varroa destructor*) appear to mediate transmission of KBV, SBV, and DWV to honey bees *Apis mellifera* (Shen et al., 2005). As a consequence, the interaction between parasites and their hosts is also likely to lead to the transfer of viruses. We investigated the virome of mosquitoes and their infesting mites to reveal mosquito-mite interactions. A key result was that a number of viruses were commonly present in both mosquitoes and mite samples (**Figure 4.6**), indicating that the transfer of viruses is likely to occur when parasitic mites feed on dipteran hosts (Dolja & Koonin, 2018). Further research is needed to assess whether these viruses are able to infect, replicate and spread in both arthropod hosts, as opposed to being of dietary origin or infecting components of the microbiome (**Figure S4.3**) (Obbard, 2018b). In the latter case, the presence of partitiviruses, such as Ormpussen galbut-like virus and Hebron partiti−like virus, both widely distributed among mite and mosquito samples, suggest the presence of common fungal and/or protozoan microbiota (**Figure 4.2B and 4.7**). Nonetheless, we cannot definitively exclude possible contamination with microorganisms present on the surface of arthropods or derived from sample processing.

The repertoire of putative viruses identified in this study spanned different viral families previously reported in mites and mosquitoes (**Table 4.1, Figure 4.1–4.4**) (Chang et al., 2021; Junglen & Drosten, 2013; Li et al., 2015; Shi et al., 2016). In particular, prior to this work there were a relatively limited number of viruses recorded in mites that parasitize other arthropods, representing the *Chuviridae*, *Dicistroviridae*, *Iflaviridae* and *Rhabdoviridae* families (Dietzgen et al., 2014; Niu et al., 2019; Shen et al., 2005). Our results expand this virus diversity to include viruses within the orders *Sobelivirales*, and *Mononegavirales* and the family *Flaviviridae* (**Figure 4.1B, 4.3B, and 4.7**). The occurrence of flaviviruses has been previously recorded in acarid ectoparasites parasitizing natural bird populations (Kovalev & Yakimenko, 2021; Santillán et al., 2015). In contrast, the highly divergent Hede virus (likely partial RdRp) found in mites was most closely related to arthropod-specific long genome flaviviruses (**Table 4.1, Figure 4.3B, Figure 4.5**) (Paraskevopoulou et al., 2021), with related viruses documented in ticks (Tokarz et al., 2014; Xu et al., 2022). Therefore, the presence of highly divergent viruses in mites suggests a hidden diversity in Acari.

A key outcome of this study was the presence of viruses restricted to either mosquito or mite libraries, which we hypothesized to correspond to host-specific viruses or those associated with the host microbiota. For example, the presence of the Baggbo virus at abundant levels across several K samples (**Table 4.1, Figure 4.7**), suggests that these arthropods might serve as natural carriers of this virus. In contrast, Kvarnon virus, which was found at low abundance in the majority of the M and MK libraries, yet not the K libraries, is presumed to derive from symbionts in *A. communis* mosquitoes (**Figure S4.3**). However, due to the small sample size, caution is needed in the formulation of definitive virus-host associations.

Insect-specific viruses (ISVs) in mosquitoes represent the *Bunyaviridae*, *Sedoreoviridae*, *Iflaviridae, Mononegavirales* and *Flaviviridae* (Roundy et al., 2017). Bro virus (*Xinmoviridae*), Nor ifla-like virus (*Iflaviridae*) and Berg reo-like virus (*Sedoreoviridae*), described here may also constitute ISVs as they were found in high abundance and grouped with other viruses reported

in mosquitoes (**Figure 4.1–4.5**) (Pettersson et al., 2019). In contrast, the novel totiviruses are more likely to have a fungal or protozoan origin (Coatsworth et al., 2021; Fauver et al., 2016; Wu et al., 2020) (**Figure S4.3**). Finally, among the set of newly discovered viruses, the broad distribution of the Buska virus across most libraries (**Figure 4.7**) is consistent with that reported for its closest match – Zhee mosquito virus in *Coquillettidia richardii* and *Aedes* spp. (Öhlund et al., 2019). Together with the substantial variation in the abundance of Buska virus between mite and mosquito libraries, we hypothesized that this bunyavirus might infect and replicate well in *A. communis* mosquitoes.

Given the occurrence of Kagbo partiti-like virus in a single MK library and the lack of detection in K and M samples, its true host-association is difficult to assign (**Figure 4.2B and 4.7**). Despite our thorough examination of the samples, it is possible that small remnants of mouthparts contaminated with fungi or protozoa were still present in the M/MK libraries, which may explain the sporadic occurrence of the novel Kagbo partiti-like virus. Conversely, the presence of contaminant viral sequences might offer an alternative explanation (Porter et al., 2021). It has previously been shown that contaminant viruses can not only be derived from multiple sources, including specimen surface contamination, reagents, controls and cell culture, but can also be introduced at any step in sample preparation and sequencing (Batson et al., 2021a; Cobbin et al., 2021; Porter et al., 2021a). Interestingly, the closest relative to Kagbo partiti-like virus (**Figure 4.2B**), has been reported in *Culex modestus, Culex. vishnui, Culex. tritaeniorhynchus*, *Culex. Quinquefasciatus*, *Culex. pipiens* and *Culex. torrentium* and *A. aegypti* from different geographic globally (Faizah et al., 2020a; Öhlund et al., 2019; Pettersson et al., 2019; C. Shi et al., 2019; L. Wang et al., 2021) .

Previous work on arthropod viral transcriptomes strongly suggests that galbut and chaq virus are associated with a satellite–helper virus system or are part of the same segmented RNA virus (Batson et al., 2021; Cross et al., 2020; Shi et al., 2018; Webster et al., 2015), although key aspects of the system are still poorly understood. In this context, the co-occurrence of the novel partitiviruses Ormpussen galbut-like virus and Hundmyran chaq-like virus distantly related to

galbut virus and chaq virus-like sequences, respectively, which further supports the notion of an existing relationship between these viruses, extending this association to more distant viral relatives. It is important to note that we detected a similar pattern of co-occurrence for Hebron partiti-like virus (**Figure 4.7**). This observation agrees with previous studies reporting the presence of multiple partitiviruses in samples (Faizah et al., 2020b; Webster et al., 2015a), although, based on the available data, we were unable to determine whether Hebron partiti–like virus is specifically associated with Ormpussen galbut-like virus and Hundmyran chaq-like sequences.

It is also important to consider the co-occurrence of multiple virus taxa within libraries regardless of the arthropod host (**Figure 4.7**). Accordingly, our analysis revealed a heterogeneous diversity (i.e. composition and abundance) of RNA viruses within libraries. These differences might reflect underlying interactions at the host–parasite–virus interface. Indeed viral infection can shape host-parasite relationships by impacting ectoparasite virulence and imposing differential selective pressures on the hosts in question (Di Prisco et al., 2016; Parratt & Laine, 2016a). However, the interactions between viruses carried by the parasite host and the base host have been poorly explored (Díaz-Muñoz, 2019), and viruses can also interact with host microbiota (Altinli et al., 2021; Hahn et al., 2020; Jagdale & Joshi, 2018). Understanding the implications of such symbiotic relationships in arthropods is therefore of importance (Altinli et al., 2021).

4.6 Significance, limitations, and future directions

Our findings provide preliminary baseline evidence for understanding the structure of the RNA virome in mosquitoes and their parasitic mites. A holistic understanding will require research addressing open questions on such major topics as host associations and competence, as well as the effect of virus infection on host biology. This study extends the current diversity of RNA viruses in arthropods and provide high-resolution insights into the RNA viral metagenome in the context of host-parasite interactions. Future research efforts should be

addressed to determine the impact of these viruses on host-parasite relationships as well as the ecological and evolutionary implications for this and other tripartite systems.

4.7 Supplementary Material

Supplementary data are available at *Virus Evolution* online.

Table S4.1. Summary of sample collection and pooling of mosquitoes and mites in this study.

Table S4.2. Summary of taxonomic assignment for the unclassified viruses identified in this study using the RdRp-scan resource.

Table S4.3. Overview of completeness and quality of the viral sequences identified in this study.

Figure S4.1. Cross-reference between mite-detached mosquitoes (MK5-MK8) and mite (K5-K8) pools. This information was only available for a limited number of pools. Links between groups indicate the source of the mites used for pooling. Shared viruses are shown with colored circles at intersections (see legend).

Figure S4.2. Comparison of *Cox-1* gene abundance levels among libraries using different reference sequences of *Mideopsis sp*.

Figure S4.3. Metagenomic characterization of group samples based on the number contigs taxonomically assigned to the most common biological groups. Break points in the y-axis show different scales for easy visualization of disparate values (S. Xu et al., 2021).

4.7 References

Altinli, M., Schnettler, E., & Sicard, M. (2021). Symbiotic interactions between mosquitoes and mosquito viruses. *Frontiers in Cellular and Infection Microbiology*, *11*, 818.

Andrews, S. (2010). *FastQC: a quality control tool for high throughput sequence data.*

Atwa, A. A., Bilgrami, A. L., & Al-Saggaf, A. I. M. (2017). Host–parasite interaction and impact of mite infection on mosquito population. *Revista Brasileira de Entomologia*, *61*(2), 101–106. https://doi.org/10.1016/j.rbe.2017.03.005

Auld, S. K. J. R., Searle, C. L., & Duffy, M. A. (2017). Parasite transmission in a natural multihost–multiparasite community. *Philosophical Transactions of the Royal Society B: Biological Sciences*, *372*(1719), 20160097. https://doi.org/10.1098/rstb.2016.0097

Batson, J., Dudas, G., Haas-Stapleton, E., Kistler, A. L., Li, L. M., Logan, P., Ratnasiri, K., & Retallack, H. (2021). Single mosquito metatranscriptomics identifies vectors, emerging pathogens and reservoirs in one assay. *ELife*, *10*. https://doi.org/10.7554/eLife.68353

Becker, N., Petric, D., Zgomba, M., Boase, C., Madon, M., Dahl, C., & Kaiser, A. (2010). *Mosquitoes and their control* (Second edi). Springer Science & Business Media. https://doi.org/10.1007/978-3-540-92874-4

Buchmann, J. P., & Holmes, E. C. (2020). Collecting and managing taxonomic data with NCBI-taxonomist. *Bioinformatics*, *36*(22–23), 5548–5550.

Campbell, G. L., Eldridge, B. F., Reeves, W. C., & Hardy, J. L. (1991). Isolation of Jamestown Canyon virus from boreal *Aedes* mosquitoes from the Sierra Nevada of California. *American Journal of Tropical Medicine and Hygiene*, *44*(3), 244–249. https://doi.org/10.4269/ajtmh.1991.44.244

Chang, T., Hirai, J., Hunt, P., & Suttle, C. (2021). Arthropods and the evolution of RNA viruses. *BioRxiv*. https://doi.org/10.1101/2021.05.30.446314

Charon, J., Buchmann, J. P., Sadiq, S., & Holmes, E. C. (2022). RdRp-scan: A Bioinformatic resource to identify and annotate divergent RNA viruses in metagenomic sequence data. *BioRxiv*.

Coatsworth, H., Bozic, J., Carrillo, J., Buckner, E. A., Rivers, A. R., Dinglasan, R. R., & Mathias, D. K. (2021). Intrinsic variation in the vertically transmitted insect-specific core virome of Aedes aegypti. *BioRxiv*.

Cobbin, J. C., Charon, J., Harvey, E., Holmes, E. C., & Mahar, J. E. (2021). Current challenges to virus discovery by meta-transcriptomics. *Current Opinion in Virology*, *51*, 48–55.

Cross, S. T., Maertens, B. L., Dunham, T. J., Rodgers, C. P., Brehm, A. L., Miller, M. R., Williams, A. M., Foy, B. D., & Stenglein, M. D. (2020). Partitiviruses infecting *Drosophila melanogaster* and *Aedes aegypti* exhibit efficient biparental vertical transmission. *Journal of Virology*, *94*(20). https://doi.org/10.1128/jvi.01070-20

Dainat, B., Ken, T., Berthoud, H., & Neumann, P. (2009). The ectoparasitic mite *Tropilaelaps mercedesae* (Acari, Laelapidae) as a vector of honeybee viruses. *Insectes Sociaux*, *56*(1), 40–43. https://doi.org/10.1007/s00040-008-1030-5

Di Prisco, G., Annoscia, D., Margiotta, M., Ferrara, R., Varricchio, P., Zanni, V., Caprio, E., Nazzi, F., & Pennacchio, F. (2016). A mutualistic symbiosis between a parasitic mite and a pathogenic virus undermines honey bee immunity and health. *Proceedings of the National Academy of Sciences of the United States of America*, *113*(12), 3203–3208. https://doi.org/10.1073/PNAS.1523515113

Díaz-Muñoz, S. L. (2019). Uncovering virus-virus interactions by unifying approaches and harnessing high-throughput tools. *MSystems*, *4*(3).

Dietzgen, R. G., Kuhn, J. H., Clawson, A. N., Freitas-Astúa, J., Goodin, M. M., Kitajima, E. W., Kondo, H., Wetzel, T., & Whitfield, A. E. (2014). Dichorhavirus: a proposed new genus for *Brevipalpus* mite-transmitted, nuclear, bacilliform, bipartite, negative-strand RNA plant viruses. *Archives of Virology*, *159*(3), 607–619. https://doi.org/10.1007/S00705-013-1834-0

Dolja, V. v., & Koonin, E. v. (2018). Metagenomics reshapes the concepts of RNA virus evolution by revealing extensive horizontal virus transfer. *Virus Research*, *244*, 36–52.

dos Santos, E. B., Favretto, M. A., dos Santos Costa, S. G., & Navarro-Silva, M. A. (2016). Mites (Acari: Trombidiformes) parasitizing mosquitoes (Diptera: Culicidae) in an Atlantic Forest area in southern Brazil with a new mite genus country record. *Experimental and Applied Acarology*, *69*(3), 323–333. https://doi.org/10.1007/s10493-016-0045-2

Ewels, P., Magnusson, M., Lundin, S., & Käller, M. (2016). MultiQC: summarize analysis results for multiple tools and samples in a single report. *Bioinformatics*, *32*(19), 3047–3048. https://doi.org/10.1093/bioinformatics/btw354

Faizah, A. N., Kobayashi, D., Isawa, H., Amoa-Bosompem, M., Murota, K., Higa, Y., Futami, K., Shimada, S., Kim, K. S., Itokawa, K., Watanabe, M., Tsuda, Y., Minakawa, N., Miura, K., Hirayama, K., & Sawabe, K. (2020a). Deciphering the virome of *Culex vishnui* subgroup mosquitoes, the major vectors of Japanese encephalitis, in Japan. *Viruses*, *12*(3), 264. https://doi.org/10.3390/v12030264

Fauver, J. R., Grubaugh, N. D., Krajacich, B. J., Weger-Lucarelli, J., Lakin, S. M., Fakoli, L. S., Bolay, F. K., Diclaro, J. W., Dabiré, K. R., Foy, B. D., Brackney, D. E., Ebel, G. D., & Stenglein, M. D. (2016). West African Anopheles gambiae mosquitoes harbor a taxonomically diverse virome including new insect-specific flaviviruses, mononegaviruses, and totiviruses. *Virology*, *498*, 288–299.

Finn, R. D., Clements, J., & Eddy, S. R. (2011). HMMER web server: interactive sequence similarity searching. *Nucleic Acids Research*, *39*(Web Server issue), W29. /pmc/articles/PMC3125773/

Gubler, D. J. (1988). The arboviruses: epidemiology and ecology. *Dengue*, *II*, 223–260.

Gurevich, A., Saveliev, V., Vyahhi, N., & Tesler, G. (2013). QUAST: quality assessment tool for genome assemblies. *Bioinformatics*, *29*(8), 1072–1075. https://doi.org/10.1093/bioinformatics/btt086

Hahn, M. A., Rosario, K., Lucas, P., & Dheilly, N. M. (2020). Characterization of viruses in a tapeworm: phylogenetic position, vertical transmission, and transmission to the parasitized host. *ISME Journal*, *14*(7), 1755–1767. https://doi.org/10.1038/S41396-020-0642-2

Harvey, E., & Holmes, E. C. (2022). Diversity and evolution of the animal virome. *Nature Reviews Microbiology*, *0123456789*. https://doi.org/10.1038/s41579-021-00665-x

Hoang, D. T., Chernomor, O., Von Haeseler, A., Minh, B. Q., & Vinh, L. S. (2018). UFBoot2: Improving the ultrafast bootstrap approximation. *Molecular Biology and Evolution*, *35*(2), 518–522. https://doi.org/10.1093/molbev/msx281

Hubálek, Z., Rudolf, I., & Nowotny, N. (2014). Arboviruses pathogenic for domestic and wild animals. In *Advances in virus research* (Vol. 89, pp. 201–275). https://doi.org/10.1016/B978-0-12-800172-1.00005-7

Hyatt, D., Locascio, P. F., Hauser, L. J., & Uberbacher, E. C. (2012). Gene and translation initiation site prediction in metagenomic sequences. *Bioinformatics*, *28*(17), 2223–2230.

Jagdale, S. S., & Joshi, R. S. (2018). Enemies with benefits: mutualistic interactions of viruses with lower eukaryotes. *Archives of Virology*, *163*(4), 821–830.

Jones, P., Binns, D., Chang, H.-Y., Fraser, M., Li, W., McAnulla, C., McWilliam, H., Maslen, J., Mitchell, A., Nuka, G., Pesseat, S., Quinn, A. F., Sangrador-Vegas, A., Scheremetjew, M., Yong, S.-Y., Lopez, R., & Hunter, S. (2014). InterProScan 5: genome-scale protein function classification. *Bioinformatics (Oxford, England)*, *30*(9), 1236–1240. https://doi.org/10.1093/bioinformatics/btu031

Junglen, S., & Drosten, C. (2013). Virus discovery and recent insights into virus diversity in arthropods. *Current Opinion in Microbiology*, *16*(4), 507–513. https://doi.org/10.1016/J.MIB.2013.06.005

Kaplan, I., & Eubanks, M. D. (2005). Aphids alter the community-wide impact of fire ants. *Ecology*, *86*(6), 1640–1649. https://doi.org/10.1890/04-0016

Kassambara, A. (2020). *ggpubr: 'ggplot2' based publication ready plots*. https://cran.r-project.org/package=ggpubr

Kassambara, A. (2021). *rstatix: pipe-friendly framework for basic statistical tests*. https://rpkgs.datanovia.com/rstatix/

Kopylova, E., Noé, L., & Touzet, H. (2012). SortMeRNA: fast and accurate filtering of ribosomal RNAs in metatranscriptomic data. *Bioinformatics*, *28*(24), 3211–3217.

Kovalev, S. Y., & Yakimenko, V. V. (2021). Kama virus (KAMV) is an atypical representative of the seabird tick-borne flaviviruses. *Virus Genes*, *57*(4), 395–399.

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. (2015). Unprecedented genomic diversity of RNA viruses in arthropods reveals the ancestry of negative-sense RNA viruses. *ELife*, *4*. https://doi.org/10.7554/eLife.05378

Li, D., Liu, C.-M., Luo, R., Sadakane, K., & Lam, T.-W. (2015). MEGAHIT: an ultra-fast single-node solution for large and complex metagenomics assembly via succinct de Bruijn graph. *Bioinformatics*, *31*(10), 1674–1676. https://doi.org/10.1093/bioinformatics/btv033

Liu, S., Chen, Y., & Bonning, B. C. (2015). RNA virus discovery in insects. *Current Opinion in Insect Science*, *8*, 54–61. https://doi.org/10.1016/J.COIS.2014.12.005

Lundström, J., Schäfer, M., Hesson, J., Blomgren, E., Lindström, A., Wahlqvist, P., Halling, A., Hagelin, A., Ahlm, C., Evander, M., Broman, T., Forsman, M., & Vinnersten, T. (2013). The geographic distribution of mosquito species in Sweden. *Journal of the European Mosquito Control Association*, *31*.

Lvov, D. K., Shchelkanov, M. Y., Alkhovsky, S. V., & Deryabin, P. G. (2015). *Zoonotic viruses of Northern Eurasia: taxonomy and ecology*. Academic Press.

Marcelino, V. R., Clausen, P. T. L. C., Buchmann, J. P., Wille, M., Iredell, J. R., Meyer, W., Lund, O., Sorrell, T. C., & Holmes, E. C. (2020). CCMetagen: Comprehensive and accurate identification of

eukaryotes and prokaryotes in metagenomic data. *Genome Biology*, *21*(1), 1–15. https://genomebiology.biomedcentral.com/articles/10.1186/s13059-020-02014-2

Medvedev, S. G., Aibulatov, S. V., & Panyukova, E. V. (2011). Ecological features and distribution of the mosquito *Aedes communis* (De Geer, 1776) in the Northwestern European Russia. *Entomological Review*, *91*(9), 1108–1121. https://doi.org/10.1134/S0013873811090041

Mullen, G. R., & O'Connor, B. M. (2019). Mites (Acari). *Medical and Veterinary Entomology*, 533–602. https://doi.org/10.1016/B978-0-12-814043-7.00026-1

Nayfach, S., Camargo, A. P., Schulz, F., Eloe-Fadrosh, E., Roux, S., & Kyrpides, N. C. (2020). CheckV assesses the quality and completeness of metagenome-assembled viral genomes. *Nature Biotechnology 2020 39:5*, *39*(5), 578–585. https://doi.org/10.1038/s41587-020-00774-7

Nguyen, L.-T., Schmidt, H. A., von Haeseler, A., & Minh, B. Q. (2015). IQ-TREE: a fast and effective stochastic algorithm for estimating maximum-likelihood phylogenies. *Molecular Biology and Evolution*, *32*(1), 268–274. https://doi.org/10.1093/molbev/msu300

Niu, J., Zhang, W., Sun, Q. Z., & Wang, J. J. (2019). Three novel RNA viruses in the spider mite Tetranychus urticae and their possible interactions with the host RNA interference response. *Journal of Invertebrate Pathology*, *166*, 107228. https://doi.org/10.1016/J.JIP.2019.107228

Obbard, D. J. (2018). Expansion of the metazoan virosphere: progress, pitfalls, and prospects. *Current Opinion in Virology*, *31*, 17–23. https://doi.org/10.1016/J.COVIRO.2018.08.008

Obbard, D. J., Shi, M., Roberts, K. E., Longdon, B., & Dennis, A. B. (2020). A new lineage of segmented RNA viruses infecting animals. *Virus Evolution*, *6*(1), 61. https://academic.oup.com/ve/article/6/1/vez061/5708932

Öhlund, P., Hayer, J., Lundén, H., Hesson, J. C., & Blomström, A.-L. (2019). Viromics reveal a number of novel RNA viruses in Swedish mosquitoes. *Viruses*, *11*(11), 1027. https://doi.org/10.3390/v11111027

Paraskevopoulou, S., Kä, S., Zirkel, F., Donath, A., Petersen, M., Liu, S., Zhou, X., Drosten, C., Misof, B., & Junglen, S. (2021). Viromics of extant insect orders unveil the evolution of the flavi-like superfamily. *Virus Evolution*, *7*(1), 30. https://doi.org/10.1093/ve/veab030

Parratt, S. R., & Laine, A. L. (2016). The role of hyperparasitism in microbial pathogen ecology and evolution. *The ISME Journal 2016 10:8*, *10*(8), 1815–1822. https://doi.org/10.1038/ismej.2015.247

Peng, P., Blamires, S. J., Agnarsson, I., Lin, H. C., & Tso, I. M. (2013). A color-mediated mutualism between two arthropod predators. *Current Biology*, *23*(2), 172–176. https://doi.org/10.1016/J.CUB.2012.11.057

Pettersson, J. H. O., Shi, M., Bohlin, J., Eldholm, V., Brynildsrud, O. B., Paulsen, K. M., Andreassen, Å., & Holmes, E. C. (2017). Characterizing the virome of *Ixodes ricinus* ticks from Northern Europe. *Scientific Reports 2017 7:1*, *7*(1), 1–7. https://doi.org/10.1038/s41598-017-11439-y

Pettersson, J. H. O., Shi, M., Eden, J. S., Holmes, E. C., & Hesson, J. C. (2019). Meta-transcriptomic comparison of the RNA viromes of the mosquito vectors *Culex pipiens* and *Culex torrentium* in Northern Europe. *Viruses 2019, Vol. 11, Page 1033*, *11*(11), 1033. https://doi.org/10.3390/V11111033

Poinar, G., & Poinar, R. (1998). Parasites and pathogens of mites. *Annual Review of Entomology*, *43*(1), 449–469. https://doi.org/10.1146/annurev.ento.43.1.449

Porter, A. F., Cobbin, J., Li, C.-X., Eden, J.-S., & Holmes, E. C. (2021a). Metagenomic identification of viral sequences in laboratory reagents. *Viruses*, *13*(11), 2122. https://doi.org/10.3390/v13112122

Pozojević, I., Juršić, L., Vučković, N., Dorić, V., Gottstein, S., Ternjej, I., & Mihaljević, Z. (2019). Is the spatial distribution of lentic water mite assemblages (Acari: Hydrachnidia) governed by prey availability? *Experimental and Applied Acarology*, *77*(4), 487–510.

R Core Team. (2021). *R: A language and environment for statistical computing*.

Rice, P., Longden, L., & Bleasby, A. (2000). EMBOSS: the European Molecular Biology Open Software Suite. *Trends in Genetics : TIG*, *16*(6), 276–277. https://doi.org/10.1016/S0168-9525(00)02024-2

Roundy, C. M., Azar, S. R., Rossi, S. L., Weaver, S. C., & Vasilakis, N. (2017). Insect-specific viruses: A historical overview and recent developments. *Advances in Virus Research*, *98*, 119–146. https://doi.org/10.1016/BS.AIVIR.2016.10.001

Santillán, M., Grande, J. M., Liébana, M. S., Martínez, P., Díaz, L. A., Bragagnolo, L. A., Solaro, C., Galmes, M. A., & Sarasola, J. H. (2015). New hosts for the mite *Ornithonyssus bursa* in

Argentina. *Medical and Veterinary Entomology*, *29*(4), 439–443. https://doi.org/10.1111/MVE.12129

Shen, M., Yang, X., Cox-Foster, D., & Cui, L. (2005). The role of varroa mites in infections of Kashmir bee virus (KBV) and deformed wing virus (DWV) in honey bees. *Virology*, *342*(1), 141–149.

Shi, C., Beller, L., Deboutte, W., Yinda, K. C., Delang, L., Vega-Rúa, A., Failloux, A.-B., & Matthijnssens, J. (2019). Stable distinct core eukaryotic viromes in different mosquito species from Guadeloupe, using single mosquito viral metagenomics. *Microbiome*, *7*(1), 121. https://doi.org/10.1186/s40168-019-0734-2

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*, *540*(7634), 539–543. https://doi.org/10.1038/nature20167

Shi, M., Lin, X.-D., Vasilakis, N., Tian, J.-H., Li, C.-X., Chen, L.-J., Eastwood, G., Diao, X.-N., Chen, M.-H., Chen, X., Qin, X.-C., Widen, S. G., Wood, T. G., Tesh, R. B., Xu, J., Holmes, E. C., & Zhang, Y.-Z. (2016a). Divergent viruses discovered in arthropods and vertebrates revise the evolutionary history of the Flaviviridae and related viruses. *Journal of Virology*, *90*(2), 659–669.

Shi, M., White, V. L., Schlub, T., Eden, J.-S., Hoffmann, A. A., & Holmes, E. C. (2018). No detectable effect of *Wolbachia* wMel on the prevalence and abundance of the RNA virome of *Drosophila melanogaster*. *Proceedings of the Royal Society B: Biological Sciences*, *285*(1883), 20181165. https://doi.org/10.1098/rspb.2018.1165

Shi, M., Zhang, Y. Z., & Holmes, E. C. (2018a). Meta-transcriptomics and the evolutionary biology of RNA viruses. *Virus Research*, *243*, 83–90. https://doi.org/10.1016/j.virusres.2017.10.016

Tokarz, R., Williams, S. H., Sameroff, S., Sanchez Leon, M., Jain, K., & Lipkin, W. I. (2014). Virome analysis of Amblyomma americanum, Dermacentor variabilis, and Ixodes scapularis ticks reveals novel highly divergent vertebrate and invertebrate viruses. *Journal of Virology*, *88*(19), 11480–11492.

Valiente Moro, C., Chauve, C., & Zenner, L. (2005). Vectorial role of some dermanyssoid mites (Acari, Mesostigmata, Dermanyssoidea). *Parasite*, *12*(2), 99–109. https://doi.org/10.1051/parasite/2005122099

Vasquez, A. A., Kabalan, B. A., Ram, J. L., & Miller, C. J. (2020). The biodiversity of water mites that prey on and parasitize mosquitoes. *Diversity 2020, Vol. 12, Page 226*, *12*(6), 226. https://doi.org/10.3390/D12060226

Wang, L., Rosales Rosas, A. L., De Coninck, L., Shi, C., Bouckaert, J., Matthijnssens, J., & Delang, L. (2021). Establishment of *Culex modestus* in Belgium and a glance into the virome of Belgian mosquito species. *MSphere*, *6*(2). https://doi.org/10.1128/msphere.01229-20

Weaver, S. C., & Reisen, W. K. (2010). Present and future arboviral threats. *Antiviral Research*, *85*(2), 328–345. https://doi.org/10.1016/j.antiviral.2009.10.008

Webster, C. L., Waldron, F. M., Robertson, S., Crowson, D., Ferrari, G., Quintana, J. F., Brouqui, J.-M., Bayne, E. H., Longdon, B., Buck, A. H., Lazzaro, B. P., Akorli, J., Haddrill, P. R., & Obbard, D. J. (2015a). The discovery, distribution, and evolution of viruses associated with *Drosophila melanogaster*. *PLoS Biology*, *13*(7), e1002210. https://doi.org/10.1371/journal.pbio.1002210

Werblow, A., Martin, P., Dörge, D. D., Koch, L. K., Mehlhorn, H., Melaun, C., & Klimpel, S. (2015a). Hyperparasitism of mosquitoes by water mite larvae. *Parasitology Research*, *114*(7), 2757–2765. https://doi.org/10.1007/s00436-015-4482-3

Wickham, H. (2016). ggplot2*: elegant graphics for data analysis*. Springer-Verlag New York.

Wu, H., Pang, R., Cheng, T., Xue, L., Zeng, H., Lei, T., Chen, M., Wu, S., Ding, Y., Zhang, J., Shi, M., & Wu, Q. (2020). Abundant and diverse RNA viruses in insects revealed by RNA-Seq analysis: ecological and evolutionary implications. *MSystems*, *5*(4).

Xu, Z., Feng, Y., Chen, X., Shi, M., Fu, S., Yang, W., Liu, W. J., Gao, G. F., & Liang, G. (2022). Virome of bat-infesting arthropods: highly divergent viruses in different vectors. *Journal of Virology*, *96*(4).

Yu, G., Smith, D. K., Zhu, H., Guan, Y., & Lam, T. T. (2017). *ggtree* : an *r* package for visualization and annotation of phylogenetic trees with their covariates and other associated data. *Methods in Ecology and Evolution*, *8*(1), 28–36. https://doi.org/10.1111/2041-210X.12628

Yu, X. -j., & Tesh, R. B. (2014a). The role of mites in the transmission and maintenance of Hantaan virus (*Hantavirus: Bunyaviridae*). *Journal of Infectious Diseases*, *210*(11), 1693–1699. https://doi.org/10.1093/infdis/jiu336

Zhang, Y.-Z., Shi, M., & Holmes, E. C. (2018). Using metagenomics to characterize an expanding virosphere. *Cell*, *172*(6), 1168–1172. https://doi.org/10.1016/j.cell.2018.02.043

Chapter 5

RNA virome diversity and *Wolbachia* infection in individual *Drosophila simulans* flies

Ayda Susana Ortiz-Baez, Mang Shi, Ary A. Hoffmann, Edward C. Holmes

JOURNAL OF
GENERAL VIROLOGY

Article

CHAPTER 5 RNA virome diversity and *Wolbachia* infection in individual *Drosophila simulans* flies

Ayda Susana Ortiz-Baez[1], Mang Shi[1], Ary A. Hoffmann[2*], Edward C. Holmes[1*]

[1] Marie Bashir Institute for Infectious Diseases and Biosecurity, School of Life and Environmental Sciences and School of Medical Sciences, The University of Sydney, Sydney, New South Wales 2006, Australia.

[2] School of BioSciences, Bio21 Institute, University of Melbourne, Parkville, Victoria 3010, Australia.

***** Correspondence: edward.holmes@sydney.edu.au

Received: 26 April 2021; Accepted: 1 July 2021; Published: 27 October 2021;
https://doi.org/10.1099/jgv.0.001639

The following version of this study may have slight differences from the original publication.

5.1 Abstract

The endosymbiont bacterium *Wolbachia* is associated with multiple mutualistic effects on insect biology, including nutritional and antiviral properties. *Wolbachia* naturally occurs in *Drosophila* fly species, providing an operational model host to study how virome composition may be impacted by its presence. *Drosophila simulans* populations can carry a variety of *Wolbachia* strains, with the *w*Au strain associated with strong antiviral protection under experimental conditions. We used *D. simulans* sampled from the Perth Hills, Western Australia, to investigate the potential virus protective effect of the *Wolbachia w*Au strain on individual wild-caught flies. Our data revealed no appreciable variation in virus composition and abundance between *Wolbachia* infected/uninfected individuals associated with the presence/absence of *w*Au. However, it remains unclear whether *w*Au might impact viral infection and host survival by increasing

tolerance rather than inducing complete resistance. These data also provide new insights into the natural virome diversity of *D. simulans*. Despite the small number of individuals sampled, we identified a repertoire of RNA viruses, including nora virus, galbut virus, thika virus and La Jolla virus, that have been identified in other *Drosophila* species. Chaq virus-like sequences associated with galbut virus were also detected. In addition, we identified five novel viruses from the families *Reoviridae*, *Tombusviridae*, *Mitoviridae* and *Bunyaviridae.* Overall, this study highlights the complex interaction between *Wolbachia* and RNA virus infections and provides a baseline description of the natural virome of *D. simulans*.

Keywords: Drosophila simulans , evolution , meta-transcriptomics , phylogeny , RNA virome and Wolbachia

---

5.2 Introduction

The alpha-proteobacterium *Wolbachia* (order *Rickettsiales*) is a widespread endosymbiont of arthropods and nematodes (i.e. filarial and plant-parasitic nematodes) that can establish interactions with their hosts ranging from parasitic to mutualistic (Ross et al., 2019; Zug & Hammerstein, 2015b). The genetic diversity of *Wolbachia* is substantial and currently represented by 11 distinctive supergroups (denoted A-J), although the majority of *Wolbachia* strains belong to supergroups A and B (Ros et al., 2009) that are estimated to have diverged around 50 million years ago (Scholz et al., 2020). Although these bacteria are commonly found in reproductive tissues and the germline of their hosts, they have also been found in somatic tissues such as the brain, salivary glands and gut (Dobson et al., 1999; Frydman et al., 2006; Strunov & Kiseleva, 2016; Tsai et al., 2004; Zouache et al., 2009), such that understanding infection dynamics in detail is not a trivial matter (Frydman et al., 2006). *Wolbachia* primarily spread by vertical inheritance through transovarian transmission. However, the presence of *Wolbachia* in a diverse range of host species suggests that horizontal transmission, likely through antagonistic interactions (i.e.

herbivory, parasitism and predation), also contributes to the dissemination of the bacteria in nature (Scholz et al., 2020; Turelli et al., 2018).

The occurrence of *Wolbachia* bacteria in insects is often associated with their ability to manipulate host reproductive mechanisms and induce a range of alterations, including parthenogenesis, feminization, cytoplasmic incompatibility and sex-ratio distortion (O'Neill et al., 1997). Among these, cytoplasmic incompatibility is the most common phenotypic effect, and as such represents an appealing approach for vector population control. In this case, embryonic lethality is contingent on the infection status and the strain type harboured by males and females (Ross et al., 2019). In addition, the study of *Wolbachia*-host interactions has revealed a variety of mutualistic effects on host biology (Iturbe-Ormaetxe & O'Neill, 2007; Zug & Hammerstein, 2015). For instance, in filarial nematodes and the parasitoid wasp *Asobara tabida*, the presence of some *Wolbachia* strains has been positively associated with developmental processes, fertility and host viability (Dedeine et al., 2001; Hoerauf et al., 1999; Iturbe-Ormaetxe & O'Neill, 2007). Furthermore, nutritional mutualism between *Wolbachia* and the bedbug *Cimex lectularius* as well as *Wolbachia*-infected planthoppers, has been suggested as a means to explain B vitamin supplementation (Hosokawa et al., 2010; Ju et al., 2020; Nikoh et al., 2014).

Arguably the most important outcome of *Wolbachia* infection in insects is its potential for virus-blocking, which also provides a basis for intervention strategies based on the control of arbovirus transmission. This seemingly antiviral effect of *Wolbachia* has been well documented in some species of insects, including flies and mosquitoes. A striking example involves the transinfection of *Aedes aegypti* mosquitoes with the *Wolbachia* strain infecting *Drosophila melanogaster* (*w*Mel). *A. aegypti* is the primary vector of a number of important arboviruses, including members of the species *Dengue virus*, *zika virus* and *chikungunya virus*, and the establishment of the *w*Mel strain in wild mosquito populations represents a powerful and promising approach to decrease virus

transmission (Moreira et al., 2009; Tantowijoyo et al., 2020). Although the underlying mechanisms remain to be fully determined, it has been suggested that *Wolbachia* can modify the host environment or boost basal immunity to viruses by pre-stimulating the immune response of their hosts (Rancès et al., 2012). Potential antiviral mechanisms impacted by *Wolbachia* include gene expression of the Toll pathway, RNA interference, and modification of the host oxidative environment that likely trigger an antiviral immune response and hence limit infection (Rancès et al., 2012; Terradas et al., 2017; Zug & Hammerstein, 2015a).

Unlike *A. aegypti* mosquitoes, *Wolbachia* naturally occur in *Drosophila* species, providing a valuable model system to study *Wolbachia*-related virus protection (Martinez et al., 2014; Teixeira et al., 2008). Natural populations of *Drosophila* can carry a diverse array of insect-specific viruses belonging to the families *Picornaviridae*, *Dicistroviridae*, *Bunyaviridae*, *Reoviridae* and *Iflaviridae* amongst others (Webster et al., 2015b). The co-occurrence of *Wolbachia* in *D. melanogaster* has been associated with increased survival and different levels of resistance to laboratory viral infections in fly stocks under experimental conditions (Hedges et al., 2008; Teixeira et al., 2008). For example, *Wolbachia*-infected flies containing the dicistrovirus Drosophila C virus (DCV) showed a delay in mortality compared to *Wolbachia*-free flies (Hedges et al., 2008). In contrast, other studies found no or limited effect of *Wolbachia* on viral protection, as well as on virus prevalence and abundance in field-collected flies (Shi et al., 2018; Webster et al., 2015b). Such contrasting data emphasize the need of further research efforts to characterize the effect of *Wolbachia* strains on virus composition in *Drosophila* in nature.

Although the origin of *D. simulans* is thought to have been in East Africa or Madagascar, this species now has a cosmopolitan distribution (Lachaise et al., 1988). In Australia, *D. simulans* has been recorded along both east and west coasts as well as Tasmania, with the earliest record dating to 1956 (Mather, 1960). Human mobility and human-mediated activities have been associated with the introduction and spread of both

D. simulans and *Wolbachia* into Australia, where wild fly populations occur near human settlements, feeding and breeding on a variety of horticultural crops (Kriesner et al., 2013; Parsons & Bock, 1979). Several *Wolbachia* strains from supergroups A and B can naturally occur in populations of *D. simulans* (e.g. *w*Au, *w*Ri, *w*Ha, *w*Ma and *w*No) (Casiraghi et al., 2005; Osborne et al., 2009b). From these, *w*Au is associated with strong antiviral protection against Flock House virus (FHV) (*Nodaviridae*) and DCV (*Dicistroviridae*) under experimental conditions (Osborne et al., 2009b). The *w*Au infection in Australia was one of the first *Wolbachia* infections identified as showing no cytoplasmic incompatibility, despite being widespread at a low to intermediate frequency (Hoffmann et al., 1996). *w*Au increased in frequency along the east coast of Australia until it was replaced by *w*Ri that exhibits cytoplasmic incompatibility. However, unlike *w*Au, *w*Ri has not yet reached the Australian west coast (Kriesner et al., 2013). In this study, we used a meta-transcriptomic (i.e. RNA shotgun sequencing) approach to determine the virome diversity of individual field-collected *D. simulans* flies from Western Australia, and investigated how this virome diversity might be impacted by the presence of the *w*Au strain of *Wolbachia*.

5.3 Materials and Methods

5.3.1 D. simulans collection and taxonomic identification

Flies used for the virus work performed here were collected at Raeburn Orchards in the Perth Hills in Western Australia (Long. 116.0695, Lat. -32.1036) in July 2018 using banana bait. The *Wolbachia* frequency at two other locations in the area (Roleystone, Long. 116.0701, Lat. -32.1396; Cannington, Long. 115.9363, Lat. -32.0243) was also established with additional samples. Taxonomic identification to the species level was conducted based on the morphology of reproductive traits of males and via DNA barcoding (*cox1* gene marker). Field-collected flies were maintained at 19ºC under standard laboratory conditions until F1 offspring were raised. Parental and F1 generations were then stored at -80ºC until molecular processing.

5.3.2 Wolbachia detection

Wolbachia infection of field females was determined using F1 offspring from each field female. Note that *w*Au is transmitted at 100% from field females to the F1 laboratory generation (Hoffmann et al., 1996). DNA extraction from heads was performed using the Chelex 100 Resin (Bio-Rad Laboratories, Hercules, CA, USA) (Endersby et al., 2005) as adapted in Shi *et al*. (Shi et al., 2018). Screening of natural *Wolbachia* infection was conducted using a real-time PCR/ high-resolution melt assay (RT/HRM) and strain-specific primers targeting a 340-bp region of the surface protein of Wolbachia (*wsp)* gene for *w*Ri and *w*Au strains. The assay was run following the protocol of Kriesner *et al*. (Kriesner et al., 2013). In addition, reads were mapped to reference *Wolbachia wsp* gene sequences for *w*Ri (CP001391.1) and *w*Au (LK055284.1) with BBMap v.37.98 (minid=0.95) (available at https://sourceforge.net/projects/bbmap/).

5.3.3 RNA extraction and meta-transcriptome sequencing

We screened a total of 16 individual flies to assess the effect of *Wolbachia* infection on virome composition in *D. simulans*. Specimens were rinsed three times in RNA and DNA-free PBS solution (GIBCO). Total RNA from individual flies was extracted using the RNeasy Plus Mini Kit (Qiagen) following the manufacturer's instructions. RNA-seq libraries were constructed using a TruSeq total RNA Library Preparation Kit (Illumina). Host ribosomal depletion was performed using a Ribo-Zero Gold rRNA Removal Kit (Human/Mouse/Rat) (Illumina) and paired-end transcriptome sequencing was performed on the HiSeq2500 platform (Illumina). Libraries from *Wolbachia*-negative and positive infected flies were run in two I lanes.

5.3.4 De novo meta-transcriptome assembly and viral genome annotation

The overall quality assessment of reads was conducted in FastQC and Trimmomatic (Bolger et al., 2014). A *de novo* assembly of RNA-Seq data was performed using MEGAHIT

v.1.1.3, with default parameters (Grabherr et al., 2011a). Assembled contigs were then annotated through comparisons against the NCBI nonredundant (NCBI-nr) database using DIAMOND v2.0.4 (Buchfink et al., 2015), with a cut-off e-value <1e-05. To identify protein-encoding sequences, open reading frames (ORFs) were predicted in positive and reverse-complement strands, with a minimum length of 600 nt between two stop codons using the GetOrf program (EMBOSS) (Rice et al., 2000). Functional annotation was carried out using InterProScan v5.39-77.0 (Zdobnov & Apweiler, 2001), and the HMMer software (http://hmmer.org/) was used to perform sequence-profile searches against the Pfam HMM database. To expand the *de novo* assembled contigs of known viruses, the reads were mapped against reference genomic sequences. Provisional virus names were derived from geographic locations in the Perth Hills, Western Australia.

5.3.5 Estimates of viral abundance

Viral abundance was assessed using the number of reads per million (RPM). This metric quantifies the number of reads per million mapped to a given contig assembly over the total number of reads. RPM values lower than 0.1% of the highest count for each virus across samples were presumed to be index-hopping artifacts and excluded from the remaining analyses (Le Lay, Shi, et al., 2020). To compare abundance levels, reads were mapped to reference ribosomal and mitochondrial genes from *Wolbachia* (*16S* and *cox1*), *D. simulans* (*rpl32* and *cox1*), as well as against all the RNA viruses identified in the annotation analyses. Mapping was performed using BBMap v.37.98 (minid=0.95) (available at https://sourceforge.net/projects/bbmap/).

5.3.6 Sequence alignment and phylogenetic analysis

RNA viral sequences identified in D. simulans were compared with homologous reference sequences retrieved from the NCBI GenBank database and aligned with MAFF v7.450 (E-INS-I algorithm) (Katoh & Standley, 2013c). Phylogenetic trees on these data were then

inferred using sequences of the conserved RNA-dependent RNA polymerase (RdRp) gene. To this end, both the best-fit model of amino acid substitution and phylogenetic relationships were estimated using the Maximum Likelihood (ML) (Felsenstein, 1981) approach implemented in IQ-TREE v1.6.12 (Nguyen et al., 2015b). Nodal support was estimated combining the SH-like approximate likelihood ratio test (SH-aLRT) and the Ultrafast Bootstrap Approximation (Ufboot) (Hoang et al., 2018b). Redundant contigs with over 99% amino acid similarity were excluded. For those libraries containing viruses that were unlikely to be associated with Drosophila, taxonomic profiling and read mapping to components of fly microbiome/diet were conducted using the CCMetagen software (default settings) (Marcelino et al., 2020) and BBMap v.37.98 (minid=0.95).

5.3.7 Statistical analysis

The assumption of data normality was assessed by visual inspection and using Kolmogorov-Smirnov (K-S) and Shapiro-Wilk's tests. As the data was not normally distributed, a Mann-Whitney-Wilcoxon test was used to compare the RNA virome composition with respect to the presence/absence of *Wolbachia*. Comparisons were made using raw and log-transformed data corresponding to RPM values (i.e. viral abundance) for each library. All analyses were performed using R software package rstatix (available at https://rpkgs.datanovia.com/rstatix/).

5.4 Results

A total of 272 female flies were wild-caught in the Perth Hills, Western Australia and tested for *Wolbachia* infection through their F1s. The overall prevalence of *Wolbachia* was 63.6% (173/272), with frequencies at the three sampled locations varying from 54.8% (Raeburn Orchard, N = 73) to 63.8% (Roleystone, N = 130) and 72.5% (Cannington, N = 69). We randomly selected 16 flies from the Raeburn Orchard field females for individual

sequencing and RNA virus screening, representing eight *Wolbachia*-positive and eight *Wolbachia*-negative specimens.

We identified the *Wolbachia* strain in *D. simulans* using sequence-specific primers targeting the *wsp* gene. We further confirmed the occurrence of *Wolbachia* by mapping the reads back to the *w*Ri and *w*Au *wsp* genes. Most of the *Wolbachia*-infected flies showed a median coverage >100 reads, number of mapping reads >40, and coverage percentage >90% to the reference *w*Au strain, confirming that infected flies harbor *w*Au rather than wRi. No reads mapped to the *wsp* gene for library RAPP88 (**Table S5.1**) despite the positive infection status determined using a *Wolbachia* specific qPCR assay.

For comparison of virus diversity among libraries we mapped the reads of each library to stably expressed genes: *16S* and *cox1* in *Wolbachia* and *rpl32* and *cox1* in *D. simulans*. This provided an internal control to identify any effect on viral abundance due to potential biases introduced during RNA extraction or library preparation. Although, as expected, there was moderate variation in the abundance values, expression levels of reference maker genes were relatively stable across libraries in both *Wolbachia* and *D. simulans* (**Figure 5.1**).

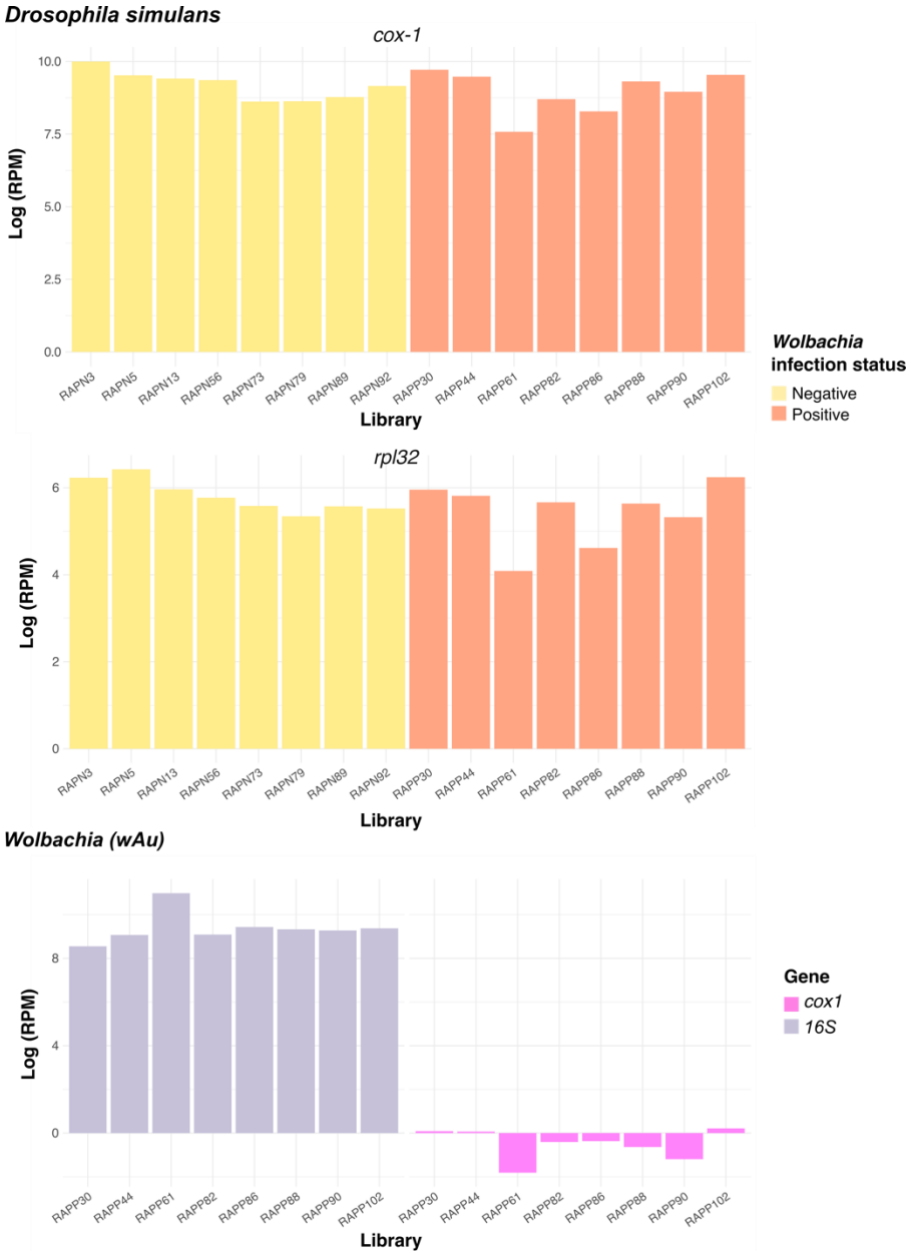


Figure 5.1 Comparison of the abundance levels of reference genes in *Wolbachia*-positive and *Wolbachia*-negative individual *D. simulans (rpl32 and cox-1)* and *Wolbachia* sp. (*16S* and *cox-1)*.

Overall, we detected nine viruses in the 16 individual *D. simulans* studied here, five of which were novel (**Figure 5.2**). Specifically, four viruses shared high sequence identity at the amino acid level (> 96%, e-value = 0.00E+00 - 4.2E-41) to the RdRp of known RNA viruses, whereas the newly discovered viruses shared only between 32.6% to 62.6% amino

acid identity to the best viral hit (e-value = 0.00E+00 - 1.4E-06) (**Table 5.1, Table S5.4**). Similarly, phylogenetic analysis of the known virus sequences identified revealed close relationships with known *Drosophila*-associated viruses: galbut virus (*Partitiviridae*), La Jolla virus (*Iflaviridae*), thika virus (*Picornaviridae*) and nora virus (*Picornaviridae*) (**Figure 5.3**). In addition, we identified contigs related to "chaq virus-like" sequences (>85% amino acid sequence similarity). The novel viruses identified, that did not share close phylogenetic relationships to known viruses, were: Raeburn bunya-like virus (*Bunyaviridae*), Araluen mito-like virus (*Mitoviridae*), Carmel mito-like virus (*Mitoviridae*), Lesley reo-like virus (*Reoviridae*), and Cannin tombus-like virus (*Tombusviridae*) (**Figure 5.3**). Similarity searches against the NCBI/nr database showed that individual flies carried multiple invertebrate-associated viruses from different virus families. For example, up to six viruses were observed in a single *w*Au-negative library (RAPN56) (**Figure 5.4**, **Table S5.2**).

Some of the newly discovered RNA viruses identified here were likely infecting hosts other than *D. simulans*, and hence might be associated with the fly diet or microbiome. Specifically, these viruses were closely related to Phytomonas sp. TCC231 leishbunyavirus 1 (in the case of Raeburn bunya-like virus), Leptomonas pyrrhocoris RNA virus (Cannin tombus-like virus) and two mito-like viruses (Araluen mito-like virus and Carmel mito-like virus) (**Figure 5.3, Table S5.3**), that are associated with trypanosomatid protozoans and fungal hosts, respectively. In addition, taxonomic composition analyses as well as read mapping to common components of *Drosophila* diet/microbiome revealed that 0.1% and 0.05% of all non-rRNA reads mapped to fungi and trypanosomatids (cox-1 gene marker), respectively. Hence, multiple microorganisms were present within individual fly libraries which may explain the occurrence of viruses not directly associated with *Drosophila* (files available at https://doi.org/10.6084/m9.figshare.c.5466690). In contrast, Lesley reo-like virus is likely a *bona fide* arthropod virus since it grouped with viruses previously detected in odonates and mosquitoes. In addition, it exhibited ~24% nucleotide similarity with those reoviruses previously reported to be contaminants in

Drosophila cell culture (Webster et al., 2015b). This suggests that Lesley reo-like virus is not a component of known contaminants and more is likely part of the natural *D. simulans* virome. The five newly identified viruses in this study corresponded to full or nearly complete genomes (see below). However, for the majority of the known *Drosophila* viruses we only were able to identify ORFs encoding the RdRp: the exceptions were La Jolla virus and thika virus for which we also predicted structural components corresponding to coat and capsid proteins.

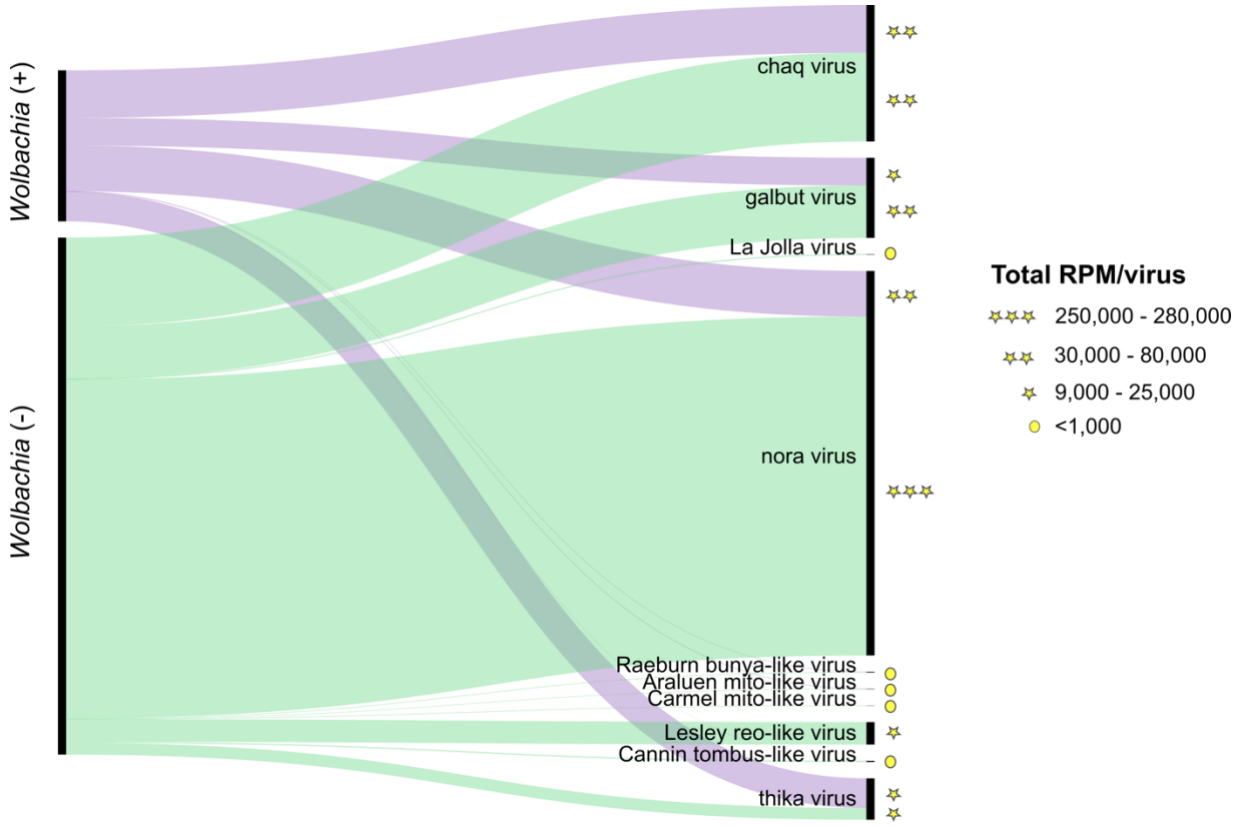
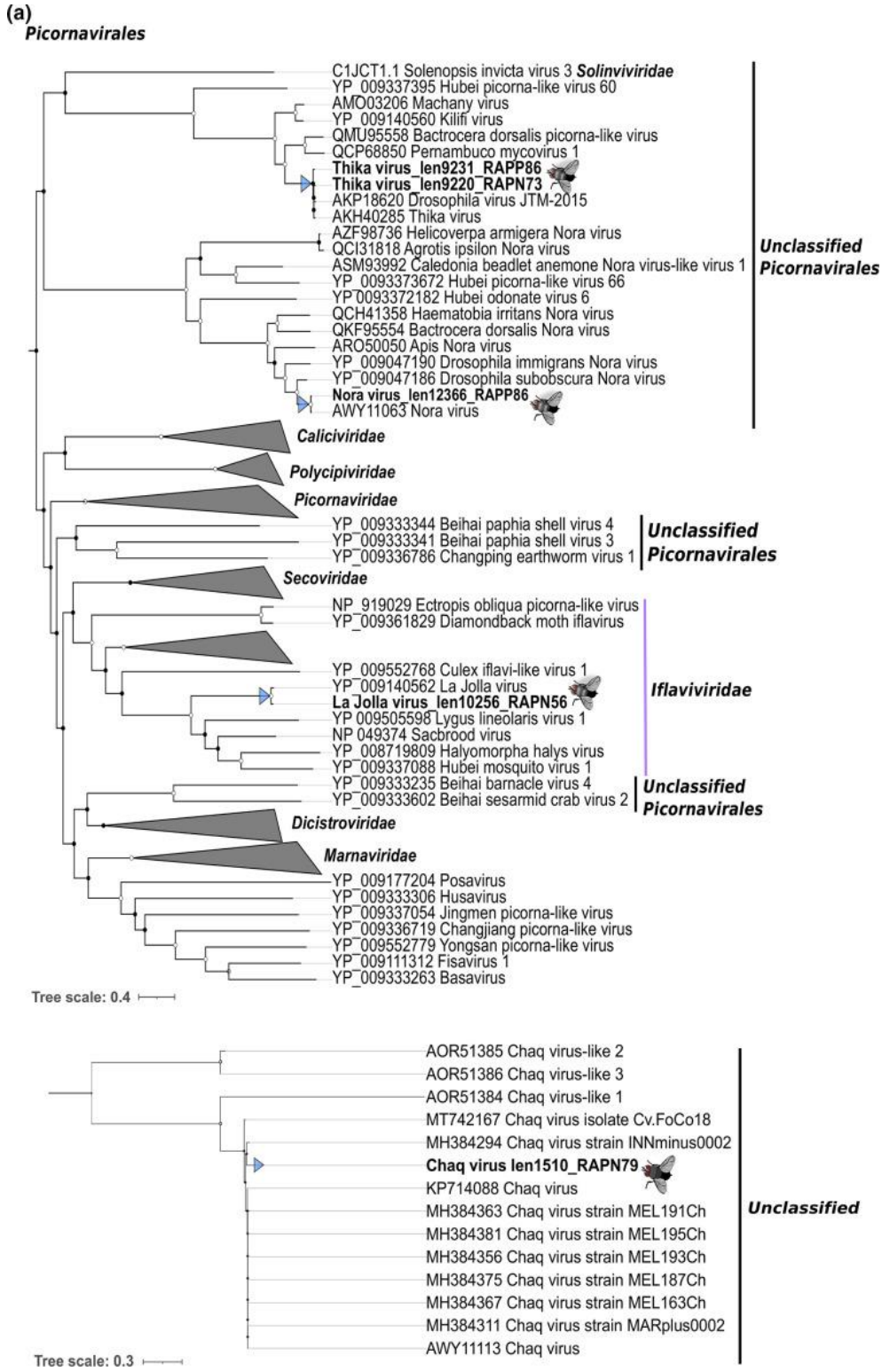
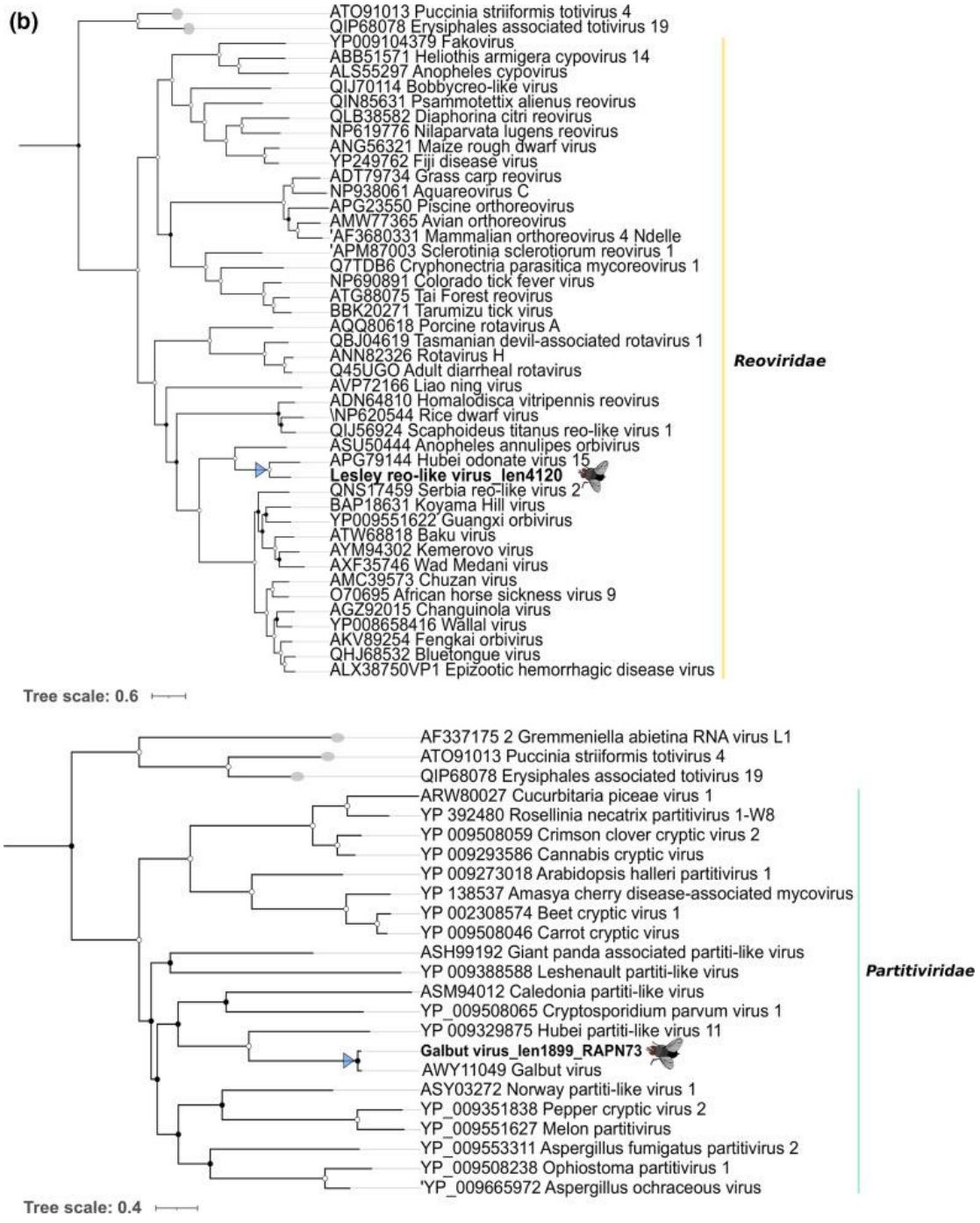


Figure 5.2 Comparison of viruses found in *Wolbachia*-positive and *Wolbachia*-negative *D. simulans*. The thickness of links is proportional to the total abundance (RPM) of each virus across the samples studied. The range of RPM values are represented with a star and circular shapes.

Table 5.1 Summary of sequence similarity searches for viruses against the NCBI non-redundant database. Viral sequences listed below correspond to those included in phylogenetic analyses.

Query sequence	Library	*Wolbachia* infection	Length (nt)	Best match against the BLAST/nr database	Similarity	e-value
k119_3301_len12366_nora virus	RAPP86	+	12366	AWY11063.1 putative replicase [Nora virus]	98.7	0.00E+00
k119_19486_len10256_La Jolla virus	RAPN56	-	10256	AWY11061.1 putative polyprotein [La Jolla virus]	98	0.00E+00
k119_20553_len9231_thika virus	RAPP86	+	9231	YP_009140561.1 putative polyprotein [Thika virus]	96.2	0.00E+00
k119_5914_len9220_thika virus	RAPN73	-	9220	YP_009140561.1 putative polyprotein [Thika virus]	97.1	0.00E+00
k119_3227_len6958_Cannin tombus-like virus	RAPN56	-	6958	ASN64756.1 putative RNA-dependent RNA polymerase, partial [Leptomonas pyrrhocoris RNA virus]	44.6	1.80E-96
k119_2329_len2049_Cannin tombus-like virus	RAPP88	+	2049	ASN64759.1 putative RNA-dependent RNA polymerase, partial [Leptomonas pyrrhocoris RNA virus]	48.4	3.80E-95
k119_4103_len1899_galbut virus	RAPN73	-	1899	AWY11176.1 putative RNA-dependent RNA polymerase [Galbut virus]	96.7	0.00E+00
k119_13353_len1510_chaq virus	RAPN79	-	1510	AWY11113.1 hypothetical protein [Chaq virus]	85.9	1.6E-153
k119_2075_len4120_ Lesley reo-like virus	RAPN73	-	4120	APG79144.1 RNA-dependent RNA polymerase [Hubei odonate virus 15]	48.6	0.00E+00
k119_10165_len2547_Carmel mito-like virus	RAPN79	-	2547	YP_009329842.1 RNA-dependent RNA polymerase [Hubei narna-like virus 24]	32.7	2.0e-76
k119_273_len2671_Araluen mito-like virus	RAPN5	-	2671	QDH87474.1 RNA-dependent RNA polymerase, partial [Mitovirus sp.]	40.3	8.0E-96
k119_22084_len2612_Araluen mito-like virus	RAPN5	-	2612	QDH87474.1 RNA-dependent RNA polymerase, partial [Mitovirus sp.]	43.2	2.3E-103
k119_14037_len2615_Araluen mito-like virus	RAPN56	-	2615	QDH87474.1 RNA-dependent RNA polymerase, partial [Mitovirus sp.]	41.7	1.7E-98
k119_14318_len2822_Araluen mito-like virus	RAPN56	-	2822	QDH87474.1 RNA-dependent RNA polymerase, partial [Mitovirus sp.]	38.1	9.7E-92


(a)
Picornavirales
C1JCT1.1 Solenopsis invicta virus 3 Solinviviridae
YP_009337395 Hubei picorna-like virus 60
AMO03206 Machany virus
YP_009140560 Kilifi virus
QMU95558 Bactrocera dorsalis picorna-like virus
QCP68850 Pernambuco mycovirus 1
Thika virus_len9231_RAPP86
Thika virus_len9220_RAPN73
AKP18620 Drosophila virus JTM-2015
AKH40285 Thika virus
AZF98736 Helicoverpa armigera Nora virus
QCI31818 Agrotis ipsilon Nora virus
ASM93992 Caledonia beadlet anemone Nora virus-like virus 1
YP_0093373672 Hubei picorna-like virus 66
YP_0093372182 Hubei odonate virus 6
QCH41358 Haematobia irritans Nora virus
QKF95554 Bactrocera dorsalis Nora virus
ARO50050 Apis Nora virus
YP_009047190 Drosophila immigrans Nora virus
YP_009047186 Drosophila subobscura Nora virus
Nora virus_len12366_RAPP86
AWY11063 Nora virus
Unclassified Picornavirales
Caliciviridae
Polycipiviridae
Picornaviridae
YP_009333344 Beihai paphia shell virus 4
YP_009333341 Beihai paphia shell virus 3
YP_009336786 Changping earthworm virus 1
Unclassified Picornavirales
Secoviridae
NP_919029 Ectropis obliqua picorna-like virus
YP_009361829 Diamondback moth iflavirus
YP_009552768 Culex iflavi-like virus 1
YP_009140562 La Jolla virus
La Jolla virus_len10256_RAPN56
YP 009505598 Lygus lineolaris virus 1
NP 049374 Sacbrood virus
YP_008719809 Halyomorpha halys virus
YP_009337088 Hubei mosquito virus 1
Iflaviviridae
YP_009333235 Beihai barnacle virus 4
YP_009333602 Beihai sesarmid crab virus 2
Unclassified Picornavirales
Dicistroviridae
Marnaviridae
YP_009177204 Posavirus
YP_009333306 Husavirus
YP_009337054 Jingmen picorna-like virus
YP_009336719 Changjiang picorna-like virus
YP_009552779 Yongsan picorna-like virus
YP_009111312 Fisavirus 1
YP_009333263 Basavirus
Tree scale: 0.4
AOR51385 Chaq virus-like 2
AOR51386 Chaq virus-like 3
AOR51384 Chaq virus-like 1
MT742167 Chaq virus isolate Cv.FoCo18
MH384294 Chaq virus strain INNminus0002
Chaq virus len1510_RAPN79
KP714088 Chaq virus
MH384363 Chaq virus strain MEL191Ch
MH384381 Chaq virus strain MEL195Ch
MH384356 Chaq virus strain MEL193Ch
MH384375 Chaq virus strain MEL187Ch
MH384367 Chaq virus strain MEL163Ch
MH384311 Chaq virus strain MARplus0002
AWY11113 Chaq virus
Unclassified
Tree scale: 0.3


(b)
ATO91013 Puccinia striiformis totivirus 4
QIP68078 Erysiphales associated totivirus 19
YP009104379 Fakovirus
ABB51571 Heliothis armigera cypovirus 14
ALS55297 Anopheles cypovirus
QIJ70114 Bobbycreo-like virus
QIN85631 Psammotettix alienus reovirus
QLB38582 Diaphorina citri reovirus
NP619776 Nilaparvata lugens reovirus
ANG56321 Maize rough dwarf virus
YP249762 Fiji disease virus
ADT79734 Grass carp reovirus
NP938061 Aquareovirus C
APG23550 Piscine orthoreovirus
AMW77365 Avian orthoreovirus
'AF3680331 Mammalian orthoreovirus 4 Ndelle
'APM87003 Sclerotinia sclerotiorum reovirus 1
Q7TDB6 Cryphonectria parasitica mycoreovirus 1
NP690891 Colorado tick fever virus
ATG88075 Tai Forest reovirus
BBK20271 Tarumizu tick virus
AQQ80618 Porcine rotavirus A
QBJ04619 Tasmanian devil-associated rotavirus 1
ANN82326 Rotavirus H
Q45UGO Adult diarrheal rotavirus
AVP72166 Liao ning virus
ADN64810 Homalodisca vitripennis reovirus
\NP620544 Rice dwarf virus
QIJ56924 Scaphoideus titanus reo-like virus 1
ASU50444 Anopheles annulipes orbivirus
APG79144 Hubei odonate virus 15
Lesley reo-like virus_len4120
QNS17459 Serbia reo-like virus 2
BAP18631 Koyama Hill virus
YP009551622 Guangxi orbivirus
ATW68818 Baku virus
AYM94302 Kemerovo virus
AXF35746 Wad Medani virus
AMC39573 Chuzan virus
O70695 African horse sickness virus 9
AGZ92015 Changuinola virus
YP008658416 Wallal virus
AKV89254 Fengkai orbivirus
QHJ68532 Bluetongue virus
ALX38750VP1 Epizootic hemorrhagic disease virus
Reoviridae
Tree scale: 0.6
AF337175 2 Gremmeniella abietina RNA virus L1
ATO91013 Puccinia striiformis totivirus 4
QIP68078 Erysiphales associated totivirus 19
ARW80027 Cucurbitaria piceae virus 1
YP 392480 Rosellinia necatrix partitivirus 1-W8
YP 009508059 Crimson clover cryptic virus 2
YP 009293586 Cannabis cryptic virus
YP 009273018 Arabidopsis halleri partitivirus 1
YP 138537 Amasya cherry disease-associated mycovirus
YP 002308574 Beet cryptic virus 1
YP 009508046 Carrot cryptic virus
ASH99192 Giant panda associated partiti-like virus
YP 009388588 Leshenault partiti-like virus
ASM94012 Caledonia partiti-like virus
YP_009508065 Cryptosporidium parvum virus 1
YP 009329875 Hubei partiti-like virus 11
Galbut virus_len1899_RAPN73
AWY11049 Galbut virus
ASY03272 Norway partiti-like virus 1
YP_009351838 Pepper cryptic virus 2
YP_009551627 Melon partitivirus
YP_009553311 Aspergillus fumigatus partitivirus 2
YP_009508238 Ophiostoma partitivirus 1
'YP_009665972 Aspergillus ochraceous virus
Partitiviridae
Tree scale: 0.4

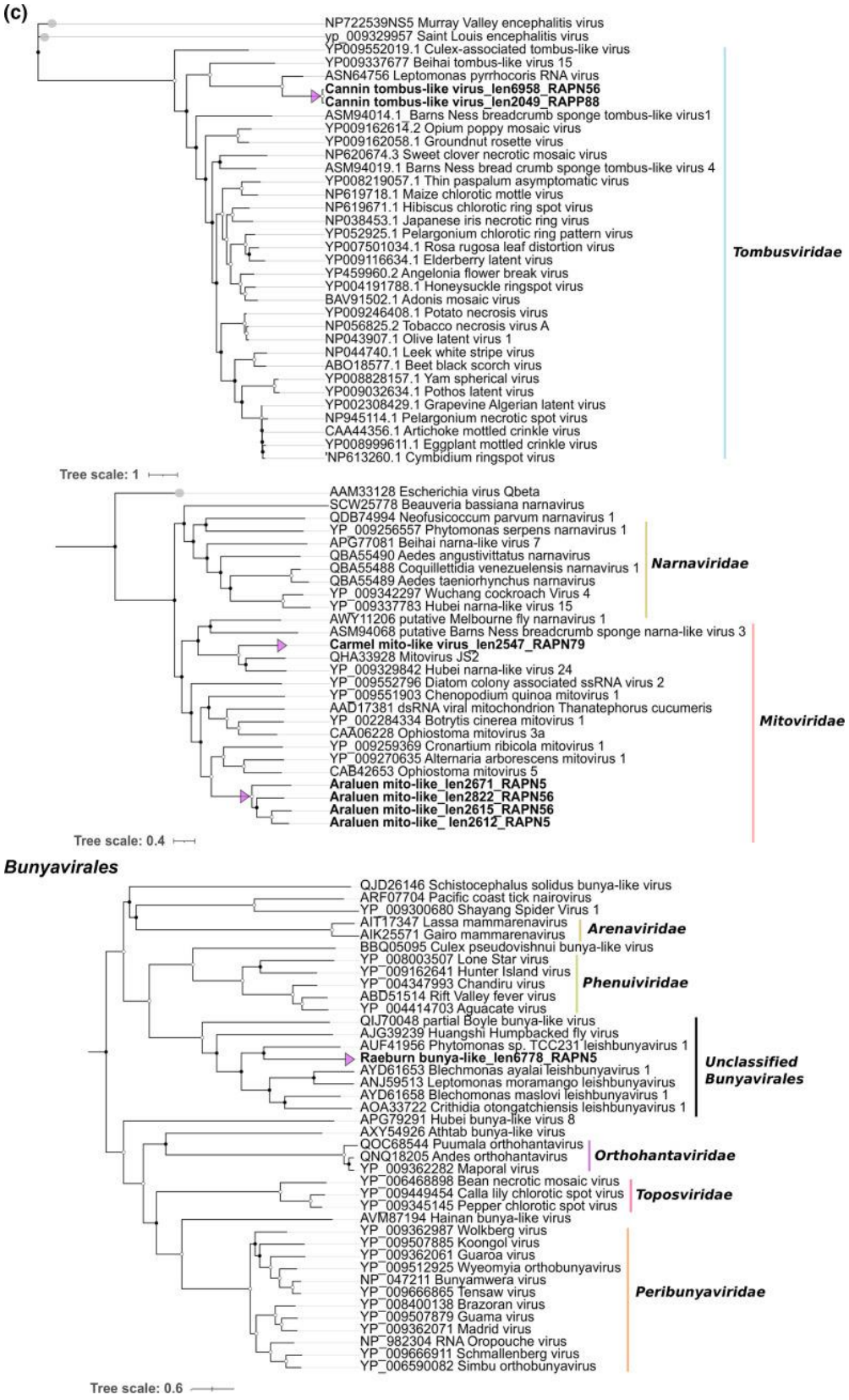

(c)
NP722539NS5 Murray Valley encephalitis virus
yp_009329957 Saint Louis encephalitis virus
YP009552019.1 Culex-associated tombus-like virus
YP009337677 Beihai tombus-like virus 15
ASN64756 Leptomonas pyrrhocoris RNA virus
Cannin tombus-like virus_len6958_RAPN56
Cannin tombus-like virus_len2049_RAPP88
ASM94014.1_Barns Ness breadcrumb sponge tombus-like virus1
YP009162614.2 Opium poppy mosaic virus
YP009162058.1 Groundnut rosette virus
NP620674.3 Sweet clover necrotic mosaic virus
ASM94019.1 Barns Ness bread crumb sponge tombus-like virus 4
YP008219057.1 Thin paspalum asymptomatic virus
NP619718.1 Maize chlorotic mottle virus
NP619671.1 Hibiscus chlorotic ring spot virus
NP038453.1 Japanese iris necrotic ring virus
YP052925.1 Pelargonium chlorotic ring pattern virus
YP007501034.1 Rosa rugosa leaf distortion virus
YP009116634.1 Elderberry latent virus
YP459960.2 Angelonia flower break virus
YP004191788.1 Honeysuckle ringspot virus
BAV91502.1 Adonis mosaic virus
YP009246408.1 Potato necrosis virus
NP056825.2 Tobacco necrosis virus A
NP043907.1 Olive latent virus 1
NP044740.1 Leek white stripe virus
ABO18577.1 Beet black scorch virus
YP008828157.1 Yam spherical virus
YP009032634.1 Pothos latent virus
YP002308429.1 Grapevine Algerian latent virus
NP945114.1 Pelargonium necrotic spot virus
CAA44356.1 Artichoke mottled crinkle virus
YP008999611.1 Eggplant mottled crinkle virus
'NP613260.1 Cymbidium ringspot virus
Tombusviridae
Tree scale: 1
AAM33128 Escherichia virus Qbeta
SCW25778 Beauveria bassiana narnavirus
QDB74994 Neofusicoccum parvum narnavirus 1
YP_009256557 Phytomonas serpens narnavirus 1
APG77081 Beihai narna-like virus 7
QBA55490 Aedes angustivittatus narnavirus
QBA55488 Coquillettidia venezuelensis narnavirus 1
QBA55489 Aedes taeniorhynchus narnavirus
YP_009342297 Wuchang cockroach Virus 4
YP_009337783 Hubei narna-like virus 15
Narnaviridae
AWY11206 putative Melbourne fly narnavirus 1
ASM94068 putative Barns Ness breadcrumb sponge narna-like virus 3
Carmel mito-like virus_len2547_RAPN79
QHA33928 Mitovirus JS2
YP_009329842 Hubei narna-like virus 24
YP_009552796 Diatom colony associated ssRNA virus 2
YP_009551903 Chenopodium quinoa mitovirus 1
AAD17381 dsRNA viral mitochondrion Thanatephorus cucumeris
YP_002284334 Botrytis cinerea mitovirus 1
CAA06228 Ophiostoma mitovirus 3a
YP_009259369 Cronartium ribicola mitovirus 1
YP_009270635 Alternaria arborescens mitovirus 1
CAB42653 Ophiostoma mitovirus 5
Araluen mito-like_len2671_RAPN5
Araluen mito-like_len2822_RAPN56
Araluen mito-like_len2615_RAPN56
Araluen mito-like_ len2612_RAPN5
Mitoviridae
Tree scale: 0.4
Bunyavirales
QJD26146 Schistocephalus solidus bunya-like virus
ARF07704 Pacific coast tick nairovirus
YP_009300680 Shayang Spider Virus 1
AIT17347 Lassa mammarenavirus
AIK25571 Gairo mammarenavirus
Arenaviridae
BBQ05095 Culex pseudovishnui bunya-like virus
YP_008003507 Lone Star virus
YP_009162641 Hunter Island virus
YP_004347993 Chandiru virus
ABD51514 Rift Valley fever virus
YP_004414703 Aguacate virus
Phenuiviridae
QIJ70048 partial Boyle bunya-like virus
AJG39239 Huangshi Humpbacked fly virus
AUF41956 Phytomonas sp. TCC231 leishbunyavirus 1
Raeburn bunya-like_len6778_RAPN5
AYD61653 Blechmonas ayalai leishbunyavirus 1
ANJ59513 Leptomonas moramango leishbunyavirus
AYD61658 Blechomonas maslovi leishbunyavirus 1
AOA33722 Crithidia otongatchiensis leishbunyavirus 1
Unclassified Bunyavirales
APG79291 Hubei bunya-like virus 8
AXY54926 Athtab bunya-like virus
QOC68544 Puumala orthohantavirus
QNQ18205 Andes orthohantavirus
YP_009362282 Maporal virus
Orthohantaviridae
YP_006468898 Bean necrotic mosaic virus
YP_009449454 Calla lily chlorotic spot virus
YP_009345145 Pepper chlorotic spot virus
Toposviridae
AVM87194 Hainan bunya-like virus
YP_009362987 Wolkberg virus
YP_009507885 Koongol virus
YP_009362061 Guaroa virus
YP_009512925 Wyeomyia orthobunyavirus
NP_047211 Bunyamwera virus
YP_009666865 Tensaw virus
YP_008400138 Brazoran virus
YP_009507879 Guama virus
YP_009362071 Madrid virus
NP_982304 RNA Oropouche virus
YP_009666911 Schmallenberg virus
YP_006590082 Simbu orthobunyavirus
Peribunyaviridae
Tree scale: 0.6

Figure 5.3 Maximum likelihood phylogenetic trees of the viruses and virus-like sequences identified from *D. simulans*. The phylogenies were inferred based on the amino acid sequences of the RdRp of six virus taxonomic groups, whereas for chaq virus-like sequences we used a protein of unknown function. Virus family trees were rooted with relevant outgroups that are indicated with grey tips. Order-level trees and the chaq virus phylogeny (for which no suitable outgroup existed) were midpoint rooted. Coloured arrow tips represent likely (A-B) *Drosophila*-associated viruses and (C) non-*Drosophila*-associated viruses (i.e. that were more likely associated with a component of fly diet or microbiome). Nodal support values greater than 80% (SH-aLRT) and 95% (UFboot) are indicated with white circular shapes at the nodes. Branch lengths are projected using scale bars below each tree.

We next characterized the virome profile present in *D. simulans* in relation to the *w*Au infection status (**Figure 5.2, Table 5.1, Table S5.4**). Accordingly, we identified a slightly higher number (n=9) of viruses in *Wolbachia*-negative flies compared to *Wolbachia*-positive flies (n=6). Among these, galbut virus, nora virus, thika virus, as well as three novel viruses identified in this study - Raeburn bunya-like virus, Araluen mito-like virus and Cannin tombus-like virus - were present in *D. simulans* regardless of *Wolbachia* infection. Likewise, “chaq virus-like” sequences were observed co-occurring with galbut virus in the two groups of *D. simulans*. In contrast, La Jolla virus, as well as the novel Carmel mito-like virus and Lesley reo-like virus, were only found in *w*Au-negative flies. Overall, assembled viral contigs displayed high sequence similarity at nucleotide and amino acid level within and between libraries and regardless of the presence/absence of *Wolbachia* (**Table S3.4**).

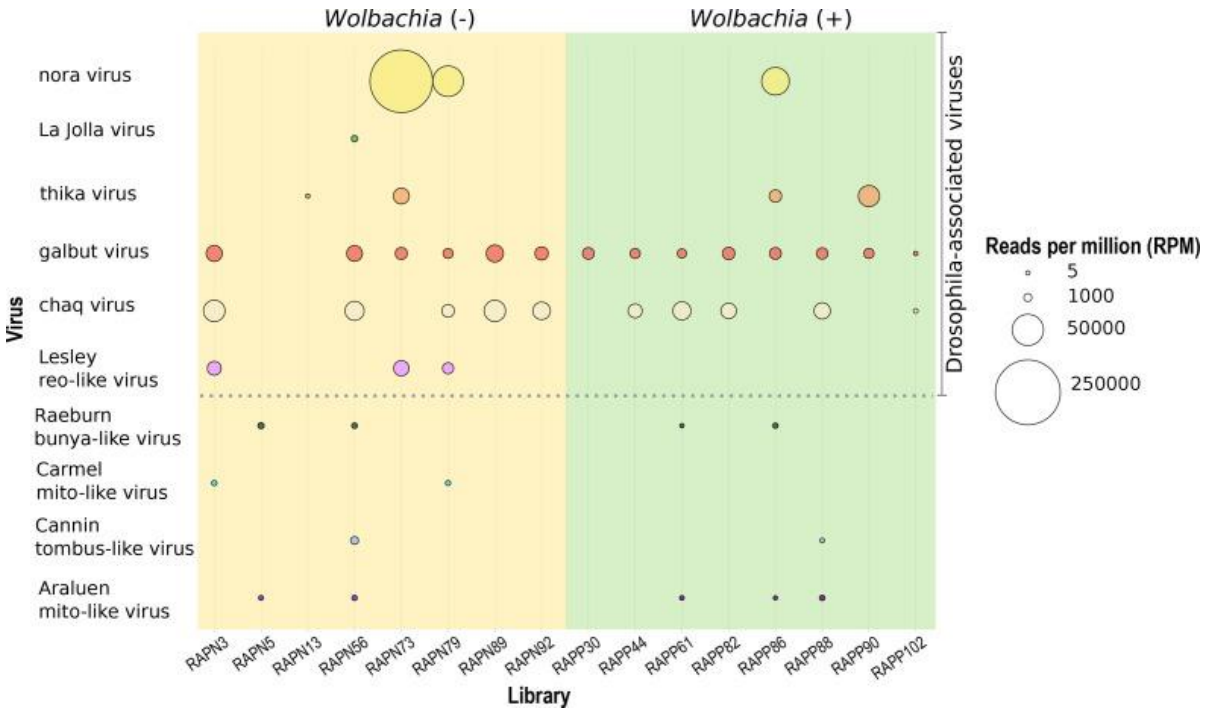


Figure 5.4 Representation of virome composition and abundance (RPM) across *Wolbachia*-positive and negative libraries. Each library represents an individual *D. simulans* fly. All reads likely due to index-hopping have been excluded.

We also assessed the potential effect of *Wolbachia* infection on the abundance of RNA viruses present in *w*Au-infected and *w*Au-uninfected flies. Overall, the number of non-rRNA reads represented ~50% of the total of reads (n= 743,389,696 pair-end reads) (**Figure S5.1**). Furthermore, the RPM values among viruses infecting *Wolbachia* negative and positive infected flies was highly heterogeneous, ranging from 47 to 232,346 and 7 to 37,688 virus RPM, respectively. With the exception of thika virus, viruses present in both *w*Au-positive and *w*Au-negative flies were 1.87 – 40.17-fold more abundant in the *w*Au-negative individuals than *w*Au-positive *D. simulans.* In contrast, the abundance of Thika virus was 0.39-fold higher in the *Wolbachia*-positive flies (**Figure 5.3, Table S5.2**). However, despite this variation in virus abundance levels between groups, there was a non-significant difference between *w*Au-negative and *w*Au-positive *D. simulans* (Mann-Whitney-Wilcoxon test; **Figure 5.5**). In the case

of the viruses only detected in the *w*Au-negative flies, La Jolla virus was present in a single library in moderate abundance (RPM = 378), whilst the newly discovered Lesley reo-like virus was detected in 4/8 libraries (RPM = 3360 - 8749) (**Table S5.2**). Although an interesting result, the limited sample size (n = 16) means that these observations should be taken with caution and that larger sample sizes are needed for corroboration.

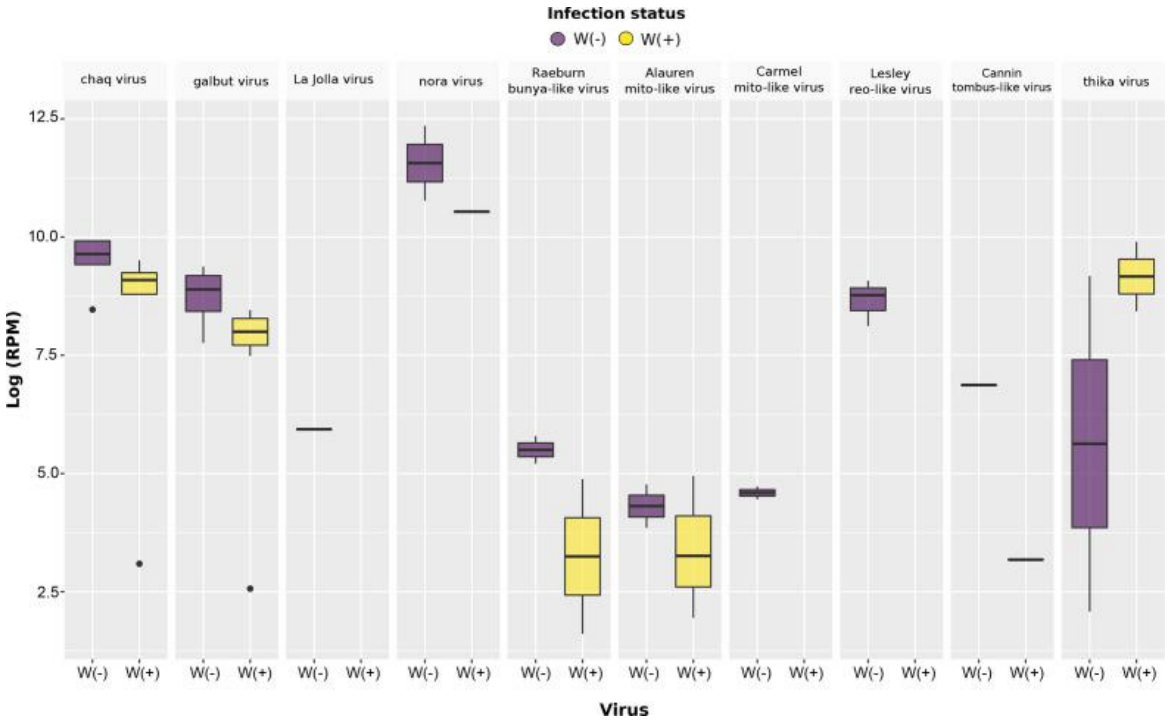


Figure 5.5 Abundance distribution of six RNA viruses and the chaq virus-like sequences identified across individual *Wolbachia*-positive and *Wolbachia*-negative *D. simulans*. A non-significant difference was observed between *Wolbachia*-infected and uninfected flies using the Mann-Whitney U test.

5.5 Discussion

The occurrence and spread of *Wolbachia* infection has been widely documented in natural populations of *Drosophila* (Kriesner et al., 2013; Turelli et al., 2018; Turelli & Hoffmann, 1991). Indeed, *D. simulans* is commonly used as an experimental model to investigate the interactions

within the tripartite *Drosophila*-*Wolbachia*-virus system. In Australia, *D. simulans* can be naturally infected with two *Wolbachia* strains from supergroup A - *w*Au and *w*Ri. While *w*Ri has been gradually displacing *w*Au in eastern Australia, reflected in the changing infection frequencies in surveyed populations since 2004, *D. simulans* from the west coast of Australia only harbor *w*Au (Kriesner et al., 2013). A simple and plausible explanation for this difference is the geographic separation of *D. simulans* populations inhabiting the east and west coasts of Australia and the challenging environmental conditions posed by the intervening desert (Kriesner et al., 2013).

We corroborated the presence of *Wolbachia* infection across samples by identifying the *wsp*, *16S* and *cox1* marker genes. The lack of reads mapping to the library RAPP88 might reflect either low levels of *wsp* RNA molecules present in the input for library preparation or high variability compared to the reference sequence. Although *Wolbachia* density was not experimentally assessed, the similar levels of *16S* and *cox-1* abundance across libraries suggest no appreciable biases in the library preparation and RNA sequencing steps.

Estimates from previous surveys showed that the frequency of the *w*Au strain in Western Australia exceeded 50% in *D. simulans* (Kriesner et al., 2013). This is consistent with the data provided here and suggests that *Wolbachia* might be present in a significant proportion of the natural fly population, at least around Perth. Although *w*Au does not cause cytoplasmic incompatibility, its spread is hypothesized to confer fitness advantages (increased survival and/or reproduction) to the host, including antiviral protection (Mancini et al., 2020; Ogunlade et al., 2020), that might favour its spread and prevent the bacteria from being eliminated from *D. simulans* populations (Correa & Ballard, 2016; Kriesner et al., 2013). However, our comparison of *Wolbachia*-infected and uninfected *D. simulans* in western Australia revealed no clear effect of *Wolbachia* infection on virome composition and viral abundance between *Wolbachia* infected/uninfected animals. Although our analysis is based on a small sample of individual flies, the apparent absence of a *Wolbachia*-mediated virus protection effect in natural *D. simulans* is compatible with previous findings on *D. melanogaster*

naturally infected with *w*Mel in eastern Australia (Shi et al., 2018), in which virus protection was not observed regardless of the *Wolbachia* infection status and *Wolbachia* density. Even so, the absence of a significant association between *w*Au infection and virus diversity does not necessarily translate into a homogeneous effect of *w*Au on the different viruses identified here. For example, it is plausible that the restricted presence of La Jolla virus and the newly identified Lesley reo-like virus in *Wolbachia*-free flies could reflect some impact of antiviral protection in *D. simulans* (Habayeb et al., 2009; Shi et al., 2018). Indeed, contrasting results were observed in *D. melanogaster*, where La Jolla virus was widely distributed across different libraries (Shi et al., 2018). Although this might provide insights into *w*Au-virus interactions, studies based on larger sample sizes are clearly needed to determine whether the apparent association between La Jolla virus and *Wolbachia*-uninfected flies observed here is an artefact due to small sample sizes. Indeed, it is notable that La Jolla virus was so rarely detected in the *D. simulans* flies studied here.

It has previously been shown that the *w*Au strain of *Wolbachia* has a protective role against virus infection in *D. simulans* when flies are challenged with FHV and DCV in a laboratory setting (Martinez et al., 2014; Osborne et al., 2009b). Moreover, the *w*Au strain is protective against the dengue (DENV) and zika (ZIKV) viruses in *Aedes aegypti* mosquitoes (Ant et al., 2018). Although our observation of an apparent lack of *Wolbachia*-mediated antiviral protection contrasts with those obtained previously, it is likely that differences may depend on *Wolbachia*-host species combinations and natural/artificial viral infections, which may also explain the contrasting results for La Jolla virus. Indeed, most of the available studies have documented the antiviral effect in transinfected insect hosts with non-natural *Wolbachia* strains/viruses under laboratory conditions, as opposed to the study of the natural virome undertaken here.

It is noteworthy that ecological variables such as temperature might impact *Wolbachia*-virus-host interactions. Here, we collected flies during the Western Australian winter (mean temperature of 21°C daylight time) and the specimens were maintained at 19°C under

laboratory conditions. Lower temperatures have been associated with an increase in viral resistance against DCV in D. melanogaster infected with *w*Mel and *w*MelCS (Chrostek et al., 2020). Similarly, variations in host developmental temperature have been associated with differences in *Wolbachia*-mediated virus blocking in natural populations (Chrostek et al., 2020). In this context, flies developed at lower temperature (18°C) exhibited a reduction in *Wolbachia*-conferred antiviral protection. On the other hand, the presence of *Wolbachia* have been hypothesized to influence host temperature preferences. For instance, *w*Ri and *w*Ho strains seem to manipulate *D. simulans* flies to seek cooler temperatures (Hague et al., 2020). Although the effect of temperature on *w*Au and *D. simulans* need to be tested, this suggests that the results observed here as well as a protective scenario might be temperature-dependent. This highlights the importance of careful future studies of the interactions within the host-virus-*Wolbachia* system along with environmental factors in natural populations (Cao et al., 2019; Fenton et al., 2011; Johnson, 2015).

As well as the small sample size, an important caveat of our work is that we explored the *Wolbachia*-mediated virus protection in terms of virus abundance levels reflected in RPM values. This provides insights into virus resistance, but not on tolerance or host survival. Thus, it is still possible that *Wolbachia* is increasing tolerance to virus infection as have been documented for DCV (Osborne et al., 2009b). In addition, although we were not able to assess *Wolbachia* density, previous studies have shown that *w*Au is maintained at high-density in *D. simulans* and has a role on virus blocking (Osborne et al., 2012). Further research is clearly needed to assess these features in natural populations to determine any link with antiviral protection.

Collectively, comparisons of the virome composition in *w*Au infected/uninfected *D. simulans* showed the presence of natural and relatively highly abundant *Drosophila* associated viruses in both groups (Palmer et al., 2018; Shi et al., 2018a; Webster et al., 2015). Consistent with previous studies we noted the co-occurrence of chaq virus-like sequences and galbut virus, supporting the idea that chaq virus might be part of a satellite-helper virus system or an

additional segment associated with galbut virus (Cross et al., 2020a; Shi et al., 2018). In addition to insect-associated viruses, we identified viruses that are likely to infect other hosts and hence were likely associated with components of *D. simulans* diet or microbiome (Ebbert et al., 2003). For instance, novel viruses from the families *Tombusviridae* and *Bunyaviridae* were related to virus in trypanosomatid protozoa (*Leptomonas* and *Leishmania*). Similarly, given their normal host range distribution, the novel viruses from the family *Narnaviridae* might be associated with fungal hosts. Evidence of trypanosomatids and fungi have been reported in the gut of several species of *Drosophila,* with effects on larvae eclosion and pupation times (Ebbert et al., 2003; Wilfert et al., 2011). This, in turn, highlights the extent to which Australian *D. simulans* can be parasitized in nature (Chandler & James, 2013; Ebbert et al., 2001; Lemaitre et al., 1996; Naranjo-Lázaro et al., 2014; Wilfert et al., 2011).

In sum, we provide a preliminary framework for assessing the effect of *w*Au strain on the virome of *D. simulans*, using a meta-transcriptomic analysis of individual *w*Au-infected and uninfected flies. In doing so we identified *Drosophila*-associated viruses along with five novel viruses likely associated with fly diet or microbiome. Although our sample size is small, we saw no detectable *Wolbachia*-associated antiviral effect on virus composition and abundance, although the approach taken prevented us from drawing conclusions on virus tolerance. Further research employing larger sample sizes over broad spatial scales, including different *Wolbachia-Drosophila* combinations, will enable a more nuanced understanding of *Wolbachi*a-virus dynamics in wild *Drosophila* populations.

5.7 Supplementary Material

One supplementary figure and four supplementary tables are available with the online version of this article.

Author Contributions: Conceptualization, M.S., A.A.H. and E.C.H.; methodology, A.S.O.-B., M.S., A.A.H. and E.C.H., formal analysis, A.S.O.-B.; investigation, A.S.O.-B. and M.S.; resources, A.A.H.

and E.C.H.; writing—original draft preparation A.S.O.-B.; writing—review and editing A.A.H., M.S and E.C.H.; visualization, A.S.O.-B.; supervision, E.C.H.; All authors have read and agreed to the published version of the manuscript.

The viral genome sequence data generated in this study have been deposited in the NCBI/GenBank database under the accession numbers MW976812–MW976882. Sequence reads are available at the public Sequence Read Archive (SRA) database under the BioProject accession PRJNA706433 (BioSample accessions: SAMN18132282–SAMN18132297).

5.6 References

Ant, T. H., Herd, C. S., Geoghegan, V., Hoffmann, A. A., & Sinkins, S. P. (2018). The *Wolbachia* strain wAu provides highly efficient virus transmission blocking in *Aedes aegypti*. *PLOS Pathogens*, *14*(1), e1006815. https://doi.org/10.1371/journal.ppat.1006815

Bolger, A. M., Lohse, M., & Usadel, B. (2014). Trimmomatic: a flexible trimmer for Illumina sequence data. *Bioinformatics*, *30*(15), 2114–2120. https://doi.org/10.1093/bioinformatics/btu170

Buchfink, B., Xie, C., & Huson, D. H. (2015). Fast and sensitive protein alignment using DIAMOND. *Nature Methods*, *12*(1), 59–60. https://doi.org/10.1038/nmeth.3176

Cao, L.-J., Jiang, W., & Hoffmann, A. A. (2019). Life history effects linked to an advantage for wAu *Wolbachia* in *Drosophila*. *Insects*, *10*(5), 126. https://doi.org/10.3390/insects10050126

Casiraghi, M., Bordenstein, S. R., Baldo, L., Lo, N., Beninati, T., Wernegreen, J. J., Werren, J. H., & Bandi, C. (2005). Phylogeny of *Wolbachia pipientis* based on *gltA*, *groEL* and *ftsZ* gene sequences: Clustering of arthropod and nematode symbionts in the F supergroup, and evidence for further diversity in the *Wolbachia* tree. *Microbiology*, *151*(12), 4015–4022. https://doi.org/10.1099/mic.0.28313-0

Chandler, J. A., & James, P. M. (2013). Discovery of trypanosomatid parasites in globally distributed *Drosophila* species. *PLoS ONE*, *8*(4), e61937. https://doi.org/10.1371/journal.pone.0061937

Chrostek, E., Martins, N., Marialva, M. S., & Teixeira, L. (2020). Wolbachia-conferred antiviral protection is determined by developmental temperature. *BioRxiv*, 2020.06.24.169169. https://doi.org/10.1101/2020.06.24.169169

Correa, C. C., & Ballard, J. W. O. (2016). Wolbachia associations with insects: Winning or losing against a master manipulator. *Frontiers in Ecology and Evolution*, *3*(JAN), 153. https://doi.org/10.3389/fevo.2015.00153

Cross, S. T., Maertens, B. L., Dunham, T. J., Rodgers, C. P., Brehm, A. L., Miller, M. R., Williams, A. M., Foy, B. D., & Stenglein, M. D. (2020). Partitiviruses infecting *Drosophila melanogaster* and *Aedes aegypti* exhibit efficient biparental vertical transmission. *Journal of Virology*, *94*(20). https://doi.org/10.1128/jvi.01070-20

Dedeine, F., Vavre, F., Fleury, F., Loppin, B., Hochberg, M. E., & Boulétreau, M. (2001). Removing symbiotic *Wolbachia* bacteria specifically inhibits oogenesis in a parasitic wasp. *Proceedings of the National Academy of Sciences of the United States of America*, *98*(11), 6247–6252. https://doi.org/10.1073/pnas.101304298

Dobson, S. L., Bourtzis, K., Braig, H. R., Jones, B. F., Zhou, W., Rousset, F., & O'Neill, S. L. (1999). *Wolbachia* infections are distributed throughout insect somatic and germ line tissues. *Insect Biochemistry and Molecular Biology*, *29*(2), 153–160. https://doi.org/10.1016/S0965-1748(98)00119-2

Ebbert, M. A., Burkholder, J. J., & Marlowe, J. L. (2001). Trypanosomatid prevalence and host habitat choice in woodland *Drosophila*. *Journal of Invertebrate Pathology*, *77*(1), 27–32. https://doi.org/10.1006/jipa.2000.4989

Ebbert, M. A., Marlowe, J. L., & Burkholder, J. J. (2003). Protozoan and intracellular fungal gut endosymbionts in *Drosophila*: prevalence and fitness effects of single and dual infections. *Journal of Invertebrate Pathology*, *83*(1), 37–45. https://doi.org/10.1016/S0022-2011(03)00033-8

Endersby, N. M., Mckechnie, S. W., Vogel, H., Gahan, L. J., Baxter, S. W., Ridland, P. M., & Weeks, A. R. (2005). Microsatellites isolated from diamondback moth, *Plutella xylostella* (L.), for studies of dispersal in Australian populations. *Molecular Ecology Notes*, *5*(1), 51–53. https://doi.org/10.1111/j.1471-8286.2004.00827.x

Felsenstein, J. (1981). Evolutionary trees from DNA sequences: A maximum likelihood approach. *Journal of Molecular Evolution*, *17*(6), 368–376. https://doi.org/10.1007/BF01734359

Fenton, A., Johnson, K. N., Brownlie, J. C., & Hurst, G. D. D. (2011). Solving the *Wolbachia* paradox: Modeling the tripartite interaction between host, *Wolbachia*, and a natural enemy. *American Naturalist*, *178*(3), 333–342. https://doi.org/10.1086/661247

Frydman, H. M., Li, J. M., Robson, D. N., & Wieschaus, E. (2006). Somatic stem cell niche tropism in *Wolbachia*. *Nature*, *441*(7092), 509–512. https://doi.org/10.1038/nature04756

Grabherr, M. G., Haas, B. J., Yassour, M., Levin, J. Z., Thompson, D. A., Amit, I., Adiconis, X., Fan, L., Raychowdhury, R., Zeng, Q., Chen, Z., Mauceli, E., Hacohen, N., Gnirke, A., Rhind, N., di Palma, F., Birren, B. W., Nusbaum, C., Lindblad-Toh, K., … Regev, A. (2011). Full-length transcriptome assembly from RNA-Seq data without a reference genome. *Nature Biotechnology*, *29*(7), 644–652. https://doi.org/10.1038/nbt.1883

Habayeb, M. S., Cantera, R., Casanova, G., Ekström, J.-O., Albright, S., & Hultmark, D. (2009). The *Drosophila* Nora virus is an enteric virus, transmitted via feces. *Journal of Invertebrate Pathology*, *101*(1), 29–33. https://doi.org/10.1016/j.jip.2009.02.003

Hague, M. T. J., Caldwell, C. N., & Cooper, B. S. (2020). Pervasive effects of *Wolbachia* on host temperature preference. *MBio*, *11*(5), 1–15. https://doi.org/10.1128/mBio.01768-20

Hedges, L. M., Brownlie, J. C., O'Neill, S. L., & Johnson, K. N. (2008). *Wolbachia* and virus protection in insects. *Science (New York, N.Y.)*, *322*(5902), 702. https://doi.org/10.1126/science.1162418

Hoang, D. T., Chernomor, O., Von Haeseler, A., Minh, B. Q., & Vinh, L. S. (2018). UFBoot2: Improving the ultrafast bootstrap approximation. *Molecular Biology and Evolution*, *35*(2), 518–522. https://doi.org/10.1093/molbev/msx281

Hoerauf, A., Nissen-Pähle, K., Schmetz, C., Henkle-Dührsen, K., Blaxter, M. L., Büttner, D. W., Gallin, M. Y., Al-Qaoud, K. M., Lucius, R., & Fleischer, B. (1999). Tetracycline therapy targets intracellular bacteria in the filarial nematode *Litomosoides sigmodontis* and results in filarial infertility. *Journal of Clinical Investigation*, *103*(1), 11–18. https://doi.org/10.1172/JCI4768

Hoffmann, A. A., Clancy, D., & Duncan, J. (1996). Naturally-occurring Wolbachia infection in *Drosophila simulans* that does not cause cytoplasmic incompatibility. *Heredity*, *76*(1), 1–8. https://doi.org/10.1038/hdy.1996.1

Hosokawa, T., Koga, R., Kikuchi, Y., Meng, X. Y., & Fukatsu, T. (2010). *Wolbachia* as a bacteriocyte-associated nutritional mutualist. *Proceedings of the National Academy of Sciences of the United States of America*, *107*(2), 769–774. https://doi.org/10.1073/pnas.0911476107

Iturbe-Ormaetxe, I., & O'Neill, S. L. (2007). *Wolbachia*-host interactions: connecting phenotype to genotype. In *Current Opinion in Microbiology* (Vol. 10, Issue 3, pp. 221–224). Elsevier Current Trends. https://doi.org/10.1016/j.mib.2007.05.002

Johnson, K. N. (2015). The impact of *Wolbachia* on virus infection in mosquitoes. *Viruses*, *7*(11), 5705–5717. https://doi.org/10.3390/v7112903

Ju, J. F., Bing, X. L., Zhao, D. S., Guo, Y., Xi, Z., Hoffmann, A. A., Zhang, K. J., Huang, H. J., Gong, J. T., Zhang, X., & Hong, X. Y. (2020). *Wolbachia* supplement biotin and riboflavin to enhance reproduction in planthoppers. *ISME Journal*, *14*(3), 676–687. https://doi.org/10.1038/s41396-019-0559-9

Katoh, K., & Standley, D. M. (2013). MAFFT Multiple Sequence Alignment Software Version 7: Improvements in Performance and Usability. *Molecular Biology and Evolution*, *30*(4), 772–780. https://doi.org/10.1093/molbev/mst010

Kriesner, P., Hoffmann, A. A., Lee, S. F., Turelli, M., & Weeks, A. R. (2013). Rapid sequential spread of two *Wolbachia* variants in *Drosophila simulans*. *PLoS Pathogens*, *9*(9), e1003607. https://doi.org/10.1371/journal.ppat.1003607

Lachaise, D., Cariou, M.-L., David, J. R., Lemeunier, F., Tsacas, L., & Ashburner, M. (1988). Historical biogeography of the *Drosophila melanogaster* species subgroup. In *Evolutionary Biology* (pp. 159–225). Springer US. https://doi.org/10.1007/978-1-4613-0931-4_4

Le Lay, C., Shi, M., Buček, A., Bourguignon, T., Lo, N., & Holmes, E. (2020). Unmapped RNA virus diversity in termites and their symbionts. *Viruses*, *12*(10), 1145. https://doi.org/10.3390/v12101145

Lemaitre, B., Nicolas, E., Michaut, L., Reichhart, J. M., & Hoffmann, J. A. (1996). The dorsoventral regulatory gene cassette spatzle/Toll/Cactus controls the potent antifungal response in *Drosophila* adults. *Cell*, *86*(6), 973–983. https://doi.org/10.1016/S0092-8674(00)80172-5

Mancini, M. V., Herd, C. S., Ant, T. H., Murdochy, S. M., & Sinkins, S. P. (2020). *Wolbachia* strain wAu efficiently blocks arbovirus transmission in *Aedes albopictus*. *PLOS Neglected Tropical Diseases*, *14*(3), e0007926. https://doi.org/10.1371/journal.pntd.0007926

Marcelino, V. R., Clausen, P. T. L. C., Buchmann, J. P., Wille, M., Iredell, J. R., Meyer, W., Lund, O., Sorrell, T. C., & Holmes, E. C. (2020). CCMetagen: Comprehensive and accurate identification of eukaryotes and prokaryotes in metagenomic data. *Genome Biology*, *21*(1), 1–15. https://genomebiology.biomedcentral.com/articles/10.1186/s13059-020-02014-2

Martinez, J., Longdon, B., Bauer, S., Chan, Y.-S., Miller, W. J., Bourtzis, K., Teixeira, L., & Jiggins, F. M. (2014). Symbionts commonly provide broad spectrum resistance to viruses in insects: A comparative analysis of *Wolbachia* strains. *PLoS Pathogens*, *10*(9), e1004369. https://doi.org/10.1371/journal.ppat.1004369

Mather, W. (1960). Additions to the Drosophila fauna of Australia. *University of Queensland Papers*, *1*(9).

Moreira, L. A., Iturbe-Ormaetxe, I., Jeffery, J. A., Lu, G., Pyke, A. T., Hedges, L. M., Rocha, B. C., Hall-Mendelin, S., Day, A., Riegler, M., Hugo, L. E., Johnson, K. N., Kay, B. H., McGraw, E. A., van den Hurk, A. F., Ryan, P. A., & O'Neill, S. L. (2009). A *Wolbachia* symbiont in *Aedes aegypti* limits infection with Dengue, Chikungunya, and Plasmodium. *Cell*, *139*(7), 1268–1278. https://doi.org/10.1016/j.cell.2009.11.042

Naranjo-Lázaro, J. M., Mellín-Rosas, M. A., González-Padilla, V. D., Sánchez-González, J. A., Moreno-Carrillo, G., & Arredondo-Bernal, H. C. (2014). Susceptibility of *Drosophila suzukii* Matsumura (Diptera: Drosophilidae) to Entomophatogenic Fungi. *Southwestern Entomologist*, *39*(1), 201–203. https://doi.org/10.3958/059.039.0119

Nguyen, L.-T., Schmidt, H. A., von Haeseler, A., & Minh, B. Q. (2015). IQ-TREE: a fast and effective stochastic algorithm for estimating maximum-likelihood phylogenies. *Molecular Biology and Evolution*, *32*(1), 268–274. https://doi.org/10.1093/molbev/msu300

Nikoh, N., Hosokawa, T., Moriyama, M., Oshima, K., Hattori, M., & Fukatsu, T. (2014). Evolutionary origin of insect-*Wolbachia* nutritional mutualism. *Proceedings of the National Academy of Sciences of the United States of America*, *111*(28), 10257–10262. https://doi.org/10.1073/pnas.1409284111

Ogunlade, S. T., Adekunle, A. I., Meehan, M. T., Rojas, D. P., & McBryde, E. S. (2020). Modeling the potential of wAu-*Wolbachia* strain invasion in mosquitoes to control *Aedes*-borne arboviral infections. *Scientific Reports*, *10*(1), 16812. https://doi.org/10.1038/s41598-020-73819-1

O'Neill, S. L., Hoffmann, A., & Werren, J. (1997). *Influential Passengers: Inherited Microorganisms and Arthropod Reproduction* (1 ed.). Oxford University Press.

Osborne, S. E., Iturbe-Ormaetxe, I., Brownlie, J. C., O'Neill, S. L., & Johnson, K. N. (2012). Antiviral protection and the importance of *Wolbachia* density and tissue tropism in *Drosophila simulans*. *Applied and Environmental Microbiology*, *78*(19), 6922–6929. https://doi.org/10.1128/AEM.01727-12

Osborne, S. E., Leong, Y. S., O'Neill, S. L., & Johnson, K. N. (2009). Variation in antiviral protection mediated by different *Wolbachia* strains in *Drosophila simulans*. *PLoS Pathogens*, *5*(11), e1000656. https://doi.org/10.1371/journal.ppat.1000656

Palmer, W. H., Medd, N. C., Beard, P. M., & Obbard, D. J. (2018). Isolation of a natural DNA virus of *Drosophila melanogaster*, and characterisation of host resistance and immune responses. *PLoS Pathogens*, *14*(6), e1007050. https://doi.org/10.1371/journal.ppat.1007050

Parsons, P. A., & Bock, I. R. (1979). The Population Biology of Australian *Drosophila*. *Annual Review of Ecology and Systematics*, *10*(1), 229–245. https://doi.org/10.1146/annurev.es.10.110179.001305

Rancès, E., Ye, Y. H., Woolfit, M., McGraw, E. A., & O'Neill, S. L. (2012). The Relative Importance of Innate Immune Priming in *Wolbachia*-Mediated Dengue Interference. *PLoS Pathogens*, *8*(2), e1002548. https://doi.org/10.1371/journal.ppat.1002548

Rice, P., Longden, L., & Bleasby, A. (2000). EMBOSS: The European Molecular Biology Open Software Suite. *Trends in Genetics*, *16*(6), 276–277. https://doi.org/10.1016/S0168-9525(00)02024-2

Ros, V. I. D., Fleming, V. M., Feil, E. J., & Breeuwer, J. A. J. (2009). How diverse is the genus *Wolbachia*? Multiple-gene sequencing reveals a putatively new *Wolbachia* supergroup recovered from spider mites (Acari: Tetranychidae). *Applied and Environmental Microbiology*, *75*(4), 1036–1043. https://doi.org/10.1128/AEM.01109-08

Ross, P. A., Turelli, M., & Hoffmann, A. A. (2019). Evolutionary Ecology of *Wolbachia* Releases for Disease Control. *Annual Review of Genetics*, *53*(1), 93–116. https://doi.org/10.1146/annurev-genet-112618-043609

Scholz, M., Albanese, D., Tuohy, K., Donati, C., Segata, N., & Rota-Stabelli, O. (2020). Large scale genome reconstructions illuminate *Wolbachia evolution*. *Nature Communications*, *11*(1), 1–11. https://doi.org/10.1038/s41467-020-19016-0

Shi, M., White, V. L., Schlub, T., Eden, J.-S., Hoffmann, A. A., & Holmes, E. C. (2018). No detectable effect of *Wolbachia* wMel on the prevalence and abundance of the RNA virome of *Drosophila melanogaster*. *Proceedings of the Royal Society B: Biological Sciences*, *285*(1883), 20181165. https://doi.org/10.1098/rspb.2018.1165

Strunov, A., & Kiseleva, E. (2016). *Drosophila melanogaster* brain invasion: Pathogenic *Wolbachia* in central nervous system of the fly. *Insect Science*, *23*(2), 253–264. https://doi.org/10.1111/1744-7917.12187

Tantowijoyo, W., Andari, B., Arguni, E., Budiwati, N., Nurhayati, I., Fitriana, I., Ernesia, I., Daniwijaya, E. W., Supriyati, E., Yusdiana, D. H., Victorius, M., Wardana, D. S., Ardiansyah, H., Ahmad, R. A., Ryan, P. A., Simmons, C. P., Hoffmann, A. A., Rancès, E., Turley, A. P., … O'Neill, S. L. (2020). Stable establishment of wMel *Wolbachia* in *Aedes aegypti* populations in Yogyakarta, Indonesia. *PLOS Neglected Tropical Diseases*, *14*(4), e0008157. https://doi.org/10.1371/journal.pntd.0008157

Teixeira, L., Ferreira, Á., & Ashburner, M. (2008). The bacterial symbiont *Wolbachia* induces resistance to RNA viral infections in *Drosophila melanogaster*. *PLOS Biology*, *6*(12), 1–11. https://doi.org/10.1371/journal.pbio.1000002

Terradas, G., Joubert, D. A., & McGraw, E. A. (2017). The RNAi pathway plays a small part in *Wolbachia*-mediated blocking of dengue virus in mosquito cells. *Scientific Reports*, *7*. https://doi.org/10.1038/srep43847

Tsai, K.-H., Lien, J.-C., Huang, C.-G., Wu, W.-J., & Chen, W.-J. (2004). Molecular (sub)grouping of endosymbiont *Wolbachia* infection among mosquitoes of Taiwan. *Journal of Medical Entomology*, *41*(4), 677–683. https://doi.org/10.1603/0022-2585-41.4.677

Turelli, M., Cooper, B. S., Richardson, K. M., Ginsberg, P. S., Peckenpaugh, B., Antelope, C. X., Kim, K. J., May, M. R., Abrieux, A., Wilson, D. A., Bronski, M. J., Moore, B. R., Gao, J. J., Eisen, M. B., Chiu, J. C., Conner, W. R., & Hoffmann, A. A. (2018). Rapid Global Spread of wRi-like *Wolbachia* across Multiple *Drosophila*. *Current Biology*, *28*(6), 963-971.e8. https://doi.org/10.1016/j.cub.2018.02.015

Turelli, M., & Hoffmann, A. A. (1991). Rapid spread of an inherited incompatibility factor in California *Drosophila*. *Nature*, *353*(6343), 440–442. https://doi.org/10.1038/353440a0

Webster, C. L., Waldron, F. M., Robertson, S., Crowson, D., Ferrari, G., Quintana, J. F., Brouqui, J.-M., Bayne, E. H., Longdon, B., Buck, A. H., Lazzaro, B. P., Akorli, J., Haddrill, P. R., & Obbard, D. J. (2015). The discovery, distribution, and evolution of viruses associated with *Drosophila melanogaster*. *PLoS Biology*, *13*(7), e1002210. https://doi.org/10.1371/journal.pbio.1002210

Wilfert, L., Longdon, B., Ferreira, A. G. A., Bayer, F., & Jiggins, F. M. (2011). Trypanosomatids are common and diverse parasites of *Drosophila*. *Parasitology*, *138*(7), 858–865. https://doi.org/10.1017/S0031182011000485

Zdobnov, E. M., & Apweiler, R. (2001). InterProScan - an integration platform for the signature-recognition methods in InterPro. *Bioinformatics*, *17*(9), 847–848. https://doi.org/10.1093/bioinformatics/17.9.847

Zouache, K., Voronin, D., Tran-Van, V., Mousson, L., Failloux, A.-B., & Mavingui, P. (2009). Persistent *Wolbachia* and cultivable bacteria infection in the reproductive and somatic tissues of the mosquito vector *Aedes albopictus*. *PLoS ONE*, *4*(7), e6388. https://doi.org/10.1371/journal.pone.0006388

Zug, R., & Hammerstein, P. (2015a). *Wolbachia* and the insect immune system: what reactive oxygen species can tell us about the mechanisms of *Wolbachia*-host interactions. *Frontiers in Microbiology*, *6*, 1201. https://doi.org/10.3389/fmicb.2015.01201

Zug, R., & Hammerstein, P. (2015b). Bad guys turned nice? A critical assessment of Wolbachia mutualisms in arthropod hosts. *Biological Reviews of the Cambridge Philosophical Society*, *90*(1), 89–111. https://doi.org/10.1111/brv.12098

Chapter 6

Substantial Viral and Bacterial Diversity at the Bat–Tick Interface

Ayda Susana Ortiz-Baez, Thomas G.T. Jaenson, Edward C. Holmes, John H.-O. Pettersson and Peter Wilhelmsson

MICROBIAL GENOMICS

Article

CHAPTER 6 Substantial Viral and Bacterial Diversity at the Bat–Tick Interface

Ayda Susana Ortiz-Baez[1], Thomas G.T. Jaenson[2], Edward C. Holmes[1], John H.-O. Pettersson[1,3,4*] and Peter Wilhelmsson[5,6,*]

[1] Sydney Institute for Infectious Diseases, School of Medical Sciences, the University of Sydney, Sydney, New South Wales 2006, Australia.

[2] Department of Organismal Biology, Evolutionary Biology Centre, Uppsala University, SE-752 36 Uppsala, Sweden.

[3] Zoonosis Science Center, Department of Medical Biochemistry and Microbiology, University of Uppsala, SE-751 23 Uppsala, Sweden.

[4] Clinical Microbiology and Hospital Hygiene, Uppsala University Hospital, 75237 Uppsala, Sweden.

[5] Department of Biomedical and Clinical Sciences, Division of Inflammation and Infection, Linköping University, SE-581 83 Linköping, Sweden.

[6] Department of Clinical Microbiology, Region Jönköping County, SE-553 05 Jönköping, Sweden.

* = Joint senior author

***** Correspondence: John H.-O. Pettersson john.pettersson@imbim.uu.se and Peter Wilhelmsson peter.wilhelmsson@liu.se

Microbial Genomics, Article In Press

The following version of this study may have slight differences from the original publication.

6.1 Abstract

Ticks harbour a high diversity of viruses, bacteria and protozoa. The soft tick *Carios vespertilionis* (Argasidae) is a common ectoparasite of bats in the Palearctic region and is suspected to be vector and reservoir of viruses and other microbial species in bat populations, some of which may act as zoonotic agents for human disease. The Soprano pipistrelle (*Pipistrellus pygmaeus*– Vespertilionidae) is widely distributed in Europe, where

it can be found inside or close to human habitation. We used meta-transcriptomic sequencing to determine the RNA virome and common microbiota in blood-fed *C. vespertilionis* ticks collected from a Soprano pipistrelle bat roosting site in south-central Sweden. Our analyses identified 16 viruses from 11 virus families, of which 15 viruses were novel. For the first time in Sweden we identified Issuk-Kul virus, a zoonotic arthropod-borne virus previously associated with outbreaks of acute febrile illness in humans. Likely bat-associated and tick-borne viruses were classified within the families *Nairoviridae, Caliciviridae,* and *Hepeviridae*, while other invertebrate-associated viruses included members of the *Dicistroviridae*, *Iflaviridae, Nodaviridae*, *Partitiviridae*, *Permutotetraviridae*, *Polycipiviridae* and *Solemoviridae*. Similarly, we found abundant bacteria in *C. vespertilionis,* including genera with known tick-borne bacteria, such as *Coxiella* sp*.* and *Rickettsia* sp. These findings demonstrate the remarkable diversity of RNA viruses and bacteria present in *C. vespertilionis* and highlight the importance of bat-associated ectoparasite surveillance as an effective and non-invasive means to track viruses and bacteria circulating in bats and ticks.

6.2 Impact Statement

Bats and ticks are known vectors and reservoirs of diverse pathogenic and non-pathogenic viruses, bacteria and protozoa. The Soprano pipistrelle is a common bat species in Europe that is often parasitised by the soft tick Carios vespertilionis. Given that both the bat and tick can be found in direct proximity to human habitation and are associated with disease-causing zoonotic agents, we used meta-transcriptomic sequencing to uncover the RNA virome and microbiota in ticks that had recently blood-fed off Soprano pipistrelle individuals. In addition to identifying 15 novel viruses and several abundant bacteria, we also detected Issuk-Kul virus, a zoonotic pathogen associated with human disease. Our study not only expands our knowledge of bat-tick associated viruses

and microbes, but also demonstrates the utility and importance of using ectoparasites to non-invasive survey bats for known and novel viruses and bacteria.

6.3 Introduction

The soft tick *Carios vespertilionis,* formerly known as *Argas vespertilionis* (Ixodida: Argasidae) (Mans et al., 2021), is a common ectoparasite of several bat species in Eurasia and Africa (Petney et al., 2017a; Sándor et al., 2021). This soft tick can be found inside or surrounding bat roosts within caves, burrows, wall crevices, tree cavities and other places associated with its hosts. Although, *C. vespertilionis* is a bat-specialist (Sándor et al., 2019), it can incidentally feed on birds, domestic dogs and humans, and may thus be a vector of zoonotic microorganisms and viruses (Jaenson & Wilhelmsson, 2021). During their life-cycle, the larvae attach to the infested bat for 14–31 days, while nymphs and adults feed to repletion in about half an hour (Hoogstraal, 1985).

Bat species in the family Vespertilionidae*,* the largest within the Chiroptera, are frequent hosts of *C. vespertilionis.* Among these*,* the Soprano pipistrelle (*Pipistrellus pygmaeus)* is an important host species in the Western Palearctic. *Pipistrellus pygmaeus* is widely distributed in Europe and it is known to congregate in colonies of several hundred. Buildings serve as common sites for roosting while riparian and woodland habitats are preferred for foraging (Davidson-Watts & Jones, 2006; Stone et al., 2015a). In Sweden, *P. pygmaeus* occurs in the south and south-central parts of the country, where it is often well-adapted to human habitations (Jaenson & Wilhelmsson, 2021). In the wild, the diet of pipistrelle bats largely comprises flying Diptera and Lepidoptera. In the IUCN Red List *P. pygmaeus* is classified in the Least Concern category, but roost destruction is a common threat to this bat species (Stone et al., 2015b).

Bats are common reservoirs for zoonotic agents that can potentially be transmitted by their ectoparasites (Lv et al., 2018b). *C. vespertilionis* has been recorded parasitising *P.*

pygmaeus in Sweden (Jaenson et al., 1994; Jaenson & Wilhelmsson, 2021) and, although evidence is currently lacking, are suspected to be vectors of bat-associated pathogens, including viruses, bacteria, and protozoans. For instance, *Borrelia* bacteria, including *Borrelia afzelii*, have been recorded from *C. vespertilionis* (Hubbard et al., 1998; Jaenson & Wilhelmsson, 2021; Zabashta et al., 2019). Other tick-borne microorganisms recorded from *C. vespertilionis* include *Rickettsia* spp*., Ehrlichia* spp., and *Babesia* spp. (Lv et al., 2018a). However, a lack of virome studies means that only a limited number of viruses have been detected in these ticks to date, including Issyk-Kul virus (ISKV; *Nairoviridae*), Sokuluk virus (SOKV; *Flaviviridae*) and Soft tick bunyavirus (Alkhovsky et al., 2013; Oba et al., 2016; Obsomer et al., 2013; Petney et al., 2017). Similarly, in the case of *P. pygmaeus* only a few zoonotic viruses within the families *Adenoviridae*, *Astroviridae*, *Coronaviridae* and *Herpesviridae* have been documented (Chen et al., 2014; Zhou et al., 2022).

The implementation of bulk RNA-sequencing (meta-transcriptomics) technologies has revolutionised our understanding of the virome diversity and virus-host interactions in nature (Ortiz-Baez et al., 2020; Shi et al., 2018a; Shi et al., 2016; Stark et al., 2019). In particular, the use of meta-transcriptomics has revealed an enormous diversity of RNA viruses in invertebrate species, as well as revealing ancestral evolutionary links to vertebrate RNA viruses (Li et al., 2015; Zhang et al., 2018). Since there is limited knowledge of the RNA virome of *C. vespertilionis* and *P. pygmaeus*, we investigated what proportion of viruses present in the bat-tick system is either shared between this ectoparasite and its bat host or is specific to each host type. To address this question, we used meta-transcriptomics to determine the virome, as well as common non-viral tick-borne microorganisms, associated with *C. vespertilionis* from a bat-box inhabited by *P. pygmaeus* in south-central Sweden.

6.4 Materials and Methods

6.4.1 Sample collection

Tick specimens of *C. vespertilionis* were collected in the mornings from 24 June to 4 August 2020 from a plastic tray placed on the ground below a man-made wooden bat-box housing a colony of about 250–500 adult females and juveniles of *P. pygmaeus* located in a garden at Snesslinge, province of Uppland, South-Central Sweden (60.19.567° N, 18.067° E). The nursery bat house was made with eight-chambers with dark exterior surfaces to increase attraction to bats (Tuttle et al., 2013). An electric heater was placed in a bat-restricted area of the house for use during very cold nights. Extra holes were included in the walls of the house to allow sufficient air circulation during hot summer days. The bat house was located in an open part of the garden with a mixture of spruce and broad-leaf trees. To minimise bats being attacked by predators, the house was placed on poles about 3.5 m above the ground. A total of 165 ticks, naturally detached from the bats, were collected, placed in vials containing RNA later (Thermo Fisher Scientific) and examined microscopically for ingested blood meal. Ticks were identified microscopically to species level and developmental stage (larva, nymph or adult) based on their morphological characters as previously described (Arthur, 1963; Filippova, 1966; Hillyard, 1996; Hoogstraal, 1958; Yamaguti et al., 1971). The ticks were stored in RNA later at -28 ^{o}C for 4–6 months and subsequently at -80 ^{o}C until molecular analyses.

6.4.2 Sample preparation and sequencing

Ticks were processed into 12 libraries, pooling between 3–24 individuals of different developmental stages per library (Supplementary table S1). Tick samples were homogenized using ZR BashingBead 0.1mm (Zymo Research, Irvine, CA, USA) for 180s using a bench-top homogenizer (TissueLyzer II, Qiagen). Total RNA was extracted from the homogenates using the ZymoBIOMICS DNA/RNA Miniprep Kit (Zymo Research) according to the manufacturer's instructions. Library preparation and ribosomal RNA (rRNA)

depletion was performed using the Tecan Trio RNA-seq kit (NuGEN Technologies, Inc. CA, USA), following the manufacturer's protocol. Bulk paired-end RNA sequencing was performed on the DNBseq platform by the Beijing Genomics Institute (BGI), Hong Kong.

6.4.3 Sequence data processing and assembly

Quality control of sequencing reads was performed with FASTQC (Andrews, 2010) and summarized using the MultiQC tool (Ewels et al., 2016b). Reads were *de novo* assembled into contigs using MEGAHIT v1.2.9 with default settings (D. Li et al., 2015). Assembled contigs were compared against the NCBI non-redundant database (NCBI-nr) using DIAMOND BLASTX with an e-value cut-off ≥ 1E-4 (Buchfink et al., 2015). To provide an overview of the viral and microbial composition in the ticks, taxonomic profiling was performed using CCMetagen (Marcelino et al., 2020). Open reading frame (ORF) prediction and protein translation were performed on contigs above 900 nt with the getORF program (EMBOSS). ORFs were predicted as translation regions between STOP codons (-minsize 600 -find 0). Proteins and conserved domains present the viral contigs were annotated using InterProScan v5.52-86.0 and HMMER v3.3 (*hmmscan* program), with default search parameters (Finn et al., 2011). To quantify virus abundance, we filtered out ribosomal reads from Bacteria, Archaea and Eukarya using SortmeRNA v. 2.1b (Kopylova et al., 2012), with the non-ribosomal reads then mapped to the virus contigs with BBMap v37.98. Relative contig abundance was computed as the number of Reads Per Million (RPM). To determine the prevalence of the viruses across the samples and avoid false-positives, read counts lower than 0.1% of the highest abundance for each virus were assumed as the result of index-hopping and removed. Virus abundance was put in context of host gene abundance by comparisons with the mitochondrial 12S and 16S rRNA genes that are stably expressed in *C. vespertilionis* and *P. pygmaeus.* Similarly, we used 16S and 18S rRNA genes to compare sequence abundance in bacteria and protozoa, respectively.

6.4.4 Microbiota profiling

We focused on targeting the common bacteria and protozoan microbiota found in ticks. To this end, we targeted the 16S rRNA gene for bacterial agents and 18S rRNA for protozoans. When no ribosomal RNA genes were detected, unfiltered reads (i.e. prior to rRNA filtering) were mapped against available reference sequences corresponding to *Anaplasma* (NR_044762.1), *Borrelia* (NR_170496.1), *Ehrlichia* (MF069159.1), *Escherichia* (NR_074902.1), *Francisella* (NR_074665.1), *Rickettsia* (NR_074394.1), *Delftia* (NR_116495.1), *Pseudomonas* (NR_117678.1), *Coxiella* (NR_104916.1), *Moraxella* (NR_104936.1), and *Babesia* (AB242176) as these are common tick microbiota components or known mammalian pathogens (Sándor et al., 2021). The majority consensus sequences were obtained from the most common nucleotides shared between the overlapping reads that mapped to the reference sequences. Consensus sequences were screened against the NCBI nr/nt and rRNA/ITS databases for validation. Further verification of the quality of the rRNA sequences was performed using the Ribovore v1.0.2 software (Schäffer et al., 2021). When no rRNA gene contigs or other suitable marker genes were detected, consensus sequences were only used for phylogenetic contextualization. Abundance was estimated as RPM by mapping reads to the reference sequences as described above.

6.4.5 Phylogenetic analysis

For each virus taxonomic group, amino acid sequences corresponding to the RNA-dependent RNA polymerase (RdRp) were aligned to reference sequences available in GenBank using the E-INS-I iterative refinement method implemented in MAFFT v7.487 software (Katoh & Standley, 2013). Accordingly, the 16S and 18S rRNA marker genes were used for bacteria and protozoans as noted above. The best-fit model of amino acid (coding sequences) and nucleotide (ribosomal sequences) substitution, as well as phylogenetic relationships, were inferred using the Maximum Likelihood (ML) method available in IQ-TREE v1.6.12 (Nguyen et al., 2015a). Tree node support was estimated with SH-aLRT and the ultrafast bootstrap (UFBoot) (Hoang et al., 2018a). A total of 1000 replicates were run

along the “bnni” option to limit branch support overestimation. Tree visualisation and annotation was performed using the R packages ggplot2 (Wickham, 2016b) and Inkscape v1.2 software.

6.4.6 Virus nomenclature

Novel viruses were provisionally named based on geographic locations within the area (province of Uppland) where the Soprano pipistrelle and the soft tick *C. vespertilionis* are known to occur.

6.5 Results

A total of 165 ticks (144 larvae, 12 nymphs and 9 adults) of *C. vespertilionis*, all with visible blood in their guts, were collected from the roost of *P. pygmaeus.* We used a meta-transcriptomics approach to reveal the RNA virome and bacterial components of bat-associated *C. vespertilionis* ticks. In total, we generated ~846 million reads, of which ~101 million corresponded to non-ribosomal reads. Approximately 51,000 contigs were assembled from the total of reads.

We detected a high diversity of RNA viruses and microbiota, corresponding to bacteria and parasitic protozoa in the bat-ticks analysed. Overall, we identified 16 viruses based on the identification of RdRp sequences, including 15 putative novel viruses within the families *Caliciviridae*, *Dicistroviridae*, *Hepeviridae*, *Iflaviridae, Nairoviridae*, *Nodaviridae*, *Partitiviridae*, *Permutotetraviridae*, *Polycipiviridae* and *Solemoviridae* (**Table 6.1**). Among these, we detected at least one bat-associated tick-borne arbovirus within the *Nairoviridae* (**Figure 6.1**). The most abundant families were the *Nairoviridae* and *Hepeviridae*, although the *Polycipiviridae* and *Caliciviridae* and *Solemoviridae* were moderately abundant (**Figure 6.1**). Also of note, we detected three short viral contigs (libraries D and E) that were highly similar to known bat paramyxoviruses (*Paramyxoviridae*), as shown in the blastx similarity search and an associated phylogenetic

analysis (**Supplementary Figure S1**, **Table S2**). Although we excluded all contigs shorter than 900 nt (300 aa) from the analyses, we further characterized these contigs given the likely bat origin and relevance to surveillance. Accordingly, the paramyxovirus-like sequences (381–595 nt) covered different regions in the L protein, including conserved motifs found in the RdRp [SRLF*RNIGDP] and the G-7-mTase [LSHP] domains. Similarly, the contig partially covering the RdRp was assigned with ~36% similarity and 99.9% confidence to the RdRp of the parainfluenza virus (**Supplementary Figure S1, Table S2)**. The full diversity of RNA viruses characterized in this study included two negative-sense RNA viruses (-ssRNA), 12 positive-sense RNA viruses (+ssRNA) and one double-strand RNA viruses (dsRNA). Likewise, the virus prevalence ranged from six to ten viruses detected per tick library.

Table 6.1 Summary of novel and known RNA viruses identified in this study and their closest hits in the NCBI/nr database.

Contig	Provisional name/virus name	contig length	Best hit on the NCBI/nr database	Similarity	E-value	Provisional classification	Pools
k99_1661	Harg calici-like virus	9766	UCS96400.1 hypothetical protein 1 [Riboviria sp.]	79.13	0.00E+00	*Caliciviridae*	A,B,C,E,G,H,I,J,K,L
k99_2737	Aspo dicistro-like virus	1290	QPG92983.1 polyprotein [Ohio dicistro-like virus]	65.6	1.28E-185	*Dicistroviridae*	A,B,C,G,J,K
k99_7	Valo virus	4700	QIS88064.1 polyprotein, partial [Bulatov virus]	46.62	0.00E+00	*Hepeviridae*	A,B,C,D,E,F,G,H,I,J,K,L
k99_1852	Barko ifla-virus	9948	ACH57393.1 polyprotein [Infectious flacherie virus]	36.5	0.00E+00	*Iflaviridae*	E,F,G,I
k99_2945	Ornas iflavirus	677	QKW94197.1 RNA-dependent RNA polymerase, partial [Sacbrood virus]	54	1.07E-82	*Iflaviridae*	I,J
k99_7939	Gimo ifla-like virus	706	AOY34458.1 polyprotein, partial [Rolda virus]	38.9	5.96E-35	*Iflaviridae*	B
k99_1517	Gubbo nairovirus	12421	AKC89355.1 RNA-dependent RNA-polymerase, partial [Artashat orthonairovirus]	50.5	0.00E+00	*Nairoviridae*	A,D,E,F,G,H,I,J,K,L
k99_1658	Issyk-Kul virus $^{\varphi}$	12288	AKI29982.1 RNA-dependent RNA-polymerase protein [Issyk-Kul virus]	99.7	0.00E+00	*Nairoviridae*	A,B,C,D,E,F,G,H,I,J,K,L
k99_2267	Gravol virus	1074	YP_009337883.1 RNA-dependent RNA polymerase [Hubei orthoptera virus 4]	42.8	1.29E-85	*Nodaviridae*	F,G
k99_1814	Bolka virus	811	NP_077730.1 RNA dependent RNA polymerase protein A [Nodamura virus]	65.9	2.33E-114	*Nodaviridae*	A,E
k99_1453	Agalma virus	751	YP_009342458.1 RdRp [Wuhan fly virus 5]	78.5	6.61E-135	*Partitiviridae*	E
k99_43	Snesslinge virus	1299	BBE15516.1 RNA-dependent RNA polymerase [Osugoroshi virus 1]	73.9	7.22E-200	*Partitiviridae*	B
k99_2789	Ladskar virus	909	AOC55066.1 polyprotein, partial [Niehaus virus]	70.8	1.64E-127	*Permutotetraviridae*	K
k99_543	Graso virus	10048	QGA87336.1 polyprotein, partial [Hammarskog picorna-like virus]	25.5	1.96E-119	*Polycipiviridae*	A,B,C,E,G,H,I,J,K,L
k99_1507	Dudero virus	919	QHA33683.1 polyprotein [Cacaos virus]	45.2	2.73E-68	*Polycipiviridae*	B,F,G
k99_3888	Ed virus	2639	QEM39297.1 RNA-dependent RNA polymerase [Humaita-Tubiacanga virus]	51	7.36E-149	*Solemoviridae*	A,B,C,E,G,H,I,J,K,L

$^{\varphi}$= known virus

6.5.1 Likely tick-borne and bat-associated virus families

We identified two viruses within the *Nairoviridae*, including one novel virus. The novel virus was tentatively named as Gubbo nairovirus (GUBV) and exhibited the three segments typical to nairoviruses. GUVB shared a limited amino acid sequence in similarity with Artashat orthonairovirus based on comparison with the viral polymerase (aa %id = 50.5) (**Table 6.1**). We also detected virus contigs corresponding to the large protein (L segment), glycoprotein (M segment) and nucleoprotein (S segment) of the bat-associated Issyk-Kul virus (ISKV) (RdRp aa %id = 99.7) (**Figure 6.2**). Both nairoviruses were detected in > 80% of the samples at high abundance levels (**Figure 6.1**). As expected, these viruses grouped phylogenetically with other known tick-borne and bat-associated viruses (**Figure 6.3**). In particular, GUBV was closely related to bat nairovirus and Berlin bat nairovirus detected in organ tissues from European vespertilionid bats. However, the short available sequences for these viruses (127–147 aa) made it difficult to assign with certainty that these correspond to GUBV.

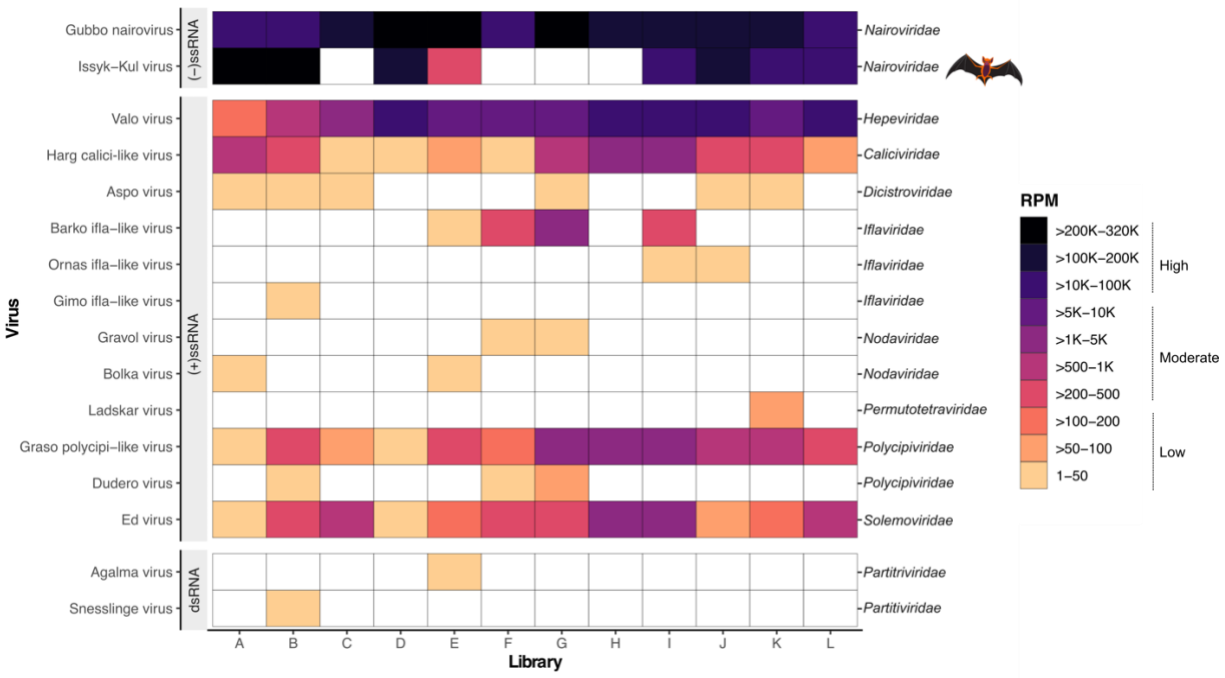


Figure 6.1 Overview of virus abundance across each bat-associated tick library. Abundance is quantified as the number of reads per million (RPM). RNA viruses are grouped according to the Baltimore classification. Levels of virus abundance are categorized as high, moderate, and low, as shown in the

legend. The bat silhouette indicates whether a virus has previously been identified in bats and only applies to Issyk−Kul virus.

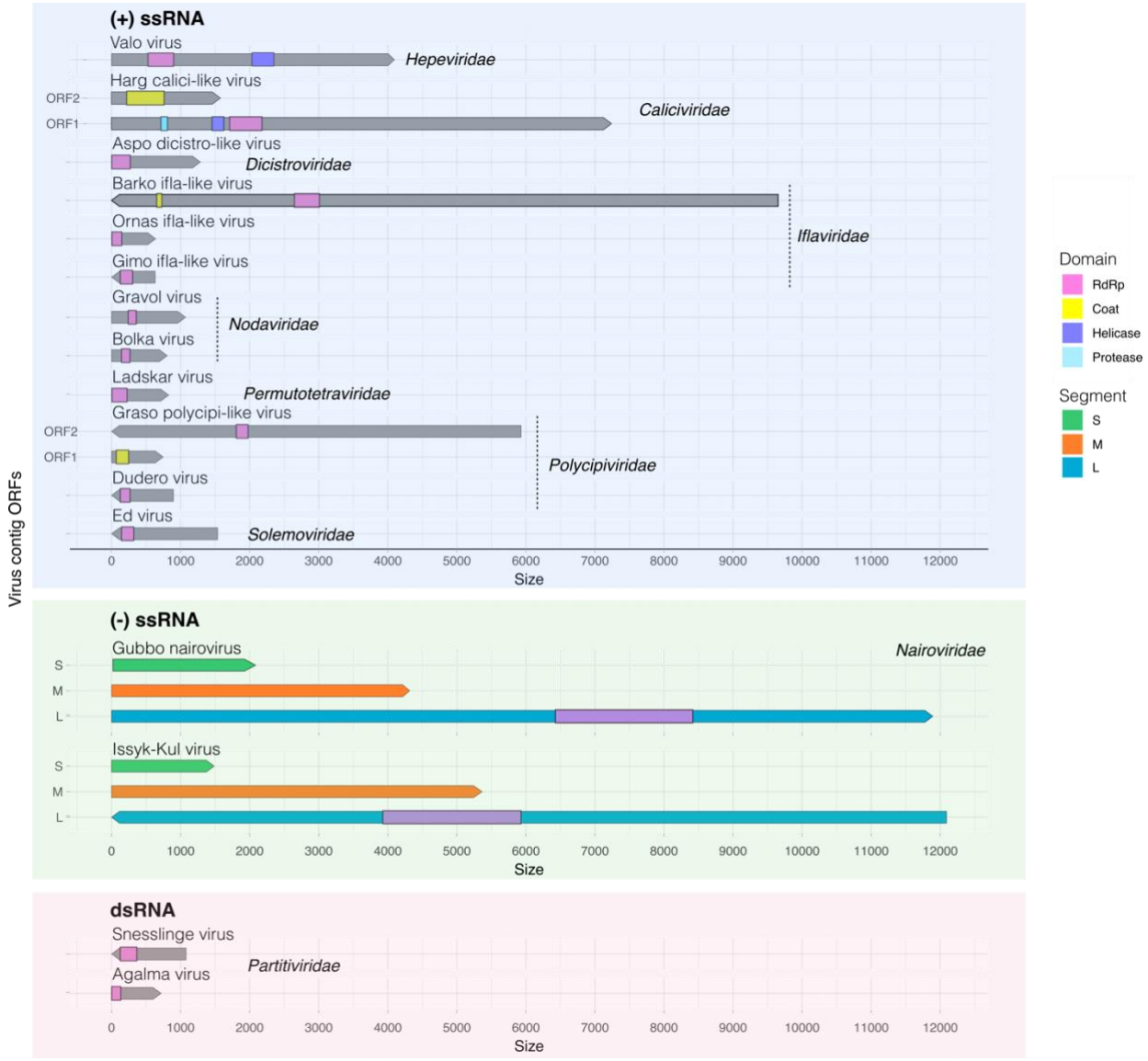


Figure 6.2 Schematic representation of the open reading frames (ORFs) found for the RNA viruses identified in the bat-tick libraries analysed. ORFs are shown as arrow-shaped boxes whose orientation depends on the frames they were identified. Domains and segments are indicated as shown in the legend. Only the RdRp domain is shown in the L segment of nairoviruses.

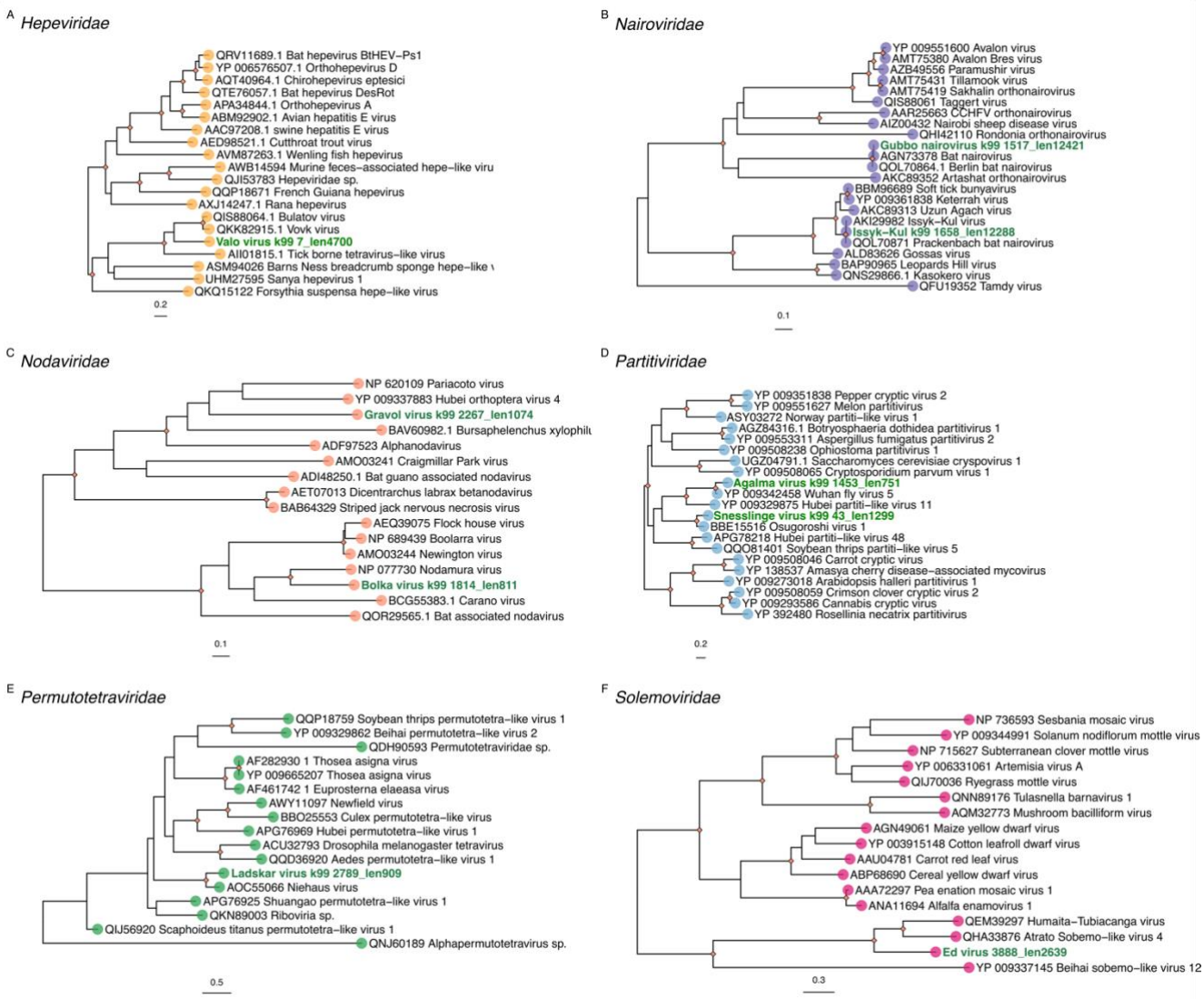


Figure 6.3 Phylogenetic relationships among the viruses identified in this study and representative background sequences from relevant families of RNA viruses. (A) *Hepeviridae* (B) *Nairoviridae*, (C) *Nodaviridae,* (D) *Partitiviridae*, (E) *Permutotetraviridae,* (F) *Solemoviridae*. The viruses obtained here are indicated with green. In each case maximum likelihood trees are mid-point rooted for clarity and were constructed based on the amino acid sequences of the RdRp. Nodal support values corresponding to SH-aLRT >= 80% and UFboot >= 95% are displayed with orange diamonds on nodes. The scale bars are shown at the bottom of each tree and represent the number of amino acid substitutions per site.

Similarly, we identified one novel member from the *Caliciviridae* provisionally referred to as Harg calici-like virus (HCAV), which was present in all the libraries at low to moderate abundance levels (**Figure 6.1**). Phylogenetic analyses showed that HCAV grouped with the unclassified Riboviria sp. virus and Clinch calicivirus (**Figure 6.4, Supplementary Figure S2**), exhibiting above 79% similarity for the RdRp protein (**Table 6.1**, **Figure 6.4**) and 93.3% similarity for the VP1 protein, respectively. Notably these viruses form a clade basal to taxa of different genera in the *Caliciviridae*. For HCAV we identified the nearly complete genome (~ 9 kb), including two ORFs encoding the RdRp and the major capsid protein VP1, respectively. Finally, among the most abundant viral families, we identified one novel virus – Valo virus (VALV) – belonging to the *Hepeviridae* that was well represented in all the libraries (RPM = 200–100K) (**Figure 6.1**). Phylogenetically, VALV grouped with Bulatov virus and Vovk virus, previously identified in ticks, although it only exhibited 42% aa sequence similarity to Bulatov virus in the RdRp region as the closest blast hit (**Table 6.1**, **Figure 6.2–6.3**). As a caveat, abundance levels might be underestimated for partial or shorter virus contigs since RPM estimates are influenced by contig length.

6.5.2 Likely arthropod and tick microbiome-associated viruses

Among the newly discovered +ssRNA viruses in the *Picornavirales*, we identified three iflaviruses (Barko virus, Ornas virus and Gimo virus), two polycipiviruses (Graso virus and Dudero virus), and one dicistrovirus (Aspo dicistro-like virus). Moreover, we identified two members of the *Nodaviridae* (Gravol virus and Bolka virus), one permutotretavirus (Ladskar virus) and one solemovirus (Ed virus). For all the viruses with the exception of Graso polycipi-like virus and Barko virus, we only detected the viral RdRp gene (**Figure 6.2**). These viruses were most closely related to other arthropod-associated viruses in the different viral families (**Table 6.1, Figure 6.3–6.4**), and were present in low to moderate abundance in the tick libraries analysed. Barko virus, Graso polycipi−like virus and Ed virus were found in higher abundance, while only Graso polycipi−like virus and Ed virus were present in all the libraries (**Figure 6.1**). With respect to the dsRNA viruses, we identified two novel partitiviruses corresponding to Agalma virus and Snesslinge virus based on the presence of a viral RdRp signal (**Table 6.1**,

Figure 6.2). Both viruses were present in a limited number of tick libraries (2/12) at low abundance levels (**Figure 6.1**). The closest relatives were partitiviruses previously found in insects, including the Wuhan fly virus 5 and Osugoroshi virus 1 (aa %id = 73.9–78.5) (**Table 6.1, Figure 6.3**).

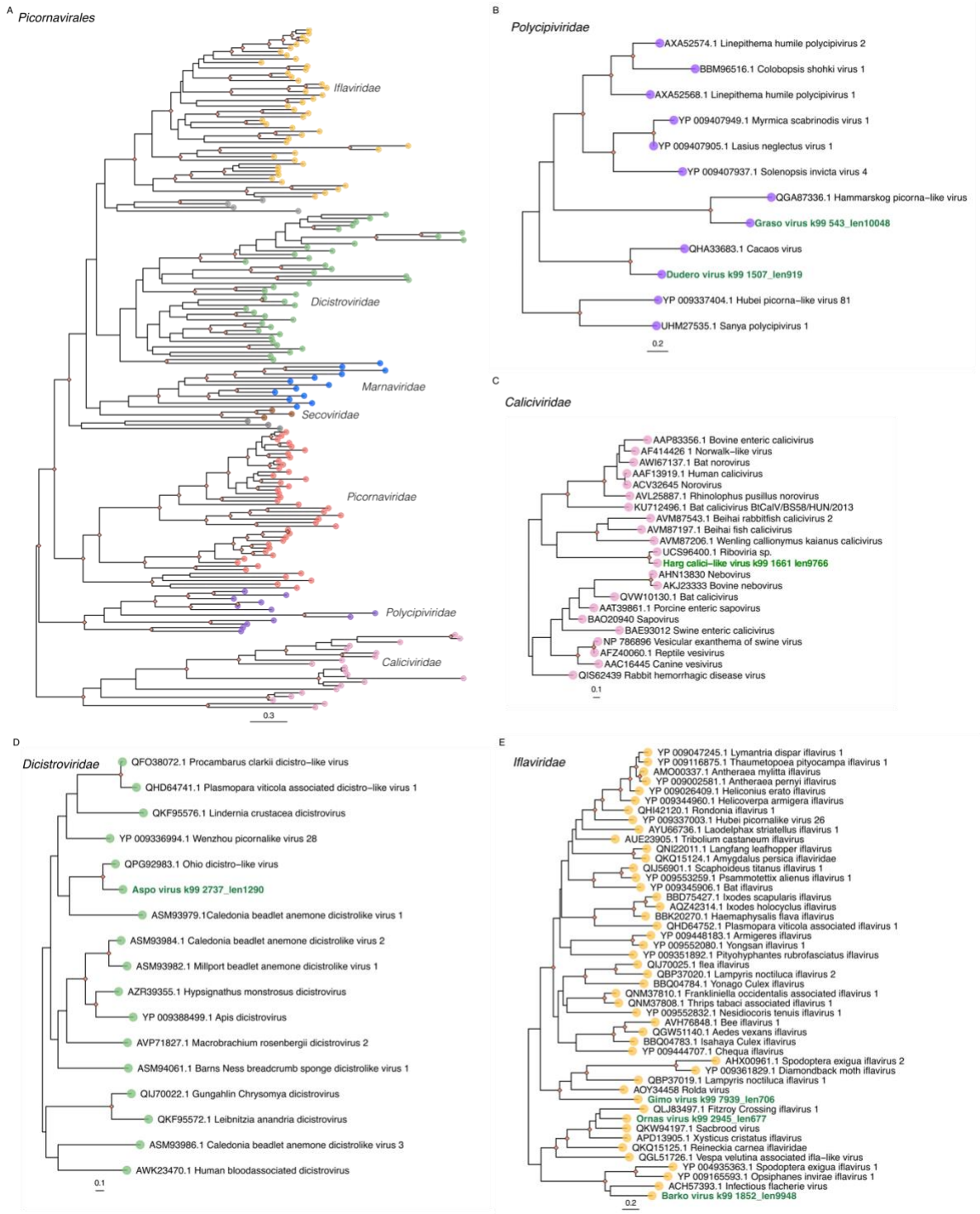


Figure 6.4 Phylogenetic relationships among the viruses identified in this study and representative background sequences within the *Picornavirales*. The family clades extracted from the order-level tree correspond to the (A) *Polycipiviridae* (B) *Caliciviridae*, (C) *Discistroviridae,* (D) *Iflaviridae*. The viruses obtained here are indicated with green-tip labels. In each case maximum likelihood trees are mid-point rooted for clarity and was constructed based on the amino acid sequences of the RdRp. Nodal support values corresponding to SH-aLRT >= 80% and UFboot >= 95% are displayed with orange diamonds on nodes. The scale bars are shown at the bottom of each tree and represent the number of amino acid substitutions per site.

6.5.3 Common microbiota in C. vespertilionis

An analysis of the microbial composition of *C. vespertilionis* revealed the presence of highly abundant tick-borne bacteria genera (~35–66% of total contigs; taxonomy profiles available at figshare: 10.6084/m9.figshare.21550899), including members of the *Rickettsia, Delftia* and *Coxiella*, which were present in all the libraries screened (**Figure 6.5, Supplementary Figure S6.3**). These bacteria exhibited > 97% similarity at the 16S rRNA gene to *Rickettsia conorii, Delftia lacustris* and *Coxiella burnetii*, respectively (**Supplementary Table S6.3, Figure 6.6A**). In particular, the *Rickettsia* identified here grouped with *Rickettsia* species classified in the spotted fever group (SFG), including *R. conorii*, *R. africae, R. slovaca* and *R. parkeri.* The phylogenetic analysis based on the outer membrane protein A (ompA) gene suggested a close relationship to *R. parkeri* within the spotted fever group (~ 99% nt similarity) (**Figure 6.6B**). In the case of *Coxiella*, we observed close relationships with other microbiota in *Ornithodoros capensis* and *Carios capensis* ticks (**Figure 6.6**). *C. vespertilionis* ticks also harboured other highly prevalent bacteria similar to *Escherichia fergusonii* and *Moraxella osloensis*, although these were detected at much lower abundance levels, and placed as divergent taxa in the phylogenetic tree (**Figure 6.5–6.6, Supplementary Table S6.3**). We did not detect members of the genera *Anaplasma, Borrelia, Ehrlichia, Francisella* and *Babesia* that were also included in the preliminary screen.

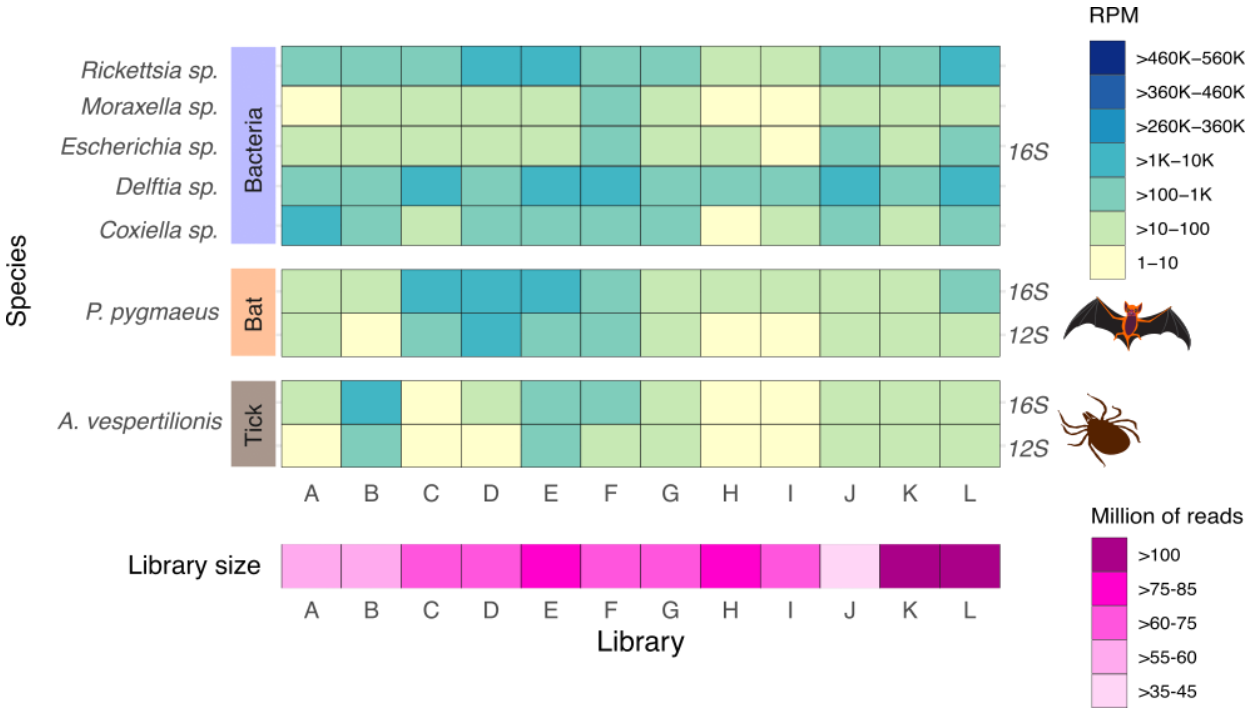


Figure 6.5 Overview of bacteria abundance across each bat-associated tick library. Abundance is quantified as the number reads per million (RPM) based on the 16S gene. Host expression was assessed using the genes 16S and 12S *C. vespertilionis* and *P. pygmaeus*, as indicated with the animal silhouettes. The bottom panel shows the size across each tick library.

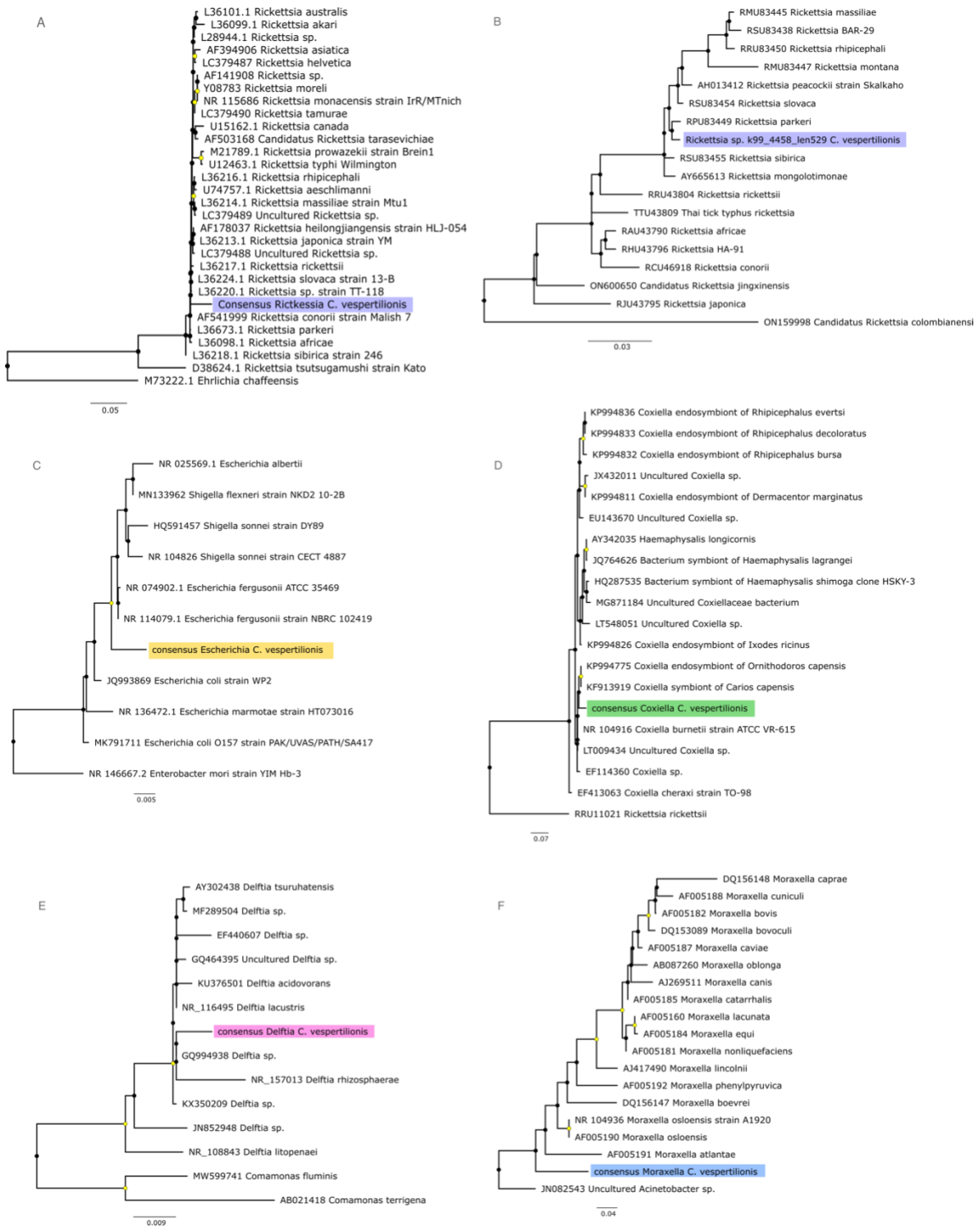


Figure 6.6 Phylogenetic relationships among the bacterial sequences identified in this study and representative background sequences. The phylogenetic placement of *Rickettsia* was assessed by comparing 16S rRNA (A) and ompA (B) genes, whereas 16S rRNA consensus sequences were used for (C)

Escherichia, (D) *Coxiella*, (E) *Delftia* and (F) *Moraxella.* Bacteria consensus sequences are highlighted in each tree. In each case maximum likelihood trees are outgroup rooted. Nodal support values corresponding to SH-aLRT >= 80% and UFboot >= 95% are displayed with yellow circles on nodes. The scale bars are shown at the bottom of each tree and represent the number of nucleotide substitutions per site.

6.6 Discussion

Ticks naturally harbour a highly diverse array of viruses, bacteria and protozoans. Since ticks are obligately hematophagous, these parasitic arthropods might also carry the viruses and microbiota of their hosts acquired during the blood meal (Allan et al., 2010). The natural history traits of bat ticks raise important questions on how the viral and bacterial diversity of ticks is shaped by bat blood meals. In addition, ticks parasitising bats are of particular interest given that bats are thought to be a natural reservoir for pathogens of veterinary and public health concern (Letko et al., 2020b). Consequently, ticks might also act as vectors of pathogens circulating in bats, posing a risk for the health of animal populations, including humans. Thus, investigating the diversity of RNA viruses and bacteria in bat-associated ticks could provide a strategy for regular active surveillance of bat-borne zoonoses.

Our analysis of the bat-tick *C. vespertilionis* virome revealed the family *Nairoviridae* (*Bunyavirales*) as the most abundant and prevalent in the libraries of recently blood fed ticks (**Figure 6.1**). Similar findings have been reported in recent metagenomic studies on different tick and host species across a variety of geographic locations (Blomström et al., 2020; Z. Liu et al., 2022; Wille et al., 2020; Z. Xu et al., 2022), suggesting that ticks might be competent hosts and vectors for the replication and transmission of nairoviruses in nature. Among the members of the *Nairoviridae* found here, we identified ISKV (Atkinson et al., 2015a; Lvov et al., 1973), a zoonotic virus associated with outbreaks of acute febrile illness in humans. ISKV virus was originally discovered in Central Asia in the 1970s, but has recently also been recorded in Germany (Brinkmann et al., 2020; Kohl et al., 2021a; Lvov et al., 1973). The virus was first identified in a *Nyctalus noctula* bat, although its host range has been expanded to other bat

species (Brinkmann et al., 2020; Kohl et al., 2021a; Lvov, 2019; Lvov et al., 1973; Walker et al., 2015). Similarly, there are reports of ISKV detected in *C. vespertilionis* (Lvov et al., 1973; Walker et al., 2015). Herein, we demonstrate for the first time the presence of ISKV at high abundance levels in bat-ticks in Sweden (**Figure 6.1**). From our current knowledge, there are no reports of ISKV in *P. pygmaeus* bats. However, the presence of ISKV in recently blood fed *C. vespertilionis* bat-ticks, as well as in other vespertilionid bats, make it plausible that this virus also occurs in *P. pygmaeus*.

The recent detection of the novel Gubbo nairovirus (GUBV) at similar abundance levels to ISKV is compatible with the notion that it might be both a tick-borne and bat-associated virus (**Figure 6.1**). This is also supported by the close relationship of GUBV to other nairoviruses isolated from European bats, suggesting that it might also be able to infect *P. pygmaeus* bats. Since GUBV is distantly related to Artashat orthonairovirus, it might represent a new species within the *Nairoviridae* together with bat nairovirus and Berlin bat nairovirus (**Table 6.1**, **Figure 6.3**). Notably, our analysis of the abundance, prevalence, and host range of GUBV is limited to a small number of tick samples (**Figure 6.1, Supplementary table S1**). Similarly, we cannot exclude the possibility of high viral loads in viraemic bat hosts. Comparative research targeting unfed questing tick and bat samples separately could help test these hypotheses more rigorously. The zoonotic potential and public health significance of GUBV for animal populations similarly merits additional investigation. In combination with previous research (Atkinson et al., 2015b; Brinkmann et al., 2020; Vargina et al., 1982), our results support the hypothesis that these ticks might serve as vectors and/or potential reservoirs for these nairoviruses.

Although we did identify paramyxovirus-related sequences in the data generated here (**Figure S6.1**, **Supplementary table S6.2**), they were not included in our analyses due to the limited length of the contigs. It should be noted, however, that paramyxoviruses have been reported in *Pipistrellus* species (Chua et al., 2001; Kohl et al., 2021a; Kohl et al., 2021b; Rizzo et al., 2017; Van Brussel & Holmes, 2022). That we only recovered a few short paramyxo-like sequences from bat-associated ticks might indicate low viral loads in the blood meal (Kurth et

al., 2012), and bat urine and faeces may be more suitable samples for the detection of these viruses (Chua et al., 2001; Kohl et al., 2021b; Rizzo et al., 2017).

The presence of newly discovered +ssRNA virus members within the *Caliciviridae* and *Hepeviridae* is consistent with previous research on bat-borne and tick-borne viruses. A few bat caliciviruses (sapoviruses and unclassified viruses) have been discovered in European vespertilionid bats (Kemenesi et al., 2014; Lazov et al., 2021), although there are no corresponding reports of viruses in *P. pygmaeus* (Chen et al., 2014; Zhou et al., 2022). The calicivirus identified in this study (HCAV) was highly divergent and unrelated to other bat caliciviruses, with its closest relative an unclassified virus found in reptile faeces (RdRp, MZ375209) (**Figure 6.4**). Notably, there is no current evidence of ticks carrying caliciviruses or playing a role on their transmission. Based on the divergent phylogenetic position of HCAV, it might represent a member of a new genus within the family *Caliciviridae*, although establishing a definitive association with vertebrate/invertebrate hosts is uncertain. In contrast, the novel hepevirus VALV is suspected to be associated with the tick virome was most closely related to Bulatov virus and Vovk virus that have been associated with the virome of *Ixodes uriae* ticks from the Antarctic peninsula (Pettersson et al., 2020; Wille et al., 2020). In addition, the high prevalence and abundance of VALV in all the libraries tentatively suggests that the virus replicates in ticks (**Figure 6.1**). Finally, although some hepeviruses have been detected circulating in bats (Drexler et al., 2012; T. Kobayashi et al., 2018; B. Wang et al., 2017), our study lacks data to assess if VALV has any association with transmission or disease in bats.

As expected, a considerable fraction of the tick virome corresponded to viruses associated with invertebrates. This included virus families such as the *Iflaviridae, Nodaviridae, Solemoviridae* and *Partitiviridae*, previously identified in the virome of different tick species (Harvey et al., 2019; D. Kobayashi et al., 2020; Pettersson et al., 2017; Vandegrift & Kapoor, 2019; Xu et al., 2022). We also found members of the *Dicistroviridae*, *Permutotetraviridae* and *Polycipiviridae* that are likely infecting the bat-ticks. Indeed, the dominance of Graso polycipi−like virus (*Polycipiviridae*) and Ed virus (*Solemoviridae*) in the tick libraries might

indicate the efficient replication of these viruses within this arthropod species, although this will need further research. Similarly, the occurrence of *Partitiviridae* in *C. vespertilionis* is compatible with studies suggesting that partitiviruses can possibly infect arthropods (Cross et al., 2020; Pettersson et al., 2017; Shi et al., 2018; Xu et al., 2022), as well as fungi and protozoa in these ectoparasites. However, we were not able to definitively determine the host of these viruses. Previous research on *C. vespertilionis* has been largely focused on targeting tick-borne viruses of public health relevance (Palomar et al., 2021; Walker et al., 2015), with the RNA virome as a whole largely unexplored. As such, our work provides a baseline for the study of RNA viruses in *C. vespertilionis*.

It has previously been shown that *C. vespertilionis* can harbour a repertoire of bacterial and protozoal species (Moustafa et al., 2022; Socolovschi et al., 2012; Zhmaeva et al., 1966). We identified sequences related to the most common microbial agents in bat-ticks, some of which are of particular interest due to their high abundance (**Figure 6.5–6.6, Supplementary Table S6.3**). For instance, our data revealed the presence of *Rickettsia* sp. in *C. vespertilionis* collected in Sweden, corroborating previous reports in Europe (Lv et al., 2018a; Socolovschi et al., 2012). Although we were unable to provide a species level classification based on the 16S rRNA and ompA genes, the close relationship to *Rickettsia* species, and in particular to R. *parkeri*, in the SFG might constitute a risk for vector-borne zoonotic disease. Rickettsial infections with some species within the SFG have been associated with pathogenicity in humans (Lv et al., 2018a; McBride et al., 2007; Nilsson et al., 2010). For instance, *R. parkeri* is an emergent tick-borne pathogen and the causative agent of *R. parkeri* rickettsiosis in America (Moo-Llanes et al., 2021; Silva-Ramos et al., 2021). Similarly, in Sweden, infections with *R. helvetica* and *R. felis* have been associated with severe clinical manifestations, including meningitis (Lindblom et al., 2010; Nilsson et al., 2010). In contrast, It has been shown that *Rickettsia* could play a role in the provision of folate in *Ixodes pacificus* ticks (Hunter et al., 2015). The range of interactions between *Rickettsia* and *C. vespertilionis* remains uncertain, as does their pathogenic potential for bat hosts (Socolovschi et al., 2012). Likewise, *Rickettsia* spp. have been reported in bat tissue samples collected from vespertilionid bats (Matei et al., 2021; S. Zhao, Yang, et al., 2020).

Therefore, whether bacterial infection can impact bat health or whether bats contribute to the maintenance of *Rickettsia* spp. in nature merits investigation (S. Zhao, Yang, et al., 2020).

We also reported the co-occurrence of bacteria such as *Delftia* sp. and *Coxiella* sp. (**Figure 6.5–6.6**). *Delftia* sp. have been reported as core bacteria in the microbiome of *Dermacentor variabilis* (Travanty et al., 2019). Given the high abundance and prevalence of *Delftia* sp. in the tick libraries, a similar situation might exist for *C. vespertilionis* (Moustafa et al., 2022). An earlier study documented the presence of *Coxiella burnetii*, the aetiological agent of Q fever, in *C. vespertilionis* ticks collected from Asia (Zhmaeva et al., 1966), although many *Coxiella* species are considered obligate and associated with nutritional and reproductive roles in ticks (Bonnet et al., 2017; Bonnet & Pollet, 2021a; Khoo et al., 2016; T. A. Smith et al., 2015; Zhong et al., 2007). General questions that remain are whether ticks act as vectors or reservoirs (or both) of all these agents, and what extent the blood meal and the environment contribute to the viral and bacterial composition in bat-ticks.

Overall, we provide new insights into the viral and bacterial diversity associated with *C. vespertilionis* ticks in Sweden. The presence of dominant and underrepresented viruses and bacteria warrants further research into the nature of bat-tick interactions and how these impact viral and microbial transmission. Additional vector competence studies are required to demonstrate that *C. vespertilionis* ticks can become infected when feeding on an infectious host and maintain the pathogen such that it is capable of being transmitted to an uninfected, susceptible host (Estrada-Peña et al., 2021). Despite the small sample size, our study demonstrates that bat-tick surveillance provides an effective and non-invasive means to detect bat and tick-borne microorganisms circulating in bat roosting habitats. These results reinforce the notion of protecting the natural environment of bats and minimizing human exposure to bat/tick habitats to prevent zoonotic spillover events (Plowright et al., 2017; Sokolow et al., 2019).

6.6 Supplementary Material

Table S6.1. Number of individuals and developmental stages of bat-ticks included in this study.

Pools	# Individuals	Stage
A	3	Adults
B	3	Adults
C	3	Adults
D	4	Nymphs
E	4	Nymphs
F	4	Nymphs
G	24	Larvae
H	24	Larvae
I	24	Larvae
J	24	Larvae
K	24	Larvae
L	24	Larvae

Table S6.2. Detection of paramyxo-like viral sequences and their closest hits in the NCBI/nr database.

Contig	Contig length	Best hit on the NCBI/nr database	Similarity	E-value	Provisional classification	Pool
k99_5075	381	AIF74192.1 polymerase, partial [Bat paramyxovirus]	78.4	1.34E-63	*Paramyxoviridae*	D
k99_3206	487	AIF74192.1 polymerase, partial [Bat paramyxovirus]	90.8	3.95e-44	*Paramyxoviridae*	D
k99_8678	595	AGU69459.1 large protein, partial [Miniopterus schreibersii paramyxovirus]	51.5	1.39E-54	*Paramyxoviridae*	E

Table S6.3. Overview of the closest BLAST hits in the NCBI/16S rRNA database for the bacterial consensus sequences generated in this study.

Best hit on the NCBI/nr database	Sequence length	Similarity	E-value
Rickettsia conorii strain Malish 7 16S ribosomal RNA, partial sequence	1498	97.39%	0.00E+00
Coxiella burnetii strain ATCC VR-615 16S ribosomal RNA, partial sequence	1465	97.14%	0.00E+00
Delftia lacustris strain 332 16S ribosomal RNA, partial sequence	1534	99.13%	0.00E+00
Moraxella osloensis strain A1920 16S ribosomal RNA, partial sequence	1522	92.78%	0.00E+00
Escherichia fergusonii ATCC 35469 16S ribosomal RNA, complete sequence	1542	99.47%	0.00E+00

Supplementary figures are available at at https://figshare.com/s/e5f49f80f6c906e519cd

Figure S6.1. Phylogenetic relationships among the putative paramyxovirus identified in this study and representative sequences within the *Paramyxoviridae*. The virus sequence detected in this study is shown in bold green in the tree and indicated in the aa alignment of the RdRp region (top-right). Only conserved positions are shown in the alignment. The maximum-likelihood tree is mid-point rooted for clarity and was constructed based on L protein amino acid sequences (source: https://talk.ictvonline.org). Protein structure homology modelling is shown in the bottom-right panel. Nodal support values corresponding to SH-aLRT >= 80% and UFboot >= 95% are displayed with orange diamonds on nodes. The scale bars are shown at the bottom of each tree and represent the number of amino acid substitutions per site.

Figure S6.2. Phylogenetic relationships among the calici-like virus identified in this study and representative sequences within the *Caliciviridae*. The novel virus obtained here is shown in bold green. The maximum-likelihood tree is mid-point rooted for clarity and was constructed based on reference sequences of the VP1 protein (source: https://talk.ictvonline.org). Nodal support values corresponding to SH-aLRT >= 80% and UFboot >= 95% are displayed with orange diamonds on nodes. The scale bars are shown at the bottom of each tree and represent the number of amino acid substitutions per site.

Figure S6.3. Map coverage plots for the 16S rRNA gene of the bacteria detected across libraries of *Carios vespertilionis* (panels A–L).

The sequencing reads and viral sequences identified in this study were deposited at the Sequence Read Archive (SRA) database under accession numbers SAMN29627891–SAMN29627902 (Bioproject: PRJNA838788) and GenBank database (OP514647–OP514662; OP804625–OP804628; OP782089-OP782093; OP857220).

5.6 References

Alkhovsky, S. V, Lvov, D. K., Shchelkanov, M. Y., Shchetinin, A. M., Deryabin, P. G., Samokhvalov, E. I., Gitelman, A. K., & Botikov, A. G. (2013). The taxonomy of the Issyk-Kul virus (ISKV, Bunyaviridae, Nairovirus), the etiologic agent of the Issyk-Kul fever isolated from bats (Vespertilionidae) and ticks Argas (Carios) vespertilionis (Latreille, 1796). *Voprosy Virusologii*, *58*(5), 11–15.

Allan, B. F., Goessling, L. S., Storch, G. A., & Thach, R. E. (2010). Blood meal analysis to identify reservoir hosts for *Amblyomma americanum* ticks. *Emerging Infectious Diseases*, *16*(3), 433–440. https://doi.org/10.3201/eid1603.090911

Andrews, S. (2010). *FastQC: a quality control tool for high throughput sequence data.*

Arthur, D. R. (1963). *British ticks*. Butterworths.

Atkinson, B., Marston, D. A., Ellis, R. J., Fooks, A. R., & Hewson, R. (2015a). Complete Genomic Sequence of Issyk-Kul Virus. *Genome Announcements*, *3*(4), 662–677. https://doi.org/10.1128/GENOMEA.00662-15

Blomström, A. L., Luz, H. R., Öhlund, P., Lukenge, M., Brandão, P. E., Labruna, M. B., & Berg, M. (2020). Novel viruses found in *Antricola* ticks collected in bat caves in the western Amazonia of Brazil. *Viruses*, *12*(1). https://doi.org/10.3390/V12010048

Bonnet, S. I., Binetruy, F., Hernández-Jarguín, A. M., & Duron, O. (2017). The tick microbiome: Why non-pathogenic microorganisms matter in tick biology and pathogen transmission. *Frontiers in Cellular and Infection Microbiology*, *7*, 236. https://doi.org/10.3389/FCIMB.2017.00236/BIBTEX

Bonnet, S. I., & Pollet, T. (2021). Update on the intricate tango between tick microbiomes and tick-borne pathogens. *Parasite Immunology*, *43*(5), 0–2. https://doi.org/10.1111/pim.12813

Brinkmann, A., Kohl, C., Radonić, A., Dabrowski, P. W., Mühldorfer, K., Nitsche, A., Wibbelt, G., & Kurth, A. (2020). First detection of bat-borne Issyk-Kul virus in Europe. *Scientific Reports 2020 10:1*, *10*(1), 1–7. https://doi.org/10.1038/s41598-020-79468-8

Buchfink, B., Xie, C., & Huson, D. H. (2015). Fast and sensitive protein alignment using DIAMOND. *Nature Methods*, *12*(1), 59–60. https://doi.org/10.1038/nmeth.3176

Chen, L., Liu, B., Yang, J., & Jin, Q. (2014). DBatVir: The database of bat-associated viruses. *Database*, *2014*. https://doi.org/10.1093/database/bau021

Chua, K. B., Wang, L. F., Lam, S. K., Crameri, G., Yu, M., Wise, T., Boyle, D., Hyatt, A. D., & Eaton, B. T. (2001). Tioman virus, a novel paramyxovirus isolated from fruit bats in Malaysia. *Virology*, *283*(2), 215–229. https://doi.org/10.1006/VIRO.2000.0882

Cross, S. T., Maertens, B. L., Dunham, T. J., Rodgers, C. P., Brehm, A. L., Miller, M. R., Williams, A. M., Foy, B. D., & Stenglein, M. D. (2020). Partitiviruses infecting *Drosophila melanogaster* and *Aedes aegypti* exhibit efficient biparental vertical transmission. *Journal of Virology*, *94*(20). https://doi.org/10.1128/jvi.01070-20

Davidson-Watts, I., & Jones, G. (2006). Differences in foraging behaviour between *Pipistrellus pipistrellus* (Schreber, 1774) and *Pipistrellus pygmaeus* (Leach, 1825). *Journal of Zoology*, *268*(1), 55–62. https://doi.org/10.1111/j.1469-7998.2005.00016.x

Drexler, J. F., Seelen, A., Corman, V. M., Tateno, A. F., Cottontail, V., Zerbinati, R. M., Gloza-Rausch, F., Klose, S. M., Adu-Sarkodie, Y., Oppong, S. K., Kalko, E. K. V., Osterman, A., Rasche, A., Adam, A., Müller, M. A., Ulrich, R. G., Leroy, E. M., Lukashev, A. N., & Drosten, C. (2012). Bats worldwide carry Hepatitis E virus-related viruses that form a putative novel genus within the family Hepeviridae. *Journal of Virology*, *86*(17), 9134. https://doi.org/10.1128/JVI.00800-12

Estrada-Peña, A., Cevidanes, A., Sprong, H., & Millán, J. (2021). Pitfalls in tick and tick-borne pathogens research, some recommendations and a call for data sharing. *Pathogens*, *10*(6), 712. https://doi.org/10.3390/PATHOGENS10060712

Ewels, P., Magnusson, M., Lundin, S., & Käller, M. (2016). MultiQC: summarize analysis results for multiple tools and samples in a single report. *Bioinformatics*, *32*(19), 3047–3048. https://doi.org/10.1093/bioinformatics/btw354

Filippova, N. A. (1966). Argasid ticks (Argasidae). *Fauna SSSR*, *4*(3), 255.

Finn, R. D., Clements, J., & Eddy, S. R. (2011). HMMER web server: interactive sequence similarity searching. *Nucleic Acids Research*, *39*(Web Server issue), W29. /pmc/articles/PMC3125773/

Harvey, E., Rose, K., Eden, J.-S., Lo, N., Abeyasuriya, T., Shi, M., Doggett, S. L., & Holmes, E. C. (2019). Extensive diversity of RNA viruses in Australian ticks. *Journal of Virology*, *93*(3). https://doi.org/10.1128/JVI.01358-18

Hillyard, P. D. (1996). *Ticks of north-west Europe.* Field Studies Council.

Hoang, D. T., Chernomor, O., Von Haeseler, A., Minh, B. Q., & Vinh, L. S. (2018). UFBoot2: Improving the ultrafast bootstrap approximation. *Molecular Biology and Evolution*, *35*(2), 518–522. https://doi.org/10.1093/molbev/msx281

Hoogstraal, H. (1958). Bat ticks of the genus *Argas* (Ixodoidea, Argasidae). The subgenus *Carios,* a redescription of *A. (C.) vespertilionis* (Latreille, 1802), and variation within an Egyptian population. *Annals of the Entomological Society of America*, *51*(1), 19–26. https://doi.org/10.1093/AESA/51.1.19

Hoogstraal, H. (1985). Argasid and nuttalliellid ticks as parasites and vectors. *Advances in Parasitology*, *24*(C), 135–238. https://doi.org/10.1016/S0065-308X(08)60563-1

Hubbard, M. J., Baker, A. S., & Cann, K. J. (1998). Distribution of *Borrelia burgdorferi* s.l. spirochaete DNA in British ticks (Argasidae and Ixodidae) since the 19th Century, assessed by PCR. *Medical and Veterinary Entomology*, *12*(1), 89–97. https://doi.org/10.1046/J.1365-2915.1998.00088.X

Hunter, D. J., Torkelson, J. L., Bodnar, J., Mortazavi, B., Laurent, T., Deason, J., Thephavongsa, K., & Zhong, J. (2015). The Rickettsia endosymbiont of *Ixodes pacificus* contains all the genes of de novo folate biosynthesis. *PLOS ONE*, *10*(12), e0144552. https://doi.org/10.1371/JOURNAL.PONE.0144552

Jaenson, T. G. T., Tälleklint, L., Lundqvist, L., Olsen, B., Chirico, J., & Mejlon, H. (1994). Geographical distribution, host associations, and vector roles of ticks (Acari: Ixodidae, Argasidae) in Sweden. *Journal of Medical Entomology*, *31*(2), 240–256. https://doi.org/10.1093/jmedent/31.2.240

Jaenson, T. G. T., & Wilhelmsson, P. (2021). First record of a suspected human-pathogenic *Borrelia* species in populations of the bat tick *Carios vespertilionis* in Sweden. *Microorganisms*, *9*(5), 1100. https://doi.org/10.3390/microorganisms9051100

Katoh, K., & Standley, D. M. (2013). MAFFT Multiple Sequence Alignment Software Version 7: Improvements in Performance and Usability. *Molecular Biology and Evolution*, *30*(4), 772–780. https://doi.org/10.1093/molbev/mst010

Kemenesi, G., Dallos, B., Görföl, T., Boldogh, S., Estók, P., Kurucz, K., Kutas, A., Földes, F., Oldal, M., Németh, V., Martella, V., Bányai, K., & Jakab, F. (2014). Molecular survey of RNA viruses in Hungarian bats: discovering novel astroviruses, coronaviruses, and caliciviruses. *Vector-Borne and Zoonotic Diseases*, *14*(12), 846–855. https://doi.org/10.1089/VBZ.2014.1637

Khoo, J. J., Chen, F., Kho, K. L., Ahmad Shanizza, A. I., Lim, F. S., Tan, K. K., Chang, L. Y., & AbuBakar, S. (2016). Bacterial community in *Haemaphysalis* ticks of domesticated animals from the Orang Asli communities in Malaysia. *Ticks and Tick-Borne Diseases*, *7*(5), 929–937. https://doi.org/10.1016/J.TTBDIS.2016.04.013

Kobayashi, D., Murota, K., Itokawa, K., Ejiri, H., Amoa-Bosompem, M., Faizah, A. N., Watanabe, M., Maekawa, Y., Hayashi, T., Noda, S., Yamauchi, T., Komagata, O., Sawabe, K., & Isawa, H. (2020). RNA virome analysis of questing ticks from Hokuriku District, Japan, and the evolutionary dynamics of tick-borne phleboviruses. *Ticks and Tick-Borne Diseases*, *11*(2), 101364. https://doi.org/10.1016/j.ttbdis.2019.101364

Kobayashi, T., Murakami, S., Yamamoto, T., Mineshita, K., Sakuyama, M., Sasaki, R., Maeda, K., & Horimoto, T. (2018). Detection of bat hepatitis E virus RNA in microbats in Japan. *Virus Genes*, *54*(4), 599. https://doi.org/10.1007/S11262-018-1577-9

Kohl, C., Brinkmann, A., Radonić, A., Dabrowski, P. W., Mühldorfer, K., Nitsche, A., Wibbelt, G., & Kurth, A. (2021a). The virome of German bats: comparing virus discovery approaches. *Scientific Reports 2021 11:1*, *11*(1), 1–18. https://doi.org/10.1038/s41598-021-86435-4

Kohl, C., Nitsche, A., & Kurth, A. (2021b). Update on potentially zoonotic viruses of European bats. *Vaccines*, *9*(7). https://doi.org/10.3390/vaccines9070690

Kopylova, E., Noé, L., & Touzet, H. (2012). SortMeRNA: fast and accurate filtering of ribosomal RNAs in metatranscriptomic data. *Bioinformatics*, *28*(24), 3211–3217.

Kurth, A., Kohl, C., Brinkmann, A., Ebinger, A., Harper, J. A., Wang, L. F., Mühldorfer, K., & Wibbelt, G. (2012). Novel Paramyxoviruses in Free-Ranging European Bats. *PLOS ONE*, *7*(6), e38688. https://doi.org/10.1371/JOURNAL.PONE.0038688

Lazov, C. M., Belsham, G. J., Bøtner, A., & Rasmussen, T. B. (2021). Full-genome sequences of alphacoronaviruses and astroviruses from Myotis and Pipistrelle bats in Denmark. *Viruses*, *13*(6), 1073. https://doi.org/10.3390/V13061073/S1

Letko, M., Seifert, S. N., Olival, K. J., Plowright, R. K., & Munster, V. J. (2020). Bat-borne virus diversity, spillover and emergence. In *Nature Reviews Microbiology* (Vol. 18, Issue 8, pp. 461–471). Nature Research. https://doi.org/10.1038/s41579-020-0394-z

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. (2015). Unprecedented genomic diversity of RNA viruses in arthropods reveals the ancestry of negative-sense RNA viruses. *ELife*, *4*. https://doi.org/10.7554/eLife.05378

Li, D., Liu, C.-M., Luo, R., Sadakane, K., & Lam, T.-W. (2015). MEGAHIT: an ultra-fast single-node solution for large and complex metagenomics assembly via succinct de Bruijn graph. *Bioinformatics*, *31*(10), 1674–1676. https://doi.org/10.1093/bioinformatics/btv033

Lindblom, A., Severinson, K., & Nilsson, K. (2010). Rickettsia felis infection in Sweden: Report of two cases with subacute meningitis and review of the literature. *Scand J Infect Dis*, *42*(11–12), 906–909. https://doi.org/10.3109/00365548.2010.508466

Liu, Z., Li, L., Xu, W., Yuan, Y., Liang, X., Zhang, L., Wei, Z., Sui, L., Zhao, Y., Cui, Y., Yin, Q., Li, D., Li, Q., Wei, F., Hou, Z., Liu, Q., Wang, Z., Liu, (Z, Xu, W., … Wei, F. (2022). Extensive diversity of RNA viruses in ticks revealed by metagenomics in northeastern China. *BioRxiv*, 2022.04.27.489762. https://doi.org/10.1101/2022.04.27.489762

Lv, J., Fernández de Marco, M. del M., Goharriz, H., Phipps, L. P., McElhinney, L. M., Hernández-Triana, L. M., Wu, S., Lin, X., Fooks, A. R., & Johnson, N. (2018a). Detection of tick-borne bacteria and babesia with zoonotic potential in *Argas (Carios) vespertilionis* (Latreille, 1802) ticks from British bats. *Scientific Reports*, *8*(1), 1865. https://doi.org/10.1038/s41598-018-20138-1

Lvov, D. K. (2019). Issyk-Kul Fever. *The Arboviruses: Epidemiology and Ecology*, 53–62. https://doi.org/10.1201/9780429280276-4

Lvov, D. K., Karas, F. R., Timofeev, E. M., Tsyrkin, Y. M., Vargina, S. G., Veselovskaya, O. V., Osipova, N. Z., Grebenyuk, Y. I., Gromashevski, V. L., Steblyanko, S. N., & Fomina, K. B. (1973). “Issyk-Kul” virus, a new arbovirus isolated from bats and *Argas (Carios) vespertilionis* (Latr., 1802) in the Kirghiz S.S.R. *Archiv Für Die Gesamte Virusforschung 1973 42:2*, *42*(2), 207–209. https://doi.org/10.1007/BF01270841

Mans, B. J., Kelava, S., Pienaar, R., Featherston, J., de Castro, M. H., Quetglas, J., Reeves, W. K., Durden, L. A., Miller, M. M., Laverty, T. M., Shao, R., Takano, A., Kawabata, H., Moustafa, M. A. M., Nakao, R., Matsuno, K., Greay, T. L., Evasco, K. L., Barker, D., & Barker, S. C. (2021). Nuclear (18S-28S rRNA) and mitochondrial genome markers of *Carios (Carios) vespertilionis* (Argasidae)

support Carios Latreille, 1796 as a lineage embedded in the Ornithodorinae: re-classification of the Carios sensu Klompen and Oliver (1993) clade . *Ticks and Tick-Borne Diseases*, *12*(4), 101688. https://doi.org/10.1016/J.TTBDIS.2021.101688

Marcelino, V. R., Clausen, P. T. L. C., Buchmann, J. P., Wille, M., Iredell, J. R., Meyer, W., Lund, O., Sorrell, T. C., & Holmes, E. C. (2020). CCMetagen: Comprehensive and accurate identification of eukaryotes and prokaryotes in metagenomic data. *Genome Biology*, *21*(1), 1–15. https://genomebiology.biomedcentral.com/articles/10.1186/s13059-020-02014-2

Matei, I. A., Corduneanu, A., Sándor, A. D., Ionică, A. M., Panait, L., Kalmár, Z., Ivan, T., Papuc, I., Bouari, C., Fit, N., & Mihalca, A. D. (2021). *Rickettsia* spp. in bats of Romania: high prevalence of *Rickettsia monacensis* in two insectivorous bat species. *Parasites and Vectors*, *14*(1), 1–8. https://doi.org/10.1186/S13071-021-04592-X/TABLES/3

McBride, W. J. H., Hanson, J. P., Miller, R., & Wenck, D. (2007). Severe spotted fever group rickettsiosis, Australia. *Emerging Infectious Diseases*, *13*(11), 1742. https://doi.org/10.3201/EID1311.070099

Moo-Llanes, D. A., de Oca-Aguilar, A. C. M., Romero-Salas, D., & Sánchez-Montes, S. (2021). Inferring the Potential Distribution of an Emerging Rickettsiosis in America: The Case of Rickettsia parkeri. *Pathogens*, *10*(5). https://doi.org/10.3390/PATHOGENS10050592

Moustafa, M. A. M., Mohamed, W. M. A., Lau, A. C. C., Chatanga, E., Qiu, Y., Hayashi, N., Naguib, D., Sato, K., Takano, A., Matsuno, K., Nonaka, N., Taylor, D. M., Kawabata, H., & Nakao, R. (2022). Novel symbionts and potential human pathogens excavated from argasid tick microbiomes that are shaped by dual or single symbiosis. *Computational and Structural Biotechnology Journal*, *20*, 1979–1992. https://doi.org/10.1016/J.CSBJ.2022.04.020

Nguyen, L.-T., Schmidt, H. A., von Haeseler, A., & Minh, B. Q. (2015). IQ-TREE: a fast and effective stochastic algorithm for estimating maximum-likelihood phylogenies. *Molecular Biology and Evolution*, *32*(1), 268–274. https://doi.org/10.1093/molbev/msu300

Nilsson, K., Elfving, K., & Påhlson, C. (2010). *Rickettsia helvetica* in patient with meningitis, Sweden, 2006. *Emerging Infectious Diseases*, *16*(3), 490–492. https://doi.org/10.3201/EID1603.090184

Oba, M., Omatsu, T., Takano, A., Fujita, H., Sato, K., Nakamoto, A., Takahashi, M., Takada, N., Kawabata, H., Ando, S., & Mizutani, T. (2016). A novel bunyavirus from the soft tick, *Argas*

vespertilionis, in Japan. In *Journal of Veterinary Medical Science* (Vol. 78, Issue 3, pp. 443–445). Japanese Society of Veterinary Science. https://doi.org/10.1292/jvms.15-0536

Obsomer, V., Wirtgen, M., Linden, A., Claerebout, E., Heyman, P., Heylen, D., Madder, M., Maris, J., Lebrun, M., Tack, W., Lempereur, L., Hance, T., & Van Impe, G. (2013). Spatial disaggregation of tick occurrence and ecology at a local scale as a preliminary step for spatial surveillance of tick-borne diseases: General framework and health implications in Belgium. *Parasites and Vectors*, *6*(1), 190. https://doi.org/10.1186/1756-3305-6-190

Ortiz-Baez, A. S., Cousins, K., Eden, J. S., Chang, W. S., Harvey, E., Pettersson, J. H. O., Carver, S., Polkinghorne, A., Šlapeta, J., Rose, K., & Holmes, E. C. (2020). Meta-transcriptomic identification of *Trypanosoma* spp. in native wildlife species from Australia. *Parasites and Vectors*, *13*(1), 447. https://doi.org/10.1186/s13071-020-04325-6

Palomar, A. M., Veiga, J., Portillo, A., Santibáñez, S., Václav, R., Santibáñez, P., Oteo, J. A., & Valera, F. (2021). Novel genotypes of nidicolous *Argas* ticks and their associated microorganisms from Spain. *Frontiers in Veterinary Science*, *8*, 243. https://doi.org/10.3389/FVETS.2021.637837/BIBTEX

Petney, T. N., Jaenson, T. G. T., & Faffle, M. P. (2017a). *Argas vespertilionis* (Latreille, 1796). In: Ticks of Europe and North Africa. In *Ticks of Europe and North Africa* (pp. 33–40). Springer International Publishing. https://doi.org/10.1007/978-3-319-63760-0

Pettersson, J. H. O., Ellström, P., Ling, J., Nilsson, I., Bergström, S., González-Acuña, D., Olsen, B., & Holmes, E. C. (2020). Circumpolar diversification of the Ixodes uriae tick virome. *PLoS Pathogens*, *16*(8). https://doi.org/10.1371/JOURNAL.PPAT.1008759

Pettersson, J. H. O., Shi, M., Bohlin, J., Eldholm, V., Brynildsrud, O. B., Paulsen, K. M., Andreassen, Å., & Holmes, E. C. (2017). Characterizing the virome of *Ixodes ricinus* ticks from Northern Europe. *Scientific Reports 2017 7:1*, *7*(1), 1–7. https://doi.org/10.1038/s41598-017-11439-y

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. *Nature Reviews Microbiology 2017 15:8*, *15*(8), 502–510. https://doi.org/10.1038/nrmicro.2017.45

Rizzo, F., Edenborough, K. M., Toffoli, R., Culasso, P., Zoppi, S., Dondo, A., Robetto, S., Rosati, S., Lander, A., Kurth, A., Orusa, R., Bertolotti, L., & Mandola, M. L. (2017). Coronavirus and

paramyxovirus in bats from Northwest Italy. *BMC Veterinary Research*, *13*(1), 1–11. https://doi.org/10.1186/S12917-017-1307-X/FIGURES/3

Sándor, A. D., Corduneanu, A., Péter, Á., Mihalca, A. D., Barti, L., Csosz, I., Szoke, K., & Hornok, S. (2019). Bats and ticks: Host selection and seasonality of bat-specialist ticks in Eastern Europe. *Parasites and Vectors*, *12*(1), 1–10. https://doi.org/10.1186/s13071-019-3861-5

Sándor, A. D., Mihalca, A. D., Domşa, C., Péter, Á., & Hornok, S. (2021). Argasid ticks of Palearctic bats: distribution, host selection, and zoonotic importance. *Frontiers in Veterinary Science*, *8*, 684737. https://doi.org/10.3389/fvets.2021.684737

Schäffer, A. A., McVeigh, R., Robbertse, B., Schoch, C. L., Johnston, A., Underwood, B. A., Karsch-Mizrachi, I., & Nawrocki, E. P. (2021). Ribovore: ribosomal RNA sequence analysis for GenBank submissions and database curation. *BMC Bioinformatics*, *22*(1), 1–29. https://doi.org/10.1186/S12859-021-04316-Z/TABLES/11

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*, *540*(7634), 539–543. https://doi.org/10.1038/nature20167

Shi, M., White, V. L., Schlub, T., Eden, J.-S., Hoffmann, A. A., & Holmes, E. C. (2018). No detectable effect of *Wolbachia* wMel on the prevalence and abundance of the RNA virome of *Drosophila melanogaster*. *Proceedings of the Royal Society B: Biological Sciences*, *285*(1883), 20181165. https://doi.org/10.1098/rspb.2018.1165

Shi, M., Zhang, Y. Z., & Holmes, E. C. (2018a). Meta-transcriptomics and the evolutionary biology of RNA viruses. *Virus Research*, *243*, 83–90. https://doi.org/10.1016/j.virusres.2017.10.016

Silva-Ramos, C. R., Hidalgo, M., & Faccini-Martínez, Á. A. (2021). Clinical, epidemiological, and laboratory features of Rickettsia parkeri rickettsiosis: A systematic review. *Ticks and Tick-Borne Diseases*, *12*(4), 101734. https://doi.org/10.1016/J.TTBDIS.2021.101734

Smith, T. A., Driscoll, T., Gillespie, J. J., & Raghavan, R. (2015). A *Coxiella*-like endosymbiont is a potential vitamin source for the Lone Star tick. *Genome Biology and Evolution*, *7*(3), 831–838. https://doi.org/10.1093/GBE/EVV016

Socolovschi, C., Kernif, T., Raoult, D., & Parola, P. (2012). *Borrelia*, *Rickettsia*, and *Ehrlichia* species in bat ticks, France, 2010. *Emerging Infectious Diseases*, *18*(12), 1966. https://doi.org/10.3201/EID1812.111237

Sokolow, S. H., Nova, N., Pepin, K. M., Peel, A. J., Pulliam, J. R. C., Manlove, K., Cross, P. C., Becker, D. J., Plowright, R. K., McCallum, H., & De Leo, G. A. (2019). Ecological interventions to prevent and manage zoonotic pathogen spillover. *Philosophical Transactions of the Royal Society B*, *374*(1782). https://doi.org/10.1098/RSTB.2018.0342

Stark, R., Grzelak, M., & Hadfield, J. (2019). RNA sequencing: the teenage years. *Nature Reviews Genetics*, *20*(11), 631–656. https://doi.org/10.1038/s41576-019-0150-2

Stone, E., Zeale, M. R. K., Newson, S. E., Browne, W. J., Harris, S., & Jones, G. (2015a). Managing conflict between bats and humans: The response of Soprano Pipistrelles (*Pipistrellus pygmaeus*) to exclusion from roosts in houses. *PLOS ONE*, *10*(8), e0131825. https://doi.org/10.1371/journal.pone.0131825

Travanty, N. V., Ponnusamy, L., Kakumanu, M. L., Nicholson, W. L., & Apperson, C. S. (2019). Diversity and structure of the bacterial microbiome of the American dog tick, *Dermacentor variabilis*, is dominated by the endosymbiont *Francisella*. *Symbiosis*, *79*(3), 239–250. https://doi.org/10.1007/S13199-019-00642-2/FIGURES/7

Tuttle, M. D., Kiser, M., & Kiser, S. (2013). *The bat house builder's handbook*. University of Texas Press.

Van Brussel, K., & Holmes, E. C. (2022). Zoonotic disease and virome diversity in bats. In *Current Opinion in Virology* (Vol. 52, pp. 192–202). Elsevier B.V. https://doi.org/10.1016/j.coviro.2021.12.008

Vandegrift, K. J., & Kapoor, A. (2019). The ecology of new constituents of the tick virome and their relevance to public health. *Viruses 2019, Vol. 11, Page 529*, *11*(6), 529. https://doi.org/10.3390/V11060529

Vargina, S. G., Kuchuk, L. A., Gershtein, V. I., & Karas, F. R. (1982). Transmission of Issyk Kul virus by *Argas vespertilionis* ticks in experiment. *Sborn Nauch Tr Inst Virus Im Ivanov Akad Med Nauk SSSR*, 123–127.

Walker, P. J., Widen, S. G., Firth, C., Blasdell, K. R., Wood, T. G., Travassos Da Rosa, A. P. A., Guzman, H., Tesh, R. B., & Vasilakis, N. (2015). Genomic characterization of yogue, kasokero, issyk-kul, keterah, gossas, and thiafora viruses: nairoviruses naturally infecting bats, shrews, and ticks. *American Journal of Tropical Medicine and Hygiene*, *93*(5), 1041–1051. https://doi.org/10.4269/AJTMH.15-0344

Wang, B., Yang, X. Lou, Li, W., Zhu, Y., Ge, X. Y., Zhang, L. B., Zhang, Y. Z., Bock, C. T., & Shi, Z. L. (2017). Detection and genome characterization of four novel bat hepadnaviruses and a hepevirus in China. *Virology Journal*, *14*(1). https://doi.org/10.1186/S12985-017-0706-8

Wickham, H. (2016). ggplot2*: elegant graphics for data analysis*. Springer-Verlag New York.

Wille, M., Harvey, E., Shi, M., Gonzalez-Acuña, D., Holmes, E. C., & Hurt, A. C. (2020). Sustained RNA virome diversity in Antarctic penguins and their ticks. *The ISME Journal*, *14*(7), 1768. https://doi.org/10.1038/S41396-020-0643-1

Xu, Z., Feng, Y., Chen, X., Shi, M., Fu, S., Yang, W., Liu, W. J., Gao, G. F., & Liang, G. (2022). Virome of bat-infesting arthropods: highly divergent viruses in different vectors. *Journal of Virology*, *96*(4).

Yamaguti, N., Tipton, V. J., Keegan, H. L., & Toshioka, S. (1971). Ticks of Japan, Korea, and the Ryukyu islands. *Brigham Young University Science Bulletin, Biological Series*, *15*(1), 1.

Zabashta, M. V., Orlova, M. V., Pichurina, N. L., Khametova, A. P., Romanova, L. V., Borodina, T. N., & Zabashta, A. V. (2019). Participation of Bats (Chiroptera, Mammalia) and Their Ectoparasites in Circulation of Pathogens of Natural Focal Infections in the South of Russia. *Entomological Review 2019 99:4*, *99*(4), 513–521. https://doi.org/10.1134/S0013873819040110

Zhang, Y. Z., Wu, W. C., Shi, M., & Holmes, E. C. (2018). The diversity, evolution and origins of vertebrate RNA viruses. *Current Opinion in Virology*, *31*, 9–16. https://doi.org/10.1016/j.coviro.2018.07.017

Zhao, S., Yang, M., Liu, G., Hornok, S., Zhao, S., Sang, C., Tan, W., & Wang, Y. (2020). Rickettsiae in the common pipistrelle *Pipistrellus pipistrellus* (Chiroptera: Vespertilionidae) and the bat soft tick *Argas vespertilionis* (Ixodida: Argasidae). *Parasites and Vectors*, *13*(1), 1–6.

Zhmaeva, Z. M., Pchelkina, A. A., & Belashova, V. S. (1966). Spontaneous infection of *Argas vespertilionis* with *Rickettsia burnetii* in the south of Kazakhstan. *Meditsinskaia parazitologiia i parazitarnye bolezni*, *35*(5), 595–596.

Zhong, J., Jasinskas, A., & Barbour, A. G. (2007). Antibiotic treatment of the tick vector *Amblyomma americanum* reduced reproductive fitness. *PloS One*, *2*(5), e405. https://doi.org/10.1371/journal.pone.0000405

Zhou, S., Liu, B., Han, Y., Wang, Y., Chen, L., Wu, Z., & Yang, J. (2022). ZOVER: the database of zoonotic and vector-borne viruses. *Nucleic Acids Research*, *50*(D1), D943–D949. https://doi.org/10.1093/NAR/GKAB862

Chapter 7

Meta-transcriptomic identification of *Trypanosoma* spp. in native wildlife species from Australia

*Ayda Susana Ortiz-Baez, Kate Cousins, John-Sebastian Eden, Wei-Shan Chang, Erin Harvey, John H-O Pettersson, Scott Carver, Adam Polkinghorne, Jan Šlapeta, Karrie Rose and Edward C. Holmes**

Parasites & Vectors

Article

CHAPTER 7 Meta-transcriptomic identification of *Trypanosoma* spp. in native wildlife species from Australia

Ayda Susana Ortiz-Baez[1], Kate Cousins[1], John-Sebastian Eden[1,2], Wei-Shan Chang[1], Erin Harvey[1], John H.-O. Pettersson[1,3], Scott Carver[4], Adam Polkinghorne[5,6], Jan Šlapeta[7], Karrie Rose[8]and Edward C. Holmes[1*]

[1] Marie Bashir Institute for Infectious Diseases and Biosecurity, School of Life & Environmental Sciences and School of Medical Sciences, The University of Sydney, Sydney, NSW, Australia.

[2] Centre for Virus Research, Westmead Institute for Medical Research, Westmead, NSW, Australia.

[3] Zoonosis Science Center, Department of Medical Biochemistry and Microbiology, Uppsala University, Uppsala, Sweden.

[4] Department of Biological Sciences, University of Tasmania, Hobart, TAS, Australia.

[5] Department of Microbiology and Infectious Diseases, NSW Health Pathology, Nepean Hospital, Penrith, NSW, Australia.

[6] The University of Sydney Medical School, Nepean Clinical School, Faculty of Medicine and Health, University of Sydney, Penrith, NSW, Australia.

[7] Laboratory of Veterinary Parasitology, Sydney School of Veterinary Science, The University of Sydney, Sydney, NSW, Australia.

[8] Australian Registry of Wildlife Health, Taronga Conservation Society Australia, Mosman, NSW, Australia.

***** Correspondence: edward.holmes@sydney.edu.au

Received: 2 July 2020; Accepted: 30 August 2020; Published: 5 September 2020;

https://doi.org/10.1186/s13071-020-04325-6

The following version of this study may have slight differences from the original publication.

7.1 Abstract

Background: Wildlife species carry a remarkable diversity of trypanosomes. The detection of trypanosome infection in native Australian fauna is central to understanding their diversity and host-parasite associations. The implementation of total RNA sequencing (meta-transcriptomics) in trypanosome surveillance and diagnosis provides a powerful methodological approach to better understand the host species distribution of this important group of parasites.

Methods: We implemented a meta-transcriptomic approach to detect trypanosomes in a variety of tissues (brain, liver, lung, skin, gonads) sampled from native Australian wildlife, comprising four marsupials (koala, *Phascolarctos cinereus*; southern brown bandicoot, *Isoodon obesulus*; swamp wallaby, *Wallabia bicolor*; bare-nosed wombat, *Vombatus ursinus*), one bird (regent honeyeater, *Anthochaera phrygia*) and one amphibian (eastern dwarf tree frog, *Litoria fallax*). Samples corresponded to both clinically healthy and diseased individuals. Sequencing reads were *de novo* assembled into contigs and annotated. The evolutionary relationships among the trypanosomatid sequences identified were determined through phylogenetic analysis of *18S* rRNA sequences.

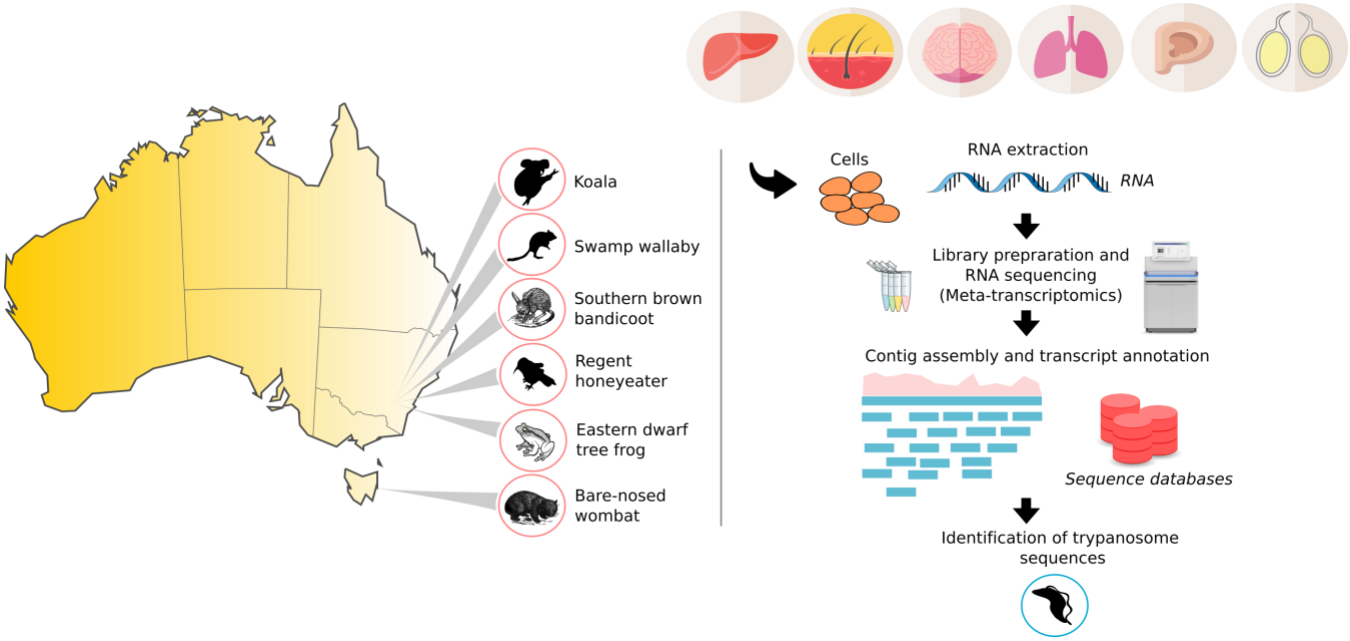
Results: We detected trypanosome sequences in all six species of vertebrates sampled, with positive samples in multiple organs and tissues confirmed by PCR. Phylogenetic analysis indicated that the trypanosomes infecting marsupials were related to those previously detected in placental and marsupial mammals, while the trypanosome in the regent honeyeater grouped with avian trypanosomes. In contrast, we provide the first evidence for a trypanosome in the eastern dwarf tree frog that was phylogenetically distinct from those described in other amphibians.

Conclusions: To our knowledge, this is the first meta-transcriptomic analysis of trypanosomes in native Australian wildlife, expanding the known genetic diversity of these important parasites. We demonstrated that RNA sequencing is sufficiently sensitive to detect low numbers of *Trypanosoma* transcripts and from diverse hosts and tissue types, thereby representing an effective means to detect trypanosomes that are divergent in genome sequence.

Keywords: Trypanosoma; Australia; Native fauna; Genetic diversity; Meta-transcriptomics; sequencing

---

7.2. Graphical abstract



7.3 Introduction

Trypanosomes are haemoprotozoan parasites that infect a wide range of animal taxa (Jakes et al., 2001; Mackie et al., 2017; Pinto et al., 2006). Endemic Australian fauna is a susceptible target for trypanosome infection, and several studies have revealed a remarkable diversity of trypanosomes in Australian wildlife (Barbosa et al., 2017; Cooper et al., 2018; Hamilton et al., 2005; Jakes et al., 2001; Thompson et al., 2014). This includes more than 15 species of exotic and endemic trypanosomes as well as several unclassified species (Cooper et al., 2017; Thompson et al., 2014). While some *Trypanosoma* species are associated with serious disease (Botero et al., 2013; Godfrey et al., 2018), others play an undetermined role in the health of their hosts. For instance, the native trypanosomes *Trypanosoma copemani* and *T. vegrandis* have been associated with population declines of woylies (*Bettongia penicillata*) in Western Australia (WA) (Botero et al., 2013; Godfrey

et al., 2018). It is likely that a similar phenomenon extends to other marsupial species, highlighting the need for continued surveillance (Cooper et al., 2017; McInnes et al., 2011).

To date, most trypanosome surveillance has been directed toward screening Australian mammals (i.e. bats, marsupials, monotremes, and rodents). Marsupials, in particular, have been widely screened, allowing the identification of several trypanosome species (e.g. *T. copemani, T. irwini, T. gilletti*) (McInnes et al., 2009, 2011; Paparini et al., 2011; Thompson et al., 2014). However, trypanosome infection has also been detected in other Australian vertebrate wildlife such as amphibians, birds, fish and reptiles (Cooper et al., 2017; Mackerras & Mackerras, 1961; O'Donoghue & Adlard, 2000). Moreover, trypanosomes have been detected in hematophagous invertebrates that become infected while feeding on infected vertebrate hosts and which may act as parasite vectors (Kreier, 2013; Spodareva et al., 2018). For example, in Australia, trypanosomes have been found in both aquatic leeches and ticks (Hamilton et al., 2005; Harvey 2019; Krige et al., 2019). Other invertebrates such as lice, culicid mosquitoes, sand-flies, and tabanid flies are also believed to be potential trypanosome vectors (Argañaraz et al., 2001; Bartlett-Healy et al., 2009b; Fermino et al., 2019b; Ferreira et al., 2008; Nuttall, 1908; Svobodová et al., 2017; Svobodová & Rádrová, 2018). However, because incidental infection during feeding is not necessarily associated with vector competence, further research is needed to determine the role of these haematophagous invertebrates in trypanosome infection and transmission (Cooper et al., 2017; Krige et al., 2019).

Multiple trypanosome species have been documented in Australian wildlife. For example, surveillance in marsupials recorded up to five species (*T. irwini, T. gilletti, T. copemani, T. vegrandis* and *T. noyesi*) in koalas (Barbosa et al., 2017), with similar results in woylies and the southern brown bandicoot (Godfrey et al., 2018; Paparini et al., 2011). In addition, the monitoring of Australian mammals has shown that *Trypanosoma* spp. are

present in animals sampled on the east and west coasts of Australia, as well as Tasmania (Thompson et al., 2014). Despite this, there are clear gaps in sampling, and it is likely that trypanosomes are widespread across the Australian continent and in mammalian species (Thompson et al., 2014).

Diagnosis of *Trypanosoma* infection largely relies on microscopy and a variety of molecular techniques (Hutchinson & Stevens, 2018). PCR-based Sanger sequencing of genetic markers constitutes the gold-standard for molecular diagnosis of *Trypanosoma*, including the *18S* rRNA gene in the small subunit rRNA (SSU), and the region encoding the glycosomal glyceraldehyde phosphate dehydrogenase (gGAPDH), an enzyme involved in the glycolytic pathway (Hamilton et al., 2004). In recent years, a number of studies have implemented amplicon-based next-generation sequencing (NGS) to reveal the genetic diversity of trypanosomes in Australian marsupials (Barbosa et al., 2017; Cooper et al., 2018). In comparison with conventional methods, NGS is able to detect low copy number of trypanosome sequences and target multiple genes with both high-throughput and accuracy. In addition, the development of meta-transcriptomics (i.e. bulk RNA sequencing) has enabled the detection and quantification of the transcripts expressed in the intra- and extracellular environments, including those derived from trypanosomes and other pathogens (Galen et al., 2020a), and hence represents an increasingly valuable diagnostic tool (Shakya et al., 2019; Stark et al., 2019c; Z. Wang et al., 2009).

Herein, we employed, for the first-time, a meta-transcriptomics approach as a method for the identification and surveillance of *Trypanosoma* in wildlife, screening different tissues from a variety of native Australian species. From this, we identified trypanosomes in several vertebrate groups from New South Wales (NSW) and Tasmania (TAS), including the identification of a divergent *Trypanosoma* in an amphibian species.

7.4 Materials and Methods

7.4.1 Sample collection

Most samples in this study were collected by the Australian Registry for Wildlife Health (ARWH) during monitoring surveys of wildlife, as well as from road-kill cases in NSW. The bare-nosed wombats were derived from road-kill in southern Tasmania. Following dissection, all tissue samples were stored at -80 °C until molecular analysis (**Table 7.1**). In total, we analysed 17 samples from different Australian native animal species, including four marsupials (koala, Phascolarctos cinereus; southern brown bandicoot, Isoodon obesulus; swamp wallaby, Wallabia bicolor; bare-nosed wombat, Vombatus ursinus), one bird (regent honeyeater, Anthochaera phrygia) and one amphibian (eastern dwarf tree frog, Litoria fallax). The amphibian specimen corresponded to a male diagnosed with severe, multisystemic, chronic trypanosomiasis (**Additional file 1: Figure S7.1**) and presumptive testicular Myxobolus-like infection. All individuals were identified to the lowest taxonomic level. Our sample set contained both healthy and diseased individuals (**Table 7.1**).

7.4.2 Sample processing

In brief, total RNA was extracted from a variety of sample tissues (**Table 7.1**) using the RNeasy® Mini Kit (Qiagen) according to the manufacturer's instructions. Sequencing libraries were generated using the TruSeq Stranded Total RNA Library Preparation protocol (Illumina) with host ribosomal RNA (rRNA) depletion (RiboZero Gold – Epidemiology). Subsequently, paired-end (100 bp) sequencing of the cDNA libraries was performed using the Illumina HiSeq 2000 system targeting at least 20M paired reads per library. All library preparation and sequencing were carried out by the Australian Genome Research Facility (AGRF).

7.4.3 Meta-transcriptomic analysis

Sequence reads were trimmed for quality using the Trimmomatic tool (Bolger et al., 2014a) and assembled *de novo* into contigs using Trinity v. 2.5.1 (Grabherr et al., 2011) with default parameter settings. The relative abundance of transcripts was quantified as the number of transcripts per kilobase million (TPM). In short, this metric normalizes transcript abundance by transcript length and sequencing depth. For sequence identification, particularly of trypanosomes, the assembled contigs were compared against the NCBI GenBank nucleotide (nt) and non-redundant protein (nr) databases using BLASTN and DIAMOND v.0.9.32 (Buchfink et al., 2015) (**Table 7.2 and 7.3; Additional file 2: Table S7.1**). Those contigs that exhibited matches to known trypanosome sequences with an e-value > $1x10^{-70}$ were retained for downstream analyses. Further, contigs corresponding to the stably expressed host mitochondrial marker, cytochrome C oxidase subunit 1 (*Cox1*), were identified based on sequence alignments using DIAMOND. All contigs were aligned to reference sequences using BBMap v.37.98 and cross-validated to DIAMOND results to verify that the matches correspond to the vertebrate host. Abundance was quantified as the sum of relative abundances of contigs for the marker. Sequence contigs were annotated as follows: (i) to find conserved domains and classify protein families, sequences were compared against the Conserved Domain Database (CDD) (Marchler-Bauer et al., 2015) and InterProScan (http://www.ebi.ac.uk/interpro/), (ii) for gene assignment, all putative trypanosome contigs were aligned against a custom reference sequence database (genome assembly ASM21029v1) using DIAMOND (Buchfink et al., 2015).

7.4.4 Confirmatory PCR

All samples included in this study were screened for Trypanosoma infection via PCR assays using primers targeting 2136-bp (outer) and 320-bp (nested) fragments of the 18S rRNA (**Additional file 3: Table S7.2**). In general, the cDNA was synthesised from up to 100 ng of

total RNA using random hexamers and SuperScript™ VILO™ (Invitrogen, CA, USA). The RT-PCR reactions proceeded as follows: 10 min of random priming at 25 °C, 20 min of extension at 50 °C, and 5 min of RT denaturation at 85 °C. Similarly, the PCR reactions with Platinum™ SuperFi™ (Invitrogen) were performed as follows: 1 min of hot start at 98 °C, followed by 40 cycles consisting of denaturation at 98 °C for 10 s, primer annealing for 10 s, and then extension at 72 °C according to conditions described in **Additional file 3: Table S7.2**. A final elongation step was run at 72 °C for 1 min. PCR products were visualized by electrophoresis on a 1.5% agarose gel stained with ethidium bromide. Controls were included to identify potential cross-contamination in reagents.

7.4.5 Phylogenetic analysis

The trypanosome contigs obtained here were compared with homologous sequences retrieved from GenBank, using 18S rRNA as a key phylogenetic marker (**Table 3, Additional file 4: Table S7.3**). Multiple sequence alignment (n = 81) was conducted using the E-INS-i algorithm in MAFFT v7.450. The best-fit model of nucleotide substitution (i.e. GTR+F+I+G4) was determined by using the Akaike Information Criterion (AIC) in the ModelFinder program (Kalyaanamoorthy, Minh, Wong, Von Haeseler, et al., 2017) implemented in IQ-TREE v1.6.7 (Nguyen et al., 2015b). Phylogenetic relationships were then inferred using the maximum likelihood method (Felsenstein, 1981) available in IQ-TREE v1.6.7 (Nguyen et al., 2015b). Nodal support values were also assessed by using a SH-like approximate Likelihood Ratio Test (SH-aLRT) and 1000 ultrafast bootstrap (UFBoot) replicates (Guindon et al., 2010).

Table 7.1 Characterization of samples from Australian vertebrates that tested positive for trypanosome infection.

Library	Host	No. of individuals examined	Tissue	Health status	Location	PCR result (*n*)
Vert1	Swamp wallaby (*Wallabia bicolor*) #1	1	Brain	Severe pulmonary congestion and oedema	Pittwater	Positive
Vert11	Regent honeyeater (*Anthochaera phrygia*)	1	Lung	Unknown	Sydney basin	Positive
Vert18	Bare-nosed wombat (*Vombatus ursinus*)	5	Liver	Healthy	Southern Tasmania	Positive (*n* = 3); negative (*n* = 2)
Vert21	Southern brown bandicoot (*Isoodon obesulus*)	1	Tail skin	Proliferative to ulcerative skin lesions	Sydney basin	Positive
Vert22	Koala (*Phascolarctos cinereus*)	7	Liver	Healthy	Sydney basin	Positive (*n* = 5); negative (*n* = 2)
Vert43	Eastern dwarf tree frog (*Litoria fallax*)	1	Testes and liver	Diseased	Kooragang island (NSW)	Positive
Vert48	Swamp wallaby (*Wallabia bicolor*) #2	1	Liver and ear	Lumpy jaw and ear lesions	Mimosa National Park (NSW)	Positive

Note: Libraries are indicated using alphanumeric codes and represent the collection of RNA fragments generated per sample for RNA sequencing

Abbreviation: n, number of samples

7.5 Results

7.5.1 Detection of Trypanosoma in screened samples

Using a meta-transcriptomic approach, we successfully identified trypanosome transcripts in six Australian species sampled in NSW and TAS, corresponding to the animal classes Amphibia, Aves and Mammalia. Trypanosome transcripts were detected in 60% (3 out of 5) of bare-nosed wombats, 71.43% (5 out of 7) of koalas, in both of the swamp wallaby samples, reagent honeyeater (n = 1), southern brown bandicoot tail (n = 1), and the eastern dwarf tree frog (n = 1). In total, trypanosomes were detected in 76.47% (13/17) of the individuals screened. With respect to target tissues, we detected trypanosome transcripts across a variety of tissues in infected individuals (**Table 7.1**), and positive samples were collected from both apparently healthy and diseased individual animals.

Despite the widespread presence of Trypanosoma in the samples characterized, we observed marked variation in the abundance and number of de novo assembled contigs among libraries. In general, the host cox1 transcripts were ~60% to ~99% more abundant than trypanosome transcripts (**Table 7.2**). Since samples showing high abundance of host cox1 also exhibited variable levels of abundance for trypanosome transcripts, these results suggest that the variation in abundance levels among samples was not due to biases in sampling processing. In addition, most transcripts were detected in the swamp wallaby #2 sample (n = 314, i.e. 0.05% of total transcripts per library) followed by the eastern dwarf tree frog (n = 149, i.e. 0.03% of total transcripts per library), whereas the lowest number of transcripts was identified in the regent honeyeater (n = 3, i.e. 0.0008% of total transcripts per library) (**Table 7.3; Additional file 2: Table S7.1**). Top BLAST hits ranged from 241 bp to 2258 bp, targeting regions corresponding to the transcribed spacers (ITS1, ITS2) and the 5.8S rRNA, 18S rRNA and 28S rRNA of the large subunit of the ribosome.

Similarly, we recovered hits against uncharacterized proteins, the surface protease GP63, and the heat shock proteins (HSPs) of Trypanosoma.

Table 7.2 Contigs with Blast hits to the small subunit (SSU) *18S* rRNA in the nt/nr database

Host	Contig accession	Length	TPM Tryp	e-value	Hit	Gene *SSU*	TPM *cox*1
Swamp wallaby (*Wallabia bicolor*) #1	VERT1_DN159759_c0_g1_i1*	299	3.27	9E-152	*Trypanosoma* sp. TL.AQ.22	*18S* rRNA	30192.26
	VERT1_DN215626_c0_g1_i1*	318	3.06	3E-162	*Trypanosoma* sp. TL.AQ.45	*18S* rRNA	
Regent honeyeater (*Anthochaera phrygia*)	VERT11_DN10127_c0_g1_i1*	666	2.82	0.00E+00	*Trypanosoma thomasbancrofti*	*18S* rRNA	512.02
Bare-nosed wombat (*Vombatus ursinus*)	VERT18_DN14693_c0_g1_i1*	615	2.94	0.00E+00	*Trypanosoma* sp.	*18S* rRNA	3805.74
	VERT18_DN33207_c0_g1_i1	241	3.42	1E-118	*Trypanosoma* sp. AB-2013	*18S* rRNA	
	VERT18_DN9224_c0_g1_i1	491	1.87	0.00E+00	*Trypanosoma* sp.	*18S* rRNA	
Southern brown bandicoot (*Isoodon obesulus*)	VERT21_DN254377_c0_g1_i1*	411	0.64	0.00E+00	*Trypanosoma* sp. LM-2010	*18S* rRNA	577.36
Koala (*Phascolarctos cinereus*)	VERT22_DN394953_c0_g1_i1*	241	0.86	3E-118	*Trypanosoma irwini*	*18S* rRNA	2622.98
Eastern dwarf tree frog (*Litoria fallax*)	VERT43_DN68004_c3_g3_i2*	1728	46.71	0.00E+00	*Trypanosoma* sp. 858	*18S* rRNA	1258.51
Swamp wallaby (*Wallabia bicolor*) #2	VERT48_DN150018_c0_g6_i1	718	55.01	0.00E+00	*Trypanosoma pestanai* LEM 110	*18S* rRNA	2152.22
	VERT48_DN190740_c0_g1_i1*	433	3.59	0.00E+00	*Trypanosoma* sp. H26	*18S* rRNA	
	VERT48_DN367248_c0_g1_i1*	890	743.4	0.00E+00	*Trypanosoma* sp. LM-2010	*18S* rRNA	

***** Contigs used for phylogenetic analysis based on the composition chi-square test performed by IQ-TREE

Table 7.3 Summary of top *Trypanosoma* hits from BLAST in the nt/nr database

Host	No. of contigs with hits for *Trypanosoma*	Length of best hit contig	Best BLAST hits against nr (DIAMOND)	Region	Best hit e-value	Gene	Best BLAST hits against nt/nr	Best hit e-value	Region
Swamp wallaby (*Wallabia bicolor*) #1	8	513	*T. theileri*	Uncharacterized protein	6.10E-49	TM35_00006 3140	*T. minasense*	0.00E+00	18S rRNA, ITS1, 5.8S rRNA, ITS2, 28S rRNA
Regent honeyeater (*Anthochaera phrygia*)	3	421	*T. theileri*	Uncharacterized protein	5.50E-48	TM35_00006 3130	*T. minasense*	0.00E+00	18S rRNA, ITS1, 5.8S rRNA, ITS2, 28S rRNA,
Bare-nosed wombat (*Vombatus ursinus*)	5	539	*T. theileri*	Uncharacterized protein	9.30E-32	TM35_00006 3140	*T. pestanai*	0.00E+00	28S rRNA
Southern brown bandicoot (*Isoodon obesulus*)	7	703	*T. theileri*	Uncharacterized protein	7.30E-37	TM35_00006 3140	*T. rangeli*	0.00E+00	28S rRNA
Koala (*Phascolarctos cinereus*)	24	241	*T. theileri*	Uncharacterized protein	5.80E-34	TM35_00006 3130	*T. theileri*	4e-106	Uncharacterized protein
Eastern dwarf tree frog (*Litoria fallax*)	149	1267	*T. cruzi*	Heat-shock protein 85, putative, partial	1.80E-195	Tco025E_097 08	*T. conorhini*	0.00E+00	Heat-shock protein 90
Swamp wallaby (*Wallabia bicolor*) #2	314	2258	*T. cruzi*	PWU95505.1 putative surface protease GP63	1.7e-143	TM35_00006 3130	*Trypanosoma grayi* surface protease GP63 partial mRNA	1e-60	Surface protease GP63

To place trypanosome sequences into a phylogenetic context (see below), and hence achieve taxonomic assignment, we identified the contigs targeting the 18S rRNA of the SSU. Abundance levels of 18S rRNA contigs ranged from 0.64 to 743.40 TPM. The highest cumulative abundances were identified in the eastern dwarf tree frog (TPM = 46.71) and the swamp wallaby #2 (TPM = 802) (**Table 7.2**), while the Southern brown bandicoot showed the lowest values (TPM = 0.64). In comparison, the host reference gene cox1 was abundantly expressed across samples (TPM: 512.02–30,192.26), with the highest levels observed in the swamp wallaby #1 sample (TPM = 30,192.26).

To validate these results, we used PCR assays and generic primers targeting the 18S rRNA gene (**Additional file 3: Table S7.2**) to detect trypanosome infection in all samples analyzed. Samples comprised a number of organs and tissues, including brain (n = 1), ear (n = 1), liver (n = 14), lung (n = 1), tail (n = 1), and testes (n = 1). A 320-bp nested fragment corresponding to the 18S rRNA was amplified in all samples containing trypanosomes, as previously identified by meta-transcriptomics (**Table 7.1**).

7.5.1 Phylogenetic analysis of Trypanosoma-positive samples

Phylogenetic analysis revealed that trypanosomes infecting the Australian native species covered in our study were generally closely related to known trypanosome species (**Figure 7.1**). We identified trypanosome sequences in the specimens of the swamp wallaby that fell into two separate clades associated with placental and marsupial mammals. However, most samples grouped with different trypanosomes identified from marsupials, forming a group that we term the “Marsupialia” clade (**Figure 7.1)**. This clade can be further divided into two groups: the first includes trypanosomes from the wallaby and the southern brown bandicoot, while the second group contained trypanosomes from the wallaby and bare-nosed wombat. Strikingly, the trypanosome from the koala fell into a different clade that is related to T. gennarii (nucleotide sequence similarity of 81.30%) and T. freitassi (82.04%) identified in South American marsupials (Monodelphis spp.), T. bennetti (92.56%) in birds (Falco sparverius) and T. irwini (98.75%) in

koalas. Moreover, we identified a trypanosome species in the regent honeyeater that is closely related to the avian trypanosomes T. thomasbancrofti and T. avium that share ~100% and 97% sequence similarity, respectively. Sequence comparisons against avian genotypes 1-4, classification sensu (Šlapeta et al., 2016), showed a perfect match with genotype 1 of T. thomasbancrofti (**Additional file 5: Table S7.4**), indicating that the regent honeyeater trypanosome likely belongs to that species.

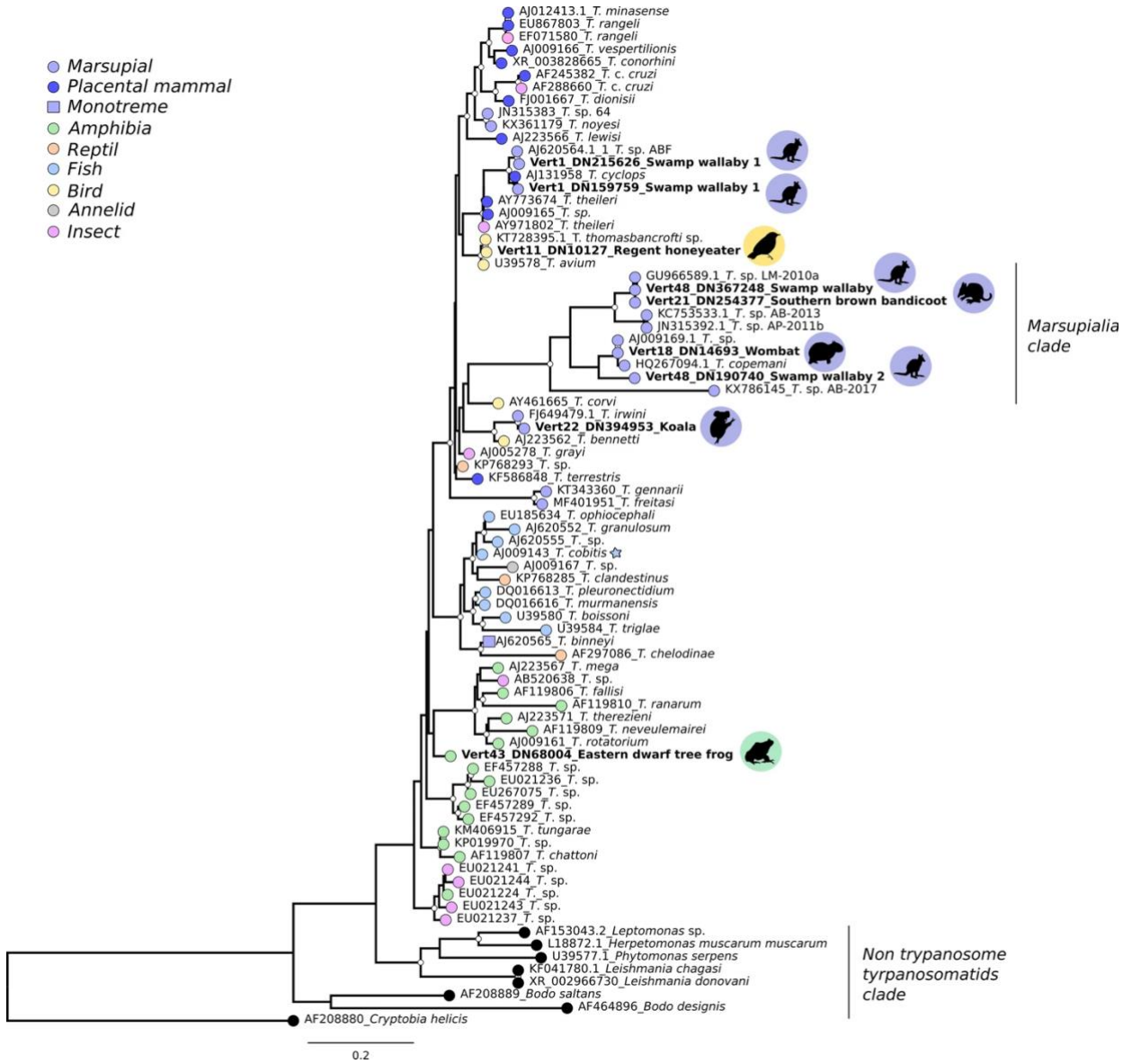


Figure 7.1 Maximum likelihood phylogenetic tree depicting the evolutionary relationships among trypanosomes sampled here (branch labels in bold) and background representative sequences. Branch

tips are colored according to the host of sampling. Trypanosomes detected in fish and annelids are indicated by a star. Animal silhouettes represent the hosts that tested positive for trypanosome infection. Node support values (SH-aLRT > 80% and UFBoot > 95%) are indicated with white circle node shapes in the tree. Trypanosome *T*. sp. ABF was also identified in a specimen from NSW.

In addition to the trypanosomes related to mammals and birds, we identified a trypanosome species infecting the eastern dwarf tree frog that was divergent from other trypanosomes in amphibians (**Additional file 1: Figure S7.1**). Notably, this amphibian trypanosome was related to those present in other amphibians, reptile and insect species, although it fell in a phylogenetically divergent position in the clade (with relatively strong support; SH-aLRT 89.6%; UFBoot 76%) and hence represents a novel lineage. The position of the dwarf tree frog sequence remained unchanged following additional analyses including a broader range of fish, reptile and leech transcriptomes (Spodareva et al., 2018), indicating that it is not an artefact due to biases in taxon sampling (**Additional file 6: Figure S7.2**).

7.6 Discussion

We have, to our knowledge for the first time, implemented a meta-transcriptomic approach for detecting *Trypanosoma* spp., investigating a variety of wildlife species indigenous to Australia. Unlike conventional methods for trypanosome diagnosis (cellular culture, PCR assays, and Sanger sequencing) (Noyes et al., 1999), meta-transcriptomics represents an unbiased approach for the detection of parasite diversity within samples, only requiring sufficient levels of gene expression (Galen et al., 2020a). To date, only a few surveillance studies have applied NGS technologies for the detection of trypanosomes in wildlife, although this approach is able to identify mixed trypanosome infections in marsupials and effectively screen their ectoparasites (Barbosa et al., 2017; Cooper et al., 2018). Using total RNA sequencing we identified trypanosomes in four marsupials, one bird and one amphibian species, highlighting the ability of this approach to detect parasites in a range of host species and target tissues (**Table 7.1**). Hence, meta-transcriptomics enables the detection of trypanosomes in a broad

range of samples that might include symptomatic and subclinical infections, different stages of disease, as well as variable levels of parasitemia.

Most of the trypanosome transcripts identified in the hosts analyzed were associated with genes encoding ribosomal components, suggesting that ribosome biogenesis and protein synthesis have a central role in the infection process (**Tables 7.2, 7.3**). In the case of the heat-shock protein 90 (Hsp90) identified in the eastern dwarf tree frog, the presence of this molecular chaperone has been associated with transitions across trypanosome life-cycle stages (Pallavi et al., 2010). Hsp90 synthesis induction has also been related to stress responses in *T. cruzi*, reflecting the change in temperature when the parasite moves from the vector to the mammalian host (G. Palmer et al., 1995; Pérez-Morales et al., 2012). Hsp90 is also known to play an essential role in protein folding and degradation under normal conditions (Dunn, n.d.; Hoter et al., 2018). The major surface protease GP63 identified in swamp wallaby #2 is a highly immunogenic antigen involved in macrophage-parasite interaction encoded by a multi-copy gene that also occurs in *Leishmania* (Donelson et al., 1998; LaCount et al., 2003). Differential expression of GP63 is associated with the parasite life-cycle, with genetic variation facilitating immune evasion and colonization (Donelson et al., 1998; Guerbouj et al., 2001).

Previous studies have suggested that trypanosomes often have deleterious effects on the health of the infected hosts (Barbosa et al., 2017; Botero et al., 2013; Godfrey et al., 2018; McInnes et al., 2011; Thompson et al., 2014a). As the trypanosomes described here were detected in both healthy and diseased individuals, we are unable to make inferences on their capacity to cause disease (**Table 7.1**). Indeed, many of the health conditions manifest in the animals studied were unspecific or prone to be associated with other sort of infections. For instance, the pulmonary congestion and oedema in the swamp wallaby #1 sample may be consistent with orbivirus infection symptoms (family *Reoviridae*) (K. Rose et al., 2012), while the pox-like lesions in the southern brown bandicoot have been previously associated with infection by the Bandicoot papillomatosis and carcinomatosis virus (BPCV2) (*Polyomaviridae*) in the western barred bandicoot (*Perameles bougainville*) (Woolford et al., 2007). Similarly,

although the ear lesions in the swamp wallaby 2 could be attributed to the trypanosome infection, other causative pathogens could be associated with the lumpy jaw and emaciation (Keane et al., 1977; McLelland, 2019). In addition, the eastern dwarf tree frog was co-infected with *Trypanosoma* and *Myxobolus*, confounding the association of disease with any etiological agent. Because our study was limited to vertebrates, it does not provide insights into the potential vector involved in parasite transmission. However, as suggested in previous studies, it is possible that both ticks and dipterans (i.e. flies and mosquitoes) are vectors of these parasites as they can feed on a large variety of hosts including mammals, birds and amphibians (Barbosa et al., 2017; Ferreira et al., 2008; Harvey 2019; Kato et al., 2010; Krige et al., 2019; Muzari, 2010; Svobodová et al., 2017). Some hemipterans might also play a vectorial role in the transmission of trypanosomes in sylvatic and peridomestic settings, as documented in the Americas (Buitrago et al., 2016; Cortez et al., 2006; Kjos et al., 2013). Clearly, more research is needed to clarify the vectors and the mode of trypanosome transmission in Australian wildlife (Cooper et al., 2017; Harvey et al., 2019; Krige et al., 2019).

Phylogenetic analysis revealed that the trypanosomes identified in native Australian fauna fell into different lineages that were largely concordant with that of the host species from which they were sampled, although we were unable to make taxonomic assignments to the species level. Notably, we identified three distinct clades of marsupial trypanosomes (**Figure 7.1**). The trypanosome species detected in the swamp wallaby that fell outside the Marsupialia clade was closely related to *Trypanosoma* sp. ABF previously described in the swamp wallaby in NSW (Cooper et al., 2017), and to *T. cyclops*, an exotic trypanosome isolated from the monkey *Macaca nemestrina* and related to *T. theileri*-related trypanosomes in ruminants and tabanids. The relatedness among these trypanosome species raises concerns over the potential susceptibility of Australian vectors and vertebrates to infection by exotic trypanosomes and hence the establishment of a zoonotic transmission cycle (Cooper et al., 2017; Thompson et al., 2014). In addition, although most marsupial trypanosomes analyzed fell into the Australian Marsupialia clade, trypanosome species infecting these mammals did not form a monophyletic group, indicative of a history of cross-species transmission (Hamilton et al., 2007).

Among the trypanosome species infecting marsupials, *T. irwini, T. gilletti, and T. copemani, T. vegrandis, T. noyesi* and *T.* sp. AB-2017 have been described in koalas (Barbosa et al., 2017; McInnes et al., 2009; C. K. Thompson et al., 2014). Our results indicated that *T.* sp detected in the koala was closely related to *T. irwini* and the avian exotic trypanosome *T. bennetti.* Given than the former has been also identified in koalas, the trypanosome detected in the sampled koala likely corresponds to *T. irwini*. The close relationship between the *T. irwini* and *T. bennetti* has been previously documented (Cooper et al., 2017; McInnes et al., 2009) and is compatible with the hypothesis that hosts sharing similar environments and vectors are susceptible to related parasites (i.e. "host-fitting") (Cooper et al., 2017; Dario et al., 2017). This provides an explanation for the relationship between trypanosomes infecting arboreal fauna inhabiting distant regions.

The trypanosome sequence we identified in the regent honeyeater likely belongs to *T. thomasbancrofti* (genotype 1), and *T. thomasbancrofti* was originally described in the regent honeyeater (Šlapeta et al., 2016). This trypanosome species has been suggested to be a culicid-vectored parasite and has been detected in healthy captive and wild regent honeyeaters (Šlapeta et al., 2016). In contrast, *T. avium* was identified in the rook (*Corvus frugilegus*) and associated with serious disease and death in birds, with suggestions that it is transmitted by blackflies (*Simulium* spp.) (Tarello, 2005; Votýpka et al., 2002) and phlebotomine sandflies (Svobodová & Rádrová, 2018). Hence, our data corroborated the presence of *T. thomasbancrofti* in the regent honeyeater and highlight the importance of parasitological surveillance in the wild for this species classified as critically endangered (CR) (*sensu* IUCN).

Of particular interest was the case of the trypanosome detected in the eastern dwarf tree frog that was related to those identified in amphibians, reptile, and insect species. Since this amphibian trypanosome fell in a divergent and basal position within the clade it might represent a new trypanosome species and hence merits further characterization (**Additional file 1: Figure S7.1; Additional file 6: Figure S7.2**). Interestingly, considering the clinical diagnosis

of the frog sampled (see Methods) as well as its transcript abundance (**Table 7.3**), it is possible that this trypanosome species or the synergistic infection by *Trypanosoma* with *Myxobolus* might have detrimental effects on amphibian health. This clearly merits further investigation. To our knowledge, this is the first report of a trypanosome in the eastern dwarf tree frog (**Additional file 1: Figure S7.1**), although amphibians are known to be parasitized by different trypanosomes species (Bardsley & Harmsen, 1973; Johnston, 1916; O'Donoghue & Adlard, 2000; Spodareva et al., 2018; Werner & Walewski, 1976) and some have been documented in Australian amphibians (Cleland & Johnston, 1910; Johnston, 1916; O'Donoghue & Adlard, 2000). That the clade containing the eastern dwarf tree frog sequence also contains a trypanosome infecting sand flies tentatively suggests that dipterans or other invertebrates could play a role vectoring trypanosome transmission (Kato et al., 2010).

While our study was focused on samples collected from multiple organs and tissues, meta-transcriptomics has been shown to be an efficient approach for characterizing blood parasites, even when they are at low abundance (Cassin-Sackett, 2020; Galen et al., 2020a). In addition, the technique has also been used to detect trypanosome sequence in the blood meals of *Ixodes holocyclus* and *Aedes camptorhynchus* (Harvey et al., 2019; Shi et al., 2017a). Hence, when combined with more traditional approaches, meta-transcriptomics offers a promising way to shed new light on the ecology and epidemiological surveillance of parasites in nature, although the approach is costly, requires extensive computational resources and may be unable to detect genes that are not expressed to sufficient levels (Cassin-Sackett, 2020).

7.7 Conclusions

To our knowledge, this is the first meta-transcriptomic analysis of trypanosomes in native Australian wildlife, expanding the known genetic diversity of these important parasites. Our findings highlight the diversity of trypanosomes infecting an important spectrum of Australian native fauna. We also demonstrated that RNA sequencing is sufficiently sensitive to

detect low levels of *Trypanosoma* transcripts from diverse hosts and tissues types, and hence represents an effective means to detect trypanosomes that are divergent in genome sequence.

7.8 Supplementary Material

The following are available online at https://doi.org/10.1186/s13071-020-04325-6.

Additional file 1: Figure S7.1. Light microphotograph of the promastigote phase of *T.* sp. in giemsa-stained blood film from *Litoria fallax.* Scale bar represents 10µm.
Additional file 2: Table S7.1. Summary of contigs with hits against trypanosome sequences at the NCBI nucleotide (nt) database. Relative abundance was calculated for each contig as transcripts per million (TPM).
Additional file 3: Table S7.2. List of PCR primers used in this study for confirmation of trypanosome infection.
Additional file 4: Table S7.3. List of sequences used for phylogenetic analysis.
Additional file 5: Table S4. Pairwise sequence identity among SSU 18S rRNA sequences of avian trypanosomes belonging to genotypes 1-4 and the putative *T. thomasbancrofti* identified in this study. Genotype classification sensu Šlapeta *et al*. 2016.
Additional file 6: Figure S7.2. Maximum likelihood tree showing phylogenetic relationships among trypanosomes within the aquatic clade based on the SSU *18S rRNA* gene. The trypanosome identified in *Litoria fallax* is indicated in blue. The hosts of trypanosomes are indicated with colour-coded tips.

Availability of data and materials

All data generated or analysed during this study are included in this published article and its additional information files. The newly generated contig sequences were deposited in the GenBank database under the accession numbers MT732373-MT732384. All new sequence reads are available at the NCBI Sequence Read Archive (SRA) database under the BioProject accession PRJNA626677 (BioSample accessions: SAMN15401543 - SAMN1540159). The dataset

supporting the conclusions of this article is available in the figshare repository, https://figshare.com/s/d9c281ada61d8a8ed884.

7.9 References

Argañaraz, E. R., Hubbard, G. B., Ramos, L. A., Ford, A. L., Nitz, N., Leland, M. M., Vandeberg, J. L., & Teixeira, A. R. L. (2001). Blood-sucking lice may disseminate Trypanosoma cruzi infection in baboons. *Revista Do Instituto de Medicina Tropical de Sao Paulo*, *43*(5), 271–276. https://doi.org/10.1590/S0036-46652001000500007

Barbosa, A. D., Gofton, A. W., Paparini, A., Codello, A., Greay, T., Gillett, A., Warren, K., Irwin, P., & Ryan, U. (2017). Increased genetic diversity and prevalence of co-infection with *Trypanosoma* spp. in koalas (*Phascolarctos cinereus*) and their ticks identified using next-generation sequencing (NGS). *PLoS ONE*, *12*(7), 1–20. https://doi.org/10.1371/journal.pone.0181279

Barbosa, A., Reiss, A., Jackson, B., Warren, K., Paparini, A., Gillespie, G., Stokeld, D., Irwin, P., & Ryan, U. (2017). Prevalence, genetic diversity and potential clinical impact of blood-borne and enteric protozoan parasites in native mammals from northern Australia. *Veterinary Parasitology*, *238*, 94–105. https://doi.org/10.1016/j.vetpar.2017.04.007

Bardsley, J. E., & Harmsen, R. (1973). The Trypanosomes of Anura. *Advances in Parasitology*, *11*(C), 1–73. https://doi.org/10.1016/S0065-308X(08)60184-0

Bartlett-Healy, K., Crans, W., & Gaugler, R. (2009). Vertebrate Hosts and Phylogenetic Relationships of Amphibian Trypanosomes from a Potential Invertebrate Vector, Culex territans Walker (Diptera: Culicidae). *Journal of Parasitology*, *95*(2), 381–387. https://doi.org/10.1645/ge-1793.1

Bolger, A. M., Lohse, M., & Usadel, B. (2014). Trimmomatic: a flexible trimmer for Illumina sequence data. *Bioinformatics (Oxford, England)*, *30*(15), 2114–2120. https://doi.org/10.1093/bioinformatics/btu170

Botero, A., Thompson, C. K., Peacock, C. S., Clode, P. L., Nicholls, P. K., Wayne, A. F., Lymbery, A. J., & Thompson, R. C. A. (2013). Trypanosomes genetic diversity, polyparasitism and the population decline of the critically endangered Australian marsupial, the brush tailed bettong or woylie

(*Bettongia penicillata*). *International Journal for Parasitology: Parasites and Wildlife*, *2*, 77–89. https://doi.org/10.1016/J.IJPPAW.2013.03.001

Buchfink, B., Xie, C., & Huson, D. H. (2015). Fast and sensitive protein alignment using DIAMOND. *Nature Methods*, *12*(1), 59–60. https://doi.org/10.1038/nmeth.3176

Buitrago, R., Bosseno, M. F., Depickère, S., Waleckx, E., Salas, R., Aliaga, C., Barnabé, C., & Brenière, S. F. (2016). Blood meal sources of wild and domestic *Triatoma infestans* (Hemiptera: Reduviidae) in Bolivia: Connectivity between cycles of transmission of *Trypanosoma cruzi*. *Parasites and Vectors*, *9*(1). https://doi.org/10.1186/s13071-016-1499-0

Cassin-Sackett, L. (2020). Promising protocols for parasites: Metatranscriptomics improves detection of hyperdiverse but low abundance communities. *Molecular Ecology Resources*, *20*(1), 8–10. https://doi.org/10.1111/1755-0998.13122

Cleland, J., & Johnston, T. (1910). The haematozoa of Australian batrachians. *Journal and Proceedings of the Royal Society of New South Wales*, *44*, 252–261.

Cooper, C., Clode, P. L., Peacock, C., & Thompson, R. C. A. (2017). Host–Parasite Relationships and Life Histories of Trypanosomes in Australia. *Advances in Parasitology*, *97*, 47–109. https://doi.org/10.1016/BS.APAR.2016.06.001

Cooper, C., Keatley, S., Northover, A., Gofton, A. W., Brigg, F., Lymbery, A. J., Pallant, L., Clode, P. L., & Thompson, R. C. A. (2018). Next generation sequencing reveals widespread trypanosome diversity and polyparasitism in marsupials from Western Australia. *International Journal for Parasitology: Parasites and Wildlife*, *7*(1), 58–67. https://doi.org/10.1016/J.IJPPAW.2018.01.005

Cortez, M. R., Pinho, A. P., Cuervo, P., Alfaro, F., Solano, M., Xavier, S. C. C., D'Andrea, P. S., Fernandes, O., Torrico, F., Noireau, F., & Jansen, A. M. (2006). *Trypanosoma cruzi* (Kinetoplastida Trypanosomatidae): Ecology of the transmission cycle in the wild environment of the Andean valley of Cochabamba, Bolivia. *Experimental Parasitology*, *114*(4), 305–313. https://doi.org/10.1016/j.exppara.2006.04.010

Dario, M. A., Lisboa, C. V., Costa, L. M., Moratelli, R., Nascimento, M. P., Costa, L. P., Leite, Y. L. R., Llewellyn, M. S., Xavier, S. C. das C., Roque, A. L. R., & Jansen, A. M. (2017). High *Trypanosoma* spp. diversity is maintained by bats and triatomines in Espírito Santo state, Brazil. *PLOS ONE*, *12*(11), e0188412. https://doi.org/10.1371/journal.pone.0188412

Donelson, J. E., Hill, K. L., & El-Sayed, N. M. A. (1998). Multiple mechanisms of immune evasion by African trypanosomes. *Molecular and Biochemical Parasitology*, *91*(1), 51–66. https://doi.org/10.1016/S0166-6851(97)00209-0

Dunn, B. M. (n.d.). *Frontiers in protein and peptide sciences. Volume 1*.

Felsenstein, J. (1981). Evolutionary trees from DNA sequences: A maximum likelihood approach. *Journal of Molecular Evolution*, *17*(6), 368–376. https://doi.org/10.1007/BF01734359

Fermino, B. R., Paiva, F., Viola, L. B., Rodrigues, C. M. F., Garcia, H. A., Campaner, M., Takata, C. S. A., Sheferaw, D., Kisakye, J. J., Kato, A., Jared, C. A. G. S., Teixeira, M. M. G., & Camargo, E. P. (2019). Shared species of crocodilian trypanosomes carried by tabanid flies in Africa and South America, including the description of a new species from caimans, Trypanosoma kaiowa n. sp. *Parasites & Vectors*, *12*(1), 225. https://doi.org/10.1186/s13071-019-3463-2

Ferreira, R. C., De Souza, A. A., Freitas, R. A., Campaner, M., Takata, C. S. A., Barrett, T. V., Shaw, J. J., & Teixeira, M. M. G. (2008). A phylogenetic lineage of closely related trypanosomes (Trypanosomatidae, Kinetoplastida) of anurans and sand flies (Psychodidae, Diptera) sharing the same ecotopes in Brazilian Amazonia. *Journal of Eukaryotic Microbiology*, *55*(5), 427–435. https://doi.org/10.1111/j.1550-7408.2008.00342.x

Galen, S. C., Borner, J., Williamson, J. L., Witt, C. C., & Perkins, S. L. (2020). Metatranscriptomics yields new genomic resources and sensitive detection of infections for diverse blood parasites. *Molecular Ecology Resources*, *20*(1), 14–28. https://doi.org/10.1111/1755-0998.13091

Godfrey, S. S., Keatley, S., Botero, A., Thompson, C. K., Wayne, A. F., Lymbery, A. J., Morris, K., & Thompson, R. C. A. (2018). Trypanosome co-infections increase in a declining marsupial population. *International Journal for Parasitology: Parasites and Wildlife*, *7*(2), 221–227. https://doi.org/10.1016/J.IJPPAW.2018.06.002

Grabherr, M. G., Haas, B. J., Yassour, M., Levin, J. Z., Thompson, D. A., Amit, I., Adiconis, X., Fan, L., Raychowdhury, R., Zeng, Q., Chen, Z., Mauceli, E., Hacohen, N., Gnirke, A., Rhind, N., di Palma, F., Birren, B. W., Nusbaum, C., Lindblad-Toh, K., … Regev, A. (2011). Full-length transcriptome assembly from RNA-Seq data without a reference genome. *Nature Biotechnology*, *29*(7), 644–652. https://doi.org/10.1038/nbt.1883

Guerbouj, S., Victoir, K., Guizani, I., Seridi, N., Nuwayri-Salti, N., Belkaid, M., Ben Ismail, R., Ray, D. L. E., & Dujardin, J. C. (2001). *Gp63* gene polymorphism and population structure of Leishmania donovani complex: Influence of the host selection pressure? *Parasitology*, *122*(1), 25–35. https://doi.org/10.1017/S0031182000007125

Guindon, S., Dufayard, J.-F., Lefort, V., Anisimova, M., Hordijk, W., & Gascuel, O. (2010). New Algorithms and Methods to Estimate Maximum-Likelihood Phylogenies: Assessing the Performance of PhyML 3.0. *Systematic Biology*, *59*(3), 307–321. https://doi.org/10.1093/sysbio/syq010

Hamilton, P. B., Gibson, W. C., & Stevens, J. R. (2007). Patterns of co-evolution between trypanosomes and their hosts deduced from ribosomal RNA and protein-coding gene phylogenies. *Molecular Phylogenetics and Evolution*, *44*(1), 15–25. https://doi.org/10.1016/j.ympev.2007.03.023

Hamilton, P. B., Stevens, J. R., Gaunt, M. W., Gidley, J., & Gibson, W. C. (2004). Trypanosomes are monophyletic: evidence from genes for glyceraldehyde phosphate dehydrogenase and small subunit ribosomal RNA. *International Journal for Parasitology*, *34*(12), 1393–1404. https://doi.org/10.1016/j.ijpara.2004.08.011

Hamilton, P. B., Stevens, J. R., Gidley, J., Holz, P., & Gibson, W. C. (2005). A new lineage of trypanosomes from Australian vertebrates and terrestrial bloodsucking leeches (*Haemadipsidae*). *International Journal for Parasitology*, *35*(4), 431–443. https://doi.org/10.1016/J.IJPARA.2004.12.005

Harvey, E., Rose, K., Eden, J. S., Lo, N., Abeyasuriya, T., Shi, M., Doggett, S. L., & Holmes, E. C. (2019). Extensive diversity of RNA viruses in australian ticks. *Journal of Virology*, *93*(3). https://doi.org/10.1101/386573

Hoter, A., El-Sabban, M. E., & Naim, H. Y. (2018). The HSP90 family: Structure, regulation, function, and implications in health and disease. In *International Journal of Molecular Sciences* (Vol. 19, Issue 9). MDPI AG. https://doi.org/10.3390/ijms19092560

Hutchinson, R., & Stevens, J. R. (2018). Barcoding in trypanosomes. In *Parasitology* (Vol. 145, Issue 5, pp. 563–573). Cambridge University Press. https://doi.org/10.1017/S0031182017002049

Jakes, K. A., O'Donoghue, P. J., & Adlard, R. D. (2001). Phylogenetic relationships of *Trypanosoma chelodina* and *Trypanosoma binneyi* from Australian tortoises and platypuses inferred from small subunit rRNA analyses. *Parasitology*, *123*(5), 483–487. https://doi.org/10.1017/S0031182001008721

Johnston, T. H. (1916). *A census of the endoparasites recorded as occurring in Queensland, arranged under their hosts*. Royal Society of Queensland.

Kalyaanamoorthy, S., Minh, B. Q., Wong, T. K. F., Von Haeseler, A., & Jermiin, L. S. (2017). ModelFinder: Fast model selection for accurate phylogenetic estimates. *Nature Methods*, *14*(6), 587–589. https://doi.org/10.1038/nmeth.4285

Kato, H., Uezato, H., Sato, H., Bhutto, A. M., Soomro, F. R., Baloch, J. H., Iwata, H., & Hashiguchi, Y. (2010). Natural infection of the sand fly Phlebotomus kazeruni by Trypanosoma species in Pakistan. *Parasites & Vectors*, *3*(1), 10. https://doi.org/10.1186/1756-3305-3-10

Keane, C., Taylor, M. R. H., Wilson, P., Smith, L., Cunningham, B., Devine, P., & English, L. F. (1977). Bacteroides ruminicola as a possible cause of 'lumpy-jaw' in Bennett's wallabies. *Veterinary Microbiology*, *2*(2), 179–183. https://doi.org/10.1016/0378-1135(77)90007-4

Kjos, S. A., Marcet, P. L., Yabsley, M. J., Kitron, U., Snowden, K. F., Logan, K. S., Barnes, J. C., & Dotson, E. M. (2013). Identification of bloodmeal sources and *Trypanosoma cruzi* infection in Triatomine bugs (Hemiptera: Reduviidae) from residential settings in Texas, the United States. *Journal of Medical Entomology*, *50*(5), 1126–1139. https://doi.org/10.1603/me12242

Kreier, J. P. (2013). *Parasitic Protozoa* (10th ed.). Elsevier Science.

Krige, A.-S., Thompson, R. C. A., & Clode, P. L. (2019). Hang on a tick – Are ticks really the vectors for Australian trypanosomes? *Trends in Parasitology*, *35*(8), 596–606. https://doi.org/10.1016/j.pt.2019.05.008

LaCount, D. J., Gruszynski, A. E., Grandgenett, P. M., Bangs, J. D., & Donelson, J. E. (2003). Expression and function of the *Trypanosoma brucei* major surface protease (GP63) genes. *Journal of Biological Chemistry*, *278*(27), 24658–24664. https://doi.org/10.1074/jbc.M301451200

Mackerras, M., & Mackerras, I. (1961). The haematozoa of Australian frogs and fish. *Australian Journal of Zoology*, *9*(1), 123. https://doi.org/10.1071/ZO9610123

Mackie, J., Stenner, R., Gillett, A., Barbosa, A., Ryan, U., & Irwin, P. (2017). Trypanosomiasis in an Australian little red flying fox (*Pteropus scapulatus*). *Australian Veterinary Journal*, *95*(7), 259–261. https://doi.org/10.1111/avj.12597

Marchler-Bauer, A., Derbyshire, M. K., Gonzales, N. R., Lu, S., Chitsaz, F., Geer, L. Y., Geer, R. C., He, J., Gwadz, M., Hurwitz, D. I., Lanczycki, C. J., Lu, F., Marchler, G. H., Song, J. S., Thanki, N., Wang, Z., Yamashita, R. A., Zhang, D., Zheng, C., & Bryant, S. H. (2015). CDD: NCBI's conserved domain database. *Nucleic Acids Research*, *43*(D1), D222–D226. https://doi.org/10.1093/nar/gku1221

McInnes, L. M., GIillett, A., Hanger, J., Reid, S. A., & Ryan, U. M. (2011). The potential impact of native Australian trypanosome infections on the health of koalas (*Phascolarctos cinereus*). *Parasitology*, *138*(7), 873–883. https://doi.org/10.1017/S0031182011000369

McInnes, L. M., Gillett, A., Ryan, U. M., Austen, J., Campbell, R. S. F., Hanger, J., & Reid, S. A. (2009). *Trypanosoma irwini* n. sp (Sarcomastigophora: Trypanosomatidae) from the koala (*Phascolarctos cinereus*). *Parasitology*, *136*(8), 875–885. https://doi.org/10.1017/S0031182009006313

McLelland, D. (2019). *Macropod progressive periodontal disease ('lumpy-jaw')* (p. 796).

Muzari, M. O. (2010). *Tabanid flies and potential transmission of* Trypanosoma evansi *in Queensland*.

Nguyen, L.-T., Schmidt, H. A., von Haeseler, A., & Minh, B. Q. (2015). IQ-TREE: a fast and effective stochastic algorithm for estimating maximum-likelihood phylogenies. *Molecular Biology and Evolution*, *32*(1), 268–274. https://doi.org/10.1093/molbev/msu300

Noyes, H. a, Stevens, J. R., Teixeira, M., Phelan, J., & Holz, P. (1999). A nested PCR for the ssrRNA gene detects *Trypanosoma binneyi* in the platypus and *Trypanosoma* sp. in wombats and kangaroos in Australia. *International Journal for Parasitology*, *29*(2), 331–339. https://doi.org/10.1016/S0020-7519(98)00167-2

Nuttall, G. H. F. (1908). The transmission of trypanosoma lewisi by fleas and lice. *Parasitology*, *1*(4), 296–301. https://doi.org/10.1017/S0031182000003589

O'Donoghue, P. J., & Adlard, R. D. (2000). *Catalogue of protozoan parasites recorded in Australia*.

Pallavi, R., Roy, N., Nageshan, R. K., Talukdar, P., Pavithra, S. R., Reddy, R., Venketesh, S., Kumar, R., Gupta, A. K., Singh, R. K., Yadav, S. C., & Tatu, U. (2010). Heat shock protein 90 as a drug target against protozoan infections: Biochemical characterization of HSP90 from plasmodium

falciparum and *Trypanosoma evansi* and evaluation of its inhibitor as a candidate drug. *Journal of Biological Chemistry*, *285*(49), 37964–37975. https://doi.org/10.1074/jbc.M110.155317

Palmer, G., Louvion, J. F., Tibbetts, R. S., Engman, D. M., & Picard, D. (1995). *Trypanosoma cruzi* heat-shock protein 90 can functionally complement yeast. *Molecular and Biochemical Parasitology*, *70*(1–2), 199–202. https://doi.org/10.1016/0166-6851(95)00007-N

Paparini, A., Irwin, P. J., Warren, K., McInnes, L. M., de Tores, P., & Ryan, U. M. (2011). Identification of novel trypanosome genotypes in native Australian marsupials. *Veterinary Parasitology*, *183*(1–2), 21–30. https://doi.org/10.1016/J.VETPAR.2011.07.009

Pérez-Morales, D., Lanz-Mendoza, H., Hurtado, G., Martínez-Espinosa, R., & Espinoza, B. (2012). *Proteomic analysis of* Trypanosoma cruzi *epimastigotes subjected to heat shock*.

Pinto, C. M., Ocaña-Mayorga, S., Lascano, M. S., & Grijalva, M. J. (2006). Infection by trypanosomes in marsupials and rodents associated with human dwellings in Ecuador. *Journal of Parasitology*, *92*(6), 1251–1255. https://doi.org/10.1645/GE-886R.1

Rose, K., Kirkland, P., Davis, R., Cooper, D., Blumstein, D., Pritchard, L., Newberry, K., & Lunt, R. (2012). Epizootics of sudden death in tammar wallabies (*Macropus eugenii*) associated with an orbivirus infection. *Australian Veterinary Journal*, *90*(12), n/a-n/a. https://doi.org/10.1111/j.1751-0813.2012.00993.x

Shakya, M., Lo, C.-C., & Chain, P. S. G. (2019). Advances and challenges in metatranscriptomic analysis. *Frontiers in Genetics*, *10*, 904. https://doi.org/10.3389/FGENE.2019.00904

Shi, M., Neville, P., Nicholson, J., Eden, J.-S., Imrie, A., & Holmes, E. C. (2017). High-resolution metatranscriptomics reveals the ecological dynamics of mosquito-associated RNA viruses in Western Australia. *Journal of Virology*, *91*(17), 1–17. https://doi.org/10.1128/jvi.00680-17

Šlapeta, J., Morin-Adeline, V., Thompson, P., McDonell, D., Shiels, M., Gilchrist, K., Votýpka, J., & Vogelnest, L. (2016). Intercontinental distribution of a new trypanosome species from Australian endemic Regent Honeyeater (Anthochaera phrygia). *Parasitology*, *143*(8), 1012–1025. https://doi.org/10.1017/S0031182016000329

Spodareva, V. V., Grybchuk-Ieremenko, A., Losev, A., Votýpka, J., Lukeš, J., Yurchenko, V., & Kostygov, A. Y. (2018b). Diversity and evolution of anuran trypanosomes: insights from the

study of European species. *Parasites & Vectors*, *11*(1), 447. https://doi.org/10.1186/s13071-018-3023-1

Stark, R., Grzelak, M., & Hadfield, J. (2019). RNA sequencing: the teenage years. *Nature Reviews Genetics*, *20*(11), 631–656. https://doi.org/10.1038/s41576-019-0150-2

Svobodová, M., Dolnik, O. V., Čepička, I., & Rádrová, J. (2017). Biting midges (Ceratopogonidae) as vectors of avian trypanosomes. *Parasites & Vectors*, *10*(1), 224. https://doi.org/10.1186/s13071-017-2158-9

Svobodová, M., & Rádrová, J. (2018). Phlebotomine sandflies – potential vectors of avian trypanosome. *Acta Protozoologica*, *57*(1), 53–59. https://doi.org/10.4467/16890027AP.18.005.8399

Tarello, W. (2005). *Trypanosoma avium* incidence, pathogenicity and response to melarsomine in falcons from Kuwait. *Parasite*, *12*(1), 85–87. https://doi.org/10.1051/parasite/2005121085

Thompson, C. K., Godfrey, S. S., & Thompson, R. C. A. (2014). Trypanosomes of Australian mammals: A review. *International Journal for Parasitology: Parasites and Wildlife*, *3*(2), 57–66. https://doi.org/10.1016/J.IJPPAW.2014.02.002

Thompson, C. K., Wayne, A. F., Godfrey, S. S., & Thompson, R. (2014a). Temporal and spatial dynamics of trypanosomes infecting the brush-tailed bettong (*Bettongia penicillata*): a cautionary note of disease-induced population decline. *Parasites & Vectors*, *7*(1), 169. https://doi.org/10.1186/1756-3305-7-169

Votýpka, J., Oborník, M., Volf, P., Svobodová, M., & Lukeš, J. (2002). *Trypanosoma avium* of raptors (Falconiformes): Phylogeny and identification of vectors. *Parasitology*, *125*(3), 253–263. https://doi.org/10.1017/S0031182002002093

Wang, Z., Gerstein, M., & Snyder, M. (2009). RNA-Seq: a revolutionary tool for transcriptomics. *Nature Reviews Genetics*, *10*(1), 57–63. https://doi.org/10.1038/nrg2484

Werner, J. K., & Walewski, K. (1976). Amphibian Trypanosomes from the McCormick Forest, Michigan. *The Journal of Parasitology*, *62*(1), 20. https://doi.org/10.2307/3279035

Woolford, L., Rector, A., Van Ranst, M., Ducki, A., Bennett, M. D., Nicholls, P. K., Warren, K. S., Swan, R. A., Wilcox, G. E., & O'Hara, A. J. (2007). A novel virus detected in papillomas and carcinomas of the endangered Western barred bandicoot (*Perameles bougainville*) exhibits genomic

features of both the Papillomaviridae and Polyomaviridae. *Journal of Virology*, *81*(24), 13280–13290. https://doi.org/10.1128/jvi.01662-07

Chapter 8

General discussion

CHAPTER 8 General discussion

Over the last decade, our knowledge of the RNA virosphere has been vividly transformed by innovations in genome sequencing and computational biology, as well as increased sampling efforts globally. The discovery of RNA viruses described in Chapters 2 to 6 not only expands the reaches of the known RNA virosphere and the sequence databases that harbour and describe virus diversity, but it also highlights the importance of covering a broader and more diverse repertoire of hosts (Harvey & Holmes, 2022; Junglen & Drosten, 2013b). For instance, the discovery of *Lauta virus* (Chapter 2) is the first report of an articulavirus in reptiles, although viruses within this order had previously been detected in other vertebrate lineages, including fish and birds. In this manner, we are also able to unmask a notable fraction of the hidden diversity of RNA viruses to the extent that sampling biases in virus discovery projects are beginning to be minimized (Obbard, 2018a). The expansion of the known RNA virosphere has also become evident given the recent trends in the vast quantities of novel viruses found at large-scale virome studies (Edgar et al., 2022; Li et al., 2015; Neri et al., 2022; Shi et al., 2018; Shi et al., 2016).

Attempts to characterize the RNA virome also face the challenge of detecting those viruses that exist within the so-called viral dark matter: sequences that are so highly divergent that they lack any detectable primary sequence similarity with known viruses in sequence databases. The implementation of more sensitive approaches for virus identification, including protein structure prediction, profile-based methods and machine learning (i.e., artificial intelligence) are promising to transform the field of virus discovery, shedding new light on the dark viral matter (Kelley et al., 2015; Khot et al., 2020; J. Park et al., 1998; Steinegger et al., 2019). With respect to protein structure prediction, the low number of hallmark protein structures such as capsids and RdRps available for comparison is currently a major limitation since the lack of highly conserved viral structures might lead to false-positive identification (e.g. misclassification) and under-detection. As discussed more in detail in Chapter 2, the enrichment of the PDB and AlphaFold databases with viral 3D protein structures will be a critical step to

identifying viruses that are currently out of reach using sequence similarity searches. Further improvements in computational methods are needed to distinguish true evolutionary homology from structural convergence, as well as more reliable statistical estimates to assign score thresholds to profile-based comparisons (Park et al., 1998).

As well as offering a broad perspective of the scale of the virosphere, the coupling of meta-transcriptomics and virus discovery pipelines has also contributed to revealing the composition and distribution of the virome of natural organisms. In this thesis, I exploited these approaches to identify numerous viruses from different taxonomic categories, including unclassified and divergent viruses within the *Riboviria* (i.e., RNA viruses). For instance, in Chapters 2 and 3 I identified novel viruses within the order *Articulavirales* by analyzing fish and reptile samples, whereas newly discovered viruses in the *Bunyavirales* were reported in a variety of invertebrate hosts, as showed in Chapters 4–6. Not only do these findings increase our understanding of virus diversity, but they also point to possible virus-host associations. However, despite clear progress is this area, accurately determining virus-host associations remains a major challenge, especially when considering holobiont hosts, since the processing of eukaryote host samples often includes a variety of symbionts and dietary components that, in turn, carry RNA viruses (Cobbin et al., 2021). Indeed, accounting for misleading host-annotation based on sequence metadata is an important contributing factor to the host-misassignment of virus sequences in public databases. In addition, the presence of viruses with a broad host-spectrum, such as some members of the *Picornavirales* and *Mononegavirales*, exacerbates the difficulty for host assignment given their association with diverse host organisms including vertebrates, invertebrates, plants, and protists. A practical example of this is addressed in Chapter 4 where I aimed to assess the virome in mosquitoes and their parasitic mites. Revealing virus-host associations for uncultured RNA viruses requires a combination of multiple but scope-limited approaches, including abundance comparison, sequence composition analyses, phylogenetic contextualization, state reconstruction, small RNA profiling, etc. (Cobbin et al., 2021; Longdon et al., 2015; Martínez-García et al., 2014). Therefore, more comprehensive methods integrating different aspects of the biology of hosts and their viruses, are required to establish precise

virus-host associations across different groups of RNA viruses identified from eukaryotic hosts (Longdon et al., 2015; Pons et al., 2021). Similarly, the development of experimental procedures and optimization of existing virological techniques that can be applied to uncultured viruses will enhance the virus discovery process and advance our understanding of virus-host interactions.

Determining virus-host affiliations goes hand in hand with the pressing need to understand virus-host interactions in nature. Despite the vertiginous expansion of the RNA virosphere, we lack essential research about the roles and impacts of RNA viruses on their hosts and environments (Roux & Emerson, 2022; Youle et al., 2012). This is also a clear limitation of present research work. Future studies investigating both the underlying dynamics of RNA viral communities, as well as their influence on the host biology, will provide a holistic insight into the ecology and evolution of virus-holobiont relationships and their impact on the surrounding environment (Youle et al., 2012). For instance, it has been suggested that viruses in soil and ocean environments participate in complex nutrient cycling process, food webs and microbial community dynamics (Emerson, 2019; Roux & Emerson, 2022). Likewise, mounting evidence pinpoints a variety of roles for RNA viruses in biological processes rather than imposing exclusively detrimental effects on hosts health. As a case in point, RNA viruses have shown to drive the antiviral response in arthropods via an RNA interference pathway and modulate the interactions between plants and insects (Schoelz & Stewart, 2018; Vogel et al., 2019). Accordingly, in Chapter 4, the identification of patterns in the RNA virus diversity at the host-parasite interface between mosquitoes and their ectoparasites implies major virus-virus and virus-host interactions that remains unexplored and should be addressed in future studies.

Trans-kingdom interactions are also central to a better understanding of broad-scale RNA virus ecology. It has been shown that microbial composition can impact virus infection and transmission in dipteran insects (Johnson, 2015). In particular, the endosymbiotic bacterium *Wolbachia* has been associated with a range of mutualistic effects on insects, including antiviral protection against RNA viruses (Cao et al., 2019; Weeks et al., 2007). Chapter 5 offers a glimpse

on the effect of the *w*Au strain of *Wolbachia* on the virome of natural populations of *Drosophila simulans* in Western Australia. The lack of detectable antiviral protection was compatible with previous research conducted on *Drosophila spp.* (Osborne et al., 2009; Shi et al., 2018a). As discussed in detail in this chapter, antiviral protection was assessed in terms of resistance rather than tolerance. It is therefore still possible that co-infection with *wAu* increases the defensive capacity of *D. simulans* by limiting the impact of RNA viruses (Schneider & Ayres, 2008). The extent of antiviral protection might also be reduced against the natural RNA virome. However, additional questions that warrant further research remain: Can environmental conditions play a key role triggering the antiviral response mediated by *w*Au *Wolbachia?* Does antiviral protection depend on *Wolbachia wAu* density? Does *Wolbachia* confers an RNA virus-specific or generalist antiviral protection?

As noted throughout this thesis, meta-transcriptomics analyses enable the scrutiny of the whole repertoire of RNA sequences present in a sample. This approach has opened new windows of opportunity for exploring the RNA virosphere as well as the microbial diversity occurring in wildlife samples (Bashiardes et al., 2016; Galen et al., 2020b; Gofton et al., 2022; Shi et al., 2018b). Taking advantage of this, in Chapters 6-7 I investigated the occurrence of targeted microbes including those of potential public health and veterinary significance such as *Rickettsia* sp., *Coxiella* sp. and *Trypanosoma* spp. As a consequence, meta-transcriptomics was implemented for the first time for the detection of *Trypanosoma* spp. parasites, which were being identified in a variety of tissue samples and biological host groups. These results provide insights into the distribution and host-range of *Trypanosoma* spp. circulating in Australian wildlife species, and raises the question of whether these can sporadically affect humans. Aside from taxonomic characterization, more research is needed to establish the nature and extent of interactions between trypanosomes and their hosts, their pathogenic potential, and their mode of transmission in natural settings (Bartlett-Healy et al., 2009a; Fermino et al., 2019a). Likewise, the identification of *Rickettsia* sequences related to species within the spotted fever group in Chapter 6 suggests that there is a risk of vector-borne rickettsioses that merits further

investigation. Overall, these findings presented in this thesis contribute to the surveillance of zoonotic agents circulating in wildlife.

Arthropod-based sampling provides a powerful and non-invasive means for pathogen surveillance in wildlife species. Bats are considered major reservoir hosts for numerous zoonotic viruses and microbial agents (Calisher et al., 2006; Letko et al., 2020a). Thus, surveys of bat-ectoparasites are convenient in terms of minimizing the handling, stress, and disturbance of captured bats and their colonies. Likewise, the exposure between humans and bats, and hence possible zoonotic disease emergence, is reduced (Letko et al., 2020a). Consequently, arthropod sampling offers a dual-purpose approach to detect infectious agents associated with ticks and their vertebrate hosts (Batson et al., 2021a; Galen et al., 2020b; Gofton et al., 2022). In Chapter 6, I explored this approach by analyzing *Carios vespertilionis* ticks naturally detached from the Soprano pipistrelle bats in Sweden. The detection of Issyk-Kul virus was of particular interest given its association with febrile outbreaks in Central Asia (Alkhovsky et al., 2013b; Atkinson et al., 2015a; Lvov, 2019). Notably, this was the first record for Sweden and the second report in Europe, suggesting that the virus might have a cryptic and broader distribution range in the continent (Brinkmann et al., 2020). Among the newly discovered viruses, the occurrence of Gubbo nairovirus at high abundance levels across all the query libraries might indicate that it is a tick-borne virus. Comparisons between engorged and unfed ticks from different developmental stages will shed light on this knowledge gap. Furthermore, it is unclear whether Gubbo nairovirus is also able of infect and replicate in bat hosts (i.e. arbovirus) as well as any potential pathogenicity to vertebrate animals (Junming et al., 2018). The study and characterization of Gubbo nairovirus through experimental approaches, including cell culture-based methods, antigen-based assays, and reverse genetics, could provide valuable insights into the pathogenicity, cell tropism and gene function of this virus.

A broad comparison between the viromes identified in the invertebrate species studied throughout Chapters 4-6, suggests a higher diversity in mosquitoes in contrast with flies, ticks, and mites. The fact that mosquitoes can harbor a large fraction of insect-specific viruses (ISV) as

well as arthropod borne viruses (arboviruses), contributes to our understanding their virome composition, although it does not explain the origins nor the ecological drivers of this diversity (Batson et al., 2021a; de Almeida et al., 2021). A systematic comparison of the RNA virome, including endogenous viruses, and the immune response pathways to viral infections, could reveal potential patterns and determinants of virus diversity across Arthropoda.

In contrast, in Chapter 5 I described the detection of viruses such as nora virus, galbut virus, thika virus and La Jolla virus in *Drosophila simulans* that could represent core components of the *Drosophila* virome, which in turn suggests that these RNA viruses have long-term associations with their dipteran hosts (Shi et al., 2018a; Webster et al., 2015, 2016). This is clear an issue that needs to be considered further. For instance, expanding the characterization of RNA virome of *Drosophila* across different geographies could help test this hypothesis.

Although the vast majority of RNA viruses identified along Chapters 4-6 are likely associated with vertebrate and invertebrate hosts, the presence of viruses in the families *Mitoviridae*, *Totiviridae, Narnaviridae,* and *Tymoviridae* suggest likely associations with fungi and protists host as well as dietary components. While virus composition can reflect underlaying trophic interactions between hosts and symbionts in food webs, the possible impact of these viruses on the outcome of fungal or protist infections in the base host is unclear and similarly merits additional research (Ferrandon et al., 2007; Parratt & Laine, 2016). For example, mycoviruses of fungal pathogens parasitizing plants have been associated with reduced virulence (i.e. hypovirulence) in host plants (Nuss, 2005). Similarly, we routinely ignore the occurrence of virus horizontal transfer between co-infecting symbionts (Dolja & Koonin, 2018) although, for example, there is evidence for virus host switching between *Leishmania* and *Blechomonas* parasites in fleas (Grybchuk et al., 2018). Moreover, the presence of mitoviruses in the nuclear genome and mitochondria of plants support the idea of ancient virus transfer among plants and fungi (Bruenn et al., 2015; Roossinck, 2019). Clearly, greater information on these and other aspects of hyperparasitic infections will add important details to the viral tapestry of holobiont hosts.

It is noteworthy that many the novel viruses presented in Chapters 2-6 only represented partial viral genomes. This could be explained by several factors, including low input/quality samples and the difficulty in assembling highly variable regions in viral genomes. In theory, it is also possible that some of these partial viruses in fact have segmented genome structures and that there is heterogeneous sequence conservation across segments, resulting in the detection of segments encoding conserved proteins such as the RdRp and capsid but not more divergent segments (Obbard et al., 2020). To identify more divergent viral segments and sequences, it is necessary to employ a variety of approaches, such as screening viral sequences against the TSA database, identifying co-occurrence patterns between contigs, comparing virus abundance and coverage levels, and binning sequences (Batson et al., 2021; Obbard et al., 2020). Furthermore, even though the phylogenetic relationships for the newly discovered viruses were inferred based on the RdRp protein, it is possible that recombination or horizontal gene transfer will mean that different genes will produce incongruent phylogenetic signals (Boussau & Scornavacca, 2020). In some instances, as shown in Chapter 5, the presence of unclassified viruses will also require more extensive taxon sampling to assess their phylogenetic relationships at deep taxonomic levels. Notably, future efforts to characterize the complete genome of these viruses will further enhance our understanding of their evolution and genome organization.

In sum, recent advances in meta-transcriptomic sequencing have it made possible to explore the natural world in depth, opening up new avenues of research in virology and microbiology. In the studies presented in this thesis I used integrative approaches, including meta-transcriptomics and novel bioinformatic techniques, to investigate different aspects of the diversity, ecology and evolution of RNA viruses and targeted microbial life forms. Overall, the findings derived from this research contribute to our understanding the composition and evolution of the RNA virosphere, as well as to the surveillance of microbial agents in wildlife. As diverse research questions were addressed in this thesis, many open questions were also

identified. Continuing to fill these gaps in knowledge will help unravel the complex diversity and the eco-evolutionary dynamics of RNA viruses and microbes in nature.

References

Alkhovsky, S. v, Lvov, D. K., Shchelkanov, M. Y., Shchetinin, A. M., Deryabin, P. G., Samokhvalov, E. I., Gitelman, A. K., & Botikov, A. G. (2013). The taxonomy of the Issyk-Kul virus (ISKV, Bunyaviridae, Nairovirus), the etiologic agent of the Issyk-Kul fever isolated from bats (Vespertilionidae) and ticks Argas (Carios) vespertilionis (Latreille, 1796). *Voprosy Virusologii*, *58*(5), 11–15.

Atkinson, B., Marston, D. A., Ellis, R. J., Fooks, A. R., & Hewson, R. (2015). Complete Genomic Sequence of Issyk-Kul Virus. *Genome Announcements*, *3*(4), 662–677. https://doi.org/10.1128/genomea.00662-15

Bartlett-Healy, K., Crans, W., & Gaugler, R. (2009). Vertebrate hosts and phylogenetic relationships of amphibian trypanosomes from a potential invertebrate vector, Culex territans Walker (Diptera: Culicidae). *The Journal of Parasitology*, *95*(2), 381–387. https://doi.org/10.1645/GE-1793.1

Bashiardes, S., Zilberman-Schapira, G., & Elinav, E. (2016). Use of metatranscriptomics in microbiome research. *Bioinformatics and Biology Insights*, *10*, 19–25. https://doi.org/10.4137/BBI.S34610

Batson, J., Dudas, G., Haas-Stapleton, E., Kistler, A. L., Li, L. M., Logan, P., Ratnasiri, K., & Retallack, H. (2021). Single mosquito metatranscriptomics identifies vectors, emerging pathogens and reservoirs in one assay. *ELife*, *10*. https://doi.org/10.7554/eLife.68353

Boussau, B., & Scornavacca, C. (2020). Reconciling Gene trees with Species Trees. In *Phylogenetics in the genomic era*. https://hal.inria.fr/PGE.

Brinkmann, A., Kohl, C., Radonić, A., Dabrowski, P. W., Mühldorfer, K., Nitsche, A., Wibbelt, G., & Kurth, A. (2020). First detection of bat-borne Issyk-Kul virus in Europe. *Scientific Reports 2020 10:1*, *10*(1), 1–7. https://doi.org/10.1038/s41598-020-79468-8

Bruenn, J. A., Warner, B. E., & Yerramsetty, P. (2015). Widespread mitovirus sequences in plant genomes. *PeerJ*, *2015*(4). https://doi.org/10.7717/PEERJ.876/SUPP-5

Calisher, C. H., Childs, J. E., Field, H. E., Holmes, K. v., & Schountz, T. (2006). Bats: Important Reservoir Hosts of Emerging Viruses. *Clinical Microbiology Reviews*, *19*(3), 531–545. https://doi.org/10.1128/CMR.00017-06

Cao, L. J., Jiang, W., & Hoffmann, A. A. (2019). Life History Effects Linked to an Advantage for wAu Wolbachia in Drosophila. *Insects 2019, Vol. 10, Page 126*, *10*(5), 126. https://doi.org/10.3390/INSECTS10050126

Cobbin, J. C., Charon, J., Harvey, E., Holmes, E. C., & Mahar, J. E. (2021). Current challenges to virus discovery by meta-transcriptomics. *Current Opinion in Virology*, *51*, 48–55.

de Almeida, J. P., Aguiar, E. R., Armache, J. N., Olmo, R. P., & Marques, J. T. (2021). The virome of vector mosquitoes. In *Current Opinion in Virology* (Vol. 49, pp. 7–12). Elsevier B.V. https://doi.org/10.1016/j.coviro.2021.04.002

Dolja, V. v., & Koonin, E. v. (2018). Metagenomics reshapes the concepts of RNA virus evolution by revealing extensive horizontal virus transfer. *Virus Research*, *244*, 36–52.

Edgar, R. C., Taylor, J., Lin, V., Altman, T., Barbera, P., Meleshko, D., Lohr, D., Novakovsky, G., Buchfink, B., Al-Shayeb, B., Banfield, J. F., de la Peña, M., Korobeynikov, A., Chikhi, R., & Babaian, A. (2022). Petabase-scale sequence alignment catalyses viral discovery. *Nature 2022 602:7895*, *602*(7895), 142–147. https://doi.org/10.1038/s41586-021-04332-2

Emerson, J. B. (2019). Soil Viruses: A New Hope. *MSystems*, *4*(3). https://doi.org/10.1128/mSystems.00120-19

Fermino, B. R., Paiva, F., Viola, L. B., Rodrigues, C. M. F., Garcia, H. A., Campaner, M., Takata, C. S. A., Sheferaw, D., Kisakye, J. J., Kato, A., Jared, C. A. G. S., Teixeira, M. M. G., & Camargo, E. P. (2019). Shared species of crocodilian trypanosomes carried by tabanid flies in Africa and South America, including the description of a new species from caimans, Trypanosoma kaiowa n. sp. *Parasites & Vectors*, *12*(1). https://doi.org/10.1186/s13071-019-3463-2

Ferrandon, D., Imler, J. L., Hetru, C., & Hoffmann, J. A. (2007). The Drosophila systemic immune response: sensing and signalling during bacterial and fungal infections. *Nature Reviews Immunology 2007 7:11*, *7*(11), 862–874. https://doi.org/10.1038/nri2194

Galen, S. C., Borner, J., Williamson, J. L., Witt, C. C., & Perkins, S. L. (2020). Metatranscriptomics yields new genomic resources and sensitive detection of infections for diverse blood parasites. *Molecular Ecology Resources*, *20*(1), 14–28. https://doi.org/10.1111/1755-0998.13091

Gofton, A. W., Blasdell, K. R., Taylor, C., Banks, P. B., Michie, M., Roy-Dufresne, E., Poldy, J., Wang, J., Dunn, M., Tachedjian, M., & Smith, I. (2022). Metatranscriptomic profiling reveals diverse tick-borne bacteria, protozoans and viruses in ticks and wildlife from Australia. *Transboundary and Emerging Diseases*, *69*(5), e2389–e2407. https://doi.org/10.1111/tbed.14581

Grybchuk, D., Kostygov, A. Y., Macedo, D. H., Votýpka, J., Lukeš, J., & Yurchenko, V. (2018). Rna viruses in Blechomonas (Trypanosomatidae) and evolution of Leishmaniavirus. *MBio*, *9*(5). https://doi.org/10.1128/mbio.01932-18

Harvey, E., & Holmes, E. C. (2022). Diversity and evolution of the animal virome. *Nature Reviews Microbiology*, *0123456789*. https://doi.org/10.1038/s41579-021-00665-x

Johnson, K. N. (2015). Bacteria and antiviral immunity in insects. *Current Opinion in Insect Science*, *8*, 97–103. https://doi.org/10.1016/j.cois.2015.01.008

Junglen, S., & Drosten, C. (2013). Virus discovery and recent insights into virus diversity in arthropods. *Current Opinion in Microbiology*, *16*(4), 507–513. https://doi.org/10.1016/j.mib.2013.06.005

Kelley, L. A., Mezulis, S., Yates, C. M., Wass, M. N., & Sternberg, M. J. E. (2015). The Phyre2 web portal for protein modeling, prediction and analysis. *Nature Protocols 2015 10:6*, *10*(6), 845–858. https://doi.org/10.1038/nprot.2015.053

Khot, V., Strous, M., & Hawley, A. K. (2020). Computational approaches in viral ecology. *Computational and Structural Biotechnology Journal*, *18*, 1605–1612. https://doi.org/10.1016/j.csbj.2020.06.019

Letko, M., Seifert, S. N., Olival, K. J., Plowright, R. K., & Munster, V. J. (2020). Bat-borne virus diversity, spillover and emergence. *Nature Reviews Microbiology 2020 18:8*, *18*(8), 461–471. https://doi.org/10.1038/s41579-020-0394-z

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. (2015). Unprecedented genomic diversity of RNA viruses in arthropods reveals the ancestry of negative-sense RNA viruses. *ELife*, *4*(4). https://doi.org/10.7554/elife.05378

Longdon, B., Murray, G. G. R., Palmer, W. J., Day, J. P., Parker, D. J., Welch, J. J., Obbard, D. J., & Jiggins, F. M. (2015). The evolution, diversity, and host associations of rhabdoviruses. *Virus Evolution*, *1*(1). https://doi.org/10.1093/ve/vev014

Lvov, D. K. (2019). Issyk-Kul Fever. *The Arboviruses: Epidemiology and Ecology*, 53–62. https://doi.org/10.1201/9780429280276-4

Martínez-García, M., Santos, F., Moreno-Paz, M., Parro, V., & Antón, J. (2014). Unveiling viral–host interactions within the 'microbial dark matter.' *Nature Communications 2014 5:1*, *5*(1), 1–8. https://doi.org/10.1038/ncomms5542

Neri, U., Wolf, Y. I., Roux, S., Camargo, A. P., Lee, B., Kazlauskas, D., Chen, I. M., Ivanova, N., Zeigler Allen, L., Paez-Espino, D., Bryant, D. A., Bhaya, D., Krupovic, M., Dolja, V. v., Kyrpides, N. C., Koonin, E. v., Gophna, U., Narrowe, A. B., Probst, A. J., … Wang, Z. (2022). Expansion of the global RNA virome reveals diverse clades of bacteriophages. *Cell*. https://doi.org/10.1016/j.cell.2022.08.023

Nuss, D. L. (2005). Hypovirulence: Mycoviruses at the fungal–plant interface. *Nature Reviews Microbiology 2005 3:8*, *3*(8), 632–642. https://doi.org/10.1038/nrmicro1206

Obbard, D. J. (2018). Expansion of the metazoan virosphere: progress, pitfalls, and prospects. *Current Opinion in Virology*, *31*, 17–23. https://doi.org/10.1016/j.coviro.2018.08.008

Obbard, D. J., Shi, M., Roberts, K. E., Longdon, B., & Dennis, A. B. (2020). A new lineage of segmented RNA viruses infecting animals. *Virus Evolution*, *6*(1), 61. https://academic.oup.com/ve/article/6/1/vez061/5708932

Osborne, S. E., Leong, Y. S., O'Neill, S. L., & Johnson, K. N. (2009). Variation in Antiviral Protection Mediated by Different Wolbachia Strains in Drosophila simulans. *PLOS Pathogens*, *5*(11), e1000656. https://doi.org/10.1371/journal.ppat.1000656

Park, J., Karplus, K., Barrett, C., Hughey, R., Haussler, D., Hubbard, T., & Chothia, C. (1998). Sequence comparisons using multiple sequences detect three times as many remote homologues as pairwise methods. *Journal of Molecular Biology*, *284*(4), 1201–1210. https://doi.org/10.1006/jmbi.1998.2221

Parratt, S. R., & Laine, A. L. (2016). The role of hyperparasitism in microbial pathogen ecology and evolution. *The ISME Journal 2016 10:8*, *10*(8), 1815–1822. https://doi.org/10.1038/ismej.2015.247

Pons, J. C., Paez-Espino, D., Riera, G., Ivanova, N., Kyrpides, N. C., & Llabrés, M. (2021). VPF-Class: taxonomic assignment and host prediction of uncultivated viruses based on viral protein families. *Bioinformatics*, *37*(13), 1805–1813. https://doi.org/10.1093/bioinformatics/btab026

Roossinck, M. J. (2019). Evolutionary and ecological links between plant and fungal viruses. *New Phytologist*, *221*(1), 86–92. https://doi.org/10.1111/nph.15364

Roux, S., & Emerson, J. B. (2022). Diversity in the soil virosphere: to infinity and beyond? *Trends in Microbiology*, *30*(11), 1025–1035. https://doi.org/10.1016/j.tim.2022.05.003

Schneider, D. S., & Ayres, J. S. (2008). Two ways to survive infection: what resistance and tolerance can teach us about treating infectious diseases. *Nature Reviews Immunology 2008 8:11*, *8*(11), 889–895. https://doi.org/10.1038/nri2432

Schoelz, J. E., & Stewart, L. R. (2018). The Role of Viruses in the Phytobiome. *5*, 93–111. https://doi.org/10.1146/annurev-virology-092917-043421

Shi, J., Hu, Z., Deng, F., & Shen, S. (2018). Tick-Borne Viruses. *Virologica Sinica 2018 33:1*, *33*(1), 21–43. https://doi.org/10.1007/S12250-018-0019-0

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 2018 556:7700*, *556*(7700), 197–202. https://doi.org/10.1038/s41586-018-0012-7

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*, *540*(7634), 539–543. https://doi.org/10.1038/nature20167

Shi, M., White, V. L., Schlub, T., Eden, J. S., Hoffmann, A. A., & Holmes, E. C. (2018a). No detectable effect of Wolbachia wMel on the prevalence and abundance of the RNA virome of Drosophila melanogaster. *Proceedings of the Royal Society B: Biological Sciences*, *285*(1883). https://doi.org/10.1098/rspb.2018.1165

Shi, M., Zhang, Y. Z., & Holmes, E. C. (2018b). Meta-transcriptomics and the evolutionary biology of RNA viruses. *Virus Research*, *243*, 83–90. https://doi.org/10.1016/j.virusres.2017.10.016

Steinegger, M., Meier, M., Mirdita, M., Vöhringer, H., Haunsberger, S. J., & Söding, J. (2019). HH-suite3 for fast remote homology detection and deep protein annotation. *BMC Bioinformatics*, *20*(1), 473. https://doi.org/10.1186/s12859-019-3019-7

Vogel, E., Santos, D., Mingels, L., Verdonckt, T. W., & Broeck, J. vanden. (2019). RNA interference in insects: Protecting beneficials and controlling pests. *Frontiers in Physiology*, *10*(JAN), 1912. https://doi.org/10.3389/fphys.2018.01912

Webster, C. L., Longdon, B., Lewis, S. H., & Obbard, D. J. (2016). Twenty-five new viruses associated with the drosophilidae (Diptera). *Evolutionary Bioinformatics*, *12*, 13–25. https://doi.org/10.4137/ebo.s39454

Webster, C. L., Waldron, F. M., Robertson, S., Crowson, D., Ferrari, G., Quintana, J. F., Brouqui, J.-M., Bayne, E. H., Longdon, B., Buck, A. H., Lazzaro, B. P., Akorli, J., Haddrill, P. R., & Obbard, D. J. (2015). The discovery, distribution, and evolution of viruses associated with *Drosophila melanogaster*. *PLoS Biology*, *13*(7), e1002210. https://doi.org/10.1371/journal.pbio.1002210

Weeks, A. R., Turelli, M., Harcombe, W. R., Reynolds, K. T., & Hoffmann, A. A. (2007). From Parasite to Mutualist: Rapid Evolution of Wolbachia in Natural Populations of Drosophila. *PLOS Biology*, *5*(5), e114. https://doi.org/10.1371/journal.pbio.0050114

Youle, M., Haynes, M., & Rohwer, F. (2012). Scratching the surface of biology's dark matter. *Viruses: Essential Agents of Life*, *9789400748996*, 61–81. https://doi.org/10.1007/978-94-007-4899-6_4

Chapter 9

Appendix

CHAPTER 9 Appendix

List of additional publications

1. Cunha, M. D. P., Duarte-Neto, A. N., Pour, S. Z., **Ortiz-Baez, A. S.**, Černý, J., Pereira, B. B. S., Braconi, C. T., Ho, Y. L., Perondi, B., Sztajnbok, J., Alves, V. A. F., Dolhnikoff, M., Holmes, E. C., Saldiva, P. H. N., & Zanotto, P. M. A. (2019). Origin of the São Paulo Yellow Fever epidemic of 2017-2018 revealed through molecular epidemiological analysis of fatal cases. *Scientific reports*, *9*(1), 20418. https://doi.org/10.1038/s41598-019-56650-1. *I contributed to the bioinformatic analysis and data curation.*

2. Campbell, S. J., Ashley, W., Gil-Fernandez, M., Newsome, T. M., Di Giallonardo, F., **Ortiz-Baez, A. S.**, ... & Geoghegan, J. L. (2020). Red fox viromes in urban and rural landscapes. *Virus Evolution*, *6*(2), veaa065. https://doi.org/10.1093/ve/veaa065. *I analysed part of the data, reviewed and edited the manuscript.*

3. Geoghegan, J. L., Di Giallonardo, F., Wille, M., **Ortiz-Baez, A. S.**, Costa, V. A., Ghaly, T., ... & Holmes, E. C. (2021). Virome composition in marine fish revealed by meta-transcriptomics. *Virus evolution*, *7*(1), veab005. https://doi.org/10.1093/ve/veab005. *I analysed part of the data, reviewed, and edited the manuscript.*

4. Van Brussel, K., Mahar, J. E., **Ortiz-Baez, A. S.**, Carrai, M., Spielman, D., Boardman, W. S., ... & Holmes, E. C. (2022). Faecal virome of the Australian grey-headed flying fox from urban/suburban environments contains novel coronaviruses, retroviruses and sapoviruses. *Virology*, *576*, 42-51. https://doi.org/10.1016/j.virol.2022.09.002. *I contributed with the formal analysis, writing – review & editing, and preparation of the manuscript.*

SCIENTIFIC REPORTS
nature research

OPEN

Origin of the São Paulo Yellow Fever epidemic of 2017–2018 revealed through molecular epidemiological analysis of fatal cases

Marielton dos Passos Cunha[1*], Amaro Nunes Duarte-Neto[2], Shahab Zaki Pour[1], Ayda Susana Ortiz-Baez[1,9], Jiří Černý[3], Bárbara Brito de Souza Pereira[1], Carla Torres Braconi[1], Yeh-Li Ho[4], Beatriz Perondi[5], Jaques Sztajnbok[6], Venancio Avancini Ferreira Alves[2], Marisa Dolhnikoff[2], Edward C. Holmes[7], Paulo Hilário Nascimento Saldiva[2] & Paolo Marinho de Andrade Zanotto[1,8*]

The largest outbreak of yellow fever of the 21st century in the Americas began in 2016, with intense circulation in the southeastern states of Brazil, particularly in sylvatic environments near densely populated areas including the metropolitan region of São Paulo city (MRSP) during 2017–2018. Herein, we describe the origin and molecular epidemiology of yellow fever virus (YFV) during this outbreak inferred from 36 full genome sequences taken from individuals who died following infection with zoonotic YFV. Our analysis revealed that these deaths were due to three genetic variants of sylvatic YFV that belong the South American I genotype and that were related to viruses previously isolated in 2017 from other locations in Brazil (Minas Gerais, Espírito Santo, Bahia and Rio de Janeiro states). Each variant represented an independent virus introduction into the MRSP. Phylogeographic and geopositioning analyses suggested that the virus moved around the peri-urban area without detectable human-to-human transmission, and towards the Atlantic rain forest causing human spill-over in nearby cities, yet in the absence of sustained viral transmission in the urban environment.

Yellow fever virus (YFV) is an enveloped virus of the family *Flaviviridae* (genus *Flavivirus*) with a single stranded, positive-sense RNA genome of approximately 11 kb that encodes a single polyprotein cleaved into three structural (capsid (C), membrane (M) and envelope (E)) and seven non-structural proteins (NS1-NS5)[1]. The virus comprises a single serotype with four genotypes: (*i*) East Africa, (*ii*) West Africa, (*iii*) South American I and (*iv*) South American II[2,3] that may have diverged around several thousand years before present[4,5] with a possible origin in the African continent[5,6]. Historical evidence points to a YFV introduction in the Americas around the 17th century, possibly due to the slave trade[3,5–7]. After its introduction, YFV established both urban and sylvatic cycles[7,8], and several urban outbreaks have been reported in Brazil since the 17th century[9]. The circulation of YFV in the urban cycle in the American continent was initially mitigated by curbing the infestation of *Aedes aegypti* and later with the advent of an effective vaccine in the early 20th century[7,10], with considerable success. As a consequence, the last urban outbreak of YFV was officially reported in 1942 in Brazil[9]. After the reintroduction of *A. aegypti* in the 1970's[11,12] the virus remained, until recently, largely in sylvatic environments in the Americas, infecting

[1]Laboratory of Molecular Evolution and Bioinformatics, Department of Microbiology, Biomedical Sciences Institute, University of São Paulo, São Paulo, Brazil. [2]Pathology Department, Clinical Hospital, Faculty of Medicine, University of São Paulo, São Paulo, Brazil. [3]Faculty of Tropical AgriSciences, Czech University of Life Sciences in Prague, Prague, Czech Republic. [4]Intensive Care Unit, Division of Clinical Infectious and Parasitic Diseases, Clinical Hospital, Faculty of Medicine, University of São Paulo, São Paulo, Brazil. [5]Yellow Fever Crisis Committee, Clinical Hospital, Faculty of Medicine, University of São Paulo, São Paulo, Brazil. [6]Institute of Infectology Emílio Ribas, São Paulo, Brazil. [7]School of Life and Environmental Sciences & Sydney Medical School, The University of Sydney, Sydney, Australia. [8]Scientific Platform Pasteur - USP, São Paulo, Brazil. [9]Marie Bashir Institute for Infectious Diseases and Biosecurity, Charles Perkins Centre, School of Life & Environmental Sciences and Sydney Medical School, The University of Sydney, Sydney, NSW 2006, Australia. *email: marieltondospassos@gmail.com; pzanotto@usp.br

non-human primates (NHPs) with sporadic cases in susceptible human hosts. The main vectors of YFV in the sylvatic cycle are mosquitoes of the genera *Haemagogus* and *Sabethes*[13,14].

In 2014, intense enzootic activity of YFV was detected in Mato Grosso do Sul and Goiás states that adjoin the Amazon region of Brazil[15,16]. YFV carried by infected monkeys kept moving in a general southeasterly direction, and in 2016 cases were reported in Minas Gerais, reaching epidemic proportions in 2017, during which cases were also reported in the states of Rio de Janeiro, Espírito Santo and Bahia[17–21].

Between January 2016 and January 2018, seven countries and regions of the Americas reported cases of yellow fever in their territories (Bolivia, Brazil, Colombia, Ecuador, French Guiana, Peru and Suriname), with the highest indices in Brazil. In early 2018 an unusually large increase in the number of confirmed cases was observed in the state of São Paulo[22]. A peak of notified human cases was reached in January 2018[23]. This was the largest outbreak registered in 21st century in the most populated state of Brazil, including the densely populated metropolitan region of São Paulo city (MRSP), which is the largest conurbation in the southern hemisphere with around 23 million inhabitants. Until 2018, vaccination was not generally recommended in MRSP because YFV had been absent in recent decades. Hence, most of the population in the area was susceptible to YFV and autochthonous cases were reported[24,25]. Due to the outbreak in São Paulo, vaccination campaigns were initiated for resident populations, starting in northern peri-urban settings bordering forested and rural areas, where cases of YFV were previously reported. Subsequently, vaccination was extended to the whole urban population as well as to all inhabitants of the São Paulo state as the epidemic expanded[22]. As this is the first time in the 21st century that cases of YFV have appeared in the MRSP, we sought to characterize the circulating viruses and establish their origin by studying their evolution and phylogeography based on samples taken from patients who died during the 2017–2018 outbreak.

Material and Methods

Ethical statement. The human autopsies analyzed in this study were performed after obtaining informed consent of the family members and following the protocol approved by the research ethics committee of the Clinical Hospital of the University of São Paulo School of Medicine (HCFMUSP) (CAPPesq #426.643). All the methods were performed in accordance with the relevant guidelines and regulations of the ethics committee of the HCFMUSP following the approval CAPPesq #426.643. All participating families were asked to sign a free and informed consent form, authorizing the autopsy and all experiments performed with the collected tissues. All laboratory procedures listed below were performed in a biosafety level (BSL)-2 laboratory, in accordance with the Brazilian standards of the Ministry of Health for Biological Agents Risk Classification[26].

Patients and samples. Overall, we analyzed 81 patients 67 of whom were confirmed to have died following YFV infection. We successfully acquired 36 genome sequences from the 67 yellow fever deaths, with the remaining samples being of insufficient quality to obtain YFV genomes at the necessary coverage. The suspected case definition of YFV infection was established by the Brazilian Ministry of Health and the Health Department of São Paulo State and included patients with sudden onset high fever associated with jaundice and/or hemorrhage who had lived or had visited areas with YFV epizootics (*i.e.*, clusters of infections in non-human primates (NHP) or isolation of YFV in vectors), regardless of the vaccine status for YFV, during the preceding 15 days. Confirmed cases had compatible clinical presentation and laboratory confirmation by at least one of the following methods: (*i*) serum IgM positive (MAC-ELISA); (*ii*) detection of YFV-RNA by qRT-PCR in blood samples; (*iii*) virus isolation; (*iv*) histopathology compatible with YFV hepatitis with detectable antigen in tissues by immunohistochemistry technique. All cases received the definitive laboratorial diagnosis of YFV by the Adolfo Lutz Institute (IAL), the State Reference Laboratory. Previous exposure or co-infection by Hepatitis A virus (HVA), B (HBV), C (HVC), Cytomegalovirus (CMV), Herpes virus (HSV), Dengue virus (DENV), Chikungunya virus (CHIKV), Human Immunodeficiency virus type 1 (HIV-1), leptospirosis and other non-infectious diseases etiologies for acute hepatitis were accessed and cases were excluded following clinical diagnostic methods. Epidemiological, clinical (including demographic data, preexisting medical conditions, clinical signs and symptoms and in-hospital follow-up until death) and other laboratory features were collected from the medical charts.

Autopsy protocol and tissue processing. The Service of Verification of Deaths of the Capital - USP investigated deaths due to yellow fever from December/2017 to April/2018. Autopsies were performed following the Letulle technique, where all the organs were removed *en masse* (one block), requiring dissection organ by organ to exam them individually. Briefly, the dissection was performed in the following organs: (*i*) heart; (*ii*) lung; (*iii*) brain; (*iv*) kidney; (*v*) spleen; (*vi*) pancreas; and (*vii*) liver.

Molecular characterization. Nucleic acid extraction from all collected tissues was performed using the TRIzol® reagent (Life Technologies, Carlsbad, CA, USA) and carried out according to the manufacturer's instructions. Molecular detection of YFV was performed with the use of the AgPath-ID One-Step RT-PCR Reagents (Ambion, Austin, TX, USA) with specific primers/probe previously described[27]. To identify cases of adverse vaccine response (*i.e.*, fatal cases associated with the vaccine virus) we used specific primers/probe specific for the vaccine virus[28]. qRT-PCR reactions consisted of a step of reverse transcription at 45 °C for 10 min, enzyme activation at 95 °C for 10 min, and 40 cycles at 95 °C for 15 s and 60 °C for 45 s for hybridization and extension using the ABI7500 equipment (Thermo Fisher Scientific, Waltham, MA, USA).

Sequencing and viral genome assembly. Based on the RNA viral concentration, total RNA were extracted from the liver tissues using the TRIzol® reagent (Life Technologies, Carlsbad, CA, USA). Subsequently, the RNA was purified with DNase I and concentrated using the RNA Clean and Concentrator $^{TM-5}$ kit (Zymo Research, Irvine, CA, USA) according to the manufacturer's instructions. The paired-end RNA libraries were constructed and validated using the TruSeq Stranded Total RNA HT sample prep kit (Illumina, San Diego, CA,

USA). Sequencing was done at the Core Facility for Scientific Research – University of São Paulo (CEFAP-USP/GENIAL) using the Illumina NextSeq platform. Each sample was barcoded individually, which allowed separation of reads for each patient. Short unpaired reads and low-quality bases and reads were removed using Trimmomatic version 0.36 (LEADING:20 TRAILING:20 SLIDINGWINDOW:4:25 MINLEN:36)[29]. Consensus genomes were assembled with paired-end reads using Bowtie2 v.2.3.4.3[30] using default parameters.

Data sets. All full genomic sequences available from YFV that contained information on location and date of isolation were recovered from the National Center for Biotechnology Information (NCBI) (https://www.ncbi.nlm.nih.gov/genbank/) website. Sequences were aligned to our 36 new YFV genomes (Supplementary Table 1) using Clustal Omega v.1.2.4[31]. A list of the sequences used is available in Supplementary Table 2. Recombinant sequences were screened using all algorithms implemented in RDP4 program (RDP, GENECONV, BootScan, MaxChi, Chimaera, Siscan and 3Seq) using the default settings[32]. No evidence for recombination was detected. Sequences containing long contiguous stretches of undefined nucleotides were excluded. A final alignment of complete genome sequences was manually inspected and edited using the program AliView v.1.18[33]. After preliminary phylogenetic analyses, the master alignment comprising 135 full-length, curated sequences encoding the complete viral polyprotein (dataset-1) (Supplementary Table 2) was subdivided into two data sets for further analysis: (*i*) a data set containing 98 genomes of the SA1 and SA2 genotypes from the Americas (dataset-2); and (*ii*) 74 sequences from 2017 and 2018 sampled from the states of Minas Gerais, Espírito Santo, Bahia, Rio de Janeiro and São Paulo (dataset-3) (Supplementary Tables 2 and 3). All alignments are available in the Supplementary Data and on GitHub (https://github.com/MarieltonCunha/ViralDiversity/).

Phylogenetic analysis. Phylogenetic trees of YFV based on full-length, curated coding sequences for all the data sets were estimated using the Maximum Likelihood (ML) method implemented in IQ-TREE 1.5.5[34] with automatic model selection by ModelFinder and using the Bayesian Information Criterion (BIC)[35]. The robustness of the groupings observed was assessed using 1,000 non-parametric bootstrap replicates. ML and Bayesian maximum clade credibility (MCC) trees (see below) were visualized and plotted using FigTree v.1.4.3[36]. All taxon labels for sequences used in this work are presented in the format: genotype/accession number/strain name/local of isolation/date of isolation. We explored the temporal signal (*i.e.*, molecular clock structure) and quality of our data set using TempEst v.1.5.1[37].

Phylodynamics and phylogeographic analysis. The spatio-temporal evolution of YFV spread was inferred within a Bayesian framework as implemented in BEAST v.1.10.1[38]. An initial descriptive summary of the demographic history of YFV was approximated using the Bayesian SkyGrid coalescent model[39] and revealed no significant variation in genetic diversity (a marker of population size) during the period of our analysis. Based on previous estimates of evolutionary dynamics of related YFV[17,40], we tested uncorrelated relaxed molecular clocks assuming a log-normal distribution, in combination with constant size, exponential and logistic growth demographic models (Supplementary Tables 4 and 5). Phylogeographic patterns and parameters were estimated using the Bayesian inference through Markov chain Monte Carlo (MCMC) run for 50 million states, sampling every 5,000 states with a 10% burn-in. Convergence and the effective sample size (ESS) > 200 were examined using Tracer v.1.7.1[41]. Likewise, the maximum clade credibility tree (MCC) was visualized and edited in FigTree v.1.4.3[36]. We recorded the time to the most recent common ancestor (tMRCA) and their 95% Bayesian credible intervals (HPD) for the MCC tree. To calculate the log marginal likelihood for molecular clock and demographic model selection, we used the path sampling (PS) and the stepping-stone (SS) sampling approaches by running 100 path steps of 1 million iterations each[42]. The spatiotemporal spread of YFV was visualized and plotted with SPREAD3[43]. XML input files for BEAST are available in the Supplementary Data and on GitHub (https://github.com/MarieltonCunha/ViralDiversity/).

Geopositioning of samples. To analyze the geographical proximity among fatal human and NHP cases we calculated the spatial distances between all cases using available geoposition information. We geopositioned only those fatal human and NHP YFV cases that occurred in the MRSP (47.0–46.2 S, 23.9–23.1 W), using the available data on patient residence and day of death. NHP cases were included only for those were coordinates for the place of where carcasses were found was available. For fatal NHP cases, the date the carcass was found was assumed to be the day of death, although death may have taken place a few days before. Distances between the human and NHP fatal YFV cases were calculated based on the available coordinates. Geographic pairwise distance matrices among all YFV cases (in kilometers) were clustered using the neighbor joining algorithm available in the PHYLIP v.3.695 package[44], this enabled us to produce a dendogram based on geoposition information.

Results

Epidemiological surveillance of YFV in São Paulo, 2017–2018. From January to August 17, 2018, the State of São Paulo reported 3028 suspected cases of yellow fever, 537 (17.7%) of which were confirmed, with 498 (92.7%) autochthonous cases and 35 (6.5%) imported from other states[45]. Of the 498 autochthonous cases, 176 died, resulting in a mortality frequency of 35.4%[45]. Despite the magnitude of the outbreak in São Paulo, little is known about the epidemiological, genetic and evolutionary characteristics of the virus circulating in the state. Accordingly, among all patients who died with suspected YFV infection between December 2017 and April 2018, we focused on 81 cases identified through the service of verification of deaths of the capital - USP (SVOC-USP) in the city of São Paulo (Fig. 1A). Our qRT-PCR results indicated that 67/81 (82.7%) individuals had been infected by YFV, while five were shown by qRT-PCR to only carry the vaccine strain YFV-17DD alone, suggesting that their death was associated with an adverse response to the vaccine as previously reported[46–48], and nine were negative for YFV infection in all tissues tested (Fig. 1B,C). All 67 confirmed YFV deaths were due to complications of fulminant yellow fever hepatitis, with hepatic encephalopathy, severe coagulopathy, bleeding (mainly

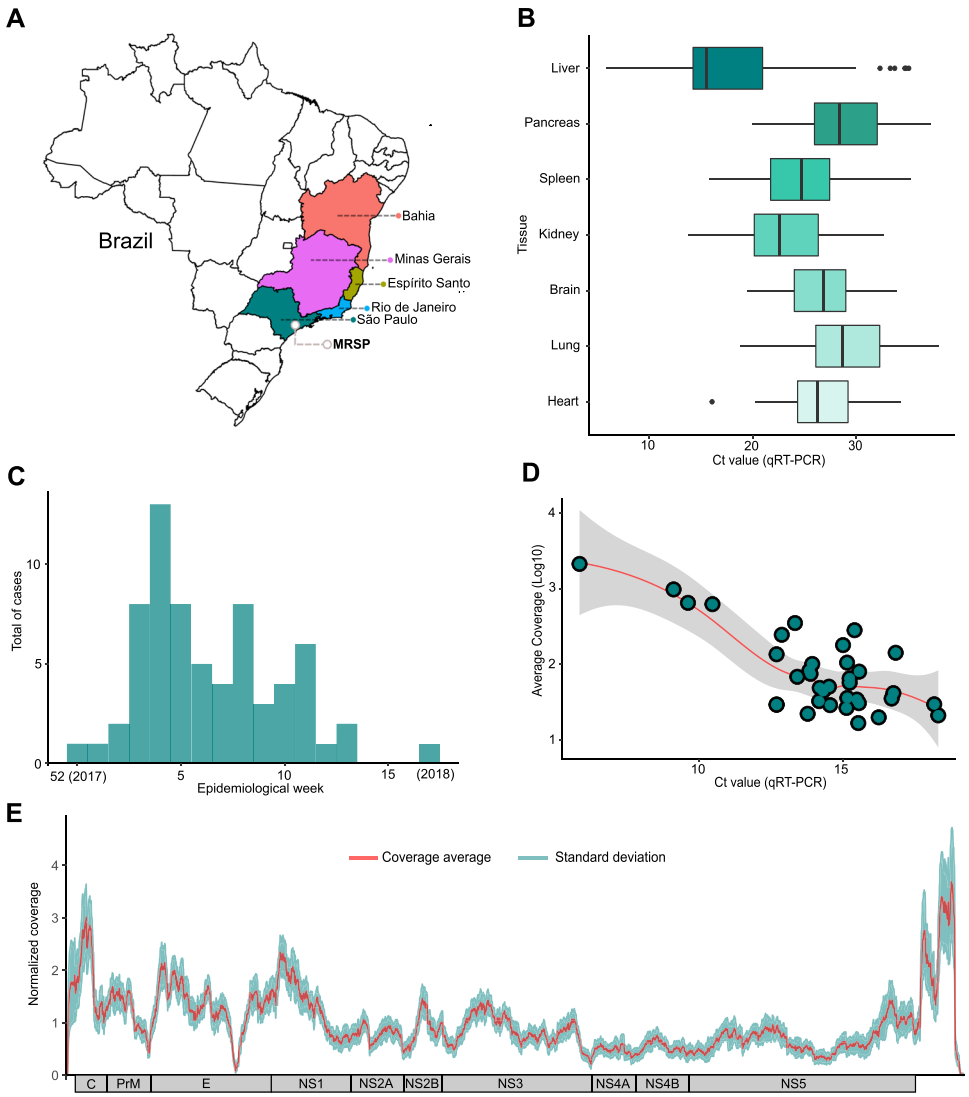


Figure 1. The current outbreak of yellow fever virus in Brazil (2016–2018). (**A**) Brazilian states with YFV cases recorded and sequenced in humans, non-human primates (NHP) and mosquitoes between 2017–2018. A grey circle marks the metropolitan region of São Paulo (MRSP). (**B**) Cycle threshold according to each of the 7 tissues analyzed for positive patients. Boxplots represent the 75th percentile, median, 25th percentile and the whiskers extend to the highest and lowest value in the 1.5x interquartile range. The different colors represent the different tissues analyzed. (**C**) Total cases recorded represented sylvatic cases of YFV (qRT-PCR positive cases) during the epidemiological weeks covered by the study (week 52 of 2017 to week 17 of 2018). (**D**) Relationship between the average coverage and the Ct values obtained for each sequenced sample. The data indicate that we obtained the expected direct inverse relationship between Ct and coverage parameters, as indicated by the trend line. (**E**) Combined coverage (normalized by the sample average) along all 36 sequenced YFV genomes generated in this study.

gastrointestinal, pulmonary and/or cerebral hemorrhages), renal dysfunction and secondary infections. We were able to successfully sequence the full YFV genome from 36 of these patient samples.

All of our cases were sampled in 17 localities in the São Paulo state, from which 16 localities had fatal cases due to YFV (Supplementary Table 6). Our molecular diagnostics indicated a peak of cases during the first epidemiological weeks of 2018, particularly at the end of January, coinciding with official cases notifications data (Fig. 1C). The median age of people with confirmed infection was 49.12 years (range 16–87) and were mainly male (82.09–55/67).

Genomic surveillance. Because detailed spatio-temporal resolution of viral evolution often relies on a few nucleotide differences among otherwise closely related viruses, complete genomes with high coverage for each base position are a prerequisite for robust inference. Therefore, to select the appropriate clinical specimens for viral sequencing, we analyzed cycle threshold (Ct) data from qRT-PCR from viral RNA in seven distinct tissues/organs (heart, lung, brain, kidney, spleen, pancreas and liver) to choose samples with the lowest possible Cts.

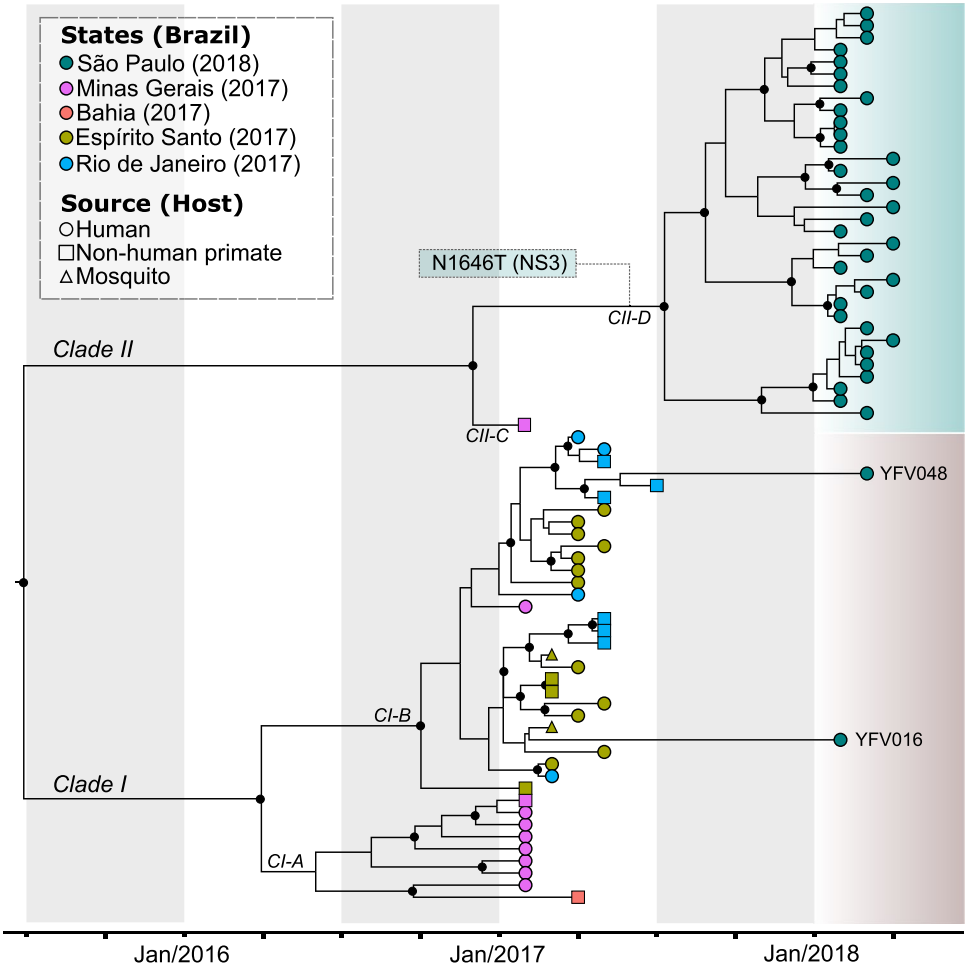


Figure 2. Time-stamped, MCC tree of YFV South American genotype I in Brazil recovered under the logistic-lognormal demographic model. The different colours indicate samples from different locations. The black circles represent posterior support upper than 0.7. The single synapomorphic change observed in Clade II [N1646T (NS3)] is shown in the box over the branch leading to Clade II-D. The three distinct introductions in the metropolitan region of São Paulo (MRSP) are shown (See also Fig. 3).

In general, all tissues had normally distributed Ct values, with the exception of the liver, which had a moderately asymmetrical distribution and a deviation to lower Ct values, and hence generally inferior to other tissues (Fig. 1B). In total, we obtained 36 complete YFV genomes from the 67 positive patients (Fig. 1D,E). All sequences of the current outbreak belonged to the South American I genotype (Supplementary Fig. 1), and were related with sequences previously isolated in neighboring states in 2017 (Fig. 2) with no evidence of recombination. Based on the phylogenetic analysis, we could infer at least three distinct introductions of YFV in the MRSP: (*i*) A major clade (34 genomes) in the northwest of the MRSP coming from Minas Gerais due to NHP movement, and likely emerging between April 2017–October 2017 (95% HPD; mean - July 2017) (Fig. 3 and Supplementary Tables 4 and 5), (*ii*) one virus lineage from a case from Espírito Santo (Patient 16), and (*iii*) one from a case from Rio de Janeiro (Patient 48) (Fig. 2). Importantly, our patient's records indicated the two single introductions were due to people visiting enzootic locations in these states and did not appear to have caused detectable additional cases in the MRSP.

Origin of the 2016–ongoing Yellow Fever virus outbreak. Phylogenetic (Supplementary Figs. 1 and 2) and phylogeographic (Figs. 2, 3 and Supplementary Tables 4 and 5) analyses of samples from the 2017–2018 YFV outbreak allowed us to reveal the origin and spread of YFV in the Southeast and Northeast region of Brazil. In particular, there was evidence of two distinct zoonotic clades (Clade I and II) that likely separated in Minas Gerais (location posterior support of 0.8) between November 2013–June 2016 (95% HPD; mean date of June 2015). The mean rate of Clade I and II migration during the whole sampled period 2017 to 2018 was approximately 3.3 km/day (95% HPD = 2.25–4.37 km/day) with a mean evolutionary rate of 9.85×10^{-4} nucleotide substitutions per site, per year (subs/site/year) (95% HPD = $6.52 \times 10^{-4} - 1.35 \times 10^{-3}$ subs/site/year). We now describe these two clades in more detail.

Clade I. Clade I divided into two smaller clades (CI-A and CI-B) in 2016 (95% HPD of divergence time = July 2015 – September 2016) and likely in Minas Gerais (location posterior support of 0.82) (Fig. 2). CI-A then

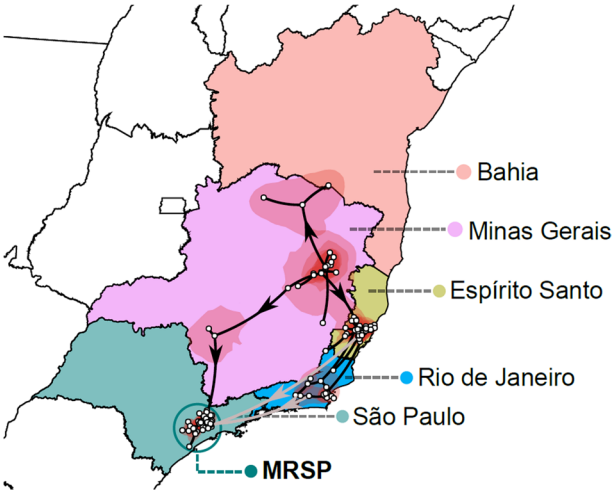


Figure 3. Highest posterior probability migration paths for the YFV Clades I and II from 2016 to 2018 towards the metropolitan region of São Paulo (MRSP), based on the analysis of 74 complete genomes. Although the sample size is small such that inferences should be made with caution, three distinct introductions in the MRSP are shown and strongly supported. The spatiotemporal spread was visualized with SPREAD3.

diversified and moved and into peri-urban and forested regions in the state of Minas Gerais, causing an outbreak after January 2017, then moving onto Bahia. In contrast, Clade CI-B likely diversified in the forest region in the border between Minas Gerais and Espírito Santo, also in 2016, and then moved to Espírito Santo and Rio de Janeiro, causing in both states an outbreak during the first part of 2017. Two YFV patients who died in 2018 and resided in the MRSP had visited Espírito Santo (Patient 16) and Rio de Janeiro (patient 48). Fittingly, the virus phylogeny showed that their posthumous viral samples were nested among isolated viruses from the areas they visited (Fig. 2). These results indicated that CI-B was circulating until early 2018.

Clade II. This clade caused the majority of the deaths in the MRSP (Fig. 2). It diverged into Clades CII-C and CII-D in the state of Minas Gerais, with a location posterior support of 0.87, near the border with São Paulo between June 2016 - January 2017 (95% HPD; mean - December 2016) (Fig. 3). Subsequently, CII-D moved towards the MRSP, causing epizootics beginning between April 2017–October 2017 (95% HPD; mean - July 2017) (Supplementary Fig. 3) in forest parks (Horto Florestal and Cantareira State Park) that form a belt around the Northern part of the MRSP (Fig. 4). It is noteworthy that our inferred dates correspond well with the reported official cases of YFV cases in NHP and humans (Fig. 4). It is also notable that CII-D is also defined by a unique synapomorphic substitution (N1646T) in the NS3 gene that is not present in CII-C and Clade I viruses (Fig. 2).

Geopositioning analysis. In total, 230 NHP carcasses were collected in the MRSP. Of these, 136 were members of the genus *Alouatta* (howler monkeys), 14 were *Callithrix* genus (marmosets), and five were *Cebus* genus (capuchin monkeys). The species identity of the remaining 75 carcasses were not determined (Fig. 4) (data provided by the Adolfo Lutz Institute). Analysis of spatio-temporal data showed that the YFV outbreak progressed in different directions in humans and NHPs (Figs. 4 and 5). While the outbreak in NHPs had a tendency to move in a south-southwest direction, in humans the outbreaks in a southeast direction (Figs. 4 and 5).

Several geographically well-defined clusters can be observed in the dendogram inferred from the pairwise geographic distances matrix among all YFV cases (Fig. 5). Two areas of intense epizootics were inferred in the north and southwest forested areas around the MRSP. We also inferred a large cluster of cases of NHP and humans in the northern region, Cantareira and Horto Florestal State parks, spreading to the nearby towns of Mairiporã and Guarulhos, where most of the human and NHP cases were reported. Another cluster represents NHP cases from the southwestern of the MRSP, around Cotia, where the second most affected NHP population was present. Hence, the most striking finding of this analysis was that most human cases occurred close to both the NPH cases and the forested belt around the MRSP.

Discussion

We describe the outbreak of YFV in the MRSP, Brazil, in 2016–2018, particularly its origin and how the virus diversified and moved around the largest conurbation in the southern hemisphere carried by NHP, killing 176 people during 2018 in the process[45]. All the isolates from São Paulo belonged to the South American I genotype and formed a single monophyletic group along with viruses (comprising Clades I and II) that also circulated in 2016–2017 in the states of Minas Gerais, Bahia, Rio de Janeiro and Espírito Santo[17,19,40]. Several synapomorphic mutational changes in different genes were previously reported by our group[19], and here we report a synapomorphy (N1646T) in the protease NS3 gene shared by all CII-D. The mean evolutionary rate for all the YFV sequences of the Brazilian outbreak (2017–2018) was 9.85×10^{-4} subs/site/year, and hence compatible with those previously estimated for YFV and for other flaviviruses[4,5,49].

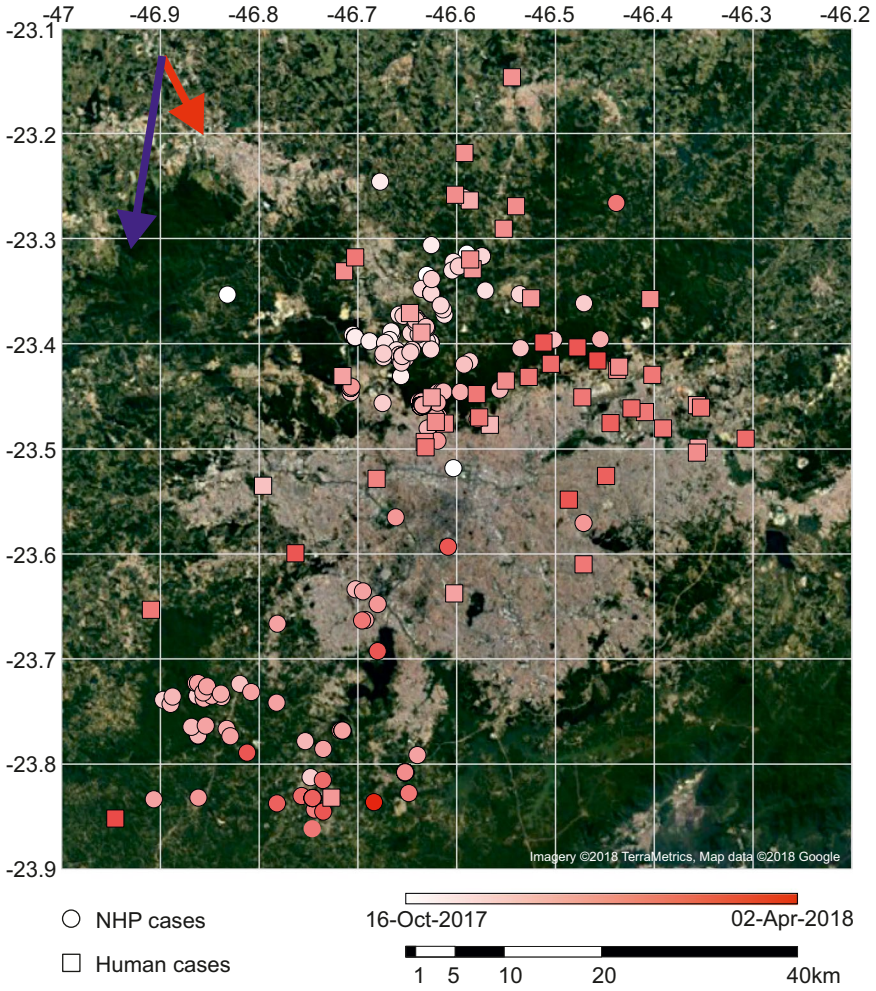


Figure 4. Spatial distribution of YFV deaths through time in non-human primates (NHP) and humans. Arrows indicate the general trend of movement around the metropolitan region of São Paulo (MRSP) estimated from distance matrices (see Fig. 5). The earliest cases in NHP are shown in the north, and later in the south and northeast of the MRSP. Most human cases are near sites with reported deaths of NPH, confirmed to be caused by YFV. The outbreak appears to have been confined mostly near the forested belt around the MRSP, contrasting with the almost empty, heavily urbanized center. Cardinal points are aligned according to the main axis of the page, (*e.g.*, top being north, etc.). The figure was created by plotting the coordinates of reported cases to a satellite image available from Google Maps (google.com/maps) as background.

The current Brazilian outbreak began in the state of Minas Gerais in June 2015, with all viruses sampled from 2017 belonging to a single monophyletic group that diverged into two main clades (Clade I and II), and indicative of a single introduction of the virus in the region. These observations are supported by other molecular epidemiological studies conducted in Brazil from 2016[17–19,40,50]. Although previous evolutionary studies point to an origin of the virus in Venezuela[40], epidemiological monitoring carried out by the Brazilian Ministry of Health suggest a likely origin of the Brazilian outbreak in 2014, with confirmed epizootics in the transitional area between the Amazon and the Cerrado biomes (with most of the confirmed cases occurring in the states of Goiás and Mato Grosso do Sul)[15,16,21]. This region was the probable link between the Amazon basin and the state of Minas Gerais, located in southeastern Brazil. It is likely that the numbers of human cases in this region were not high due to the vaccine coverage there[51]. The viral invasion into southeast Brazil, associated with the rapid spatial spread of the virus (estimated here at a mean of rate 3.3 km/day), caused the virus to circulate in important fragments of the Atlantic Forest near the peri-urban areas of the main Brazilian megacities (notably São Paulo and Rio de Janeiro), and led to a marked increase in the number of cases during the outbreak. In the MRSP, the virus (Clade CII-D) was introduced, maintained and spread in the sylvatic transmission cycle, with occasional cases of infection in humans between April 2017 and October 2017, with the interstate border between São Paulo and Minas Gerais as the route of introduction. In São Paulo state, the routes of viral dispersion included only interconnected forested, corridors linked to peri-urban regions. The patients studied here were mainly unvaccinated adult males that had contact with the sylvatic environment or lived nearby. No autochthonous cases were documented in the central region of the city of São Paulo. Importantly, the MRSP cases reduced in numbers as the populations of NHP collapsed and with vaccination campaigns in areas classified as at risk[52].

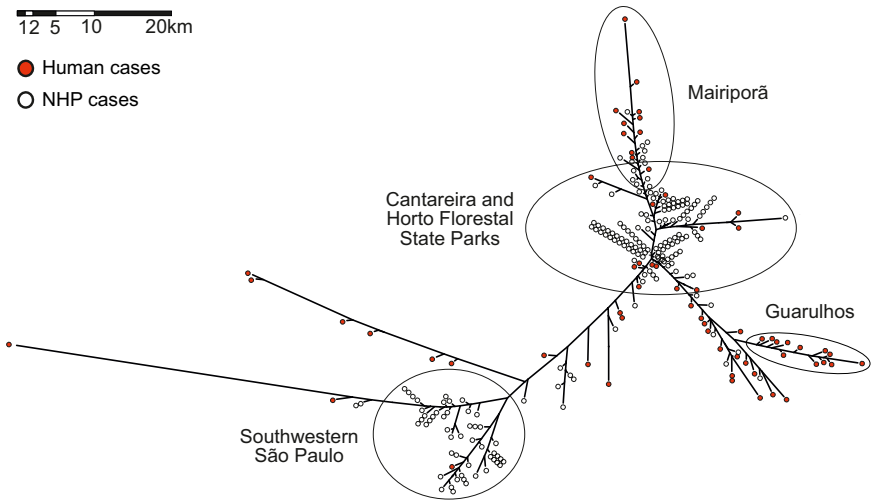


Figure 5. Neighbor joining tree calculated from pairwise geoposition distances among all the non-human primates and human cases available from the metropolitan region of São Paulo (MRSP).

The introduction and establishment of the YFV Clade II-D in the state of São Paulo can be further explained by environmental factors, including: (*i*) mosquitoes of the genus *Haemagogus* are abundant in the forested areas of the state of São Paulo[53,54] and were the primary vectors in the YFV outbreak occurred in Brazil, 2016–2018[55]; (*ii*) NHPs are found in areas of the Atlantic Forest and are susceptible and responsible for the maintenance of the virus in the sylvatic cycle[56–58]; and (*iii*) the regions affected by the current outbreak had low vaccine coverage[51]. Our findings support previous work indicating that the outbreak of 2016–2018 (sampled in the states of Minas Gerais, Bahia, Espírito Santo and Rio de Janeiro), occurred in a sylvatic environment with occasional infections in humans[17].

Importantly, we also recorded two introductions of YFV Clade I-B detected in patients who travelled to Espírito Santo and Rio de Janeiro - both states that experienced significant circulation of this virus lineage in 2018. In both these states an increase in the number of YFV notifications was reported in 2017 across successive epidemic periods, showcasing the maintenance of epizootic YFV. In addition, we highlighted the extent of viral movement, such as observed in cases imported from Brazil by other countries[59], largely facilitated by rapid human movement such as those resulting from air travel[60].

In contrast to other arboviruses in Brazil such as dengue virus, in which continuous reintroductions are responsible for keeping the virus circulating in the urban cycle[61–63], YFV is dependent on epizootics to cause cases in humans. The South American I genotype belongs to a "modern lineage", that has been circulating in America since 1995 and that perhaps originated in Trinidad and Tobago[40]. It is believed that from there the virus spread to South American countries, especially Venezuela and Brazil[40], carried mainly by NHP and sylvatic mosquitoes, moving along forested corridors and perhaps promoted by a series of interlocked epizootics involving the exchange of viruses among infected and susceptible individuals[64,65]. Epizootics among social animals, such as New World arboreal primates, may be reduced by self-exclusion of infected individuals[66]. For instance, it is in theory possible that social avoidance, changes in group size, group isolation and several other behaviors may have evolved due to reduce pathogen transmission. Nevertheless, in the case of vector-borne diseases any isolation mechanism is efficient only at distances that minimize transmission[66]. Howlers were the most affected monkey species in the forested belt around the MRSP[52]. As in several other previous YFV epizootics[64], the high overall fatality rate in howlers led to almost the complete extinction of these monkeys in most areas around Sao Paulo[52].

It has been assumed that the decline in the numbers of howler monkeys and the severe reduction of several species of NHP from around the MRSP had a significant effect on ending the outbreak. Although perhaps due to poor sampling of monkeys in that locality, it is possible that Clade II-D could have caused a limited number of human-to-human transmission cases, as suggested by a cluster of human cases in Guarulhos (Fig. 5). Critically, however, a key factor that differentiates the current outbreaks of YFV in the Americas and Africa is that there is no clear evidence for urban cycles of YFV in the Americas has been observed since the first half of the 20th Century. A possible, although untested, explanation is that the former *A. aegypti* colonizing the Americas was from Africa (Senegalese strain), while the *A. aegypti* reintroduced in the early 1970's is Asiatic, where no urban spread of YFV is observed[67].

Received: 22 April 2019; Accepted: 11 December 2019;
Published online: 31 December 2019

References

1. Chambers, T. J., Hahn, C. S., Galler, R. & Rice, C. M. Flavivirus Genome Organization, Expression, and Replication. *Annu. Rev.* **44**, 649–688 (1990).
2. Barrett, A. D. T. & Higgs, S. Yellow Fever: A Disease that Has Yet to be Conquered. *Annu. Rev. Entomol.* **52**, 209–229 (2007).
3. Monath, T. P. & Vasconcelos, P. F. C. Yellow fever. *J. Clin. Virol.* **64**, 160–173 (2015).

4. Zanotto, P. M. *et al*. Population dynamics of flaviviruses revealed by molecular phylogenies. *Proc. Natl. Acad. Sci.* **93**, 548–553 (1996).
5. Bryant, J. E., Holmes, E. C. & Barrett, A. D. T. Out of Africa: A Molecular Perspective on the Introduction of Yellow Fever Virus into the Americas. *PLoS Pathog.* **3**, e75 (2007).
6. Nunes, M. R. T. *et al*. Genomic and phylogenetic characterization of Brazilian Yellow Fever virus strains. *J. Virol*., 1–33, https://doi.org/10.1128/JVI.00565-12 (2012).
7. Vasconcelos, P. F. & da, C. Yellow Fever. *Rev. Soc. Bras. Med. Trop.* **36**, 275–293 (2003).
8. Soper, F. L. *et al*. Yellow Fever without Aedes Aegypti. Study of a rural epidemic in the Valle Do Chanaan, Espirito Santo, Brazil, 1932. *Am. J. Hyg.* **18**, 555–587 (1933).
9. Consoli, R. A. G. B. & Oliveira, R. L. *Principais mosquitos de importância sanitária no Brasil*. (1994).
10. Theiler, M. & Smith, H. H. The Use of Yellow Fever Virus Modified By *in Vitro* Cultivation for Human Immunization. *J. Exp. Med.* **65**, 787–800 (1937).
11. Kraemer, M. U. G. *et al*. The global distribution of the arbovirus vectors Aedes aegypti and Ae. Albopictus. *Elife* **4**, 1–18 (2015).
12. Braga, I. A. & Valle, D. Aedes aegypti: History of Control in Brazil. *Epidemiol. e Serviços Saúde* **16**, 113–118 (2007).
13. Rodaniche, Ede & Galindo, P. Isolation of Yellow Fever virus from Haemagogus mesodentatus, H. equinus and Sabethes chloropterus captured in Guatemala in 1956. *Am. Soc. Trop. Med. Hyg.* **6**, 232–237 (1957).
14. Rodaniche, E., de, Galindo, P. & Johnson, C. M. Isolation of Yellow Fever Virus from Haemagogus Lucifer, H. Equinus, H. Spegazzinii Falco, Sabethes Chloropterus and Anopheles Neivai Captured in Panama in the Fall of 1956. *Am. Soc. Trop. Med. Hyg.* **6**, 681–685 (1957).
15. Possas, C. *et al*. Yellow fever outbreak in Brazil: the puzzle of rapid viral spread and challenges for immunisation. *Mem. Inst. Oswaldo Cruz* **113**, e180278 (2018).
16. da Saúde, Ministério Reemergência da Febre Amarela Silvestre no Brasil, 2014/2015: situação epidemiológica e a importância da vacinação preventiva e da vigilância intensificada no período sazonal. *Bol. Epidemiológico* **46**, 1–10 (2015).
17. Faria, N. R. *et al*. Genomic and epidemiological monitoring of yellow fever virus transmission potential. *Science (80-)*. **7115**, 1–12 (2018).
18. Rezende, I. Mde *et al*. Persistence of Yellow fever virus outside the Amazon Basin, causing epidemics in Southeast Brazil, from 2016 to 2018. *PLoS Negl. Trop. Dis.* **12**, 1–12 (2018).
19. Barbosa, C. M. *et al*. Yellow Fever Virus RNA in Urine and Semen of Convalescent Patient, Brazil. *Emerg. Infect. Dis.* **24**, 176–178 (2018).
20. Figueiredo, P. O. *et al*. Detection and Molecular Characterization of Yellow Fever Virus, 2017, Brazil. *Ecohealth* **1942**, 1–7 (2018).
21. Delatorre, E. *et al*. Distinct YFV lineages co-circulated in the Central-Western and Southeastern Brazilian regions from 2015 to 2018. *Front. Microbiol.* **10**, 1–12 (2018).
22. da Saúde, Ministério Monitoramento do Período Sazonal da Febre Amarela Brasil-2017/2018. *Bol. Epidemiológico* **27**, 1–14 (2018).
23. Saúde, S. & do E. da. Boletim epidemiológico Febre Amarela -28/12/2018. *Bol. Epidemiológico* 1–9 (2018).
24. da Saúde, M. Monitoramento do Período Sazonal da Febre Amarela Brasil-2017/2018. *Bol. Epidemiológico* **27**, 1–14 (2018).
25. Moreno, E. S. & Barata, RdeC. B. Municipalities of higher vulnerability to sylvatic Yellow Fever occurence in the São Paulo state, Brazil. *Rev. Inst. Med. Trop. Sao Paulo* **53**, 335–339 (2011).
26. da Saúde, M. Classificação de Risco dos Agentes Biológicos. *Ministério da Saúde* 1–50 Available at, http://bvsms.saude.gov.br/bvs/publicacoes/classificacao_risco_agentes_biologicos_3ed.pdf, (Accessed: 2nd September 2019) (2017).
27. Diallo, D. *et al*. Patterns of a sylvatic yellow fever virus amplification in southeastern Senegal, 2010. *Am. J. Trop. Med. Hyg.* **90**, 1003–1013 (2014).
28. Avelino-Silva, V. I. *et al*. Yellow fever vaccine viremia following ablative BM suppression in AML. *Bone Marrow Transplant.* **48**, 1008–1009 (2013).
29. Bolger, A. M., Lohse, M. & Usadel, B. Trimmomatic: A flexible trimmer for Illumina sequence data. *Bioinformatics* **30**, 2114–2120 (2014).
30. Langmead, B. & Salzberg, S. L. Fast gapped-read alignment with Bowtie 2. *Nat. Methods* **9**, 357–359 (2012).
31. Larkin, M. A. *et al*. Clustal W and Clustal X version 2.0. *Bioinformatics* **23**, 2947–2948 (2007).
32. Martin, D. P., Murrell, B., Golden, M., Khoosal, A. & Muhire, B. RDP4: Detection and analysis of recombination patterns in virus genomes. *Virus Evol.* **1**, 1–5 (2015).
33. Larsson, A. AliView: a fast and lightweight alignment viewer and editor for large datasets. *Bioinformatics* **30**, 3276–3278 (2014).
34. Nguyen, L.-T., Schmidt, H. A., Haeseler, Avon & Minh, B. Q. IQ-TREE: A Fast and Effective Stochastic Algorithm for Estimating Maximum-Likelihood Phylogenies. *Mol. Biol. Evol.* **32**, 268–274 (2015).
35. Kalyaanamoorthy, S., Minh, B. Q., Wong, T. K. F., Von Haeseler, A. & Jermiin, L. S. ModelFinder: Fast model selection for accurate phylogenetic estimates. *Nat. Methods* **14**, 587–589 (2017).
36. Rambaut, A. & Drummond, A. *FigTree* (2009).
37. Rambaut, A., Lam, T. T., Max Carvalho, L. & Pybus, O. G. Exploring the temporal structure of heterochronous sequences using TempEst (formerly Path-O-Gen). *Virus Evol.* **2**, vew007 (2016).
38. Suchard, M. A. *et al*. Bayesian phylogenetic and phylodynamic data integration using BEAST 1.10. *Virus Evol.* **4**, 1–5 (2018).
39. Gill, M. S. *et al*. Improving bayesian population dynamics inference: A coalescent-based model for multiple loci. *Mol. Biol. Evol.* **30**, 713–724 (2013).
40. Mir, D. *et al*. Phylodynamics of Yellow Fever Virus in the Americas: New insights into the origin of the 2017 Brazilian outbreak. *Sci. Rep.* **7**, 1–9 (2017).
41. Rambaut, A. & Drummond, A. J. Tracer v1.4. Available from, http://beast.bio.ed.ac.uk/Tracer, *Available from*, http://beast.bio.ed.ac.uk/Tracer (2007).
42. Baele, G., Li, W. L. S., Drummond, A. J., Suchard, M. A. & Lemey, P. Accurate model selection of relaxed molecular clocks in Bayesian phylogenetics. *Mol. Biol. Evol.* **30**, 239–243 (2013).
43. Bielejec, F. *et al*. SpreaD3: Interactive Visualization of Spatiotemporal History and Trait Evolutionary Processes. *Mol. Biol. Evol.* **33**, 2167–2169 (2016).
44. Retief, J. D. Phylogenetic analysis using PHYLIP. In *Bioinformatics methods and protocols* **132**, 243–258 (2000).
45. da Saúde, M. Boletim epidemiológico - Febre Amarela - 17/08/2018. *Bol. Epidemiológico* 1–10 (2018).
46. Vasconcelos, P. F. C. *et al*. Serious adverse events associated with yellow fever 17DD vaccine in Brazil: A report of two cases. *Lancet* **358**, 91–97 (2001).
47. Lindsey, N. P. *et al*. Adverse event reports following yellow fever vaccination. *Vaccine* **26**, 6077–6082 (2008).
48. Whittembury, A. *et al*. Viscerotropic disease following yellow fever vaccination in Peru. *Vaccine* **27**, 5974–5981 (2009).
49. Jenkins, G. M., Rambaut, A., Pybus, O. G. & Holmes, E. C. Rates of molecular evolution in RNA viruses: A quantitative phylogenetic analysis. *J. Mol. Evol.* **54**, 156–165 (2002).
50. Moreira-Soto, A. *et al*. Evidence for multiple sylvatic transmission cycles during the 2016–2017 yellow fever virus outbreak, Brazil. *Clin. Microbiol. Infect.* **24**, 1019.e1–1019.e4 (2018).
51. Shearer, F. M. *et al*. Existing and potential infection risk zones of yellow fever worldwide: a modelling analysis. *Lancet Glob. Heal.* **6**, e270–e278 (2018).
52. Fioravanti, C. *O alarme dos macacos*. *Pesquisa Fapesp* 263 (2018).

53. Camargo-Neves, V. L. Fde *et al*. Entomological investigation of a sylvatic yellow fever area in São Paulo State, Brazil. *Cad. Saude Publica* **21**, 1278–1286 (2005).
54. Mucci, L. F. *et al*. Haemagogus leucocelaenus and Other Mosquitoes Potentially Associated With Sylvatic Yellow Fever In Cantareira State Park In the São Paulo Metropolitan Area, Brazil. *J. Am. Mosq. Control Assoc.* **32**, 329–332 (2016).
55. Abreu, F. V. Sde *et al*. *Haemagogus leucocelaenus* and *Haemagogus janthinomys* are the primary vectors in the major yellow fever outbreak in Brazil, 2016–2018. *Emerg. Microbes Infect.* **8**, 218–231 (2019).
56. Chiarello, A. G. Home Range of the Brown Howler Monkey, Alouatta fusca, in a Forest Fragment of Southeastern Brazil. *Folia Primatol.* **60**, 173–175 (1993).
57. Pinto, L. P. S., Costa, C. M. R., Strier, K. B. & Fonseca, G. A. Bda Habitat, Density and Group Size of Primates in a Brazilian Tropical Forest. *Folia Primatol.* **61**, 135–143 (1993).
58. Bonvicino, C. R. & Viana, M. C. Genetic Diversity of Alouatta (Primates) from Brazilian Atlantic Forest. *J. Primatol.* **04**, 2 (2015).
59. Hamer, D. H. *et al*. Fatal Yellow Fever in Travelers to Brazil, 2018. *Morb. Mortal. Wkly. Rep.* **67**, 340–341 (2018).
60. Nunes, M. R. T. *et al*. Air Travel Is Associated with Intracontinental Spread of Dengue Virus Serotypes 1–3 in Brazil. *PLoS Negl. Trop. Dis*. **8** (2014).
61. Cunha, M. *et al*. Phylodynamics of DENV-1 reveals the spatiotemporal co-circulation of two distinct lineages in 2013 and multiple introductions of dengue virus in Goiás, Brazil. *Infect. Genet. Evol.* **43**, 130–134 (2016).
62. Ortiz-baez, A. S. *et al*. Origin, tempo, and mode of the spread of DENV-4 Genotype IIB across the state of São Paulo, Brazil during the 2012-2013 outbreak. *Mem. Inst. Oswaldo Cruz* **114**, 1–8 (2019).
63. Nunes, M. R. T. *et al*. Emergence and potential for spread of Chikungunya virus in Brazil. *BMC Med*. **13** (2015).
64. Vargas-Mendez, O. & Elton, N. W. Naturally acquired Yellow Fever in wild monkeys of Costa Rica. *Am. J. Trop. Med. Hyg.* **2**, 850–863 (1952).
65. Vasconcelos, P. F. C. *et al*. Genetic Divergence and Dispersal of Yellow Fever Virus, Brazil. **10** (2004).
66. Loehle, C. Social Barriers to Pathogen Transmission in Wild Animal Populations. *Ecology* **76**, 326–335 (1995).
67. Powell, J. R., Gloria-Soria, A. & Kotsakiozi, P. Recent History of Aedes aegypti: Vector Genomics and Epidemiology Records. *Bioscience* **68**, 854–860 (2018).

Acknowledgements

We thank Renato de Souza and Fernando L. Macedo from the Adolfo Lutz Institute for sharing dates and places of confirmed cases of non-human primate (NHP) deaths. We thank the Core Facility for Scientific Research – University of São Paulo (CEFAP-USP/GENIAL) for excellent technical assistance. PMAZ was funded by the Brazilian National Council of Scientific and Technological Development (CNPq) (process no. 441105/2016-5), by the Fiocruz/Pasteur/Aucani-FUSP (process no. 314502) and by the São Paulo Research Foundation (FAPESP) (process no. 2017/23281-6). PHNS is funded by the FAPESP (process no. 2013/21728-2). ECH is funded by an ARC Australian Laureate Fellowship (FL170100022). CTB is funded by the CNPq (process no. 405691/2018-1). MPC and ASOB received FAPESP grants: no. 2016/08204-2 and 2013/25434-3, respectively. CTB received a CAPES fellowship grant no. 1796450. JC received Czech Academy of Sciences grant no. MSM200961703.

Author contributions

Conceptualization: M.d.P.C., A.N.D.N., J.C., P.H.N.S., E.C.H. and P.M.d.A.Z. Medical care for patients during hospitalization: Y.L.H. and J.S. Responsible by the Yellow Fever crisis committee – HCFMUSP: B.P. Autopsies: A.N.D.N., V.A.F.A., M.D. and P.H.N.S. Sample collection: M.d.P.C., and S.Z.P. Molecular analysis: M.d.P.C., S.Z.P., B.B.S.P. and C.T.B. Bioinformatics analysis: M.d.P.C., A.S.O.B. and J.C. Data curation: M.d.P.C., A.N.D.N., A.S.O.B., J.C., E.C.H. and P.M.d.A.Z. Funding acquisition: C.T.B., P.H.S.N., E.C.H. and P.M.d.A.Z. Project administration: M.d.P.C., and P.M.d.A.Z. Manuscript preparation and revision – M.d.P.C., E.C.H. and P.M.d.A.Z. All authors revised, edited and approved the final version of the manuscript.

Competing interests

The authors declare no competing interests.

Additional information

Supplementary information is available for this paper at https://doi.org/10.1038/s41598-019-56650-1.

Correspondence and requests for materials should be addressed to M.d.P.C. or P.M.d.A.Z.

Reprints and permissions information is available at www.nature.com/reprints.

Publisher's note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

Virus Evolution, 2020, 6(2): veaa065

doi: 10.1093/ve/veaa065
Research Article

Red fox viromes in urban and rural landscapes

Sarah J. Campbell,[1] Wilbur Ashley,[1] Margarita Gil-Fernandez,[1] Thomas M. Newsome,[2] Francesca Di Giallonardo,[3] Ayda Susana Ortiz-Baez,[4,†] Jackie E. Mahar,[4,‡] Alison L. Towerton,[5] Michael Gillings,[1] Edward C. Holmes,[4,§] Alexandra J.R. Carthey,[1] and Jemma L. Geoghegan[1,6,7,*,¶]

[1]Department of Biological Sciences, Macquarie University, Sydney, New South Wales 2109, Australia, [2]School of Life and Environmental Sciences, The University of Sydney, Sydney, New South Wales 2006, Australia, [3]The Kirby Institute, University of New South Wales, Sydney, New South Wales 2052, Australia, [4]Marie Bashir Institute for Infectious Diseases and Biosecurity, School of Life and Environmental Sciences and School of Medical Sciences, The University of Sydney, Sydney, New South Wales 2006, Australia, [5]Greater Sydney Local Land Services, Sydney, New South Wales 2750, Australia, [6]Department of Microbiology and Immunology, University of Otago, Dunedin 9016, New Zealand and [7]Institute of Environmental Science and Research, Wellington 5018, New Zealand

*Corresponding author: E-mail: jemma.geoghegan@otago.ac.nz

[†]https://orcid.org/0000-0003-0080-0633

[‡]https://orcid.org/0000-0003-2687-9810

[§]https://orcid.org/0000-0001-9596-3552

[¶]https://orcid.org/0000-0003-0970-0153

Abstract

The Red fox (*Vulpes vulpes*) has established large populations in Australia's urban and rural areas since its introduction following European settlement. The cryptic and highly adaptable nature of foxes allows them to invade cities and live among humans whilst remaining largely unnoticed. Urban living and access to anthropogenic food resources also influence fox ecology. Urban foxes grow larger, live at higher densities, and are more social than their rural counterparts. These ecological changes in urban red foxes are likely to impact the pathogens that they harbour, and foxes could pose a disease risk to humans and other species that share these urban spaces. To investigate this possibility, we used a meta-transcriptomic approach to characterise the virome of urban and rural foxes across the Greater Sydney region in Australia. Urban and rural foxes differed significantly in virome composition, with rural foxes harbouring a greater abundance of viruses compared to their urban counterparts. We identified ten potentially novel vertebrate-associated viruses in both urban and rural foxes, some of which are related to viruses associated with disease in domestic species and humans. These included members of the *Astroviridae*, *Picobirnaviridae*, *Hepeviridae*, and *Picornaviridae* as well as rabbit haemorrhagic disease virus-2. This study sheds light on the viruses carried by urban and rural foxes and emphasises the need for greater genomic surveillance of foxes and other invasive species at the human–wildlife interface.

Key words: Vulpes vulpes ; carnivore ; predator ; canine ; exotic species ; urban ; virus ; metagenomics.

1. Introduction

Red foxes (*Vulpes vulpes*) have the largest natural distribution of any wild terrestrial carnivore (Schipper et al. 2008), extending through Eurasia and North America (Statham et al. 2014). Introduced to Australia in the mid-1800s, they rapidly expanded across the continent. Red foxes exploit a wide range of habitats with varying climates, from alpine to desert, and are considered one of the most adaptable species on the planet. They are broadly distributed across natural and forested landscapes as well as highly urbanised, human dominated areas (Saunders, Gentle, and Dickman 2010; Bateman and Fleming 2012). Red fox home ranges vary depending on resource availability and land-use type. In Australia, home ranges for foxes in arid regions can reach at least 120 km^2 (Newsome, Spencer, and Dickman 2017), between 5 and 7km^2 in rural areas (Coman, Robinson, and Beaumont 1991) and <1 km^2 in urban centres (Marks and Bloomfield 2006).

Foxes have recently established a large presence in major metropolitan centres (Marks and Bloomfield 1999; Saunders, Gentle, and Dickman 2010). Urban areas support surprisingly high densities of foxes. For example, there are up to sixteen individuals per km^2 in Melbourne (Marks and Bloomfield 1999), compared to just 0.2 individuals per km^2 in rural areas (Saunders, Gentle, and Dickman 2010). In Bristol city in the UK, densities reach as high as thirty-five foxes per km^2 (Baker et al. 2001).

Predation by red foxes is a key threat to Australian native fauna (EPBC 1999). Due to this threat to vulnerable wildlife and Australian biodiversity, fox populations are actively controlled. Whilst poison baiting is common and cost-effective in rural areas (Saunders, Gentle, and Dickman 2010), risks to pets and humans restrict control methods in urban areas to trapping and shooting (Marks et al. 1996). These methods are both relatively expensive and difficult to apply at large scales, making urban fox control challenging.

Red foxes are both cryptic and nocturnal, often remaining unnoticed in urban areas despite their high abundance (Phillips and Catling 1991; Doncaster and Macdonald 1997). They thrive on anthropogenic resources and may develop distinct behaviours through urban living (Contesse et al. 2004; Bateman and Fleming 2012; Stepkovitch 2017). Other urban carnivores such as coyotes (*Canis latrans*) display increased boldness and decreased human aversion by comparison to rural counterparts (Bateman and Fleming 2012; Robertson 2018; Breck et al. 2019). Urban carnivores often become larger in size, which may have positive effects on fitness and fecundity (Bateman and Fleming 2012; Stepkovitch et al. 2019). Abundant food can decrease carnivore home ranges, support higher densities, and increase conspecific encounter rates (Bateman and Fleming 2012; Newsome et al. 2015; Dorning and Harris 2019). Urban fox family group sizes are often larger than rural ones, as juvenile females may forego dispersal to assist with cub rearing (Macdonald 1979, 1983; Marks and Bloomfield 1999). Thus, urban environments may enhance conspecific tolerance and social behaviours in foxes (Macdonald 1979, 1983; Marks and Bloomfield 1999; Dorning and Harris 2019).

Although red foxes are known to harbour a diversity of viruses (Bodewes et al. 2013; Lojkić et al. 2016), it is unknown whether urban and rural foxes have different viral compositions. High-density living and increased contact can increase pathogen transmission rates among hosts (Nunn et al. 2015). As such, a high-density population of cryptic urban foxes living in proximity to largely unsuspecting humans could pose an important pathogen risk. Foxes interact with human refuse, including compost and rubbish bins, and consume food scraps from surfaces such as outdoor barbeques and furniture, eat from pet bowls and wildlife feeding stations, and defaecate nearby, increasing the potential for pathogen transfer (Contesse et al. 2004). In addition, as urban animals often habituate to humans (Bateman and Fleming 2012), we might predict an increase in fox–human interactions with the potential for diseased transmission between the two species.

Using a meta-transcriptomic approach we describe, for the first time, the virome of the introduced Australian red fox sampled from urban and rural regions. We hypothesised that foxes in urban areas could harbour a greater viral diversity and abundance compared to rural foxes, due to higher population densities and increased conspecific interactions in urban areas. Whilst there is limited information on fox social dynamics in Australia, we also postulated that females could harbour a greater diversity and abundance of viruses than males due to particular social behaviours reported for female foxes in their native ranges, such as cooperative cub rearing (Macdonald 1979, 1983). To this end, samples (liver, faecal, and ectoparasite) were collected from foxes around the Greater Sydney region, Australia, including in urban and rural areas (Fig. 1). Due to diet and organ function, we hypothesised that these tissues comprised very different viromes and together provided a more comprehensive view of the red fox virome. Samples were pooled (based on sampling location, tissue type and sex) and subject to RNA sequencing to reveal viral diversity, evolution, and abundance.

2. Materials and methods

2.1 Sample collection

The current project was part of a larger research program into urban foxes in partnership with Greater Sydney Local Land Services, a New South Wales State Government organisation responsible for management of pest species across the region. We collected fresh carcases from independent licenced trappers and shooters who were actively controlling foxes in the Greater Sydney region (see Fig. 1 for sample locations). To minimise degradation of RNA, samples were taken as soon as possible after death (03:19:00 ± 02:59:00 h post-mortem, $n = 27$). One carcase had been frozen for approximately 1 week and one carcase had been dead for an unknown amount of time. The foxes used for this study were either trapped in cages and shot, or tracked and shot. One individual was obtained as recent roadkill. Foxes killed by poison baits were excluded.

Whole fox carcases were collected and transported to the laboratory where they were immediately dissected to collect faecal, liver, and ectoparasite samples. All samples were individually stored in RNALater at −80 °C. We sampled a total of twenty-nine individual foxes; thirteen males and sixteen females. For this study, foxes were classified as juvenile if their body mass and body length were less than 3.3 kg and 51 cm, respectively. These values were chosen as the body mass of an

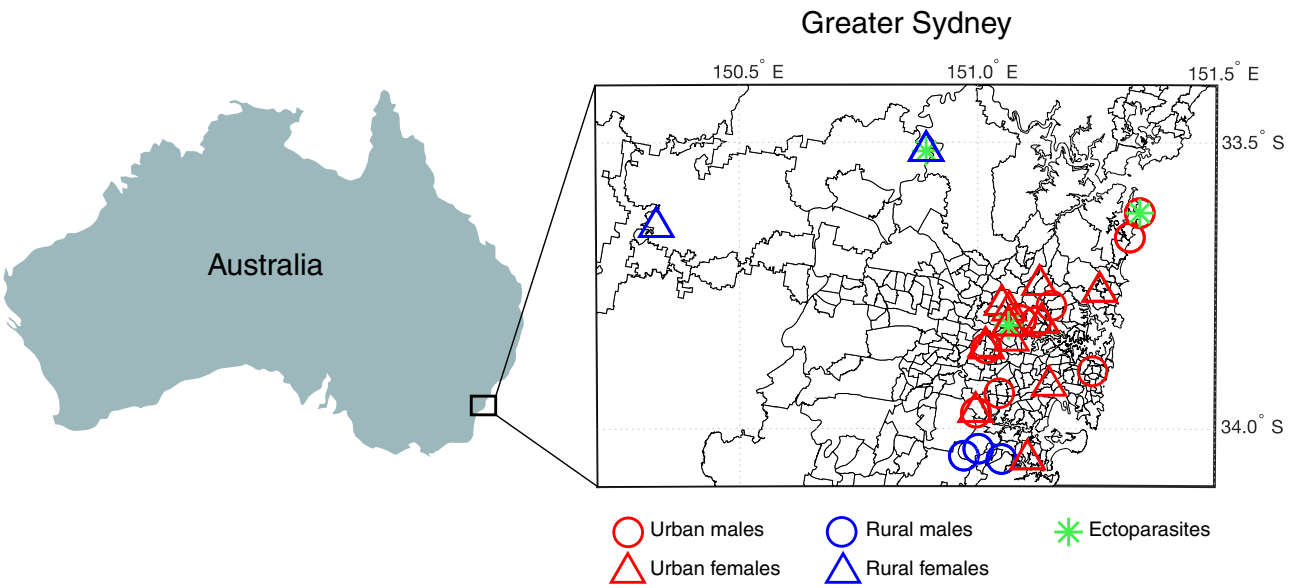


Figure 1. Map of the Greater Sydney region showing fox sampling locations of urban (red) and rural (blue) fox carcases, identified as male (circle) or female (triangle), as well as those harbouring ectoparasites (green asterisk).

adult red fox can range between 3.3 and 8.2 kg, whilst body length can range between 51 and 78 cm (when measured from the tip of the nose to the first vertebra of the tail) (Cavallini 1995). Based on this assessment, twenty-five foxes were classified as adults (twelve males and thirteen females) and four as juveniles (one male and three females).

2.2 Sampling in urban and rural areas

Fox sampling relied on coordination with professional pest control operators who focus control efforts in specific locations in accordance with local control initiatives. For this reason, a representative sample across a land-use gradient from urban to rural was not possible. Sufficiently fresh rural and bushland fox samples were also difficult to obtain since poison baiting is the principal control method in these areas. Therefore, 'rural' was broadly defined as any natural bushland, national park, mostly agricultural, or sparsely populated region outside the central urban districts, with a human population density of fewer than 500 people per km^2. Similarly, 'urban' was defined as built-up areas inside the central urban district (including parks, gardens, and golf courses) with a population density of more than 500 people per km^2 either in the area sampled or in the immediate surrounding areas. Human population density information was obtained from the Australian Bureau of Statistics (2016 census data) (Australian Bureau of Statistics 2016a). Central urban districts were defined by the Urban Centres and Localities statistical classification (Australian Bureau of Statistics 2016b). Land-use classification and human population density cut-offs were loosely based on work by Stepkovitch et al. (2019).

2.3 RNA extraction and whole-transcriptome sequencing

Qiagen RNeasy Plus Mini Kits were used to extract RNA from liver, faecal, and ectoparasite samples from collected red fox carcases. Thawed samples were transferred to a lysis buffer solution containing 1 per cent β-mercaptoethanol and 0.5 per cent Reagent DX. Samples were homogenised and centrifuged. DNA was removed from the supernatant via gDNA eliminator spin column and RNA was eluted via RNeasy spin column. RNA concentration and purity were measured using the Thermo Fisher Nanodrop. Samples were pooled based on land-use category (urban or rural), sex, and sample type (liver, faecal, or ectoparasite), resulting in nine representative sample pools (Table 1). Adults and juveniles were pooled as only two juveniles were sampled. Ectoparasites included fleas (*Siphonaptera*) and ticks (*Ixodida*). These were not classified below the Order level and due to the small number sampled were also pooled. The TruSeq Stranded Total RNA Ribo-Zero Gold (h/m/r) kit was used to prepare pooled samples for sequencing. Pooled samples were sequenced on the NextSeq 500 with 2× 75 bp output at the Ramaciotti Centre for Genomics at the University of New South Wales, Sydney. Sequencing resulted in nine representative data libraries (Table 1). The raw reads and virus sequences are available on NCBI's SRA database under BioProject PRJNA640177 GenBank accession numbers MT833874-MT833883.

2.4 Virus discovery

Sequencing reads were assembled *de novo* into longer sequences (contigs) based on overlapping nucleotide regions using Trinity RNA-Seq (Haas et al. 2013). Assembled contigs were assigned to a taxonomic group (virus, Bacteria, Archaea, and Eukarya) and viruses were identified to their closest species match based on sequence similarity searches against the NCBI nucleotide (nt) and non-redundant protein (nr) databases using BLASTn (Altschul et al. 1990) and Diamond (BLASTX) (Buchfink, Xie, and Huson 2015), respectively. An *e*-value threshold of 1×10^{-5} was used as a cut-off to identify positive matches. We removed non-viral hits, including host contigs with similarity to viral sequences (e.g. endogenous viral elements).

2.5 Inferring the evolutionary history of fox viruses

We inferred the phylogenetic relationships of the vertebrate-associated viruses identified in the fox samples. Vertebrate-associated viruses were defined as viruses, which shared sequence

Table 1. Breakdown of red fox representative samples, detailing land use, sex, and sample type, as well as the number of individuals pooled for RNA sequencing.

Representative sample	Land use	Sex	Sample type	Number of individual foxes pooled	Viral transcripts found?
1	Urban	Male	Liver	9	No
2	Urban	Male	Faeces	6	Yes
3	Rural	Male	Liver	3	No
4	Rural	Male	Faeces	3	Yes
5	Urban	Female	Liver	9	No
6	Urban	Female	Faeces	13	Yes
7	Rural	Female	Liver	3	Yes
8	Rural	Female	Faeces	3	Yes
9	Both	Male (1) Female (2)	Ectoparasites	3	Yes

Table 2. Vertebrate-associated viral contigs, contig length (nt), percent abundance in their respective pools, and the percent amino acid identity to their closest match on NCBI/GenBank.

Land use (sex)	Virus name (species)	Virus family	Contig length (nt)	% Relative abundance	Closest match (GenBank accession number)	% Amino acid identity	Sample type
Rural (female)	Vixey virus	*Picornaviridae*	2,427	0.007	Canine kobuvirus (AZS64124.1)	97.65	Faeces
	Wilde virus-1	*Picornaviridae*	7,236	5.66	Canine picornavirus (YP_005351240.)	89.18	Faeces
	Wilde virus-3	*Picornaviridae*	1,428	0.0004	Canine picornavirus (AMX81409.1.)	96.22	Liver
	Swiper virus	*Hepeviridae*	7,374	0.01	Elicom virus-1 (YP_009553584.)	28.92	Faeces
	Red fox-associated rabbit haemorrhagic disease virus-2	*Caliciviridae*	7,026	0.14	Rabbit haemorrhagic disease virus-2 (MF421679.1)	99.62	Faeces
Rural (male)	Tod virus-2	*Picornaviridae*	4,263	0.17	Canine picodicistrovirus (YP_007947664.)	98.53	Faeces
	Vulpix virus	*Astroviridae*	2,556	0.046	Feline astrovirus (YP_009052460.)	96.11	Faeces
Urban (female)	Tod virus-1	*Picornaviridae*	2,062	0.0004	Canine picodicistrovirus (YP_007947664.)	98.83	Faeces
	Charmer virus	Picobirnaviridae	448	0.0001	Wolf picobirnavirus (ANS53886.1)	80.27	Faeces
Urban (male)	Wilde virus-2	*Picornaviridae*	1,524	0.00058	Canine picornavirus (YP_005351240.)	73.37	Faeces

similarity to other known vertebrate viruses. Due to the high divergence of the virus transcripts, we used only the RNA-dependant RNA polymerase (RdRp) transcripts for phylogenetic analysis. First, the amino acid translations of the viral transcripts were combined with other virus protein sequences from the same virus families obtained from GenBank (Table 2). Second, the sequences were aligned using MAFFT v.3.4, employing the E-INS-I algorithm. Ambiguously aligned regions were removed using trimAl v.1.2 (Capella-Gutiérrez, Silla-Martínez, and Gabaldón 2009). To estimate phylogenetic trees, we selected the optimal model of amino acid substitution identified using the Bayesian Information Criterion as implemented in Modelgenerator v0.85 (Keane et al. 2006) and employed the maximum-likelihood approach available in PhyML v3.1 (Guindon et al. 2010) with 1,000 bootstrap replicates. For the viral transcript matching rabbit haemorrhagic disease virus-2 (RHDV2), we used a nucleotide alignment with similar viruses. New viruses were named after fictional fox characters.

2.6 Diversity and abundance analysis

Transcript abundance for all viruses (vertebrate and invertebrate associated) was estimated using RSEM within Trinity (Li and Dewey 2011). Specifically, we assessed how many short

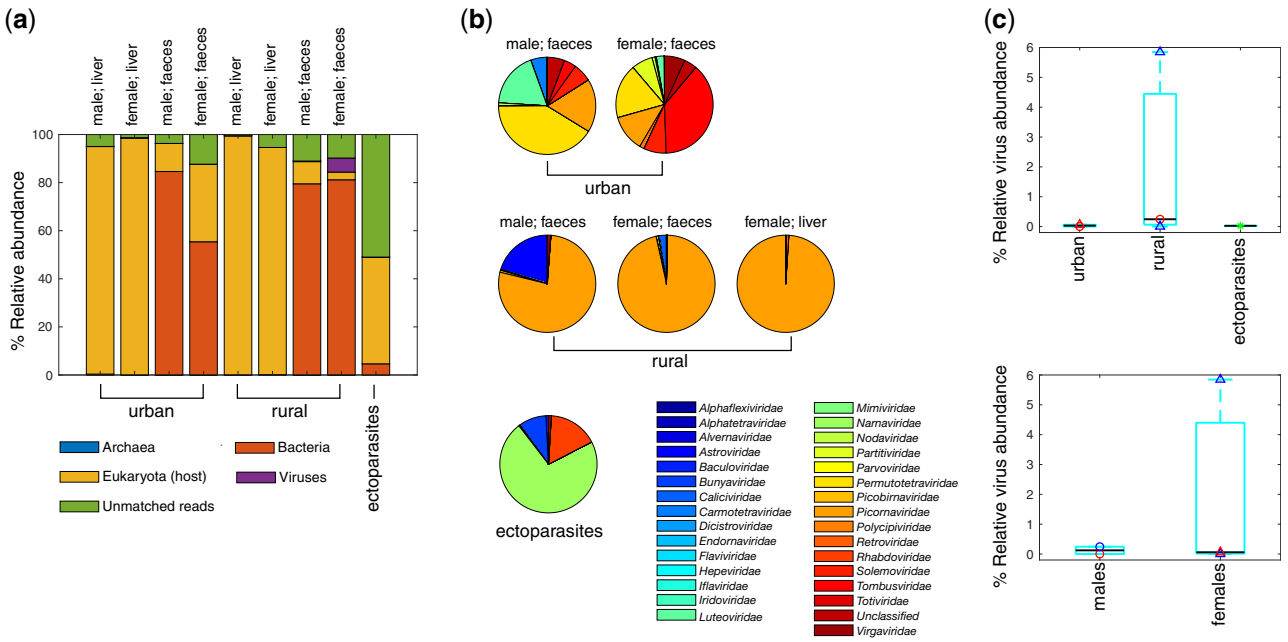


Figure 2. Overview of the red fox virome. (a) Percentage abundance of each taxonomic group identified in each respective pooled sample, standardised against the number of raw reads per pool. Due to their low abundance, archaea (0.002–0.021 per cent) and some of the viral reads (0.001–5.85 per cent) are too small to visualise. (b) Percentage abundance of (eukaryotic-associated) viral families detected in each respective pooled sample (excluding bacteriophage). (c) Boxplots showing percentage abundance of (eukaryotic-associated) viral reads in urban, rural, and ectoparasite samples and males and females. A black line indicates the median and the bottom and top edges of the box indicate the 25th and 75th percentiles, respectively. Raw abundances are superimposed, and the colour and shape of data points are as in Fig. 1.

reads within a given library mapped to a particular transcript. Raw counts were then standardised against the total number of reads within each library. Virome diversity (i.e. virus species richness) and relative abundance were compared among samples using a non-metric multidimensional scaling (nMDS) ordination in conjunction with an analysis of similarities (ANOSIM) based on Bray–Curtis dissimilarity as implemented in the vegan package in R (Oksanen et al. 2019). To determine which viral families were contributing the most to differences between samples, an 'indicator species' analysis was performed, using a point biserial coefficient of correlation within the indicspecies package in R (De Cáceres, Legendre, and Moretti 2010).

3. Results

Meta-transcriptomic sequencing of nine representative pooled samples resulted in 44–57 million paired reads per pool (593,406,706 reads in total). BLAST analyses revealed that the faecal samples were dominated by bacteria (51.17–84.61%), whilst the liver samples were dominated by eukaryotic transcripts (92.90–99.43%), largely comprising fox RNA. Viruses made up a small proportion of the four representative faecal samples (0.002–5.85%) and were detected in only one of the representative liver samples (0.001%). Archaea were detected at very low levels in faecal samples only (0.002–0.021%). The ectoparasites (fleas and ticks) differed substantially to the liver and faecal samples with 50.97 per cent of reads classed as 'unmatched' meaning they did not share sequence similarity to any known sequence. The remainder of the contigs from ectoparasite samples were from eukaryotes (44.39%), bacteria (4.64%), and viruses (0.004%). Unmatched reads in liver and faecal samples ranged between 0.52 per cent and 12.22 per cent (Fig. 2a).

Multiple novel vertebrate-associated virus transcripts were identified from both urban and rural foxes, including a hepevirus, picobirnavirus, astrovirus, and various picornaviruses (Table 2). In addition, we found virus transcripts with sequence similarity to RHDV2. Vertebrate-associated virus transcripts represented between 0.4 per cent and 98 per cent of viral reads. The remainder comprised mostly invertebrate-, plant-, and fungi-associated virus transcripts, which were most likely acquired from the foxes' diet. As no vertebrate-associated viruses were detected in the ectoparasite pool, we performed no further evolutionary analyses.

3.1 Virome composition

Urban, rural, and ectoparasite samples had distinctly different virome compositions (ANOSIM $R = 1$, $P = 0.0167$; Figs 2 and 3). Transcripts from a total of thirty distinct viral families were identified across the six pools in which viral RNA was detected (rural male faeces, rural female faeces, rural female liver, urban male faeces, urban female faeces, and ectoparasites). Overall, twenty-one viral families were identified in transcripts from urban foxes and nineteen from rural foxes. Urban foxes exhibited a slightly higher diversity of viruses compared to rural foxes; transcripts from the latter were heavily dominated by *Picornaviridae*, which made up between 77.33 and 98.97 per cent of the virome of rural foxes (Fig. 2b). Indicator species analysis suggested that *Picornaviridae* were associated with rural samples (stat = 0.978, $P = 0.0496$), whilst *Nodaviridae* were associated with urban samples (stat = 0.998, $P = 0.0498$). Viral diversity was higher in females (twenty-five distinct viral families) than in

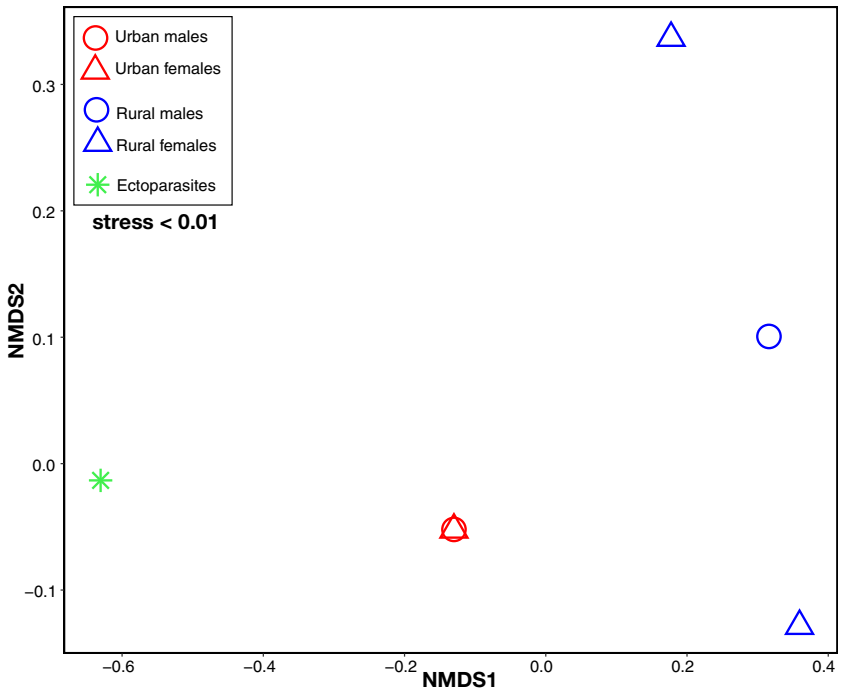


Figure 3. nMDS ordination showing differences in virome composition (at the family level) among samples according to habitat and sex. Individual points represent individual pooled samples. Points closer together have a more similar virome composition (based on Bray–Curtis dissimilarity, which incorporates both the diversity and abundance of viruses) and *vice versa* for those further apart. The stress value was <0.01 and is indicated on the figure.

males (thirteen distinct viral families). A much larger percentage of the viral transcripts identified were vertebrate associated in rural foxes (male: 98.23% and female: 97.84%) compared to urban foxes (male: 2.41% and female: 0.39%), although this percentage was higher in males in both groups. In this context, it is important to note that some virus transcripts found here may be the result of contamination by reagents.

On average, total viral abundance (including both vertebrate and non-vertebrate viruses) was higher in rural foxes ($2.03 \pm 3.31\%$, $n = 3$) than in urban foxes ($0.03 \pm 0.04\%$, $n = 2$), and in female foxes ($1.97 \pm 3.36\%$, $n = 3$) than in male foxes ($0.12 \pm 0.17\%$, $n = 2$) (Fig. 2c). However, due to the small sample size, differences may be due to some individual animals contributing more to overall abundance or diversity in their respective pool than others. For example, the rural female fox pool (comprising three individual foxes) contained an unusually high number of viruses (>5%) compared to the others. This may have inflated virus abundance counts in females when combined. Whilst virome composition was compared among a relatively small number of samples, this is balanced by the fact that each sample comprises the viromes of multiple individual foxes ($n = 3$–13 foxes per pool; Table 1).

3.2 Vertebrate-associated viruses in foxes

3.2.1 *Hepeviridae*

Hepevirus (positive-sense single-stranded RNA viruses) sequences were discovered in the rural female faecal samples. Tentatively named swiper virus, this virus transcript was very distinct in sequence, sharing only 28.92 per cent amino acid identity to its closest relative, elicom virus-1 from mussels, and had a relative abundance of 0.01 per cent (Table 2). Whilst its closest genetic relative is not from a vertebrate host suggesting it may be a diet-associated contaminant, phylogenetic analysis of the RdRp encoding region placed this hepevirus in proximity to both house mouse hepevirus and elicom virus-1, with these viruses forming a distinct monophyletic group (Fig. 4).

3.2.2 *Astroviridae*

We detected an astrovirus (positive-sense single-stranded RNA virus), tentatively named vulpix virus, in the rural male faecal samples. Notably, the sequence shared a 96.11 per cent amino acid identity with feline astrovirus D1 and had a relative abundance of 0.046 per cent (Table 2). Based on phylogenetic analysis of the RdRp, this virus clustered with other mammalian-associated viruses within the mamastroviruses (Fig. 4).

3.2.3 *Picobirnaviridae*

Picobirnavirus (double-stranded RNA viruses) sequences were detected in urban male, rural male, and urban female faecal samples. As some of the sequences represented less conserved regions of the viral genome, only one RdRp sequence (from the urban female samples) was used for phylogenetic analysis. The sequence, tentatively named charmer virus, shared an 80.27 per cent amino acid identity with a picobirnavirus found in wolves and had a relative abundance of 0.0001 per cent (Table 2). The sequence also clustered with other mammalian-associated picobirnaviruses (Fig. 4).

3.2.4 *Picornaviridae*

Several picornaviruses (positive-sense single-stranded RNA viruses) were discovered. Two kobuvirus-related sequences were discovered in the rural female faecal samples. The longer sequence, tentatively named vixey virus, shared highest amino acid identity with canine kobuvirus from a domestic dog (97.65%) and had a relative abundance of 0.007 per cent

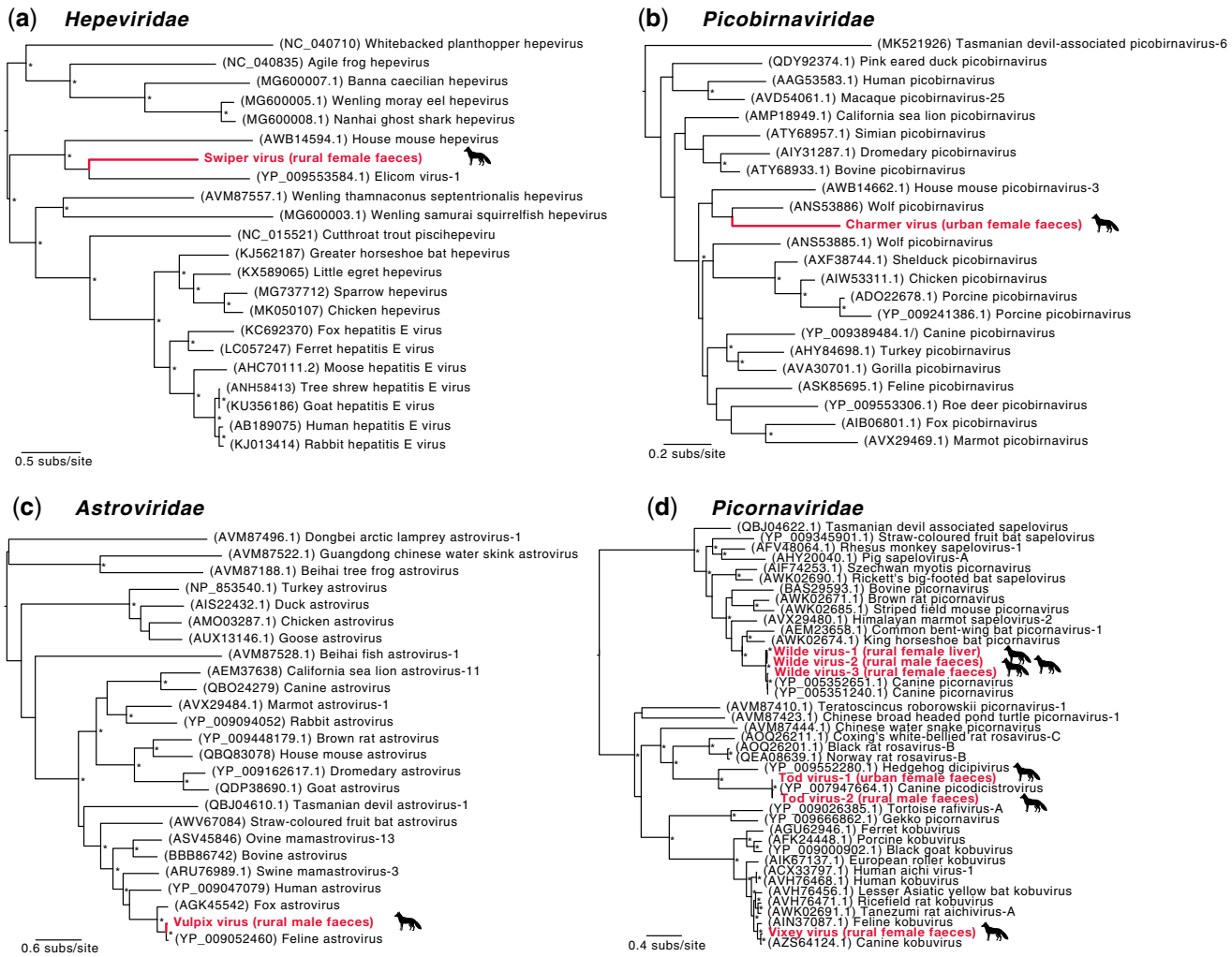


Figure 4. Phylogenetic relationships of likely vertebrate-associated viruses discovered from assembled contigs: (a) *Hepeviridae*, (b) *Picobirnaviridae*, (c) *Astroviridae*, and (d) *Picornaviridae*. The maximum-likelihood phylogenetic trees show the topological position of the newly discovered potential viruses (bold, red text), in the context of their closest relatives. All branches are scaled to the number of amino acid substitutions per site and trees were mid-point rooted for clarity only. An asterisk indicates node support of >70 per cent bootstrap support.

(Table 2). Analysis of the RdRp region showed that the sequence clustered most closely with feline kobuvirus and other mammalian kobuviruses (Fig. 4).

A number of picodicistrovirus sequences were detected in the urban male, rural male, and urban female faecal samples. Two of the sequences, tentatively named tod virus-1 and tod virus-2, both shared 98 per cent amino acid identity with canine picodicistrovirus (Table 2). Based on analysis of the RdRp region, the sequences clustered together with mammalian dicipivirus and rosaviruses as well as reptilian picornaviruses (Fig. 4).

Multiple picornavirus sequences were identified in the rural male faecal and the rural female faecal and liver samples. Three sequences, tentatively named wilde virus-1, 2, and 3, all shared between 73 and 89 per cent amino acid identity with canine picornavirus and had relative abundances of 5.66 per cent, 0.00058 per cent, and 0.0004 per cent, respectively (Table 2). These sequences clustered with other mammalian picornaviruses (Fig. 4).

3.2.5 *Caliciviridae*

One of the most striking observations was the identification of RHDV2 (a positive-sense single-stranded RNA virus) in rural female and urban male faecal samples. The viral sequence in the rural female samples shared a 99.62 per cent amino acid identity with RHDV2 isolated from rabbits between 2015 and 2016 and had a relative abundance of 0.14 per cent (Table 2) (Fig. 5). The viral sequence in the urban male samples was too short to enable phylogenetic analysis. This is the second time that RHDV2 has been found in non-rabbit hosts (Chong et al. 2019), presumably through rabbit consumption in this case.

4. Discussion

We show that Sydney's red foxes, in both urban and rural environments, harbour a wide diversity of viruses, some of which are genetically similar to those that infect domestic pets and humans. Domestic mammals tend to hold central positions in mammal viral transmission networks (Wells et al. 2020). The

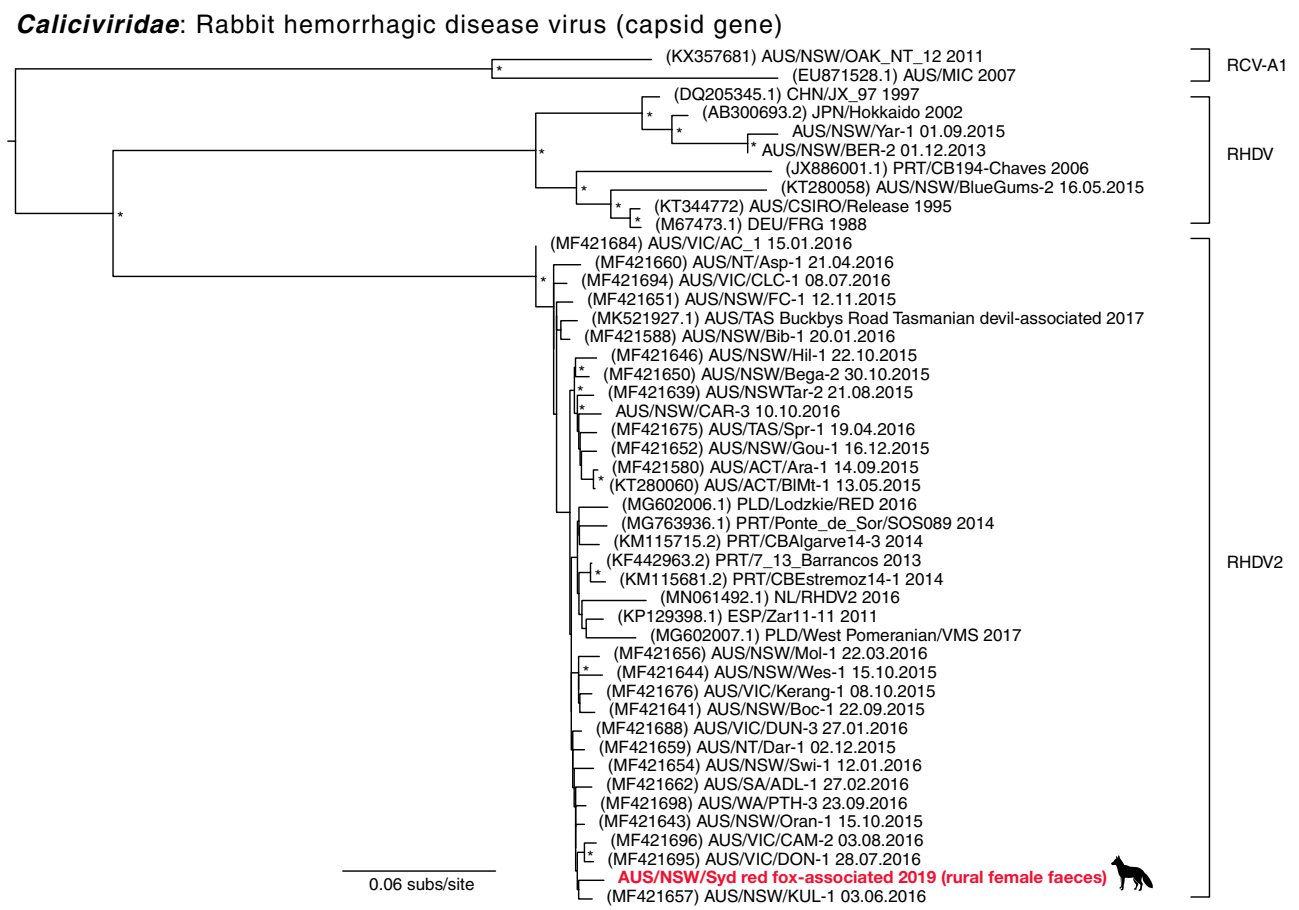


Figure 5. A maximum-likelihood phylogenetic tree showing the topological position of RHDV2 capsid gene in the red fox (bold, red text), in the context of its closest relatives. Major clades are labelled. All branches are scaled to the number of nucleotide substitutions per site and trees were mid-point rooted for clarity only. An asterisk indicates node support of >70 per cent bootstrap support.

close genetic similarity of the viruses found here to viruses frequently found in common domestic pets such as cats and dogs suggests that cross-species transmission between foxes and domestic species may have occurred. The most cited case of viral transmission between humans and domestic pets is the transmission of rabies virus (Ghasemzadeh and Namazi 2015), although other examples include noroviruses from dogs, isolated cases of influenza A(H7N2) virus from cats (Lee et al. 2017; Marinova-Petkova et al. 2017), and numerous bacterial diseases and parasites (Ghasemzadeh and Namazi 2015; O'Neil 2018). There may also be additional cases of viral sharing between humans and their pets, although these may go undiagnosed due to insufficient knowledge of the genetic variability of these viruses and their relationships with hosts.

All vertebrate-associated viruses found here were RNA viruses. Although this may in part be due to the reliance on transcript-based viral detection, RNA viruses are in general characterised by lower host specificity than DNA viruses, reflecting an increased occurrence of cross-species transmission (Geoghegan, Duchêne, and Holmes 2017; Wells et al. 2020). The opportunity for interactions between urban wildlife, pets, and humans provides likely transmission pathways for novel RNA viruses. Indeed, eukaryotic parasites are already known to infect human hosts following the wildlife–domestic pet–human transmission network (Wells et al. 2018).

We discovered viral transcripts with some sequence similarity to the *Hepeviridae* that cause hepatitis E in mammals, which has already been isolated from various domestic and wild animals including foxes in the Netherlands (Meng 2010; Bodewes et al. 2013). Confirmed zoonotic cases include transmission to humans from domestic pigs, cats, and wild rodents (Meng 2010; Dremsek et al. 2012). In contrast, the hepevirus detected here was phylogenetically distinct from the fox hepatitis E virus previously detected (Bodewes et al. 2013) and instead was more closely related to hepeviruses detected in freshwater mussels and a house mouse. Hence, although we have classed the virus as vertebrate associated, its divergent phylogenetic position could in fact mean that it results from dietary consumption.

The astrovirus transcript (vulpix virus) showed the greatest sequence similarity (96 per cent) to astroviruses from domestic cats as well as from other foxes, humans, and pigs. Astroviruses have a broad host range (Donato and Vijaykrishna 2017) and are frequently detected in the faeces of mammals, birds, and humans with gastroenteritis (Finkbeiner et al. 2009; De Benedictis et al. 2011). Astroviruses have also been associated with other diseases and disorders such as shaking syndrome in minks (Blomström et al. 2010), neurological disease in cattle (Li et al. 2013), and encephalitis in humans (Quan et al. 2010). Some human astroviruses are more closely related to those in animals than to each other, suggesting that these viruses periodically emerge from zoonotic origins (Kapoor et al. 2009). The similarity

of fox astroviruses to those found in cats indicates that these viruses may have jumped hosts in the past and highlights further the potential role of domestic pets and wildlife in virus transmission.

Picobirnaviruses are found in humans and other mammals and are thought to be linked with gastroenteritis, however their role in disease remains unclear (Malik et al. 2014; Conceição-Neto et al. 2016). The picobirnavirus-related transcript found here showed the greatest sequence similarly to a picobirnavirus found in wolves with diarrhoeic symptoms (Conceição-Neto et al. 2016). It is also similar to picobirnaviruses described as potentially zoonotic in humans with gastroenteritis (Yinda et al. 2019). There is, however, evidence that picobirnaviruses may actually be bacteriophage rather than eukaryote-associated viruses (Krishnamurthy and Wang 2018), such that the virology of these viruses is currently unclear.

We identified novel fox viruses within the *Picornaviridae* belonging to three distinct genera: kobuvirus, picodicistrovirus, and picornavirus. The *Picornaviridae* are a large and diverse family that include viruses associated with a variety of human diseases such as hand, foot and mouth disease, polio, myocarditis, hepatitis A virus, and rhinovirus (Zell 2018). All viral sequences here were most closely related to those viruses previously found in dogs. Whilst we cannot assume that these viruses cause disease, kobuviruses have been isolated from dogs and other mammals with diarrhoeic symptoms (Reuter, Boros, and Pankovics 2011; Di Martino et al. 2013). Additionally, the fox picornaviruses found here are closely related to sapeloviruses that cause encephalitis in domestic pigs (Lan et al. 2011; Schock et al. 2014; Arruda et al. 2017).

Finally, and of particular note, we identified RHDV2 in fox faeces. RHDV was initially released (or escaped) in Australia in 1995 following testing as a biological control agent for invasive rabbits. A novel variant of the disease, RHDV2, began circulating in Australia in 2015 and is presumed to be an incursion from Europe where it first emerged in 2010 (Hall et al. 2015). RHDV2 has become the dominant strain circulating in Australia's wild rabbits (Mahar et al. 2018). The virus identified here was most closely related to RHDV2 strains found in rabbits in New South Wales, Australia in 2015–6. It is likely, then, that Sydney foxes consume diseased rabbits and the virus is simply a gut contaminant with no active RHDV2 replication in the fox host. Although it is worth noting that antibodies against RHDV have been detected in red foxes in Germany, there was no evidence of illness or viral replication (Frölich, Klima, and Dedek 1998).

Urbanisation influences pathogen exposure and prevalence in wildlife. For example, the prevalence of parvovirus increases with proximity to urban areas in grey foxes (*Urocyon cinereoargenteus*) in the USA (Riley, Foley, and Chomel 2004), and dogs in urban areas in Brazil harbour more tick-borne pathogens than rural dogs (Vieira et al. 2013). In addition, the prevalence of West Nile virus in wild birds in the USA increases with proximity to urban areas and human population density (Gibbs et al. 2006). Here, we found the highest overall viral abundance in rural foxes whilst urban foxes harboured a slightly higher diversity of viruses (Fig. 2b and c). Whilst differences in overall abundance and diversity of viruses present in foxes may be a reflection of differences in diet and environment, we found rural foxes to have a much higher abundance of vertebrate-associated viruses than urban. It has previously been suggested that red foxes in highly urbanised areas experience lower exposure to canine distemper virus due to reduced movement opportunities as a result of wildlife corridors being absent in densely built-up areas (Gras et al. 2018). By comparison, exposure to canine distemper virus increased in areas with more natural habitats (Gras et al. 2018).

It is possible that urban living reduces fox susceptibility to viral infection by positively influencing host immunity. For example, an abundance of rich food sources would increase nutritional intake, positively influencing overall health and condition and hence resistance to viral infections (Beldomenico and Begon 2010). Kit foxes (*Vulpes macrotis*) in urban areas in California show less nutritional stress, increased body condition, and improved immune function when compared to foxes in a nearby nature reserve (Cypher and Frost 1999). Australian lace monitors (*Varanus varius*) consuming human refuse experience improved body condition and reduced blood parasite infection compared to those that do not subsist on anthropogenic food waste (Jessop et al. 2012). Foxes in urban Sydney grow larger and are heavier than foxes in rural areas (Stepkovitch et al. 2019), and there may be an advantage to consuming anthropogenic food sources for overall condition and pathogen resistance.

Across both rural and urban habitats, we observed that female foxes harboured a higher abundance and had almost twice the diversity of viruses found in male foxes (when including both vertebrate and non-vertebrate associated). This difference in viromes may indicate different ecologies and behaviours in male and female foxes. Whilst other studies looking at sex differences and immunity suggest that females typically display stronger immune responses and reduced pathogen load compared to males (Klein 2000), greater sociality in females (Macdonald 1979, 1983) may increase viral transmission opportunities. However, our understanding of red fox sociality in Australia is limited (Newsome 1995) and males may be more likely to be involved in aggressive encounters with conspecifics than females (White and Harris 1994). Alternatively, a combination of biological and ecological differences, such as hormones, diet, and environment, could contribute to variation in male and female viromes (Vemuri et al. 2019).

Multiple co-occurring factors could simultaneously affect viral infection in Sydney's foxes. Additional assessments of habitat structure, fox densities, movement behaviours, and social dynamics in urban and rural areas in the Greater Sydney region will help to elucidate such factors. An obvious extension to this work is to examine fox viromes across a more comprehensive urban-rural gradient, including foxes from more isolated bush habitats. This would help us to understand differences in pathogen prevalence and transmission between isolated natural habitats and more disturbed environments, and how introduced species such as foxes contribute to disease prevalence across different ecosystems. Another useful approach could compare viral transmission dynamics in red foxes between their native and introduced ranges.

Human encroachment on wild environments and the adaptation of wild animals to urban areas continues to intensify human–wildlife interactions. The effects of urbanisation on wildlife pathogen dynamics may have unexpected consequences for human and domestic animal health. Although we cannot say definitively that the viruses identified here cause disease outbreaks or spill-over events, it is clear that foxes living in Greater Sydney carry viruses that are related to those found in domestic animals and humans. Our findings indicate that foxes may be reservoirs for viral pathogens with zoonotic potential.

Conflict of interest: None declared.

Acknowledgements

This work was funded by a grant awarded by the Greater Sydney Local Land Services. A.J.R.C. is a recipient of the Macquarie University Research Fellowship. E.C.H. is supported by an ARC Australian Laureate Fellowship (FL170100022).

References

Altschul, S. F. et al. (1990) 'Basic Local Alignment Search Tool', *Journal of Molecular Biology*, 215: 403–10.

Arruda, P. H. E. et al. (2017) 'Detection of a Novel Sapelovirus in Central Nervous Tissue of Pigs with Polioencephalomyelitis in the USA', *Transboundary and Emerging Diseases*, 64: 311–5.

Australian Bureau of Statistics (2016a) *Regional Population Growth, Australia 2016–17*, Cat. No. 3218.0, Australian Bureau of Statistics.

Australian Bureau of Statistics (2016b) *Australian Statistical Geography Standard (ASGS): Volume 4—Significant Urban Areas, Urban Centres and Localities, Section of State, July 2016*, Cat. No. 1270.0.55.004, Australian Bureau of Statistics.

Baker, P. J. et al. (2001) 'Differences in the Capture Rate of Cage-Trapped Red Foxes *Vulpes vulpes* and an Evaluation of Rabies Control Measures in Britain', *Journal of Applied Ecology*, 38: 823–35.

Bateman, P. W., and Fleming, P. A. (2012) 'Big City Life: Carnivores in Urban Environments', *Journal of Zoology*, 287: 1–23.

Beldomenico, P. M., and Begon, M. (2010) 'Disease Spread, Susceptibility and Infection Intensity: Vicious Circles?', *Trends in Ecology & Evolution*, 25: 21–7.

Blomström, A.-L. et al. (2010) 'Detection of a Novel Astrovirus in Brain Tissue of Mink Suffering from Shaking Mink Syndrome by Use of Viral Metagenomics', *Journal of Clinical Microbiology*, 48: 4392–6.

Bodewes, R. et al. (2013) 'Identification of Multiple Novel Viruses, Including a Parvovirus and a Hepevirus, in Feces of Red Foxes', *Journal of Virology*, 87: 7758–64.

Breck, S. W. et al. (2019) 'The Intrepid Urban Coyote: A Comparison of Bold and Exploratory Behavior in Coyotes from Urban and Rural Environments', *Scientific Reports*, 9: 2104.

Buchfink, B., Xie, C., and Huson, D. H. (2015) 'Fast and Sensitive Protein Alignment Using DIAMOND', *Nature Methods*, 12: 59–60.

Capella-Gutiérrez, S., Silla-Martínez, J. M., and Gabaldón, T. (2009) 'trimAl: A Tool for Automated Alignment Trimming in Large-Scale Phylogenetic Analyses', *Bioinformatics (Oxford, England)*, 25: 1972–3.

Cavallini, P. (1995) 'Variation in the Body Size of the Red Fox', *Annales Zoologici Fennici*, 32: 421–7.

Chong, R. et al. (2019) 'Fecal Viral Diversity of Captive and Wild Tasmanian Devils Characterized Using Virion-Enriched Metagenomics and Metatranscriptomics', *Journal of Virology*, 93: e00205.

Coman, B. J., Robinson, J., and Beaumont, C. (1991) 'Home Range, Dispersal and Density of Red Foxes (*Vulpes vulpes* L.) in Central Victoria', *Wildlife Research*, 18: 215–23.

Conceição-Neto, N. et al. (2016) 'Reassortment among Picobirnaviruses Found in Wolves', *Archives of Virology*, 161: 2859–62.

Contesse, P. et al. (2004) 'The Diet of Urban Foxes (*Vulpes vulpes*) and the Availability of Anthropogenic Food in the City of Zurich, Switzerland', *Mammalian Biology*, 69: 81–95.

Cypher, B. L., and Frost, N. (1999) 'Condition of San Joaquin Kit Foxes in Urban and Exurban Habitats', *The Journal of Wildlife Management*, 63: 930–8.

De Benedictis, P. et al. (2011) 'Astrovirus Infections in Humans and Animals—Molecular Biology, Genetic Diversity, and Interspecies Transmissions', *Infection, Genetics and Evolution*, 11: 1529–44.

De Cáceres, M., Legendre, P., and Moretti, M. (2010) 'Improving Indicator Species Analysis by Combining Groups of Sites', *Oikos*, 119: 1674–84.

Di Martino, B. et al. (2013) 'Canine Kobuviruses in Diarrhoeic Dogs in Italy', *Veterinary Microbiology*, 166: 246–9.

Donato, C., and Vijaykrishna, D. (2017) 'The Broad Host Range and Genetic Diversity of Mammalian and Avian Astroviruses', *Viruses*, 9: 102.

Doncaster, C. P., and Macdonald, D. W. (1997) 'Activity Patterns and Interactions of Red Foxes (*Vulpes vulpes*) in Oxford City', *Journal of Zoology*, 241: 73–87.

Dorning, J., and Harris, S. (2019) 'Individual and Seasonal Variation in Contact Rate, Connectivity and Centrality in Red Fox (*Vulpes vulpes*) Social Groups', *Scientific Reports*, 9: 20095.

Dremsek, P. et al. (2012) 'Seroprevalence Study in Forestry Workers from Eastern Germany Using Novel Genotype 3-and Rat Hepatitis E Virus-Specific Immunoglobulin G ELISAs', *Medical Microbiology and Immunology*, 201: 189–200.

EPBC (1999) *Environment Protection and Biodiversity Conservation Act*, Department of the Environment and Energy, Australian Government.

Finkbeiner, S. R. et al. (2009) 'Identification of a Novel Astrovirus (Astrovirus VA1) Associated with an Outbreak of Acute Gastroenteritis', *Journal of Virology*, 83: 10836–9.

Frölich, K., Klima, F., and Dedek, J. (1998) 'Antibodies against Rabbit Hemorrhagic Disease Virus in Free-Ranging Red Foxes from Germany', *Journal of Wildlife Diseases*, 34: 436–42.

Geoghegan, J. L., Duchêne, S., and Holmes, E. C. (2017) 'Comparative Analysis Estimates the Relative Frequencies of Co-Divergence and Cross-Species Transmission within Viral Families', *PLoS Pathogens*, 13: e1006215.

Ghasemzadeh, I., and Namazi, S. H. (2015) 'Review of Bacterial and Viral Zoonotic Infections Transmitted by Dogs', *Journal of Medicine and Life*, 8: 1–5.

Gibbs, S. E. J. et al. (2006) 'Factors Affecting the Geographic Distribution of West Nile Virus in Georgia, USA: 2002–2004', *Vector-Borne and Zoonotic Diseases*, 6: 73–82.

Gras, P. et al. (2018) 'Landscape Structures Affect Risk of Canine Distemper in Urban Wildlife', *Frontiers in Ecology and Evolution*, 6: 136.

Guindon, S. et al. (2010) 'New Algorithms and Methods to Estimate Maximum-Likelihood Phylogenies: Assessing the Performance of PhyML 3.0', *Systematic Biology*, 59: 307–21.

Haas, B. J. et al. (2013) 'De Novo Transcript Sequence Reconstruction from RNA-Seq Using the Trinity Platform for Reference Generation and Analysis', *Nature Protocols*, 8: 1494–512.

Hall, R. N. et al. (2015) 'Emerging Rabbit Hemorrhagic Disease Virus 2 (RHDVb)', *Emerging Infectious Diseases*, 21: 2276–8.

Jessop, T. S. et al. (2012) 'Demographic and Phenotypic Effects of Human Mediated Trophic Subsidy on a Large Australian Lizard (*Varanus varius*): Meal Ticket or Last Supper?', *PLoS One*, 7: e34069.

Kapoor, A. et al. (2009) 'Multiple Novel Astrovirus Species in Human Stool', *The Journal of General Virology*, 90: 2965–72.

Keane, T. M. et al. (2006) 'Assessment of Methods for Amino Acid Matrix Selection and Their Use on Empirical Data Shows That

Ad Hoc Assumptions for Choice of Matrix Are Not Justified', *BMC Evolutionary Biology*, 6: 29.

Klein, S. L. (2000) 'Hormones and Mating System Affect Sex and Species Differences in Immune Function among Vertebrates', *Behavioural Processes*, 51: 149–66.

Krishnamurthy, S. R., and Wang, D. (2018) 'Extensive Conservation of Prokaryotic Ribosomal Binding Sites in Known and Novel Picobirnaviruses', *Virology*, 516: 108–14.

Lan, D. et al. (2011) 'Isolation and Characterization of the First Chinese Porcine Sapelovirus Strain', *Archives of Virology*, 156: 1567–74.

Lee, C. T., et al. (2017) 'Outbreak of Influenza A(H7N2) among Cats in an Animal Shelter with Cat-to-Human Transmission—New York City, 2016', *Clinical Infectious Diseases*, 65: 1927–9.

Li, B., and Dewey, C. N. (2011) 'RSEM: Accurate Transcript Quantification from RNA-Seq Data with or without a Reference Genome', *BMC Bioinformatics*, 12: 323.

Li, L. et al. (2013) 'Divergent Astrovirus Associated with Neurologic Disease in Cattle', *Emerging Infectious Diseases*, 19: 1385–92.

Lojkić, I. et al. (2016) 'Faecal Virome of Red Foxes from Peri-Urban Areas', *Comparative Immunology, Microbiology and Infectious Diseases*, 45: 10–5.

Macdonald, D. W. (1979) 'Helpers'in Fox Society', *Nature*, 282: 69–71.

Macdonald, D. W. (1983) 'The Ecology of Carnivore Social Behaviour', *Nature*, 301: 379–84.

Mahar, J. E. et al. (2018) 'Rabbit Hemorrhagic Disease Virus 2 (RHDV2; GI.2) Is Replacing Endemic Strains of RHDV in the Australian Landscape within 18 Months of Its Arrival', *Journal of Virology*, 92: e01374.

Malik, Y. S. et al. (2014) 'Epidemiology, Phylogeny, and Evolution of Emerging Enteric Picobirnaviruses of Animal Origin and Their Relationship to Human Strains', *BioMed Research International*, 2014: 1–13.

Marinova-Petkova, A. et al. (2017) 'Avian Influenza A(H7N2) Virus in Human Exposed to Sick Cats, New York, USA, 2016', *Emerging Infectious Diseases*, 23: 2046–9.

Marks, C. A. et al. (1996) 'Preliminary Field Assessment of a Cabergoline Baiting Campaign for Reproductive Control of the Red Fox (*Vulpes vulpes*)', *Wildlife Research*, 23: 161–8.

Marks, C. A., and Bloomfield, T. E. (1999) 'Distribution and Density Estimates for Urban Foxes (*Vulpes vulpes*) in Melbourne: Implications for Rabies Control', *Wildlife Research*, 26: 763–75.

Marks, C. A., and Bloomfield, T. E. (2006) 'Home-Range Size and Selection of Natal Den and Diurnal Shelter Sites by Urban Red Foxes (*Vulpes vulpes*) in Melbourne', *Wildlife Research*, 33: 339–47.

Meng, X. J. (2010) 'Hepatitis E Virus: Animal Reservoirs and Zoonotic Risk', *Veterinary Microbiology*, 140: 256–65.

Newsome, A. E. (1995) 'Socio-Ecological Models for Red Fox Populations Subject to Fertility Control in Australia', *Annales Zoologici Fennici*, 32: 99–110.

Newsome, T. M. et al. (2015) 'The Ecological Effects of Providing Resource Subsidies to Predators', *Global Ecology and Biogeography*, 24: 1–11.

Newsome, T. M., Spencer, E. E., and Dickman, C. R. (2017) 'Short-Term Tracking of Three Red Foxes in the Simpson Desert Reveals Large Home-Range Sizes', *Australian Mammalogy*, 39: 238–42.

Nunn, C. L. et al. (2015) 'Infectious Disease and Group Size: More than Just a Numbers Game', *Philosophical Transactions of the Royal Society B: Biological Sciences*, 370: 20140111.

Oksanen, J. et al. (2019) *vegan: Community Ecology Package*, R Package, Version 2.5-6, https://cran.r-project.org/web/packages/vegan/index.html.

O'Neil, J. (2018) 'Zoonotic Infections from Common Household Pets', *The Journal for Nurse Practitioners*, 14: 363–70.

Phillips, M., and Catling, P. C. (1991) 'Home Range and Activity Patterns of Red Foxes in Nadgee Nature Reserve', *Wildlife Research*, 18: 677–86.

Quan, P.-L. et al. (2010) 'Astrovirus Encephalitis in Boy with X-Linked Agammaglobulinemia', *Emerging Infectious Diseases*, 16: 918–25.

Reuter, G., Boros, Á., and Pankovics, P. (2011) 'Kobuviruses—A Comprehensive Review', *Reviews in Medical Virology*, 21: 32–41.

Riley, S. P. D., Foley, J., and Chomel, B. (2004) 'Exposure to Feline and Canine Pathogens in Bobcats and Gray Foxes in Urban and Rural Zones of a National Park in California', *Journal of Wildlife Diseases*, 40: 11–22.

Robertson, K. E. (2018), 'Boldness Behavior and Chronic Stress in Free-Ranging, Urban Coyotes (*Canis latrans*)', PhD thesis, The Ohio State University, Columbus, OH.

Saunders, G., Gentle, M., and Dickman, C. (2010) 'The Impacts and Management of Foxes *Vulpes vulpes* in Australia', *Mammal Review*, 40: 181–211.

Schipper, J. et al. (2008) 'The Status of the World's Land and Marine Mammals: Diversity, Threat, and Knowledge', *Science*, 322: 225–30.

Schock, A. et al. (2014) 'Investigation into an Outbreak of Encephalomyelitis Caused by a Neuroinvasive Porcine Sapelovirus in the United Kingdom', *Veterinary Microbiology*, 172: 381–9.

Statham, M. J. et al. (2014) 'Range-Wide Multilocus Phylogeography of the Red Fox Reveals Ancient Continental Divergence, Minimal Genomic Exchange and Distinct Demographic Histories', *Molecular Ecology*, 23: 4813–30.

Stepkovitch, B. (2017) 'Feeding Ecology of an Invasive Predator across an Urban Land Use Gradient', M.Res. thesis, Western Sydney University, Sydney, NSW, Australia.

Stepkovitch, B. et al. (2019) 'Urban Lifestyle Supports Larger Red Foxes in Australia: An Investigation into the Morphology of an Invasive Predator', *Journal of Zoology*, 309: 287–94.

Vemuri, R. et al. (2019) 'The Microgenderome Revealed: Sex Differences in Bidirectional Interactions between the Microbiota, Hormones, Immunity and Disease Susceptibility', *Seminars in Immunopathology*, 41: 265–75.

Vieira, T. S. W. J. et al. (2013) 'Serosurvey of Tick-Borne Pathogens in Dogs from Urban and Rural Areas from Parana State, Brazil', *Revista Brasileira de Parasitologia Veterinária*, 22: 104–9.

Wells, K. et al. (2018) 'Global Spread of Helminth Parasites at the Human–Domestic Animal–Wildlife Interface', *Global Change Biology*, 24: 3254–65.

Wells, K. et al. (2020) 'Distinct Spread of DNA and RNA Viruses among Mammals amid Prominent Role of Domestic Species', *Global Ecology and Biogeography*, 29: 470–81.

White, P. C. L., and Harris, S. (1994) 'Encounters between Red Foxes (*Vulpes vulpes*): Implications for Territory Maintenance, Social Cohesion and Dispersal', *The Journal of Animal Ecology*, 63: 315–27.

Yinda, C. K. et al. (2019) 'Gut Virome Analysis of Cameroonians Reveals High Diversity of Enteric Viruses, Including Potential Interspecies Transmitted Viruses', *mSphere*, 4: e00585.

Zell, R. (2018) 'Picornaviridae—the Ever-Growing Virus Family', *Archives of Virology*, 163: 299–317.

Virus Evolution, 2021, 7(1): veab005

doi: 10.1093/ve/veab005
Research Article

Virome composition in marine fish revealed by meta-transcriptomics

Jemma L. Geoghegan,[1,2,3,*,†] Francesca Di Giallonardo,[4] Michelle Wille,[5] Ayda Susana Ortiz-Baez,[6,‡] Vincenzo A. Costa,[2] Timothy Ghaly,[2,§] Jonathon C. O. Mifsud,[2,¶] Olivia M. H. Turnbull,[2] David R. Bellwood,[7] Jane E. Williamson,[2] and Edward C. Holmes[6,**]

[1]Department of Microbiology and Immunology, University of Otago, Dunedin 9016, New Zealand, [2]Department of Biological Sciences, Macquarie University, Sydney, NSW 2109, Australia, [3]Institute of Environmental Science and Research, Wellington 5018, New Zealand, [4]The Kirby Institute, University of New South Wales, Sydney, NSW 2052, Australia, [5]WHO Collaborating Centre for Reference and Research on Influenza, The Peter Doherty Institute for Infection and Immunity, Melbourne, VIC, Australia, [6]Marie Bashir Institute for Infectious Diseases and Biosecurity, School of Life and Environmental Sciences and School of Medical Sciences, The University of Sydney, Sydney, NSW 2006, Australia and [7]ARC Centre of Excellence for Coral Reef Studies and College of Science and Engineering, James Cook University, Townsville, QLD 4811, Australia

*Corresponding author: E-mail: jemma.geoghegan@otago.ac.nz

†https://orcid.org/0000-0003-0970-0153

‡https://orcid.org/0000-0003-0080-0633

§https://orcid.org/0000-0002-5162-4054

¶https://orcid.org/0000-0001-8934-9193

**https://orcid.org/0000-0001-9596-3552

Abstract

Revealing the determinants of virome composition is central to placing disease emergence in a broader evolutionary context. Fish are the most species-rich group of vertebrates and so provide an ideal model system to study the factors that shape virome compositions and their evolution. We characterized the viromes of nineteen wild-caught species of marine fish using total RNA sequencing (meta-transcriptomics) combined with analyses of sequence and protein structural homology to identify divergent viruses that often evade characterization. From this, we identified twenty-five new vertebrate-associated viruses and a further twenty-two viruses likely associated with fish diet or their microbiomes. The vertebrate-associated viruses identified here included the first fish virus in the *Matonaviridae* (single-strand, positive-sense RNA virus). Other viruses fell within the *Astroviridae*, *Picornaviridae*, *Arenaviridae*, *Reoviridae*, *Hepadnaviridae*, *Paramyxoviridae*, *Rhabdoviridae*, *Hantaviridae*, *Filoviridae*, and *Flaviviridae*, and were sometimes phylogenetically distinct from known fish viruses. We also show how key metrics of virome composition—viral richness, abundance, and diversity—can be analysed along with host ecological and biological factors as a means to understand virus ecology. Accordingly, these data suggest that that the vertebrate-associated viromes of the fish sampled here are predominantly shaped by the phylogenetic history

(i.e. taxonomic order) of their hosts, along with several biological factors including water temperature, habitat depth, community diversity and swimming behaviour. No such correlations were found for viruses associated with porifera, molluscs, arthropods, fungi, and algae, that are unlikely to replicate in fish hosts. Overall, these data indicate that fish harbour particularly large and complex viromes and the vast majority of fish viromes are undescribed.

Key words: fish; virome; virus evolution; metagenomics; host-jumping.

1. Introduction

Metagenomic next-generation sequencing (mNGS) has led to a revolution in virus discovery (Zhang, Shi, and Holmes 2018; Zhang et al. 2018; Shi, Zhang, and Holmes 2018b), exposing more of the diversity, scale and structure of the virosphere. However, while it is now possible to reveal host viromes en masse (Lim et al. 2015; Paez-Espino et al. 2016; Shi et al. 2016; Temmam et al. 2016; Roux et al. 2017; Tirosh et al. 2018; Vibin et al. 2018; Geoghegan et al. 2018b; Chang et al. 2019; Pettersson et al. 2019; Porter et al. 2019), we still have an incomplete understanding of the factors that structure viromes. Until recently, studies of virus evolution were largely limited to single viruses and/or single hosts, restricting our ability to explore the diverse host and environmental factors that might structure viromes as a whole. Fortunately, this is changing with the advent of mNGS, particularly total RNA sequencing. In particular, metagenomic-based studies have shown that aspects of host biology can greatly impact virus diversification (Wille et al. 2019; Wille 2020) and as such may also be key drivers of virus emergence. As a simple case in point, the behavioural ecology of host species directly affects contact rates among individuals in a population, and more frequent intra- and inter-species contacts are likely to increase the potential for viral transmission.

The marine environment is a rich source of viruses. For example, the bacteriophage in aquatic ecosystems greatly outnumber other life-forms (Maranger and Bird 1995). There is an estimated concentration of 10 billion virus particles per litre of surface water (Bergh et al. 1989; Breitbart and Rohwer 2005; Suttle 2005; Middelboe and Brussaard 2017), although abundance levels vary with such factors as ocean depth (De Corte et al. 2012; Lara et al. 2017), temperature (Coutinho et al. 2017), latitude (Gregory et al. 2019), and phytoplankton bloom development (Alarcon-Schumacher et al. 2019). In marked contrast to bacteriophage, little is known about the factors that contribute to virus diversity in aquatic vertebrate populations, even though viruses can cause large-scale disease outbreaks in farmed fish (Whittington and Reddacliff 1995; Crane and Hyatt 2011; Jarungsriapisit et al. 2020).

Fish provide an ideal model to better understand the diversity of viruses that exist in nature as well as the range of host and environmental factors that shape virome composition and abundance. Fish are the most species-rich group of vertebrates with over 33,000 species described to date (fishbase.org), the vast majority of which (~85%) are bony fish (the Osteichthyes) (Betancur-R et al. 2017). Bony fish themselves are an extremely diverse and abundant group comprising forty-five taxonomic orders, exhibiting a wide range of biological features that likely play an important role in shaping the diversity of their viromes. Initial studies indicate that fish harbour a remarkable diversity of viruses, particularly those with RNA genomes, that may exceed that seen in any other class of vertebrate (Lauber et al. 2017; Geoghegan et al. 2018a; Shi et al. 2018a). In addition, those viruses present in fish often appear to be the evolutionary predecessors of viruses infecting other vertebrate hosts, generally indicative of a pattern of virus-host associations that can date back hundreds of millions of years, although with frequent cross-species transmission. Despite the apparent diversity and ubiquity of fish viruses, they are severely under-studied compared to mammalian and avian viruses and there is little data on the factors that determine the structure of fish viromes.

To reveal more of the unexplored aquatic virosphere we sampled wild-caught ray-finned marine fish spanning twenty-three species across nine taxonomic orders and quantified a variety of host characteristics that together may impact virome composition, abundance and evolution. Specifically, we utilized meta-transcriptomics together with both sequence and protein structural homology searches of known viruses to: (1) reveal the total virome composition of fish, (2) describe the phylogenetic relationships of the novel viruses obtained, (3) determine whether, on these data, there may be associations between virome composition, abundance, richness, and diversity and particular host traits, and (4) explore whether taxonomically-related fish hosts have more similar viromes. The host characteristics initially considered here were: fish taxonomic order, swimming behaviour (i.e. solitary or schooling fish), preferred climate, mean preferred water temperature, host community diversity (i.e. multi- or single- species community), average body length, maximum life span, trophic level, and habitat depth (Supplementary Table S1).

2. Methods

2.1 Ethics

Biosafety was approved by Macquarie University, Australia (ref: 5201700856). This study involved dead fish purchased from a fish market for which no animal ethics approval was required. The pygmy goby was collected under GBRMPA permit G16/37684.1 and JCU Animal Ethics Committee #A2530.

2.2 Fish sample collection

Dead fish from twenty-three species were sampled for virome analysis (Supplementary Table S1). These included eighteen new species collected from a fish market in Sydney, Australia, together with four species from our previous sampling of the same fish market (Geoghegan et al. 2018a). These animals were caught by commercial fisheries in coastal waters in New South Wales, Australia by several different suppliers in Autumn 2018. By way of contrast, an additional species, the pygmy goby (*Eviota zebrina*), was obtained from the coral reefs of tropical northern Queensland at approximately the same time. Fish were snap frozen at -20 °C immediately upon capture. Fish obtained from the market were purchased on the day of catch. Tissues were dissected and stored in RNALater before being transferred to a -80° C freezer. To increase the likelihood of virus

discovery during metagenomic sequencing, 10 individuals from each species were pooled.

2.3 Transcriptome sequencing

mNGS was performed on fish tissue (liver and gill). Frozen tissue was partially thawed and submerged in lysis buffer containing 1 per cent ß-mercaptoethanol and 0.5 per cent Reagent DX before tissues were homogenized together with TissueRupture (Qiagen). The homogenate was centrifuged to remove any potential tissue residues, and RNA from the clear supernatant was extracted using the Qiagen RNeasy Plus Mini Kit. RNA was quantified using NanoDrop (ThermoFisher) and tissues from each species were pooled to 3 μg per pool (250 ng per individual). Libraries were constructed using the TruSeq Total RNA Library Preparation Protocol (Illumina) and host ribosomal RNA (rRNA) was depleted using the Ribo-Zero-Gold Kit (Illumina) to facilitate virus discovery. Paired-end (100 bp) sequencing of the RNA library was performed on the HiSeq 2500 platform (Illumina). All library preparation and sequencing were carried out by the Australian Genome Research Facility (AGRF).

2.4 Transcript sequence similarity searching for viral discovery

Sequencing reads were first quality trimmed then assembled *de novo* using Trinity RNA-Seq (Haas et al. 2013). The assembled contigs were annotated based on similarity searches against the NCBI nucleotide (nt) and non-redundant protein (nr) databases using BLASTn and Diamond (BLASTX) (Buchfink, Xie, and Huson 2015), and an e-value threshold of 1×10^{-5} was used as a cut-off to identify positive matches. We removed non-viral hits including host contigs with similarity to viral sequences (e.g. endogenous viral elements). To reduce the risk of incorrect assignment of viruses to a given library due to index-hoping, those viruses with a read count less than 0.1 per cent of the highest count for that virus among the other libraries was assumed to be contamination.

2.5 Protein structure similarity searching for viral discovery

To identify highly divergent viral transcripts, particularly those that might be refractory to detection using similarity searching methods such as the BLAST approach described above, we employed a protein structure-based similarity search for 'orphan' contigs that did not share sequence similarity with known sequences. Accordingly, assembled orphan contigs were translated into open reading frames (ORFs) using EMBOSS getorf program (Rice, Longden, and Bleasby 2000). ORFs were arbitrarily defined as regions between two stop codons with a minimum size of 200 amino acids in length. To reduce redundancy, amino acid sequences were grouped based on sequence identity using the CD-HIT package v4.6.5 (Li and Godzik 2006). The resulting data set was then submitted to Phyre2, which uses advanced remote homology detection methods to build 3D protein models, predict ligand binding sites, and analyse the effect of amino acid variants (Kelley et al. 2015). Virus sequences with predicted structures were selected on the basis of having confidence values ≥90 per cent. Following structure prediction, we used the associated annotations for preliminary taxonomic classification. To avoid false positives due to the limited number of available structures in the Protein Data Bank (PDB) for template modelling, the taxonomic assignment was cross-validated with the results from the Diamond (BLASTX) similarity search. Subsequently, putative viruses were aligned with reference viral protein sequences at the immediate higher taxonomic level (e.g. genus, family), using MAFFT v7.4 (E-INS-i algorithm) (Katoh and Standley 2013). Finally, we verified the similarity among sequences by careful visual inspection of the most highly conserved motifs of target proteins.

2.6 Inferring the evolutionary history of fish viruses

We inferred the evolutionary relationships of the viruses contained in the fish samples and compared them with known viruses to determine those that were likely associated with vertebrate or non-vertebrate hosts. Specifically, we assumed that viruses that grouped with other vertebrate viruses in phylogenetic trees were likely to infect the fish sampled here, while those virus that were more closely related to those usually associated with other host types (such as invertebrates, fungi and plants) were unlikely to infect and replicate in fish hosts. To achieve this, the translated viral contigs were combined with representative protein sequences within each virus family obtained from NCBI RefSeq. The sequences retrieved were then aligned with those generated here again using MAFFT v7.4 (E-INS-i algorithm) as described above. Ambiguously aligned regions were removed using trimAl v.1.2 (Capella-Gutierrez, Silla-Martinez, and Gabaldon 2009). To estimate phylogenetic trees, we selected the optimal model of amino acid substitution identified using the Bayesian Information Criterion as implemented in Modelgenerator v0.85 (Keane et al. 2006) and analysed the data using the maximum likelihood approach available in IQ-TREE (Nguyen et al. 2015) with 1000 bootstrap replicates. Phylogenetic trees were annotated with FigTree v.1.4.2. Viruses newly identified here were named reflecting the host common name.

2.7 Revealing virome abundance and diversity

Transcriptomes were quantified using RNA-Seq by Expectation-Maximization (RSEM) as implemented within Trinity (Li and Dewey 2011). We first estimated the relative abundance of a host reference gene, ribosomal protein S13 (RPS13), to assess the sequencing depth across libraries. Next, we used RSEM to estimate the relative abundance of each virus transcript in these data.

For those viruses most likely associated with fish themselves, rather than components of their diet or microbiome (see Results), we performed analyses of virome abundance and diversity using R v3.4.0 integrated into RStudio v1.0.143 and plotted using ggplot2. Both the observed virome richness and Shannon effective (i.e. alpha diversity) were calculated for each library at the virus family level using modified Rhea script sets (Lagkouvardos et al. 2017; Wille et al. 2019). We used generalized linear models (GLM) to initially evaluate the effect of host taxonomic order, swimming behaviour (solitary or schooling fish), preferred climate, mean preferred water temperature, host community diversity, average species length, trophic level and habitat depth on viral abundance and alpha diversity (see Supplementary Table S1 for all variables). Models were χ^2 tested (LRT) to assess model significance. When the number of factor levels in an explanatory variable exceeded two, we conducted Tukey posthoc testing (glht) using the *multcomp* package (Hothorn, Bretz, and Westfall 2008). Beta diversity (i.e. the diversity between samples) was calculated using the Bray Curtis dissimilarity matrix. Effects of variables on viral community composition were evaluated using permanova (Adonis Tests)

and Mantel tests with 10,000 permutations using the *vegan* package (Oksanen 2007).

To establish connectivity (i.e. sharing) among virus families that were likely associated with non-fish hosts, we generated a cord diagram by quantifying the number of fish species harbouring each virus family identified in this study. Virus families that occur in the same fish species were represented by ribbons or links in the diagram.

3. Results

We used mNGS to characterize viral transcripts from 23 marine fish spanning nine taxonomic orders: 19 species from this current study together with four from our previous work (Geoghegan et al. 2018a). We combined data from our previous fish sampling to expand our data set and to apply novel viral protein structural searching methods not used previously. For these reasons, individual viruses discovered in our previous study are not detailed here. Combined, the extracted total RNA was organized into 23 libraries for high-throughput RNA sequencing. Ribosomal RNA-depleted libraries resulted in a median of 45,690,996 (range 33,344,520–51,071,142) reads per pool.

3.1 Diversity and abundance of viruses in fish

The fish viromes characterized here contained viruses that were associated with vertebrate hosts as well as those that were more likely associated with porifera, invertebrates, fungi, and algae (Fig. 1). We primarily focused on the former since we assumed that the vertebrate-associated viruses were directly infecting the fish sampled, rather than being associated with the aquatic environment, diet or a co-infecting parasite, and hence are more informative in determining how host factors shape virus ecology and evolution.

Overall, we identified virus transcripts likely associated with vertebrate hosts that could be assigned to 11 viral families and present in a variety of fish species (Supplementary Fig. S1a). With the exception of the *Hepadnaviridae*, all were RNA viruses. Across all the fish sampled, those viral families found at relatively high abundances included the *Astroviridae* (representing 39% of all viruses discovered), *Picornaviridae* (19%), *Arenaviridae* (16%), *Reoviridae* (13%) and the *Hepadnaviridae* (9%) (Fig. 1a). Other viral families found at lower relative abundances were the *Matonaviridae* (previously the *Togaviridae*) (2%), *Paramyxoviridae* (1%), as well as the *Rhabdoviridae*, *Hantaviridae*, *Filoviridae*, and *Flaviviridae* (all <1%) (Fig. 1a). The most common vertebrate-associated viruses found in these fish were picornaviruses (eight species), astroviruses (seven species), and hepadnaviruses (six species) (Fig. 1b). The eastern sea garfish (*Hyporhamphus australis*) harboured the most diverse virome with four distinct vertebrate-associated viruses (Fig. 1b). Six fish contained no vertebrate-associated viruses, and we found no viral sequences in the yellowfin bream (*Acanthopagrus australis*) (Fig. 1c). An equivalent analysis of a host reference gene, ribosomal protein S13 (RPS13) that is stably expressed in fish, revealed similar abundances across species (0.004–0.02%), implying similar sequencing depth across libraries (Fig. 1c). RPS13 was, on average, ~55 per cent more abundant than the total virome.

We also examined viruses that were phylogenetically related to those associated with porifera, molluscs, arthropods, fungi,

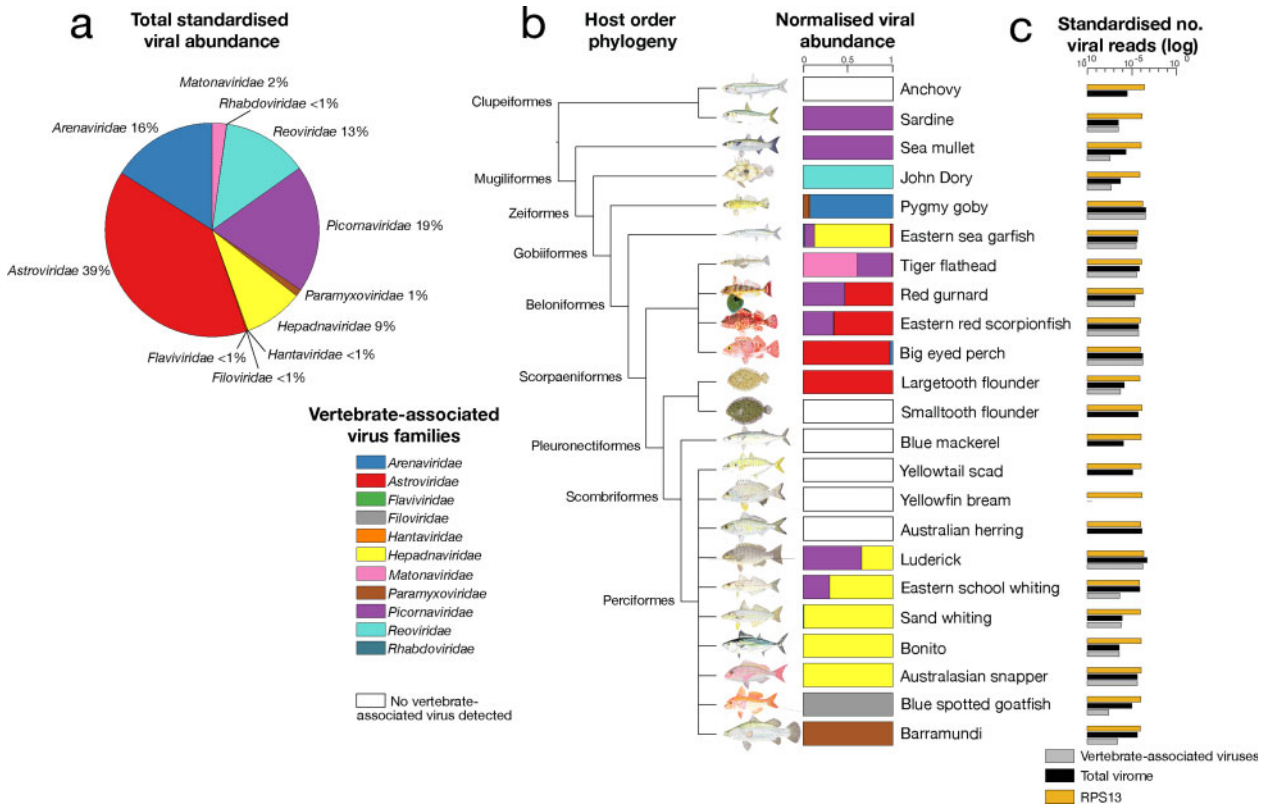


Figure 1. (A) Total standardized abundance of vertebrate-associated viruses (at the level of virus family) across the fish species examined. (B) Normalized viral abundance set out on a backbone of the fish host phylogeny at the order level. (C) Standardized number of total viral reads (black), vertebrate-associated viral reads (grey) and host reference gene ribosomal protein S13 (RPS13) (orange) in each species library.

and algae, and hence were unlikely to infect the fish themselves. Accordingly, we identified an additional 22 viruses across 11 virus families (Supplementary Fig. S1b). These viruses were found in the *Chuviridae, Hepeviridae, Narnaviridae, Nodaviridae, Partitiviridae, Picornaviridae, Solemoviridae, Tombusviridae, Totiviridae, Dicistroviridae*, and *Iflaviridae*, and are described in more detail below.

3.2 Evolutionary relationships of fish viruses

To infer stable phylogenetic relationships among the viruses sampled and to identify those that are novel, where possible we utilized the most conserved (i.e. polymerase) viral regions that comprise the RNA-dependent RNA polymerase (RdRp) or the polymerase (P) ORF in the case of the hepadnaviruses. From this, we identified 25 distinct and potentially novel vertebrate-associated virus species, in addition to the eight novel viruses described previously (Geoghegan et al. 2018a) (Supplementary Table S2). All novel vertebrate-associated viruses shared sequence similarity to other known fish viruses with the exception of those viruses found in the *Matonaviridae* and *Rhabdoviridae*, the latter of which was found using structure similarity methods (Fig. 2, Supplementary Table S3; see below). We found a further 22 viruses that clustered with viruses found in porifera, molluscs, arthropods, fungi, and algae (Supplementary Figs S2–S4).

Among the viruses identified was tiger flathead matonavirus (in *Neoplatycephalus richardsoni*) – the first fish virus found in the *Matonaviridae*. This novel viral sequence shared only 35 per cent amino acid similarity with its closest relative—Guangdong Chinese water snake rubivirus (Shi et al. 2018a). Until recently, the only other representative of this family was the distantly related human rubella virus, although additional members of this family have recently been identified in other mammalian species (Bennett et al. 2020). Given the high levels of genetic divergence in this family, it is likely that these fish-associated viruses at least constitute a discrete and novel genus.

Another divergent virus discovered in this analysis is eastern sea garfish rhabdovirus (in *Hyporhamphus australis*), which was most closely related to Fujian dimarhabdovirus sampled from an amphibian host, sharing 45 per cent amino acid RdRp sequence identity. Notably, this highly divergent virus was only identified by using protein structure homology, and forms a clade that is distinct from other fish rhabdoviruses (Fig. 2). We also identified two novel viral sequences in the *Filoviridae* in John Dory (*Zeus faber*) and the blue spotted goatfish (*Upeneichthys lineatus*). These viruses shared sequence similarity to the only other known fish filovirus, Wenling filefish filovirus (Shi et al. 2018a). With the exception of these fish viruses, all other known filoviruses including Ebola and Marburg viruses, are found in mammalian hosts, notably humans, bats, and primates.

We also found numerous viruses that cluster within established clades of fish viruses. For example, pygmy goby hantavirus (in *Eviota zebrina*) grouped with other hantaviruses recently found in fish (Fig. 2). Although they were previously only thought to infect mammals, hantaviruses have now been found to infect amphibians, jawless fish, and ray-finned fish (Shi et al. 2018a). The evolutionary history of the *Paramyxoviridae* shows two distinct fish virus lineages, of which both barramundi and pygmy goby paramyxoviruses grouped with Pacific spade-nose shark paramyxovirus and shared 50 and 45 per cent amino acid L gene sequence similarity, respectively. This group of fish viruses is phylogenetically distinct from other paramyxoviruses. We also found novel fish viruses in the *Flaviviridae*, *Arenaviridae*, and *Reoviridae*: although these grouped with other fish viruses, they greatly expand the known diversity of these virus families. Finally, as noted above, the most abundant viruses fell within the *Picornaviridae* and *Astroviridae*, and all shared sequence similarity to other fish viruses. Notably, both picornaviruses and astroviruses are single-stranded positive-sense RNA viruses that possess small icosahedral capsids with no external envelope, which may aid their preservation in harsh marine environments.

The only DNA viruses we identified were novel hepadnaviruses. Those found in bonito (*Sarda australis*), ludrick (*Girella tricuspidata*), and eastern school whiting (*Sillago flindersi*), fell into the divergent group of hepadna-like viruses, the nackednaviruses, that have been identified in a number of fish species (Lauber et al.. 2017). In contrast, sand whiting hepadnavirus (in *Sillago ciliate*) fell into the fish virus clade that is more closely related to mammalian hepatitis B viruses (Dill et al. 2016) (Fig. 2).

As expected, many of the viruses identified here were associated with marine hosts belonging to invertebrates (including porifera, molluscs, and arthropods; $n = 20$), fungi ($n = 1$), and algae ($n = 1$) as determined by their phylogenetic position and sequence similarity to viruses previously described in these taxa (Supplementary Figs S2–S4). This implies that these viruses more likely originated from host species that are associated with fish diet, fish microbiomes or the surrounding environment, rather than from the fish themselves. None of these viruses are highly divergent from other known viruses, but do help fill gaps in the phylogenetic diversity of these groups.

3.3 Assessing the impact of host biology on virome composition

Our relatively small sample of 23 fish species precluded us from performing a detailed statistical analysis of the relationship between host traits and virome composition. Rather, we provide an initial analysis that should be regarded as a framework for understanding how key host variables might impact viral ecology and evolution, and that can be extended as more species are analysed.

To this end we examined the possible association between eight host traits and viral abundance (the proportion of viral reads in each sample), alpha diversity (the diversity within each sample, measured by observed richness and Shannon diversity) and beta diversity (the diversity between samples). The host traits initially considered here were: host taxonomic order, swimming behaviour (solitary or schooling fish), preferred climate, mean preferred water temperature, community diversity, average species length, maximum life span, trophic level, and habitat depth.

We first focused on the vertebrate-associated virome. This initial analysis revealed that the phylogenetic relationships of the fish studied, as reflected in their taxonomic order, seemingly had the strongest association with the overall composition of fish viromes. This pattern was consistent when assessing viral abundance, alpha diversity, and beta diversity (Fig. 3). That is, fish order ($\chi^2 = 0.003$, df = 8, $P = 0.0049$) and mean preferred water temperature ($\chi^2 = 0.008$, df = 1, $P = 0.035$) were important predictors of viral abundance, such that Scopaeniformes (i.e. bigeye ocean perch, red gurnard, tiger flathead, and eastern red scorpionfish) had significantly higher viral abundance than Pleuronectiformes (i.e. largetooth and smalltooth flounder) (Tukey: $z = 3.766$, $P = 0.00479$), while viral abundance had a

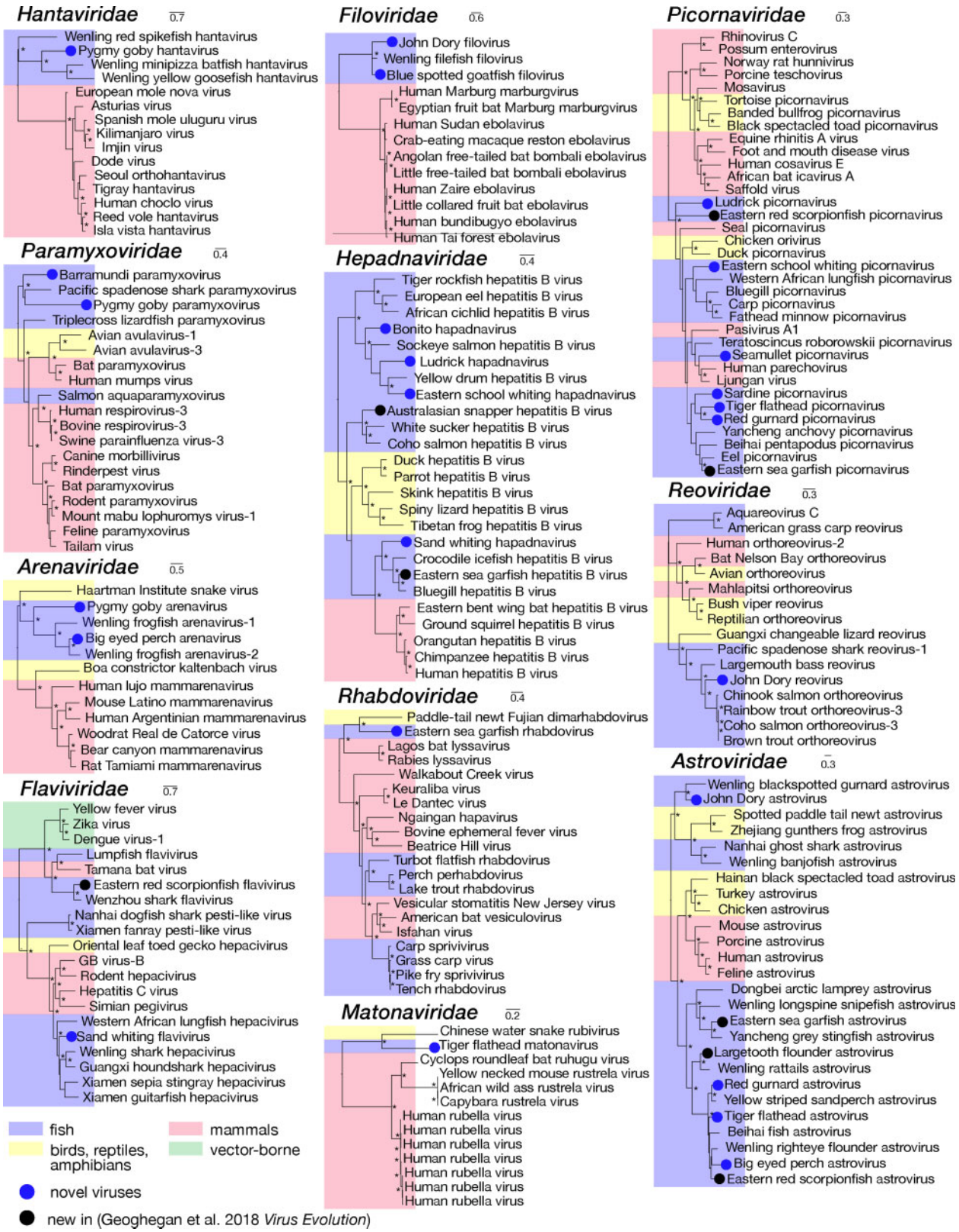


Figure 2. Phylogenetic relationships of likely vertebrate-associated viruses identified here. The maximum likelihood phylogenetic trees show the topological position of the newly discovered viruses (blue circles) and those identified in an earlier study (Geoghegan et al. 2018a), in the context of their closest phylogenetic relatives. Branches are highlighted to represent host class (fish = blue; mammals = red; birds, reptiles and amphibians = yellow; vector-borne (mammals and arthropods) = green). All branches are scaled according to the number of amino acid substitutions per site and trees were mid-point rooted for clarity only. An asterisk indicates node support of >70 per cent bootstrap support. See Supplementary Table S3 for all accession numbers.

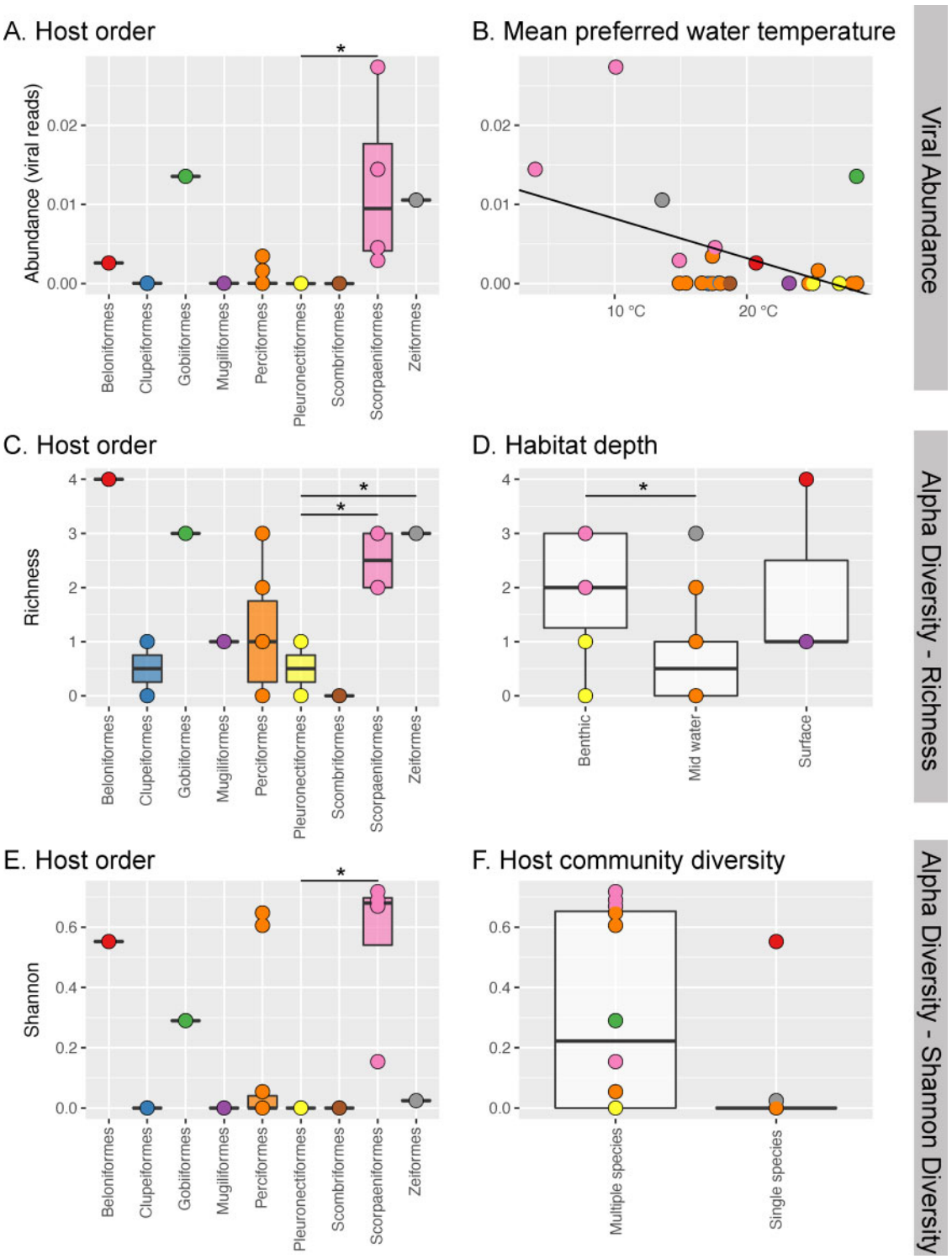


Figure 3. Significant explanatory variables in generalized linear models (GLM) for viral abundance and two measures of alpha diversity. Viral abundance is best explained by (A) fish host order and (B) mean preferred water temperature. Alpha diversity is best explained by (C) host order and (D) preferred habitat (Observed Richness) and by (E) host order and (F) host community diversity (Shannon Diversity). Stars indicate significant differences between groups determined by *post hoc* Tukey tests. Points represent different fish species and are coloured by host order.

negative relationship to mean preferred water temperature (Fig. 3). It is worth noting, however, that virus abundance within the Scopaeniformes were widely distributed and that their overall high abundance might only be due to a few species or individuals.

We applied two measures of alpha diversity to our sample set: observed richness, a count of the number of viral families, and Shannon diversity, which also incorporates abundance. Observed richness was best explained by fish order ($\chi^2 = 22.839$, df = 8, $P = 3.8^{-6}$) and habitat depth ($\chi^2 = 3.914$, df = 2, $P = 0.032$),

while Shannon diversity was best explained by fish order ($\chi^2 = 0.96$, df = 8, $P = 0.016$) and community diversity ($\chi^2 = 0.41$, df = 1, $P = 0.05$), with a larger Shannon diversity in multispecies communities compared with single species communities. As with viral abundance, there was a significant difference in alpha diversity between Scopaeniformes compared to Pleuronectiformes (Tukey Richness z = 3.039, $P = 0.0495$; Tukey Shannon z = 2.845, $P = 0.05$). Notably, in these data mid-water fish had decreased viral richness compared to benthic fish (Tukey z = -2.452, $P = 0.0338$), and fish that reside in multispecies communities had a larger Shannon diversity compared to single species communities ($\chi^2 = 0.17089$, df = 1, $P = 0.05$) (Fig. 3). Our analysis also revealed that fish order ($R^2 = 0.57215$, $P = 0.003$), swimming behaviour ($R^2 = 0.09904$, $P = 0.005$), climate ($R^2 = 0.13315$, $P = 0.012$), and mean preferred water temperature ($R^2 = 0.1005$, $P = 0.05$) were significant predictors of beta diversity.

Importantly, we repeated the above analysis on the factors associated with virome composition on those viruses ($n = 22$) that likely infected hosts other than fish. Because we can assume that these viruses do not replicate in fish (for example, because they are related to host diet), and hence should be not shaped by aspects of fish biology and ecology, this analysis effectively constitutes an internal negative control. Indeed, this analysis revealed no association between virome composition and host ecological traits (viral abundance: $P = 0.0$; alpha diversity: $P = 0.3$; Shannon diversity: $P = 0.9$; and beta diversity: $P = 0.3$), thereby adding weight to the biological associations described above in the fish viruses.

4. Discussion

The metagenomic revolution is enabling us to uncover more of a largely unknown virosphere. Here, we utilized mNGS to identify new viruses associated with fish, characterising the viromes of 23 species of marine fish that spanned nine taxonomic orders and identifying 47 novel viruses spanning 22 different virus families. This included 25 new vertebrate-associated viruses and a further 22 viruses associated with protozoans, plants, arthropods, and fungi. Interestingly, the novel viruses included the first fish virus in the *Matonaviridae* that are the closest phylogenetic relatives of the mammalian rubella viruses. We also used these data to provide an initial assessment of how aspects of host biology might impact virus diversity and evolution. Although our study was limited to 23 fish species, on these data we found that host phylogeny (taxonomy) was strongly associated with the composition of fish viromes. We also identified several other host traits that were also associated with virus abundance and/or diversity, particularly preferred mean water temperature, climate, habitat depth, community diversity and whether fish swim in schools or are solitary. That these traits were not correlated with the composition of diet and microbiome-associated viruses that do not actively replicate in fish suggests that the patterns observed in marine fish are real, although it will clearly be important to test these initial conclusions using larger numbers of fish species sampled from a diverse set of environments.

Many of the viruses identified in this study were phylogenetically related to other, recently discovered, viruses of fish (Dill et al. 2016; Lauber et al. 2017; Geoghegan et al. 2018a; Shi et al. 2018a). However, there were some notable exceptions. Tiger flathead matonavirus represents the only fish viral species in the *Matonaviridae* and forms a distinct clade with a rubivirus discovered in a Chinese water snake. The discovery of this phylogenetically distinct fish virus tentatively suggests the possibility of a fish host origin for this family, although it is clear that confirmation will require the sampling of a far wider set of hosts. Indeed, it is notable that additional rubella-like viruses have recently been identified in a range of mammalian hosts, including bats (Bennett et al. 2020). A fish origin might also be the case for other virus families such as the *Hantaviridae* and *Filoviridae*, as the fish viruses in these families often fall basal to viruses in other vertebrate hosts such as birds and mammals (also see Shi et al. 2018a). In contrast, in some other virus families such as the *Astroviridae, Picornaviridae, Flaviviridae*, and *Rhabdoviridae*, viruses associated with fish are distributed throughout the phylogeny suggestive of a past history of common host-jumping. Regardless, available data suggests that fish viruses harbour more phylogenetic diversity than the better studied mammalian and avian viruses within these families. It is also clear that the discovery of novel viruses in fish has expanded our knowledge of the diversity, evolutionary history and host range of RNA viruses in general.

Although there is often a clear phylogenetic division between those viruses likely to infect fish and those associated with diet or microbiome, in some cases this separation can be nuanced. For instance, although totiviruses were thought to only infect unicellular fungi, their known host range has now expanded to include arthropods and fish (Mikalsen, Haugland, and Evensen 2016; Mor and Phelps 2016; Løvoll et al. 2010). In particular, piscine myocarditis virus is a totivirus shown by *in situ* hybridization to infect Atlantic salmon and is associated with cardiomyopathy syndrome in salmon (Haugland et al. 2011). Similarly, viruses within the *Narnaviridae* are widespread in fungi, and have now been extended to include both invertebrates (Shi et al. 2016) and protist (Charon et al. 2019). Due to their phylogenetic position, we assume the narna-like viruses identified here are associated with fungal parasites in these samples.

As well as identifying new viruses, we sought to provisionally identify associations between host traits and the overall composition of fish viruses, although this analysis was clearly limited by the available sample size. A notable observation was that fish virome composition, reflected in measures of viral richness, abundance and diversity, is most impacted by the phylogenetic relationships (i.e. taxonomy) of the host in question. This in turn suggests that fish viruses might have co-diverged with fish hosts over evolutionary time-scales, a pattern supported by the general relationship between vertebrate host class and virus phylogeny observed for RNA viruses as a whole (Shi et al. 2018a). However, it is also clear that cross-species is also a common occurrence in virus evolution (Geoghegan, Duchêne, and Holmes 2017). Indeed, it is possible that the strong association of host taxonomy and virome composition in some cases reflects preferential host switching among fish species (otherwise known as the 'phylogenetic distance effect'; Longdon et al. 2014), perhaps because viruses spread more often between phylogenetically closely related hosts due to the use of similar cell receptors (Charleston and Robertson 2002). These competing theories could be tested by more detailed co-phylogenetic comparisons among fish species that exhibit no ecological overlap thereby precluding cross-species transmission.

Our analysis also provided some evidence that virus abundance was negatively associated with the preferred water temperature of the fish species in question. Specifically, viruses were more abundant in fish that preferred cooler temperatures compared to those that prefer warmer temperatures. In this

context it is noteworthy that virus transmission and disease outbreaks have been shown to be influenced by temperature and seasonality in farmed fish (Crane and Hyatt 2011). Moreover, for some viruses, host mortality is water temperature-dependent. For example, a highly infectious disease in fish, nervous necrosis virus, is more pathogenic at higher temperatures (Toffan et al. 2016), while infectious haematopoietic necrosis virus, which causes disease in salmonid fish such as trout and salmon, causes mortality only at low temperatures (Dixon et al. 2016). As the oceans continue to warm, it is crucial to understand the impact of increased temperatures on both marine life and virus evolution and emergence, especially as it is projected that outbreaks of marine diseases are likely to increase in frequency and severity (Karvonen et al. 2010; Dallas and Drake 2016).

Also of note was that on these data, fish living in diverse, multi-fish species communities harboured more diverse viromes at a higher abundance than fish that live in less diverse, single-species communities. Previously, host community diversity has been hypothesized to lead to a decrease in infectious disease risk through the theory of the 'dilution effect'(Schmidt and Ostfeld, 2001). This theory views an increase in host species' community diversity as likely to reduce disease risk, because encounter rates among preferred hosts are decreased, and both experimental and field studies have shown this phenomenon to occur across many host systems, particularly those involving vector-borne disease (LoGiudice et al. 2003; Keesing, Holt, and Ostfeld 2006; Ostfeld and Keesing 2012). Although it might be reasonable to assume that increased virus abundance and diversity is directly correlated with disease risk, the association between host community diversity with that of virus diversity and abundance has not previously been tested. Our results, although preliminary, indicated that high multi-species community diversity in fish may be associated with increased virus diversity and abundance. It is possible that elevated community diversity in fish simply increases the total number of hosts in the system, in turn increasing viral diversity, particularly since host jumping appears to be common in fish viruses (Geoghegan et al. 2018a).

Finally, it is noteworthy that since these fish species were market-bought rather than being directly sampled during fishing trips (with the exception of the pygmy goby), it is possible that viruses with short durations of infection were not detected. In addition, the relatively small number of individuals sampled here, and that samples were necessarily pooled to aid virus discovery, unavoidably limits some of the conclusions drawn. In particular, the host traits summarized here, such as life span, were taken at the overall species level rather than for the individuals sampled. It is therefore important to broaden sampling of fish and their viruses both geographically and seasonally, and include phenotypic data for the individuals sampled. This notwithstanding, our data again shows that fish harbour a very large number of diverse viruses (Shi et al. 2018; Lauber et al. 2017). Indeed, even the pygmy goby, one of the shortest-lived vertebrates on earth that lives for a maximum of 59 days on the reef (Depczynski and Bellwood 2005), harboured novel viruses that were assigned to three distinct virus families.

The new viruses discovered here greatly expand our knowledge of the evolutionary history of many virus families, particularly those with RNA genomes, with viruses identified in fish species that span highly diverse taxonomic orders. More broadly, the use of metagenomics coupled with a diverse multi-host, tractable system such as fish has the potential to reveal how host factors can shape the composition of viromes and that might ultimately lead to cross-species transmission and virus emergence.

Data availability

All sequence reads generated in this project are available under the NCBI Short Read Archive (SRA) under BioProject PRJNA637122 and all consensus virus genetic sequences have been deposited in GenBank under accession MT579871-MT579895.

Supplementary data

Supplementary data are available at *Virus Evolution* online.

Acknowledgements

The authors thank the New South Wales Department of Primary Industries for help sourcing fish samples. We thank efishalbum.com for fish images in Fig. 1, which were used with permission.

Funding

E.C.H. and D.R.B. were funded by ARC Australian Laureate Fellowships (FL170100022 and FL190100062, respectively). This work was partly funded by ARC Discovery grant DP200102351 to E.C.H. and J.L.G. and a Macquarie University Grant awarded to J.L.G.

Conflict of Interest: none declared.

References

Alarcon-Schumacher, T. et al. (2019) 'Elucidating Viral Communities during a Phytoplankton Bloom on the West Antarctic Peninsula', *Frontiers in Microbiology*, 10: 1014.

Bennett, A. J. et al. (2020) 'Relatives of Rubella Virus in Diverse Mammals', *Nature in Nature*, 586: 424–8.

Bergh, Ø. et al. (1989) 'High Abundance of Viruses Found in Aquatic Environments', *Nature*, 340: 467–8.

Betancur-R, R. et al. (2017) 'Phylogenetic Classification of Bony Fishes', *BMC Evolutionary Biology*, 17: 162.

Breitbart, M., and Rohwer, F. (2005) 'Here a Virus, There a Virus, Everywhere the Same Virus?', *Trends in Microbiology*, 13: 278–84.

Buchfink, B., Xie, C., and Huson, D. H. (2015) 'Fast and Sensitive Protein Alignment Using DIAMOND', *Nature Methods*, 12: 59–60.

Charon, J. et al. (2019) 'Novel RNA Viruses Associated with *Plasmodium vivax* in Human Malaria and Leucocytozoon Parasites in Avian Disease', *PLoS Pathogens*, 15: e1008216.

Capella-Gutierrez, S., Silla-Martinez, J. M., and Gabaldon, T. (2009) 'trimAl: A Tool for Automated Alignment Trimming in Large-Scale Phylogenetic Analyses', *Bioinformatics*, 25: 1972–3.

Chang, W.-S. et al. (2019) 'Metagenomic Discovery and co-Infection of Diverse Wobbly Possum Disease Viruses and a Novel Hepacivirus in Australian Brushtail Possums', *BMC* ', *One Health Outlook*, 1: 5.

Charleston, M. A., and Robertson, D. L. (2002) 'Preferential Host Switching by Primate Lentiviruses Can account for Phylogenetic Similarity with the Primate Phylogeny', *Systematic Biology*, 51: 528–35.

Coutinho, F. H. et al. (2017) 'Marine Viruses Discovered via Metagenomics Shed Light on Viral Strategies throughout the Oceans', *Nature Communications*, 8: 15955.

Crane, M., and Hyatt, A. (2011) 'Viruses of Fish: An Overview of Significant Pathogens', *Viruses*, 3: 2025–46.

Dallas, T., and Drake, J. M. (2016) 'Fluctuating Temperatures Alter Environmental Pathogen Transmission in a Daphnia-Pathogen System', *Ecology and Evolution*, 6: 7931–8.

De Corte, D. et al. (2012) 'Links between Viruses and Prokaryotes throughout the Water Column along a North Atlantic Latitudinal Transect', *The ISME Journal*, 6: 1566–77.

Depczynski, M., and Bellwood, D. R. (2005) 'Shortest Recorded Vertebrate Lifespan Found in a Coral Reef Fish', *Current Biology*, 15: R288–9.

Dill, J. A. et al. (2016) 'Distinct Viral Lineages from Fish and Amphibians Reveal the Complex Evolutionary History of Hepadnaviruses', *Journal of Virology*, 90: 7920–33.

Dixon, P. et al. (2016) 'Epidemiological Characteristics of Infectious Hematopoietic Necrosis Virus (IHNV): A Review', *Veterinary Research*, 47: 63.

Geoghegan, J. L., Duchêne, S., and Holmes, E. C. (2017) 'Comparative Analysis Estimates the Relative Frequencies of co-Divergence and Cross-Species Transmission within Viral Families', *PLoS Pathogens*, 13: e1006215.

—— et al. (2018a) 'Hidden Diversity and Evolution of Viruses in Market Fish', *Virus Evolution*, 4: vey031.

—— et al. (2018b) 'Virological Sampling of Inaccessible Wildlife with Drones', *Viruses*, 10: 300.

Gregory, A. C. et al. (2019) 'Marine DNA Viral Macro- and Microdiversity from Pole to Pole', *Cell*, 177: 1109–23.e1114.

Haas, B. J. et al. (2013) 'De Novo Transcript Sequence Reconstruction from RNA-Seq: Reference Generation and Analysis with Trinity', *Nature Protocols*, 8: 1494–512.

Haugland, O. et al. (2011) 'Cardiomyopathy Syndrome of Atlantic Salmon (*Salmo salar* L.) is Caused by a Double-Stranded RNA Virus of the *Totiviridae* Family', *Journal of Virology*, 85: 5275–86.

Hothorn, T., Bretz, F., and Westfall, P. (2008) 'Simultaneous Inference in General Parametric Models', *Biometrical Journal*, 50: 346–63.

Jarungsriapisit, J. et al. (2020) 'The Effect of Temperature on the Survival of Salmonid Alphavirus Analysed Using in Vitro and in Vivo Methods', *Aquaculture*, 516: 734647.

Karvonen, A. et al. (2010) 'Increasing Water Temperature and Disease Risks in Aquatic Systems: Climate Change Increases the Risk of Some, but Not All, Diseases', *International Journal for Parasitology*, 40: 1483–8.

Katoh, K., and Standley, D. M. (2013) 'MAFFT Multiple Sequence Alignment Software Version 7: Improvements in Performance and Usability', *Molecular Biology and Evolution*, 30: 772–80.

Keane, T. M. et al. (2006) 'Assessment of Methods for Amino Acid Matrix Selection and Their Use on Empirical Data Shows That Ad Hoc Assumptions for Choice of Matrix Are Not Justified', *BMC Evolutionary Biology*, 6: 29.

Keesing, F., Holt, R. D., and Ostfeld, R. S. (2006) 'Effects of Species Diversity on Disease Risk', *Ecology Letters*, 9: 485–98.

Kelley, L. A. et al. (2015) 'The Phyre2 Web Portal for Protein Modeling, Prediction and Analysis', *Nature Protocols*, 10: 845–58.

Lagkouvardos, I. et al. (2017) 'Rhea: A Transparent and Modular R Pipeline for Microbial Profiling Based on 16S rRNA Gene Amplicons', *PeerJ*, 5: e2836.

Lara, E. et al. (2017) 'Unveiling the Role and Life Strategies of Viruses from the Surface to the Dark Ocean', *Science Advances*, 3: e1602565.

Lauber, C. et al. (2017) 'Deciphering the origin and evolution of hepatitis B viruses by means of a family of non-enveloped fish viruses', *Cell Host and Microbe*, 22: 387–99.e6.

Li, B., and Dewey, C. N. (2011) 'RSEM: Accurate Transcript Quantification from RNA-Seq Data with or without a Reference Genome', *BMC Bioinformatics*, 12: 323.

Li, W., and Godzik, A. (2006) 'Cd-hit: A Fast Program for Clustering and Comparing Large Sets of Protein or Nucleotide Sequences', *Bioinformatics*, 22: 1658–9.

Lim, E. S. et al. (2015) 'Early Life Dynamics of the Human Gut Virome and Bacterial Microbiome in Infants', *Nature Medicine*, 21: 1228–34.

LoGiudice, K. et al. (2003) 'The Ecology of Infectious Disease: Effects of Host Diversity and Community Composition on Lyme Disease Risk', *Proceedings of the National Academy of Sciences USA*, 100: 567–71.

Longdon, B. et al. (2014) 'The Evolution and Genetics of Virus Host Shifts', *PLoS Pathogens*, 10: e1004395.

Løvoll, M. et al. (2010) 'A Novel Totivirus and Piscine Reovirus (PRV) in Atlantic Salmon (Salmo salar) with Cardiomyopathy Syndrome (CMS)', *Virology Journal*, 7: 309.

Maranger, R., and Bird, D. F. (1995) 'Viral Abundance in Aquatic Systems: A Comparison between Marine and Fresh Waters', *Marine Ecology Progress Series*, 121: 217–26.

Middelboe, M., and Brussaard, C. P. D. (2017) 'Marine Viruses: Key Players in Marine Ecosystems', *Viruses*, 9: 302.

Mikalsen, A. B., Haugland, Ø., and Evensen, Ø. (2016) 'Chapter 16 - Totiviruses of Fish', *Aquaculture Virology*, 251–8.

Mor, S. K., and Phelps, N. B. (2016) 'Molecular Detection of a Novel Totivirus from Golden Shiner (*Notemigonus Crysoleucas*) Baitfish in the USA', *Archives of Virology*, 161: 2227–34.

Nguyen, L.-T. et al. (2015) 'IQ-TREE: A Fast and Effective Stochastic Algorithm for Estimating Maximum-Likelihood Phylogenies', *Molecular Biology and Evolution*, 32: 268–74.

Oksanen, J. et al. (2007) 'Vegan: community ecology package', R v.2.2-0.

Ostfeld, R. S., and Keesing, F. (2012) 'Effects of Host Diversity on Infectious Disease', *Annual Review of Ecology, Evolution, and Systematics*, 43: 157–82.

Paez-Espino, D. et al. (2016) 'Uncovering Earth's Virome', *Nature*, 536: 425–30.

Pettersson, J. H. O. et al. (2019) 'Meta-Transcriptomic Identification of Hepatitis B Virus in Cerebrospinal Fluid in Patients with Central Nervous System Disease', *Diagnostic Microbiology and Infectious Disease*, 95: 114878.

Porter, A. F. et al. (2019) 'Diversity and Evolution of Novel Invertebrate DNA Viruses Revealed by Meta-Transcriptomics', *Viruses*, 11: 1092.

Rice, P., Longden, I., and Bleasby, A. (2000) 'EMBOSS: The European Molecular Biology Open Software Suite', *Trends in Genetics*, 16: 276–7.

Roux, S. et al. (2017) 'Ecogenomics of Virophages and Their Giant Virus Hosts Assessed through Time Series Metagenomics', *Nature Communications*, 8: 858.

Schmidt, K. A., and Ostfeld, R. S. (2001) 'Biodiversity and The Dilution Effect in Disease Ecology', *Ecology*, 82: 609–19.

Shi, M., Lin, X.-D. et al. (2016) 'Redefining the Invertebrate RNA Virosphere', *Nature*, 540: 539–43.

——, —— et al. (2018a) 'The Evolutionary History of Vertebrate RNA Viruses', *Nature*, 556: 197–202.

——, Zhang, Y. Z., and Holmes, E. C. (2018b) 'Meta-Transcriptomics and the Evolutionary Biology of RNA Viruses', *Virus Research*, 243: 83–90.

Suttle, C. A. (2005) 'Viruses in the Sea', *Nature*, 437: 356–61.

Temmam, S. et al. (2016) 'Characterization of Viral Communities of Biting Midges and Identification of Novel Thogotovirus Species and Rhabdovirus Genus', *Viruses*, 8: 77.

Tirosh, O. et al.; NISC Comparative Sequencing Program. (2018) 'Expanded Skin Virome in DOCK8-Deficient Patients', *Nature Medicine*, 24: 1815–21.

Toffan, A. et al. (2016) 'Water Temperature Affects Pathogenicity of Different Betanodavirus Genotypes in Experimentally Challenged Dicentrarchus Labrax', *Diseases of Aquatic Organisms*, 119: 231–8.

Vibin, J. et al. (2018) 'Metagenomics Detection and Characterisation of Viruses in Faecal Samples from Australian Wild Birds', *Scientific Reports*, 8: 8686.

Whittington, R. J., and Reddacliff, G. L. (1995) 'Influence of Environmental Temperature on Experimental Infection of Redfin Perch (*Perca Fluviatilis*) and Rainbow Trout (*Oncorhynchus mykiss*) with Epizootic Haematopoietic Necrosis Virus, an Australian Iridovirus', *Australian Veterinary Journal*, 72: 421–4.

Wille, M. et al. (2019) 'Virome Heterogeneity and Connectivity in Waterfowl and Shorebird Communities', *The ISME Journal*, 13: 2603–16.

—— (2020) 'Unravelling Virus Community Ecology in Bats through the Integration of Metagenomics and Community Ecology', *Molecular Ecology*, 29: 23–5.

Zhang, Y.-Z., Shi, M., and Holmes, E. C. (2018) 'Using Metagenomics to Characterize an Expanding Virosphere', *Cell*, 172: 1168–72.

—— et al. (2018) 'The Diversity, Evolution and Origins of Vertebrate RNA Viruses', *Current Opinion in Virology*, 31: 9–16.

Virology 576 (2022) 42–51

Contents lists available at ScienceDirect

Virology

journal homepage: www.elsevier.com/locate/virology

Faecal virome of the Australian grey-headed flying fox from urban/suburban environments contains novel coronaviruses, retroviruses and sapoviruses

Kate Van Brussel [a], Jackie E. Mahar [a], Ayda Susana Ortiz-Baez [a], Maura Carrai [b], Derek Spielman [c], Wayne S.J. Boardman [d], Michelle L. Baker [e], Julia A. Beatty [b], Jemma L. Geoghegan [f,g], Vanessa R. Barrs [b,h], Edward C. Holmes [a,*]

[a] Sydney Institute for Infectious Diseases, School of Life & Environmental Sciences and School of Medical Sciences, The University of Sydney, NSW, 2006, Australia
[b] Jockey Club College of Veterinary Medicine & Life Sciences, City University of Hong Kong, Kowloon Tong, People's Republic of China
[c] School of Veterinary Science, Faculty of Science, University of Sydney, Sydney, NSW, 2006, Australia
[d] School of Animal and Veterinary Sciences, Faculty of Science, Engineering and Technology, University of Adelaide, Adelaide, SA, 5371, Australia
[e] CSIRO Australian Centre for Disease Preparedness, Health and Biosecurity Business Unit, Geelong, VIC, 3220, Australia
[f] Department of Microbiology and Immunology, University of Otago, Dunedin 9010, New Zealand
[g] Institute of Environmental Science and Research, Wellington, 5022, New Zealand
[h] Centre for Animal Health and Welfare, City University of Hong Kong, Kowloon Tong, People's Republic of China

ARTICLE INFO

Keywords:
Coronavirus
Sapovirus
Retrovirus
Faecal
Mammalian
Grey-headed flying fox

ABSTRACT

Bats are important reservoirs for viruses of public health and veterinary concern. Virus studies in Australian bats usually target the families *Paramyxoviridae*, *Coronaviridae* and *Rhabdoviridae*, with little known about their overall virome composition. We used metatranscriptomic sequencing to characterise the faecal virome of grey-headed flying foxes from three colonies in urban/suburban locations from two Australian states. We identified viruses from three mammalian-infecting (*Coronaviridae, Caliciviridae, Retroviridae*) and one possible mammalian-infecting (*Birnaviridae*) family. Of particular interest were a novel bat betacoronavirus (subgenus *Nobecovirus*) and a novel bat sapovirus (*Caliciviridae*), the first identified in Australian bats, as well as a potentially exogenous retrovirus. The novel betacoronavirus was detected in two sampling locations 1375 km apart and falls in a viral lineage likely with a long association with bats. This study highlights the utility of unbiased sequencing of faecal samples for identifying novel viruses and revealing broad-scale patterns of virus ecology and evolution.

1. Introduction

Bats (order Chiroptera) are one of the largest mammalian orders with a unique physiology adapted for flight. The number of bat colonies in urban habitats has increased in recent decades, leading to more frequent interactions with humans, companion animals and livestock that have in turn facilitated outbreaks of zoonotic disease (Plowright et al., 2011). This process has been dramatically highlighted by the emergence of severe acute respiratory syndrome coronavirus 2 (SARS-CoV-2) and the detection of SARS-like coronaviruses in Asian bat populations (Temmam et al., 2022; Zhou et al., 2020, 2021; Wacharapluesadee et al., 2021; Murakami et al., 2020). In addition, bats have been associated with the emergence of Hendra virus (Halpin et al., 2000), Nipah virus (Yob et al., 2001), lyssaviruses (Botvinkin et al., 2003; Gould et al., 1998) and SARS-CoV-1 (Li et al., 2005). In turn, these outbreaks have led to an increased sampling of bat species, and the widespread use of metagenomic sequencing has enabled more detailed exploration of the bat virome (Wu et al., 2016; Hardmeier et al., 2021; Van Brussel and Holmes, 2022).

In Australia, bat species of the *Pteropus* genus are reservoir hosts for Hendra virus and Menangle virus, zoonotic pathogens of the family *Paramyxoviridae* (Halpin et al., 2000; Philbey et al., 1998), as well as Australian bat lyssavirus, a zoonotic virus of the *Rhabdoviridae* that causes rabies in mammals (Gould et al., 1998). Studies of viruses in bats in Australia have largely focused on these virus families and recently identified a new member of the *Paramyxoviridae* – Cedar virus – as well

* Corresponding author.
E-mail addresses: vanessa.barrs@cityu.edu.hk (V.R. Barrs), edward.holmes@sydney.edu.au (E.C. Holmes).

https://doi.org/10.1016/j.virol.2022.09.002
Received 7 July 2022; Received in revised form 2 September 2022; Accepted 4 September 2022
Available online 19 September 2022

as a novel genotype of Hendra virus (Wang et al., 2021; Marsh et al., 2012). Although important, these studies lack information on overall virome composition, particularly those virus families not included in targeted PCR studies.

The grey-headed flying fox (*Pteropus poliocephalus*), a member of the megabat family Pteropodidae and native to Australia, is a species of importance in the context of zoonotic viruses. Grey-headed flying foxes are distributed throughout the eastern coastline of Australia (Queensland, New South Wales and Victoria) and more recently a colony was established in Adelaide (South Australia). Grey-headed flying foxes feed on fruit, pollen and nectar and roost in large colonies, sometimes sharing roosting locations with other species of *Pteropus,* allowing intraspecies and interspecies virus transmission (Timmiss et al., 2021). Roosting sites are commonly located alongside human communities including in densely populated urban settings (Williams et al., 2006). As numerous viruses are transmitted by faeces and other excretions, the co-habitation between bats and humans likely increases the risk of zoonotic spill-over.

Herein, we used metatranscriptomic sequencing of faecal samples to describe the community of viruses present in the gastrointestinal tract of grey-headed flying foxes from three sampling locations in two Australian states – Centennial Park and Gordon in Sydney, New South Wales, and the Botanic Park, Adelaide in South Australia. Specifically, to reveal the composition and abundance of viruses in bats residing in metropolitan areas we sampled roosting sites either located in a residential setting or in parks that are frequented by humans.

2. Methods

2.1. Sample collection

Faecal samples were collected from grey-headed flying fox roosting sites in three regions of Australia: Centennial Parklands, Centennial Park New South Wales (NSW), Gordon NSW, and Botanic Park, Adelaide parklands, Adelaide, South Australia (Table 1, Fig. 1A). Sampling was conducted over two dates in 2019 for the Centennial Park and Gordon sites, while the roosting site in the Adelaide parklands was sampled over several months in 2019 (Table 1). A plastic sheet of approximately 3 × 5 m was placed under densely populated trees the night before collection. The following morning samples captured by the plastic sheet were placed into 2 mL tubes and immediately stored at −80 °C until processing. Any faecal sample touching or submerged in urine was discarded.

Table 1
Sampling overview, including number of samples allocated to sequencing pools and sequencing metadata.

Location	Sampling date	Pool no.	No. Of samples	No. Of reads	No. Of contigs
Centennial Park, NSW 33. 89999°S, 151.23592°E	5 February 2019	01	12	24,732,494	159,527
		02	9	35,835,953	147,425
		03	9	31,960,624	107,431
	26 February 2019	04	9	19,833,973	111,196
		05	11	31,410,836	136,180
		06	9	29,318,213	105,118
		07	10	19,160,704	90,339
Gordon, NSW 33.75065°S, 151.16242°E	12 March 2019	01	12	52,605,108	89,247
		02	12	48,784,843	50,574
		03	9	27,396,450	118,509
	26 March 2019	04	11	36,591,148	181,524
		05	12	36,815,461	146,466
		06	12	52,934,611	97,013
		07	10	37,980,832	156,960
Adelaide, SA 34.91571°S, 138.6068°E	2019	01	8	25,977,712	135,969
	2019	02	9	21,113,731	113,546

2.2. RNA extraction, sequencing and read processing

Faecal samples were homogenised at 5 ms−1 for 1.5 min using the Omni Bead Ruptor 4 with 1.44 mm ceramic beads (Omni international) in 600 μL lyse buffer. Total RNA was extracted from each sample individually using the RNeasy Plus Mini Kit (Qiagen) following the manufacturer's protocol. RNA was pooled in equimolar ratios and separated by sampling location, date and RNA concentration (Table 1). Ribosomal RNA was depleted followed by the construction of sequencing libraries using the Illumina Stranded Total RNA Prep with Ribo-Zero Plus (Illumina) preparation kit. Libraries were sequenced as 150 bp paired-end on the Illumina Novaseq 6000 platform at the Australian Genome Research Facility (AGRF).

Read ends with a quality score of below 25 phred and adapter sequences were removed using cutadapt v1.8.3 (Kechin et al., 2017). Sortmerna v4.3.3 was used to remove 5 S and 5.8 S, eukaryotic 18 S and 23 S, bacterial 16 S and 23 S, and archaeal 16 S and 23 S ribosomal RNA (rRNA) reads (Kopylova et al., 2012). The filtered reads were then *de novo* assembled using Megahit v1.1.3 (Li et al., 2015) and the contigs compared to the non-redundant protein database using diamond v2.0.9. The Genemark heuristic approach (Besemer and Borodovsky, 1999; Zhu et al., 2010) and information from closely related viruses were used to predict genes and annotate genomes. Intact retrovirus genomes were detected using an in-house pipeline (Chang et al., manuscript in preparation). The Geneious assembler (available in Geneious Prime version 2022.1.1) was used to reassemble megahit contigs from multiple libraries for bat faecal associated retrovirus 2 (see Results). The final sequence for bat faecal associated retrovirus 2 (see Results) was determined by mapping reads from all libraries to the reassembled genome on Geneious Prime and using a 0% (majority) threshold for the final consensus sequence. Additionally, a negative control extraction library was sequenced to help exclude viral contaminants. No viruses present in the negative control library were present in the bat faecal libraries.

2.3. Abundance estimation

Virus and host abundance were estimated by mapping non-rRNA reads from each library to assembled contigs, and to the COX1 gene (accession no. KF726143) from the *P. alecto* (Black flying fox) genome using the Bowtie2 alignment method in RSEM and expected count values (Langmead and Salzberg, 2012). The impact of index-hopping was minimised by excluding the expected count value for a contig in any library that was less than 0.01% of the highest read count for that assembled contig in any other library.

2.4. Phylogenetic analysis

Virus amino acid sequences were aligned with related sequences (i.e., representing the same virus family and/or genus) retrieved from the NCBI/GenBank database using MAFTT v7.450 (Katoh and Standley, 2013) and the E–INS–I algorithm (Katoh et al., 2005). The partial RdRp sequence of P. alecto/Aus/SEQ/2009 was retrieved from Smith et al. (2016). The gappyout method in TrimAL v1.4.1 was used to remove ambiguous regions in the alignment (Capella-Gutiérrez et al., 2009). Maximum likelihood trees of each data set were inferred using IQ-TREE v1.6.7 (Nguyen et al., 2014), employing the best-fit amino acid substitution model determined by the ModelFinder program (Kalyaanamoorthy et al., 2017) in IQ-TREE. Nodal support was accessed using 1000 ultrafast bootstrap replicates (Hoang et al., 2017). Any virus sequence in this study with over 90% nucleotide similarity to another detected here was excluded from phylogenetic analysis.

2.5. PCR validation of coronavirus, sapovirus and retrovirus

SuperScript IV One-Step RT-PCR (Invitrogen) was used to amplify bat faecal coronavirus CP07/aus/1 (RdRp), bat faecal sapovirus Ad02/

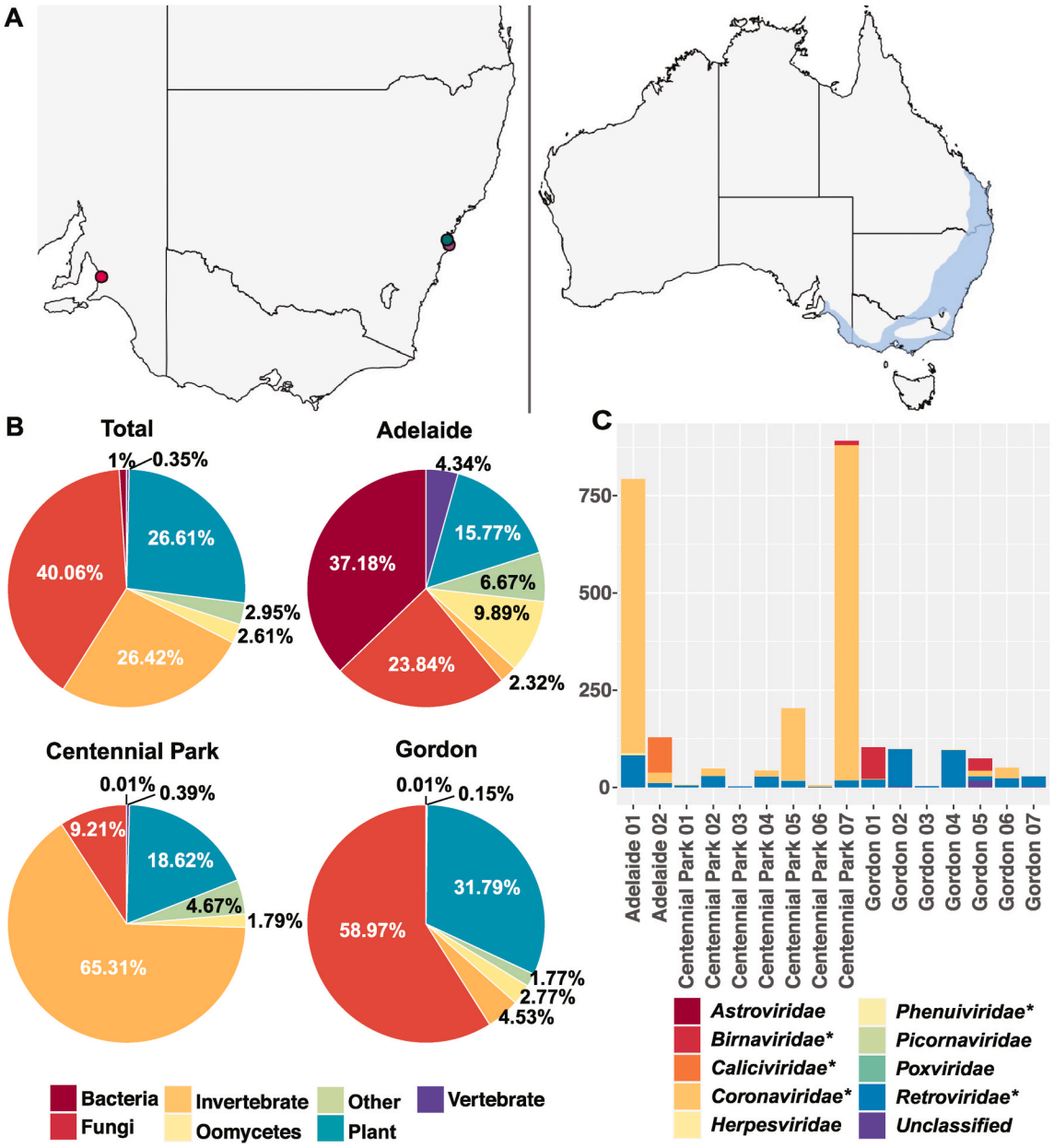


Fig. 1. Overview of sampling sites and bat faecal sample composition. (A) Sampling locations in Australia (left) and distribution map of the grey-headed flying fox (right) (IUCN, 2021). (B) The proportion of virus reads separated by likely host group based on the contig host designation of the closest relatives in the NCBI non-redundant protein database, shown as a proportion of the total virus reads in each sample location. (C) Read abundance presented as reads per million (RPM) for the vertebrate-associated virus sequences for each library and separated by virus family. The virus families discussed in this study are highlighted with an asterisk.

aus/1 (RdRp), bat faecal associated retrovirus 1 G04/aus/1 (pol) and bat faecal associated retrovirus AdCPG/aus/1 (complete genome) from total RNA from all 16 sequencing library pools (Supplementary Table 2).

3. Results

3.1. Virome overview

In total, 164 faecal samples allocated to 16 libraries underwent metatranscriptomic sequencing. This generated 19,160,704 to 52,934,611 reads per library (average of 33,278,293 reads) after read filtering (Table 1). Reads were *de novo* assembled into 50,574 to 181,524 contigs (average of 121,689 contigs) per library (Table 1). A total of 5933 contigs were assigned as of viral origin across all the libraries. The samples collected at Centennial Park, Sydney produced the most viral contigs, with 3216 identified from 65 virus families (Supplementary Fig. 1). The Gordon, NSW sample site produced 2399 virus contigs from 66 virus families, while the Adelaide site contained 318 virus contigs from 33 virus families, although this site had only two sequencing libraries comprising 17 faecal samples, compared to seven sequencing libraries for each of the other two locations (69 faecal samples from Centennial Park, 78 from Gordon) (Table 1, Supplementary Fig. 1).

Analysis of read abundance of the 5933 virus contigs identified by screening the NCBI protein database revealed that virus reads were largely associated with viruses of invertebrates (26.42% of total contigs), fungi (40.06%) and plants (26.61%), representing 79 virus families (Fig. 1B, Supplementary Fig. 1). These viruses were most likely associated with host diet and differed in frequency depending on sampling site (Fig. 1B, Supplementary Fig. 1). The plant, fungal, and oomycete-associated viruses, as well as those likely to be bacteriophage

(including the picobirnaviruses) were not considered further. Importantly, however, we also identified sequences from viruses likely associated with mammalian infection (3% overall), including near complete genomes from members of the *Coronaviridae, Caliciviridae* and *Retroviridae* (Fig. 1B).

3.1.1. Mammalian-associated viruses

We detected contigs from nine viral families likely to infect mammals (Fig. 1C). The *Coronaviridae* and *Retroviridae* were particularly abundant and present in 10 and 16 libraries, respectively (Fig. 1C). Members of the *Birnaviridae* and *Caliciviridae* were also abundant in specific libraries (Fig. 1C). The remaining mammalian-associated viral families were only detected at low abundance and the contigs were not of sufficient length for further characterisation.

3.2. Novel betacoronavirus (Coronaviridae)

A novel complete betacoronavirus genome (single-strand, positive-sense RNA virus; +ssRNA) – provisionally denoted bat faecal coronavirus CP07/aus/1 – was identified in a sequencing library sampled from Centennial Park (pool no. 07) and in a sequencing library from Adelaide (pool no. 01). These two sequences exhibited 99.8% identity over the complete viral genome indicating that they represent the same species. Additionally, three sequences with 99.2–100% sequence identity to CP07/aus/1 were identified in an additional Centennial Park library (pool no. 05).

CP07/aus/1 contains ten ORFs in the arrangement ORF1a, ORF1ab, spike, NS3, envelope, matrix, nucleocapsid, NS7a, NS7b and NS7c. Transcription Regulatory Sequences (TRS) preceeded all ORFs. Additional bat coronavirus contigs ranging from 318 to 1309 bp were detected in sequencing libraries from two Gordon sampling locations. These short contigs shared 40–95% amino acid identity to CP07/aus/1. Three of these contigs contained RdRp or spike amino acid sequences of sufficient length for phylogenetic analysis, and these were provisionally denoted bat faecal coronavirus G05/aus/1, G05/aus/2 and G05/aus/3. Based on phylogenetic analysis of the RNA-dependent RNA polymerase (RdRp) and/or spike protein, the novel betacoronaviruses detected here fell within the *Betacoronavirus* subgenus *Nobecovirus* (Fig. 2) and were most closely related to P. alecto/Aus/SEQ/2009 (for which only a partial RdRp is available) sampled from a black flying fox in south east Queensland, Australia (Smith et al., 2016) and to Pteropus rufus nobecovirus sampled from a flying fox in Madagascar (accession no. OK067319; Fig. 2) (Kettenburg et al., 2022). Pairwise comparisons revealed that CP07/aus/1 had 83% amino acid identity to Pteropus rufus nobecovirus over the complete ORF1ab replicase and 97% to P. alecto/Aus/SEQ/2009 over the partial RdRp. Amino acid identity to Pteropus rufus nobecovirus over the spike and non-structural proteins was 72% and 58%, respectively. The RdRp of G05/aus/1 shared 95% amino acid identity to CP07/aus/1, while the partial spike proteins of G05/aus/2 and G05/aus/3 shared 57% and 63% amino acid identity to CP07/aus/1, respectively. It is possible that G05/aus/1 and G05/aus/2 represent transcripts from the same virus, while G05/aus/3 represents a different species to CP07/aus/1. However, this could not be confirmed as the G05/aus/3 genome was incomplete. Regardless, it is clear from the spike protein phylogeny that at least three different coronaviruses are circulating in the bats sampled here.

3.2.1. Novel sapovirus (Caliciviridae)

A near complete genome of a novel sapovirus (*Caliciviridae*, +ssRNA virus), tentatively named bat faecal sapovirus Ad02/aus/1, was detected in a sequencing library sampled from Adelaide (pool no. 2). Nine additional bat sapovirus sequences ranging from 340 to 783 bp were detected in the same sequencing library. The nine sequences shared 66–74% nucleotide and 76–81% amino acid identity to Ad02/aus/1 over the polyprotein, suggesting the presence of additional diverse sapoviruses. The near complete Ad02/aus/1 genome is 7254 bp and contains two ORFs encoding a polyprotein (near complete with likely 45 residues missing from the 5' end), and the VP2. Ad02/aus/1 exhibited

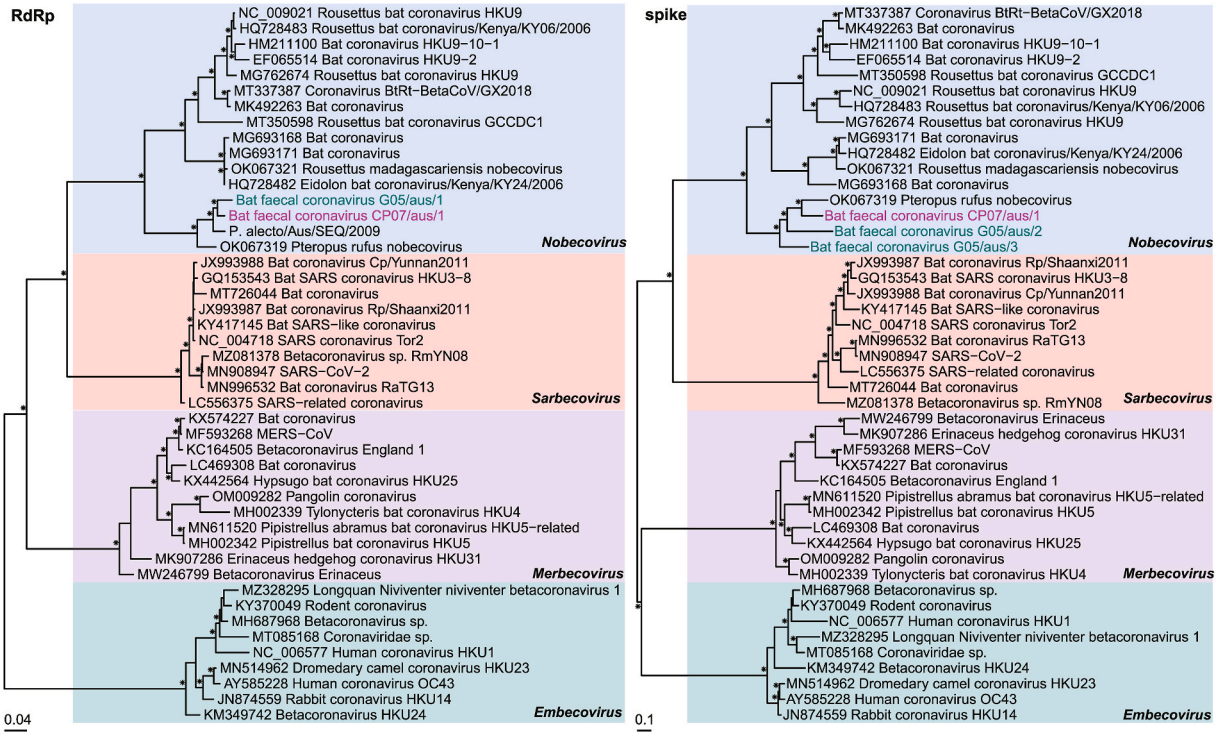


Fig. 2. Phylogenetic relationships of the novel bat betacoronaviruses based on the amino acid sequences of the RdRp and spike protein. Amino acid alignment lengths were 832 and 1092 residues for the RdRp and spike protein, respectively. Representative betacoronavirus sequences from this study are coloured by sampling location (Centennial Park, Sydney – purple, and Gordon – green) and the subgenera are highlighted. Bootstrap values > 70% are represented by the symbol shown at the branch node. The tree is rooted at midpoint for clarity and the scale bar represents the amino acid substitutions per site.

44.8% amino acid identity in the partial polyprotein to its closest relative – Bat sapovirus Bat-SaV/Limbe65/CAM/2014 (accession no. KX759620) – detected in the faeces of *Eidolon helvum* bats in Cameroon, Africa (Yinda et al., 2017). Phylogenetic analysis of the RdRp and VP1 revealed a clustering of bat sapoviruses in both trees that included the novel Australian bat sapoviruses found here (Fig. 3). Bat sapoviruses have been assigned to the putative genogroups GXIV, GXVI, GXVII, GXVIII and GXIX based on VP1 phylogeny and amino acid sequence identities. Using the same criteria, the novel sapovirus Ad02/aus/1 identified here should be assigned to its own genogroup, putatively named GXX which would also include the partial VP1 Ad02/aus/4 sequence (Supplementary Table 1, Fig. 3).

3.3. Novel birna-like virus (Birnaviridae)

Sequences related to the *Birnaviridae* (double-stranded RNA viruses; dsRNA) were detected in one Centennial Park and two Gordon libraries. All the birna-like virus sequences identified in the Centennial Park and Gordon libraries shared >99% nucleotide identity, and the complete coding region of segment B, which encodes the RdRp, was obtained from one library (Gordon 05). The *Birnaviridae* segment A that encodes the polyprotein and a small overlapping ORF was not identified in our data. Phylogenetic analysis revealed that the birna-like virus RdRp sequence, denoted G05/aus/1, was most closely related (50% amino acid identity) to the disease-causing virus Chicken proventricular necrosis virus (Fig. 4) (Guy et al., 2011), forming a distinct clade that is distantly related to the birnaviruses that infect a wide range of hosts.

3.3.1. Bat retrovirus (Retroviridae)

A near complete genome of a retrovirus was identified in Gordon library 04 and provisionally named bat faecal associated retrovirus 1 G04/aus/1. Four ORFs were observed over the 7455 bp genome and assigned as the *gag*, *pro*, *pol* and *env* genes based on the presence of conserved domains. In the pro gene we were able to identify an active site motif DTGAD predominately observed in functional retroviruses, and a helix motif GRDVL (Turnbull and Douville, 2018). We were unable to identify complete long terminal repeat (LTR) regions in the 7455 bp genome, although this may be due to incomplete assembly at the 5′ and/or 3’ end, rather than a true absence of LTRs. Importantly, as the four ORFs contained the appropriate retrovirus conserved domains and were uninterrupted by stop codons, it is possible that G04/aus/1 is potentially exogenous and functional. A BLASTn analysis of the complete G04/aus/1 genome revealed no match to any bat reference genome on NCBI/GenBank. G04/aus/1 exhibited 56% amino acid identity in the pol protein to its closest relative, Simian retrovirus 2 (accession M16605), a presumably exogenous retrovirus (Thayer et al., 1987). The abundance for this novel retrovirus in the Gordon 04 library was 90 RPM (2453 reads) (Fig. 1C).

A further near complete retroviral genome was identified by reassembling 31 partial contig sequences from 10 libraries from all three sample locations. PCR confirmed that the entire genome was present in the G07 sequencing library pool (Supplementary Table 2). This bat faecal associated retrovirus 2 AdCPG/aus/1 is 6630 bp and contains four open reading frames encoding the *gag*, *pro*, *pol* and *env* genes. It also contains the conserved domains expected in functional retroviruses, although the terminal end of the *env* gene is missing (either from true

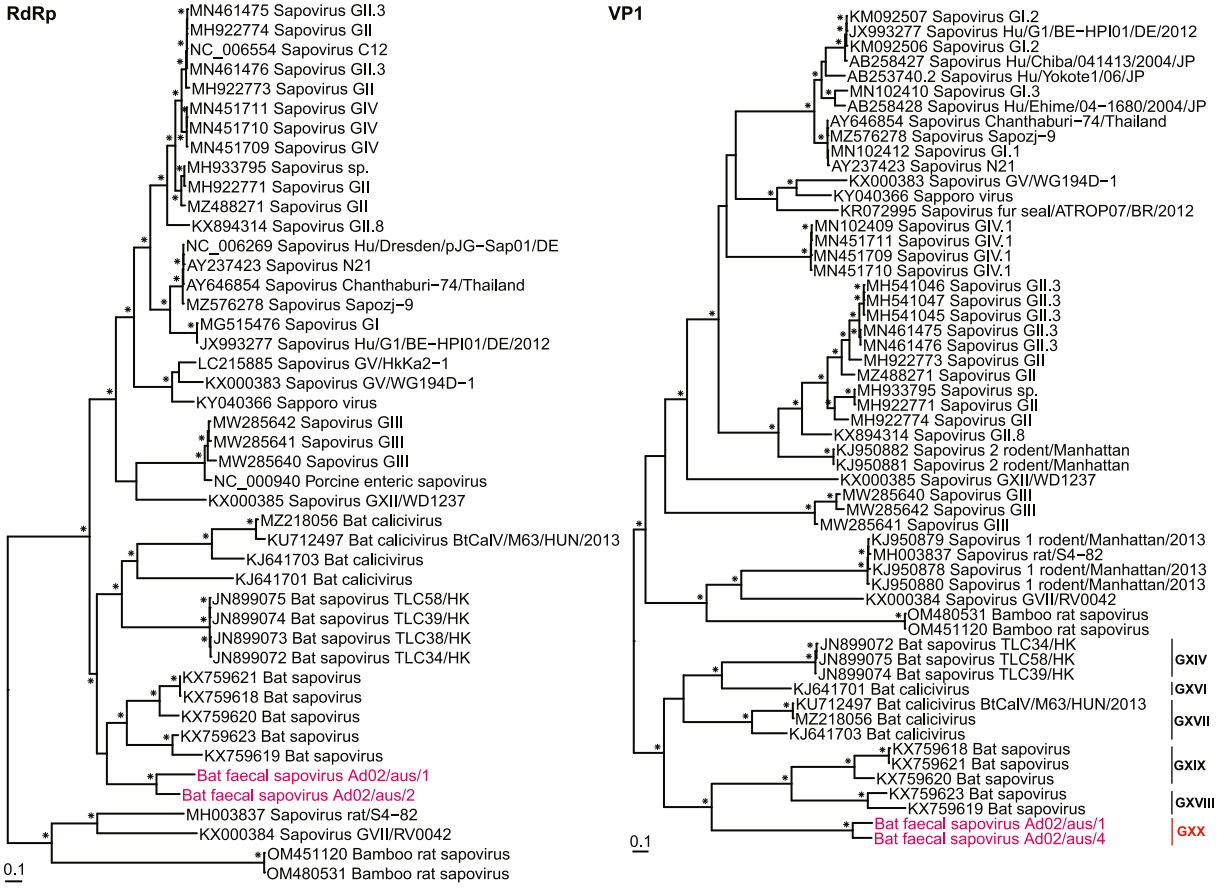


Fig. 3. Phylogenetic relationships of the novel bat sapoviruses using the amino acid sequences of the RdRp and VP1. Amino acid alignment lengths were 491 and 623 residues for the RdRp and VP1, respectively. Bat sapoviruses from this study are coloured by sampling location (Adelaide – pink) and bootstrap values > 70% are represented by the symbol shown at the branch node. The putative bat sapovirus genogroups are displayed to the right of the VP1 tree and our proposed putative genogroup is coloured in red. The trees are rooted at midpoint for clarity and the scale bar represents the amino acid substitutions per site.

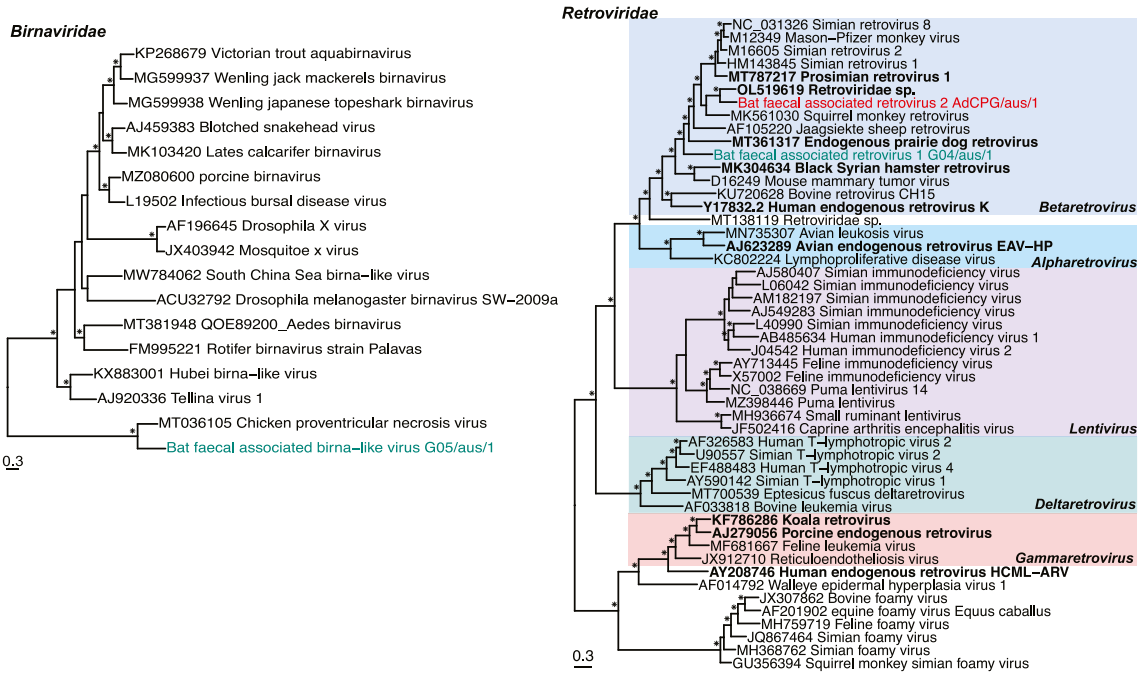


Fig. 4. Phylogenetic analysis of the birna-like virus and bat retroviruses based on the RdRp and pol amino acid sequences, respectively. The *Birnaviridae* RdRp sequence alignment was 767 amino acid resides in length while the *Retroviridae* pol alignment comprised 1356 residues. The viruses from this study are coloured by sampling location (Gordon – green) and the reassembled retrovirus sequence is in red (to indicate multiple locations). The *Retroviridae* genera are highlighted and endogenous viruses are shown in bold. The bootstrap values > 70% are represented by the symbol shown at the branch node and the tree is midpoint rooted for clarity, with the scale bar representing the amino acid substitutions per site.

truncation or incomplete assembly). AdCPG/aus/1 is most closely related to a retrovirus sampled from the lung tissue of Malayan pangolins (Ning et al., 2022). BLASTn analysis of the complete genome of AdCPG/aus/1 showed the absence of this genome in any bat reference genome on NCBI/GenBank. AdCPG/aus/1 reads were detected in 13 libraries (two Adelaide, four Centennial Park and seven Gordon) and the abundance in each library ranged from 3.7 to 68.8 RPM (127–1786 reads) (Fig. 1C). Phylogenetic analysis of the pol protein that contains the reverse transcriptase (RT) domain revealed that G04/aus/1 and AdCPG/aus/1 fell within the genus *Betaretrovirus,* clustering with both exogenous and endogenous retroviruses associated with various mammalian species (Fig. 4).

3.4. PCR confirmation

PCR confirmed that bat faecal coronavirus CP07/aus/1, bat faecal sapovirus Ad02/aus/1 and bat faecal associated retrovirus 1 G04/aus/1 were present in eight, one and three library pools, respectively (Supplementary Table 2). For all the library pools that were PCR positive, metagenomic read abundance was above the 0.01% index-hopping cut-off, although in the case of library pool CP06 read abundance was slightly below this cut-off. Library pools that had no sequence reads for these viruses were also negative by PCR (Supplementary Table 2).

3.5. Invertebrate-associated viruses

We detected likely invertebrate-associated virus sequences from seven single-strand negative-sense RNA viruses (-ssRNA), three + ssRNA virus and one dsRNA virus families, in addition to the order *Bunyavirales* (-ssRNA). The virus sequences from the *Chuviridae*, *Lispiviridae*, *Artoviridae*, *Nyamiviridae*, *Xinmoviridae*, *Qinviridae*, *Disctroviridae* and *Iflaviridae* are not discussed further, although information on positive libraries is provided (Supplementary Fig. 1) and phylogenetic analysis was performed (Supplementary Fig. 2). Virus sequences from the *Orthomyxoviridae*, *Nodaviridae*, *Reoviridae* and *Bunyavirales* are considered further as these viral groups include mammalian-infecting viruses, are important vector-borne viruses, or are able to infect mammals experimentally (*Nodaviridae*, genus *Alphanodavirus*).

Orthomyxovirus (-ssRNA virus) segments were identified in five libraries from Centennial Park. Full coding regions for two polymerase segments – PB2 and PA – and the hemagglutinin segment 2 and nucleocapsid segment 5 were present in all libraries, although a full coding region for polymerase segment PB1 was only present in a single Centennial Park library. The three polymerase proteins of Centennial Park library 06 were used for phylogenetic analysis, which revealed that this sequence was most closely related to an orthomyxovirus sampled from jumping plant lice in Australia (Fig. 5) (Käfer et al., 2019). Nodaviruses (+ssRNA virus) were detected in five Centennial Park libraries and three Gordon libraries. Both the RNA1 (RdRp) and RNA2 segments were identified, including two sequences with the complete RdRp. Nodavirus CP01/aus/1 and CP02/aus/1 were related to a nodavirus sampled from birds in China (Zhu et al., 2022) and most likely belong to the same viral species, although these fragments were only 476 and 232 amino acids, respectively. The nodavirus CP07/aus/1 RdRp segment was related to a nodavirus from arthropod hosts from China (Fig. 5) (Shi et al., 2016). Gene segments related to the *Reoviridae* (dsRNA) were present in all Centennial Park, three Gordon and one Adelaide library. The reovirus VP1 Pol segments detected here were related, albeit distantly (~40% amino acid identity) to reoviruses associated with ticks (Harvey et al., 2019; Vanmechelen et al., 2021), moths (Graham et al., 2006), bat flies (Xu et al., 2022) and the Asian citrus psyllid (Nouri et al., 2015) (Fig. 5).

Finally, bunyavirus fragments were detected in all the Adelaide and Centennial Park libraries and six Gordon libraries. Eleven RdRp coding regions were used for phylogenetic analysis which revealed that two bunyavirus sequences fell into the *Phenuiviridae* and four were basal to that family, while two sequences fell into the *Phasmaviridae*, two were basal to the *Arenaviridae* and one was basal to a grouping of five families (Fig. 6). The Adelaide bunyavirus Ad02/aus/1 was related to the plant associated genus *Tenuivirus* and the remaining 10 were related to

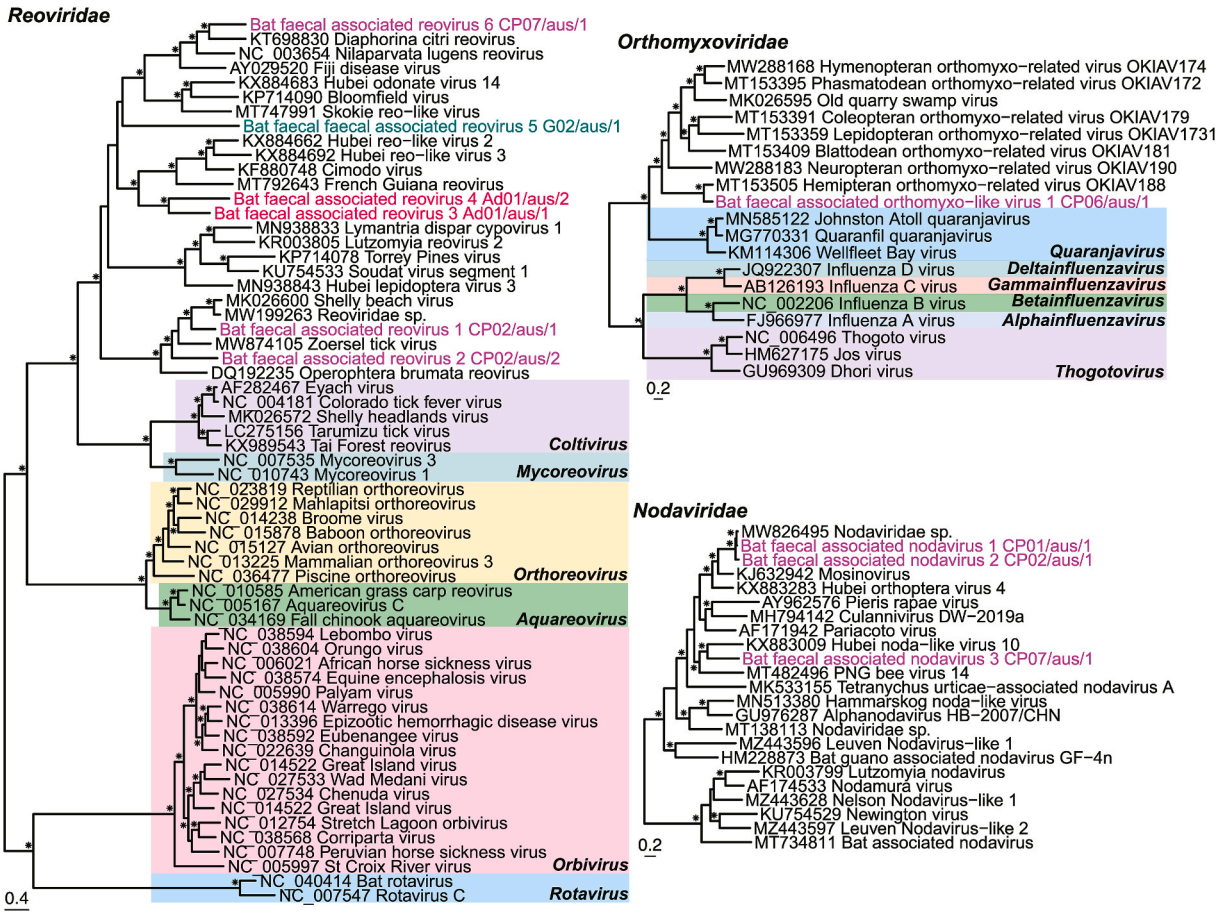


Fig. 5. Phylogenetic analysis of the invertebrate-assocaited reoviruses, orthomyxoviruses and nodaviruses based on the VP1 Pol, concatenated PB2-PB1-PA and RdRp amino acid sequences, respectively. Amino acid alignment length were 1020 residues for *Reoviridae*, 2233 residues for the *Orthomyxoviridae* and 774 residues for the *Nodaviridae*. Viruses from this study are coloured by sampling location (Adelaide – pink, Centennial Park – purple and Gordon – green) and genera are highlighted in the *Reoviridae* and *Orthomyxoviridae* tress. Bootstrap values > 70% are represented by the symbol shown at the branch node. The tree is rooted at midpoint for clarity and the scale bar represents the amino acid substitutions per site.

invertebrate hosts (Fig. 6).

4. Discussion

Virological surveillance of bats in Australia has largely focused on screening for known zoonotic viruses such as Hendra virus and Australian bat lyssavirus, although the paramyxovirus Tioman virus, for which flying foxes are the natural host, and coronaviruses are also targeted (Boardman et al., 2020; Prada et al., 2019a; Smith et al., 2016). The primary aim of these studies is to identify specific viruses using either PCR or serological data. Although such surveillance has been successful in determining the active circulation of these specific viruses, these approaches necessarily have restricted capacity to detect novel or unexpected viruses, thus providing a very limited understanding of viruses circulating in Australian bats. As bats are frequently found near human populations, they are of particular concern regarding potential zoonoses (Plowright et al., 2011; Williams et al., 2006; Halpin et al., 2000). Herein, we used metatranscriptomics to reveal the natural faecal virome of the grey-headed flying fox. Although most of the viruses identified were likely associated with bat diet, as expected from faecal sampling, we also identified viruses from three mammalian-associated families (*Coronaviridae, Caliciviridae, Retroviridae*) and one virus from the *Birnaviridae* family that may also have a mammalian association.

Both alpha- and betacoronaviruses have been identified in a variety of bat species (Smith et al., 2016; Prada et al., 2019b). Here, we characterised the complete genome of a betacoronavirus in grey-headed flying foxes that was closely related to two other betacoronaviruses sampled in flying foxes in Australia and Madagascar (Smith et al., 2016; Kettenburg et al., 2022). The current ICTV classification for coronavirus species states that less than 90% amino acid identity in the ORF1ab conserved replicase domains constitutes a new species. Although bat faecal coronavirus CP07/aus/1 shares high sequence similarity to another reported bat betacoronavirus, the P. alecto/Aus/SEQ/2009 sequence is only 146 amino acids in length, does not span the complete RdRp and is therefore difficult to classify. Accordingly, we suggest that betacoronavirus bat faecal coronavirus CP07/aus/1 represents a novel species, to which P. alecto/Aus/SEQ/2009 may also belong. The complete genome of this virus was found in both Adelaide and New South Wales (99.8% nucleotide similarity between the two genomes) and abundance counts were high in both locations (Fig. 1C), indicative of virus exchange between bat populations. Flying foxes are known to travel long distances to feed, roosting sites change depending on season, and in Australia several flying fox species share roosting sites (Timmiss et al., 2021), all of which provide opportunities for viruses to infect new individuals. Importantly, while we were only able to assemble the complete genome of one novel coronavirus, we identified partial genome fragments of at least two more diverse coronaviruses (Fig. 2), indicating that Australian bats carry a high diversity of coronaviruses as has been seen in other bat species.

This is the first report of a sapovirus in Australian bats. Previously, bat sapoviruses have been sampled from *Eidolon helvum* (Straw-coloured fruit bat) in Cameroon (Yinda et al., 2017) and Saudi Arabia (Mishra

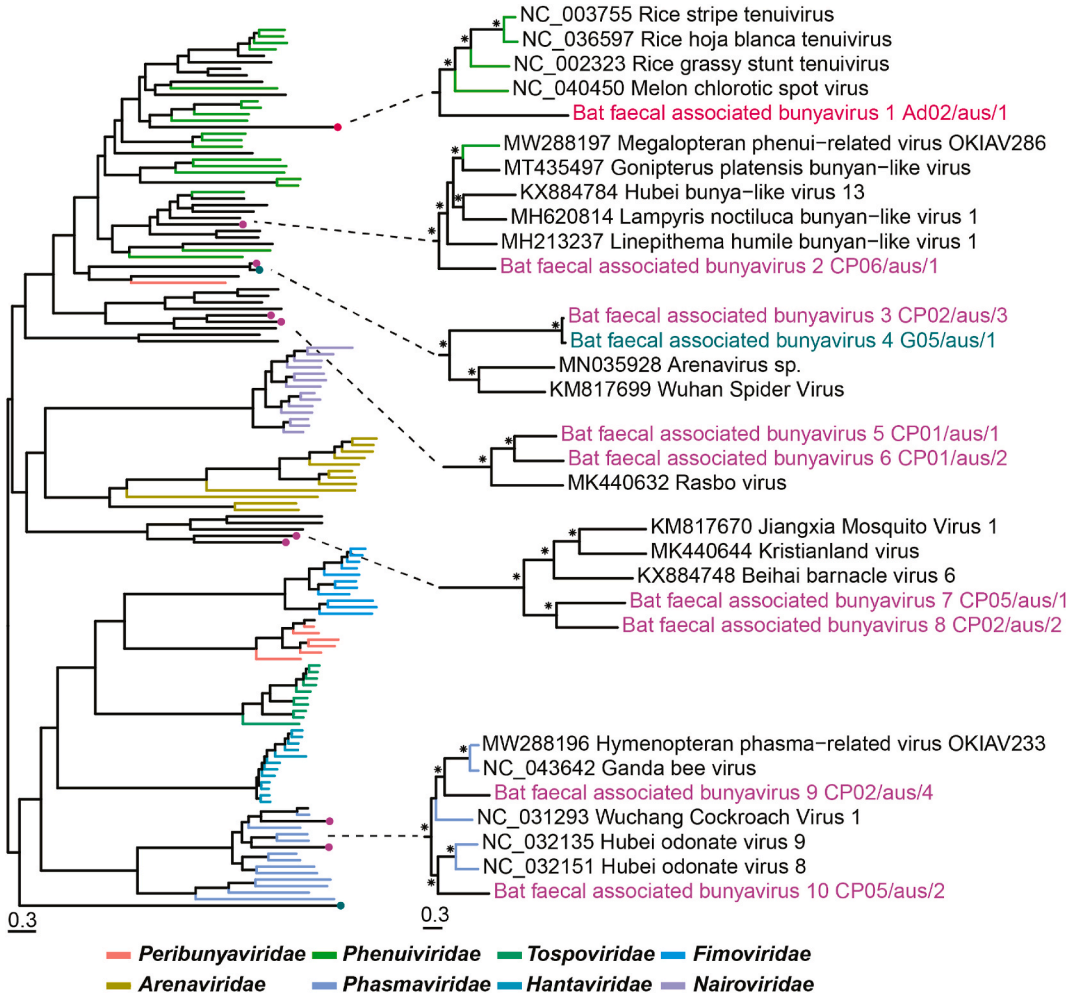


Fig. 6. Phylogenetic analysis of viruses from the order *Bunyavirales*. The RdRp amino acid sequence was used to estimate phylogenetic trees and the alignment length was 1434 amino acid residues. Viruses from this study are coloured by sampling location and bootstrap values > 70% are represented by the symbol shown at the branch node. The tree is midpoint rooted for clarity and the scale bar represents the amino acid substitutions per site.

et al., 2019) and *Hipposideros Pomona* (Pomona leaf-nosed bat) from Hong Kong (Tse et al., 2012). Currently, the bat sapoviruses characterised have been from bats with no apparent disease (Tse et al., 2012; Yinda et al., 2017; Mishra et al., 2019). Whether this is the case here is unknown because the reliance on faecal sampling meant that there was no direct interaction with individual animals. The disease potential of bat sapoviruses should be investigated further as sapoviruses have been linked to acute gastroenteritis outbreaks in humans (Oka et al., 2015) and some animal sapoviruses are closely related to those found in humans (Mombo et al., 2014; Firth et al., 2014; Martella et al., 2008).

Until the metagenomic detection of porcine birnavirus (Yang et al., 2021) and porcupine birnavirus (He et al., 2022) it was believed that the *Birnaviridae* infected fish, insects and birds exclusively (Crane et al., 2000; Da Costa et al., 2003; Chung et al., 1996; Brown and Skinner, 1996; Guy et al., 2011). We identified the segment B sequence of a novel bat faecal associated birna-like virus that was most closely related to a divergent pathogenic avian birnavirus (50% amino acid identity). Given its divergent phylogenetic position – falling basal to all other birnaviruses in a mid-point rooted tree (Fig. 4) – it is unclear whether this virus actively infects grey-headed flying foxes or is associated with a component of their diet or microbiome. While grey-headed flying foxes are not insectivores, the ingestion of insects through the consumption of fruit and nectar seems likely given the high number of invertebrate, plant and fungi viruses sequenced here (Fig. 1B, Supplementary Fig. 1). The moderate abundance values (81.6 and 31.3 RPM) cannot exclude either scenario as using a host reference gene such as COX1 for sequencing depth comparisons may not be as reliable for faecal samples as it would be for tissue samples. Further investigation is needed to determine the natural host of bat faecal associated birna-like virus and to determine what tissue types are affected.

Two intact, possibly exogenous retrovirus near complete genomes were also identified in this study and were most closely related to mammalian associated retroviruses from the genus *Betaretrovirus*. Six retroviruses have been previously characterised from Australian bat brain tissue and excretions (including faeces), all from the genus *Gammaretrovirus* (Hayward et al., 2020; Cui et al., 2012) and hence highly divergent from the viruses identified here. Although the exogenous status needs to be confirmed, it is possible that bat faecal associated retrovirus 1 G04/aus/1 and bat faecal associated retrovirus 2 AdCP-G/aus/1 constitute the first exogenous and intact betaretroviruses sampled from the faeces of bats in Australia. Unfortunately, virus identification through metatranscriptomics does not provide reliable information on whether a virus is endogenous and defective, or still functional and exogenous (Hayward et al., 2013; Hayward and Tachedjian, 2021). That the retroviruses detected here have all the necessary genes to comprise a functional virus, with undisrupted ORFs, were not detected in every library, and are not present in the bat genome, at the very least suggests that they are only recently endogenized and currently unfixed in the bat population. Further work confirming the nature of the retroviruses detected here is warranted

since bats are known to be major hosts for retroviruses (Cui et al., 2015) and their cross-species transmission across mammalian orders is commonplace (Hayward et al., 2013).

In addition to mammalian viruses, we detected virus sequences that are likely invertebrate-associated. Of particular interest were those from the *Orthomyxoviridae* and *Reoviridae* that span a wide variety of hosts including mammals and were at high abundance in some of the Centennial Park libraries. Notably, bat faecal associated reovirus 1 CP02/aus/1 groups with members of the *Reoviridae* associated with ticks. Tick-associated reoviruses from the genus *Coltivirus* – Colorado tick fever virus and Eyach virus (Goodpasture et al., 1978; Rehse-Küpper et al., 1976) – have been associated with human infection and disease such that their presence in urban wildlife merits attention.

Our study highlights the diversity of viruses in wildlife species from metropolitan areas. In this context it is notable that the bat coronaviruses identified fall within the subgenus *Nobecovirus* of betacoronaviruses. Currently, this subgenus is strongly associated with bats sampled on multiple continents, with the phylogenetic depth of the *Nobecovirus* lineage further suggesting that bats have harbored these viruses for millennia with no apparent infection of humans.

Data statement

The raw data generated for this study are available in the NCBI SRA database under the BioProject accession number PRJNA851532 and SRA accession numbers SRR19790899-SRR19790914. All genome sequences presented in phylogenetic trees are available in NCBI GenBank under the accession numbers ON872523-ON872588.

Ethics statement

Ethics approval was granted by the University of Sydney Animal Ethics Committee (AEC, 2018/1460) and the Sydney Institute for Infectious Diseases.

Funding source

The work was funded by an Australian Research Council Australian Laureate Fellowship to ECH (FL170100022) and the Sydney Institute for Infectious Diseases.

CRediT authorship contribution statement

Kate Van Brussel: Investigation, Data curation, Formal analysis, Writing – original draft, Writing – review & editing, preparation. **Jackie E. Mahar:** Formal analysis, Writing – review & editing, preparation. **Ayda Susana Ortiz-Baez:** Formal analysis, Writing – review & editing, preparation. **Maura Carrai:** Investigation, Writing – review & editing. **Derek Spielman:** Investigation. **Wayne S. J. Boardman:** Investigation, Writing – review & editing. **Michelle L. Baker:** Resources, Writing – review & editing, preparation. **Julia A. Beatty:** Investigation, Supervision. **Jemma L. Geoghegan:** Resources, Writing – review & editing. **Vanessa R. Barrs:** Conceptualisation, Methodology, Funding acquisition, Supervision, Writing – review & editing. **Edward C. Holmes:** Conceptualisation, Funding acquisition, Supervision, Writing – review & editing, Writing – original draft, Writing – review & editing.

Declaration of competing interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Appendix A. Supplementary data

Supplementary data to this article can be found online at https://doi.org/10.1016/j.virol.2022.09.002.

References

Besemer, J., Borodovsky, M., 1999. Heuristic approach to deriving models for gene finding. Nucleic Acids Res. 27, 3911–3920.
Boardman, W.S.J., Baker, M.L., Boyd, V., Crameri, G., Peck, G.R., Reardon, T., Smith, I. G., Caraguel, C.G.B., Prowse, T.A.A., 2020. Seroprevalence of three paramyxoviruses; Hendra virus, Tioman virus, Cedar virus and a rhabdovirus, Australian bat lyssavirus, in a range expanding fruit bat, the Grey-headed flying fox (*Pteropus poliocephalus*). PLoS One 15, e0232339.
Botvinkin, A.D., Poleschuk, E.M., Kuzmin, I.V., Borisova, T.I., Gazaryan, S.V., Yager, P., Rupprecht, C.E., 2003. Novel lyssaviruses isolated from bats in Russia. Emerg. Infect. Dis. 9, 1623–1625.
Brown, M.D., Skinner, M.A., 1996. Coding sequences of both genome segments of a European 'very virulent' infectious bursal disease virus. Virus Res. 40, 1–15.
Capella-Gutiérrez, S., Silla-Martínez, J.M., Gabaldón, T., 2009. trimAl: a tool for automated alignment trimming in large-scale phylogenetic analyses. Bioinformatics 25, 1972–1973.
Chung, H.K., Kordyban, S., Cameron, L., Dobos, P., 1996. Sequence analysis of the bicistronic Drosophila X virus genome segment A and its encoded polypeptides. Virology 225, 359–368.
Crane, M.S., Hardy-Smith, P., Williams, L.M., Hyatt, A.D., Eaton, L.M., Gould, A., Handlinger, J., Kattenbelt, J., Gudkovs, N., 2000. First isolation of an aquatic birnavirus from farmed and wild fish species in Australia. Dis. Aquat. Org. 43, 1–14.
Cui, J., Tachedjian, G., Wang, L.F., 2015. Bats and rodents shape mammalian retroviral phylogeny. Sci. Rep. 5, 16561.
Cui, J., Tachedjian, M., Wang, L., Tachedjian, G., Wang, L.F., Zhang, S., 2012. Discovery of retroviral homologs in bats: implications for the origin of mammalian gammaretroviruses. J. Virol. 86, 4288–4293.
Da Costa, B., Soignier, S., Chevalier, C., Henry, C., Thory, C., Huet, J.C., Delmas, B., 2003. Blotched snakehead virus is a new aquatic birnavirus that is slightly more related to avibirnavirus than to aquabirnavirus. J. Virol. 77, 719–725.
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.
Goodpasture, H.C., Poland, J.D., Francy, D.B., Bowen, G.S., Horn, K.A., 1978. Colorado tick fever: clinical, epidemiologic, and laboratory aspects of 228 cases in Colorado in 1973-1974. Ann. Intern. Med. 88, 303–310.
Gould, A.R., Hyatt, A.D., Lunt, R., Kattenbelt, J.A., Hengstberger, S., Blacksell, S.D., 1998. Characterisation of a novel lyssavirus isolated from Pteropid bats in Australia. Virus Res. 54, 165–187.
Graham, R.I., Rao, S., Possee, R.D., Sait, S.M., Mertens, P.P., Hails, R.S., 2006. Detection and characterisation of three novel species of reovirus (*Reoviridae)*, isolated from geographically separate populations of the winter moth *Operophtera brumata* (Lepidoptera: geometridae) on Orkney. J. Invertebr. Pathol. 91, 79–87.
Guy, J.S., West, A.M., Fuller, F.J., 2011. Physical and genomic characteristics identify chicken proventricular necrosis virus (R11/3 virus) as a novel birnavirus. Avian Dis. 55, 2–7.
Halpin, K., Young, P.L., Field, H.E., Mackenzie, J.S., 2000. Isolation of Hendra virus from pteropid bats: a natural reservoir of Hendra virus. J. Gen. Virol. 81, 1927–1932.
Hardmeier, I., Aeberhard, N., Qi, W., Schoenbaechler, K., Kraettli, H., Hatt, J.M., Fraefel, C., Kubacki, J., 2021. Metagenomic analysis of fecal and tissue samples from 18 endemic bat species in Switzerland revealed a diverse virus composition including potentially zoonotic viruses. PLoS One, e0252534.
Harvey, E., Rose, K., Eden, J.-S., Lo, N., Abeyasuriya, T., Shi, M., Doggett, S.L., Holmes, E. C., 2019. Extensive diversity of RNA viruses in Australian ticks. J. Virol. 93, e01358, 18.
Hayward, A., Grabherr, M., Jern, P., 2013. Broad-scale phylogenomics provides insights into retrovirus-host evolution. Proc Natl Acad Sci USA110 20146–20151.
Hayward, J.A., Tachedjian, G., 2021. Retroviruses of bats: a threat waiting in the wings? mBio 12, e0194121.
Hayward, J.A., Tachedjian, M., Kohl, C., Johnson, A., Dearnley, M., Jesaveluk, B., Langer, C., Solymosi, P.D., Hille, G., Nitsche, A., Sánchez, C.A., Werner, A., Kontos, D., Crameri, G., Marsh, G.A., Baker, M.L., Poumbourios, P., Drummer, H.E., Holmes, E.C., Wang, L.F., Smith, I., Tachedjian, G., 2020. Infectious KoRV-related retroviruses circulating in Australian bats. Proc. Natl. Acad. Sci. U.S.A. 117, 9529–9536.
He, W.T., Hou, X., Zhao, J., Sun, J., He, H., Si, W., Wang, J., Jiang, Z., Yan, Z., Xing, G., Lu, M., Suchard, M.A., Ji, X., Gong, W., He, B., Li, J., Lemey, P., Guo, D., Tu, C., Holmes, E.C., Shi, M., Su, S., 2022. Virome characterization of game animals in China reveals a spectrum of emerging pathogens. Cell 185, 1117–1129 e8.
Hoang, D.T., Chernomor, O., von Haeseler, A., Minh, B.Q., Vinh, L.S., 2017. UFBoot2: improving the ultrafast bootstrap approximation. Mol. Biol. Evol. 35, 518–522.
International Union for Conservation of Nature, 2021. Data from "Pteropus poliocephalus". The IUCN Red List of Threatened Species 2021–2023. https://www.iucnredlist.org/species/18751/22085511.
Käfer, S., Paraskevopoulou, S., Zirkel, F., Wieseke, N., Donath, A., Petersen, M., Jones, T. C., Liu, S., Zhou, X., Middendorf, M., Junglen, S., Misof, B., Drosten, C., 2019. Reassessing the diversity of negative strand RNA viruses in insects. PLoS Pathog. 15, e1008224.
Kalyaanamoorthy, S., Minh, B.Q., Wong, T.K.F., von Haeseler, A., Jermiin, L.S., 2017. ModelFinder: fast model selection for accurate phylogenetic estimates. Nat. Methods 14, 587–589.

Katoh, K., Kuma, K., Toh, H., Miyata, T., 2005. MAFFT version 5: improvement in accuracy of multiple sequence alignment. Nucleic Acids Res. 33, 511–518.
Katoh, K., Standley, D.M., 2013. MAFFT multiple sequence alignment software version 7: improvements in performance and usability. Mol. Biol. Evol. 30, 772–780.
Kechin, A., Boyarskikh, U., Kel, A., Filipenko, M., 2017. cutPrimers: a new tool for accurate cutting of primers from reads of targeted next generation sequencing. J. Comput. Biol. 24, 1138–1143.
Kettenburg, G., Kistler, A., Ranaivoson, H.C., Ahyong, V., Andrianiaina, A., Andry, S., DeRisi, J.L., Gentles, A., Raharinosy, V., Randriambolamanantsoa, T.H., Ravelomanantsoa, N.A.F., Tato, C.M., Dussart, P., Heraud, J.M., Brook, C.E., 2022. Full genome *Nobecovirus* sequences from Malagasy fruit bats define a unique evolutionary history for this coronavirus clade. Front. Public Health 10, 786060.
Kopylova, E., Noé, L., Touzet, H., 2012. SortMeRNA: fast and accurate filtering of ribosomal RNAs in metatranscriptomic data. Bioinformatics 28, 3211–3217.
Langmead, B., Salzberg, S.L., 2012. Fast gapped-read alignment with Bowtie 2. Nat. Methods 9, 357–359.
Li, D., Liu, C.M., Luo, R., Sadakane, K., Lam, T.W., 2015. MEGAHIT: an ultra-fast single-node solution for large and complex metagenomics assembly via succinct de Bruijn graph. Bioinformatics 31, 1646–1674.
Li, W., Shi, Z., Yu, M., Ren, W., Smith, C., Epstein, J.H., Wang, H., Crameri, G., Hu, Z., Zhang, H., Zhang, J., McEachern, J., Field, H., Daszak, P., Eaton, B.T., Zhang, S., Wang, L.F., 2005. Bats are natural reservoirs of SARS-like coronaviruses. Science 310, 676–679.
Marsh, G.A., de Jong, C., Barr, J.A., Tachedjian, M., Smith, C., Middleton, D., Yu, M., Todd, S., Foord, A.J., Haring, V., Payne, J., Robinson, R., Broz, I., Crameri, G., Field, H.E., Wang, L.F., 2012. Cedar virus: a novel Henipavirus isolated from Australian bats. PLoS Pathog. 8, e1002836.
Martella, V., Lorusso, E., Banyai, K., Decaro, N., Corrente, M., Elia, G., Cavalli, A., Radogna, A., Costantini, V., Saif, L.J., Lavazza, A., Di Trani, L., Buonavoglia, C., 2008. Identification of a porcine calicivirus related genetically to human sapoviruses. J. Clin. Microbiol. 46, 1907–1913.
Mishra, N., Fagbo, S.F., Alagaili, A.N., Nitido, A., Williams, S.H., Ng, J., Lee, B., Durosinlorun, A., Garcia, J.A., Jain, K., Kapoor, V., Epstein, J.H., Briese, T., Memish, Z.A., Olival, K.J., Lipkin, W.I., 2019. A viral metagenomic survey identifies known and novel mammalian viruses in bats from Saudi Arabia. PLoS One 14, e0214227.
Mombo, I.M., Berthet, N., Bouchier, C., Fair, J.N., Schneider, B.S., Renaud, F., Leroy, E. M., Rougeron, V., 2014. Characterization of a genogroup I sapovirus isolated from chimpanzees in the Republic of Congo. Genome Announc 2, e00680, 14.
Murakami, S., Kitamura, T., Suzuki, J., Sato, R., Aoi, T., Fujii, M., Matsugo, H., Kamiki, H., Ishida, H., Takenaka-Uema, A., Shimojima, M., Horimoto, T., 2020. Detection and characterization of bat sarbecovirus phylogenetically related to SARS-CoV-2, Japan. Emerg. Infect. Dis. 26, 3025–3029.
Nguyen, L.-T., Schmidt, H.A., von Haeseler, A., Minh, B.Q., 2014. IQ-TREE: a fast and effective stochastic algorithm for estimating maximum-likelihood phylogenies. Mol. Biol. Evol. 32, 268–274.
Ning, S., Dai, Z., Zhao, C., Feng, Z., Jin, K., Yang, S., Shen, Q., Wang, X., Sun, R., Zhang, W., 2022. Novel putative pathogenic viruses identified in pangolins by mining metagenomic data. J. Med. Virol. 94, 2500–2509.
Nouri, S., Salem, N., Nigg, J.C., Falk, B.W., 2015. Diverse array of new viral sequences identified in worldwide populations of the Asian citrus psyllid (*Diaphorina citri*) using viral metagenomics. J. Virol. 90, 2434–4245.
Oka, T., Wang, Q., Katayama, K., Saif, L.J., 2015. Comprehensive review of human sapoviruses. Clin. Microbiol. Rev. 28, 32–53.
Philbey, A.W., Kirkland, P.D., Ross, A.D., Davis, R.J., Gleeson, A.B., Love, R.J., Daniels, P.W., Gould, A.R., Hyatt, A.D., 1998. An apparently new virus (family *Paramyxoviridae*) infectious for pigs, humans, and fruit bats. Emerg. Infect. Dis. 4, 269–271.
Plowright, R.K., Foley, P., Field, H.E., Dobson, A.P., Foley, J.E., Eby, P., Daszak, P., 2011. Urban habituation, ecological connectivity and epidemic dampening: the emergence of Hendra virus from flying foxes (*Pteropus* spp.). Proc. Biol. Sci. 278, 3703–3712.
Prada, D., Boyd, V., Baker, M., Jackson, B., O'Dea, M., 2019a. Insights into Australian bat lyssavirus in insectivorous bats of Western Australia. Trav. Med. Infect. Dis. 4, 46.
Prada, D., Boyd, V., Baker, M.L., O'Dea, M., Jackson, B., 2019b. Viral diversity of microbats within the south west botanical province of Western Australia. Viruses 11, 1157.
Rehse-Küpper, B., Casals, J., Rehse, E., Ackermann, R., 1976. Eyach-an arthropod-borne virus related to Colorado tick fever virus in the Federal Republic of Germany. Acta Virol. 20, 339–342.
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 540, 539–543.
Smith, C.S., de Jong, C.E., Meers, J., Henning, J., Wang, L., Field, H.E., 2016. Coronavirus infection and diversity in bats in the Australasian region. EcoHealth 13, 72–82.
Temmam, S., Vongphayloth, K., Baquero, E., Munier, S., Bonomi, M., Regnault, B., Douangboubpha, B., Karami, Y., Chrétien, D., Sanamxay, D., Xayaphet, V., Paphaphanh, P., Lacoste, V., Somlor, S., Lakeomany, K., Phommavanh, N., Pérot, P., Dehan, O., Amara, F., Donati, F., Bigot, T., Nilges, M., Rey, F.A., van der Werf, S., Brey, P.T., Eloit, M., 2022. Bat coronaviruses related to SARS-CoV-2 and infectious for human cells. Nature 604, 330–336.
Thayer, R.M., Power, M.D., Bryant, M.L., Gardner, M.B., Barr, P.J., Luciw, P.A., 1987. Sequence relationships of type D retroviruses which cause simian acquired immunodeficiency syndrome. Virology 157, 317–329.
Timmiss, L.A., Martin, J.M., Murray, N.J., Welbergen, J.A., Westcott, D., McKeown, A., Kingsford, R.T., 2021. Threatened but not conserved: flying-fox roosting and foraging habitat in Australia. Aust. J. Zool. 68, 226–233.
Tse, H., Chan, W.M., Li, K.S., Lau, S.K., Woo, P.C., Yuen, K.Y., 2012. Discovery and genomic characterization of a novel bat sapovirus with unusual genomic features and phylogenetic position. PLoS One 7, e34987.
Turnbull, M.G., Douville, R.N., 2018. Related endogenous retrovirus-K elements harbor distinct protease active site motifs. Front. Microbiol. 9, 1577.
Van Brussel, K., Holmes, E.C., 2022. Zoonotic disease and virome diversity in bats. Curr. Opin. Virol. 52, 192–202.
Vanmechelen, B., Merino, M., Vergote, V., Laenen, L., Thijssen, M., Martí-Carreras, J., Claerebout, E., Maes, P., 2021. Exploration of the *Ixodes ricinus* virosphere unveils an extensive virus diversity including novel coltiviruses and other reoviruses. Virus Evol 7, veab066.
Wacharapluesadee, S., Tan, C.W., Maneeorn, P., Duengkae, P., Zhu, F., Joyjinda, Y., Kaewpom, T., Chia, W.N., Ampoot, W., Lim, B.L., Worachotsueptrakun, K., Chen, V. C., Sirichan, N., Ruchisrisarod, C., Rodpan, A., Noradechanon, K., Phaichana, T., Jantarat, N., Thongnumchaima, B., Tu, C., Crameri, G., Stokes, M.M., Hemachudha, T., Wang, L.F., 2021. Evidence for SARS-CoV-2 related coronaviruses circulating in bats and pangolins in Southeast Asia. Nat. Commun. 12, 972.
Wang, J., Anderson, D.E., Halpin, K., Hong, X., Chen, H., Walker, S., Valdeter, S., van der Heide, B., Neave, M.J., Bingham, J., O'Brien, D., Eagles, D., Wang, L.F., Williams, D. T., 2021. A new Hendra virus genotype found in Australian flying foxes. Virol. J. 18, 197.
Williams, N.S.G., McDonnell, M.J., Phelan, G.K., Keim, L.D., Van Der Ree, R., 2006. Range expansion due to urbanization: increased food resources attract Grey-headed Flying-foxes (*Pteropus poliocephalus*) to Melbourne. Austral. Ecol. 31, 190–198.
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, 609–620.
Xu, Z., Feng, Y., Chen, X., Shi, M., Fu, S., Yang, W., Liu, W.J., Gao, G.F., Liang, G., 2022. Virome of bat-infesting arthropods: highly divergent viruses in different vectors. J. Virol. 96, e0146421.
Yang, Z., He, B., Lu, Z., Mi, S., Jiang, J., Liu, Z., Tu, C., Gong, W., 2021. Mammalian birnaviruses identified in pigs infected by classical swine fever virus. Virus Evol 7, veab084.
Yinda, C.K., Conceição-Neto, N., Zeller, M., Heylen, E., Maes, P., Ghogomu, S.M., Van Ranst, M., Matthijnssens, J., 2017. Novel highly divergent sapoviruses detected by metagenomics analysis in straw-colored fruit bats in Cameroon. Emerg. Microb. Infect. 6, e38.
Yob, J.M., Field, H., Rashdi, A.M., Morrissy, C., van der Heide, B., Rota, P., bin Adzhar, A., White, J., Daniels, P., Jamaluddin, A., Ksiazek, T., 2001. Nipah virus infection in bats (order Chiroptera) in peninsular Malaysia. Emerg. Infect. Dis. 7, 439–441.
Zhou, H., Ji, J., Chen, X., Bi, Y., Li, J., Wang, Q., Hu, T., Song, H., Zhao, R., Chen, Y., Cui, M., Zhang, Y., Hughes, A.C., Holmes, E.C., Shi, W., 2021. Identification of novel bat coronaviruses sheds light on the evolutionary origins of SARS-CoV-2 and related viruses. Cell 184, 4380–4391 e14.
Zhou, P., Yang, X.L., Wang, X.G., Hu, B., Zhang, L., Zhang, W., Si, H.R., Zhu, Y., Li, B., Huang, C.L., Chen, H.D., Chen, J., Luo, Y., Guo, H., Jiang, R.D., Liu, M.Q., Chen, Y., Shen, X.R., Wang, X., Zheng, X.S., Zhao, K., Chen, Q.J., Deng, F., Liu, L.L., Yan, B., Zhan, F.X., Wang, Y.Y., Xiao, G.F., Shi, Z.L., 2020. A pneumonia outbreak associated with a new coronavirus of probable bat origin. Nature 579, 270–273.
Zhu, W., Lomsadze, A., Borodovsky, M., 2010. *Ab initio* gene identification in metagenomic sequences. Nucleic Acids Res. 38, e132.
Zhu, W., Yang, J., Lu, S., Jin, D., Pu, J., Wu, S., Luo, X.L., Liu, L., Li, Z., Xu, J., 2022. RNA virus diversity in birds and small mammals from Qinghai-Tibet plateau of China. Front. Microbiol. 13, 780651.