# Revealing the Evolution of Viruses using Phylogenetics, Metagenomics and Structural Prediction

Jonathon Charles Owen Mifsud

2025

University of Sydney
Faculty of Medicine and Health

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

This research reported in this thesis was supported by the award of a Research Training Program scholarship and fee offset to the PhD Candidate.

# Authorship Attribution Statement

Chapter 2 is published in *npj Viruses* as: **Mifsud, J.C.O**., Hall, J., Van Brussel, K., Rose, K., Parry, R.H., Holmes, E.C., Harvey, E., 2024. "A novel papillomavirus in a New Zealand fur seal (Arctocephalus forsteri) with oral lesions" *npj Viruses* 2, 10.
*I conceptualised the study, performed the laboratory work, conducted the sequence and phylogenetic analyses, and wrote the manuscript in collaboration with input from the co-authors.*

Chapter 3 is being preprepared for publication as: **Mifsud, J.C.O**., Harvey, E., Van Brussel, K., Annabelle Olsson, A., Hall, J., Alting B.F., Sadiq, S., Holmes, E.C. "Diverse pathogens in Australian canines: limited geographic structure and the first detection of rotavirus A in dingoes".
*I conceptualised the study, performed the data analysis and part of the laboratory work, and wrote the manuscript with input from the co-authors.*

Chapter 4 is published in *Journal of Virology* as: **Mifsud**, **J. C. O.**, Gallagher, R. V., Holmes, E. C., Geoghegan, J. L., 2022. "Transcriptome mining expands knowledge of RNA viruses across the plant kingdom" Journal of Virology, 96, e00260-22.
*I conceptualised the study, performed the data analysis, and wrote the manuscript with input from the co-authors.*

Chapter 5 is published in *Virus Evolution* as: **Mifsud, J.C.O**., Costa, V.A., Petrone, M.E., Marzinelli, E.M., Holmes, E.C., Harvey, E., 2023. "Transcriptome mining extends the host range of the Flaviviridae to non-bilaterians." *Virus Evolution* 9, veac124.
*I conceptualised the study, conducted the sequence and phylogenetic analyses, and wrote the manuscript in collaboration with input from the co-authors.*

Chapter 6 is published in *Nature* as: **Mifsud, J. C.O**., Lytras, S., Oliver, M. R., Toon, K., Costa, V. A., Holmes, E. C., Grove, J., 2024. "Mapping glycoprotein structure reveals Flaviviridae evolutionary history" *Nature*, 633, 695–703.
*I conceptualised the study, compiled and curated the sequence data, conducted the sequence-based benchmarking analysis and phylogenetic analyses, and wrote the manuscript in collaboration with the co-authors.*

Appendix I is published in *Virus Evolution* as: Costa, V. A., **Mifsud**, **J. C. O.**, Gilligan, D., Williamson, J. E., Holmes, E. C., Geoghegan, J. L., 2021. "Metagenomic sequencing reveals a lack of virus exchange between native and invasive freshwater fish across the Murray–Darling Basin, Australia." Virus Evolution, 7, veab034.
*I performed parts of the data analysis and helped write the manuscript in collaboration with the other authors.*

Appendix II is published in *Virus Evolution* as: Miller, A. K., **Mifsud**, **J. C. O.**, Costa, V. A., Grimwood, R. M., Kitson, J., Baker, C., Brosnahan C. L., Pande, A., Holmes, E. C., Gemmell, N. J., Geoghegan, J. L., 2021. "Slippery when wet: cross-species transmission of divergent coronaviruses in bony and jawless fish and the evolutionary history of the *Coronaviridae*." *Virus Evolution*, 7, veab050.
*I conceptualised the study in collaboration with the other authors, performed parts of the data collection and analysis, and helped write the manuscript in collaboration with the other authors.*

Appendix III is published in *Virus Evolution* as: Costa, V. A., Bellwood, D. R., **Mifsud**, **J. C.O**, Van Brussel, K., Geoghegan, J. L., Holmes, E. C., Harvey, E., 2023. "Limited cross-species virus transmission in a spatially restricted coral reef fish community." *Virus Evolution*, 9, vead011.
*I performed parts of the laboratory work and data analysis, and provided input and editing towards writing of the manuscript.*

Appendix IV is published in *Proceedings of the National Academy of Sciences* as: Petrone, M. E., Parry, R., **Mifsud**, **J. C. O.**, Van Brussel, K., Voorhees, I. E. H., Richards, Z. T., Holmes, E. C., 2023. "Evidence for an ancient aquatic origin of the RNA viral order *Articulavirales*." *Proceedings of the National Academy of Sciences*, 120, e2310529120.
*I performed parts of the bioinformatic work and provided input and editing towards writing of the manuscript.*

Appendix V is published in *Virus Evolution* as: Harvey, E., **Mifsud, J. C.O.**, Holmes, E. C., Mahar, J. E., 2023. "Divergent hepaciviruses, delta-like viruses, and a chu-like virus in Australian marsupial carnivores (dasyurids)." *Virus Evolution*, 9, vead061.
*I performed parts of the bioinformatic work and provided input and editing towards writing of the manuscript.*

Appendix VI is under review at *FEMS Microbiology Ecology* as: Costa, V. A., Bellwood, D. R., **Mifsud**, **J. C.O.**, Geoghegan, J. L., Harvey, E., Holmes, E. C., 2024. “Limited transmission of microbial species among coral reef fishes from the Great Barrier Reef, Australia.” *bioRxiv*, 2024.2002. 2024.581894.
*I performed parts of the bioinformatic work and provided input and editing towards writing of the manuscript.*

Appendix VII is published in *Proceedings of the National Academy of Sciences* as: Petrone, M. E., Grove, J., Mélade, J., **Mifsud, J. C. O.**, Parry, R. H., Marzinelli, E. M., Holmes, E. C., 2024. “A ~40-kb flavi-like virus does not encode a known error-correcting mechanism.” *Proceedings of the National Academy of Sciences*, 121, e2403805121.
I performed parts of the bioinformatic work and provided input and editing towards writing of the manuscript.

Appendix VIII is published in *Virology* as: Sadiq, S., Harvey, E., **Mifsud, J. C.O.**, Minasny, B., McBratney, A. B., Pozza, L. E., Holmes, E. C., 2024. “Australian terrestrial environments harbour extensive RNA virus diversity.” *Virology,* 593, 110007.
*I performed parts of the field work and provided input and editing towards writing of the manuscript.*

Appendix IX is published in *Current Biology* as: Costa, V. A., Ronco, F., **Mifsud, J. C.O.**, Harvey, E., Salzburger, W., Holmes, E. C., 2024. “Host adaptive radiation is associated with rapid virus diversification and cross-species transmission in African cichlid fishes.” *Current Biology*, 34, 1247-1257. e1243.
*I performed parts of the data preparation and phylogenetic analyses and helped write the manuscript in collaboration with the other authors*

Appendix X is published in *Frontiers in Virology* as: Charon, J., Olendraite, I., Forgia, M., Chong, L. C., Hillary, L. S., Roux, S., Kupczok, A., Debat, H., Sakaguchi, S., Tahzima, R., Nakagawa, S., Babaian, A., Abroi, A., Bejerman, N., Ben Mansour, K., Brown, K., Butkovic, A., Cervera, A., Charriat, F., Chen, G., Chiba, Y., De Coninck, L., Demina, T., Dominguez-Huerta, G., Dubrulle, J., Gutierrez, S., Harvey, E., Jayaraj Mallika, F. R., Karapliafis, D., Lim, S. J., Kasibhatla, S. M., **Mifsud, J. C. O.**, Nishimura, Y., Ortiz-Baez, A. S., Raco, M., Rivero, R., Sadiq, S., Saghaei, S., San, J. E., Shaikh, H. M., Sieradzki, E. T., Sullivan, M. B., Sun, Y.,

Wille, M., Wolf, Y. I., Zrelovs, N., Neri, U., 2024. “Consensus statement from the first RdRp Summit: advancing RNA virus discovery at scale across communities.” *Frontiers in Virology*, 4, 1371958.
*I wrote parts of the manuscript in collaboration with the other authors.*

Appendix XI is under review at *Journal of Virology* as: Mahar J., **Mifsud J. C. O**, Van Brussel K., Lachenauer A. E., Harvey E., Turnbull O. M. H, Bonat S., Newsome T. M., Olsson A., Chiu-Werner A., Jones M. E., Holmes E. C., Maina S. 2024. “Detection of exotic biosecurity threat ribgrass mosaic virus and novel tobamoviruses through metatranscriptomic sequencing of animal gut content”. *bioRxiv* doi:10.1101/2024.12.10.627875:2024.12.10.627875.
*I performed RNA extractions for a subset of the samples, bioinformatic work, and provided input and editing towards writing of the manuscript.*

Appendix XII is under review at *Current Biology* as: Petrone, M. E., Grove, J., Parry, R., Van Brussel, K., **Mifsud, J. C. O.**, Dindar, Z., Mei, S. Q., Shi, M., Turnbull, O. M., Marzinelli, E. M. and Holmes, E. C., 2024. “Tunicate metatranscriptomes reveal ancient virus-host co-divergence and inter-order recombination in the evolutionary history of disease-causing viruses”. *bioRxiv*, 2024-12.
*I performed parts of the bioinformatic work and provided input and editing towards writing 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.

Jonathon Charles Owen Mifsud,
28/12/2024

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

Professor Edward C. Holmes

28/12/2024

## 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.

Part of the data analysis underlying Chapter 5 was conducted during a Master of Research degree prior to my PhD. However, the majority of this chapter, including writing of the final manuscript, was conducted during my PhD.

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.

Jonathon Charles Owen Mifsud
28/12/2024

# Acknowledgements

I dedicate this thesis to my family, Mum, Dad, Pat, Lac, Callie and Squeaker who have supported me from day one and tolerated me ruining their dinners with talk of parasites, sampling and poo. Mum and Dad, thank you for your love and support. I appreciate the freedom you gave me to pursue my passions, even when I wasn't sure of myself. Nanna and Nannu, thank you for all your love, my apologises that I am not married yet. Grandma, I know it is slightly different to Hawkesbury riverboats, but I know you'd be proud, thank you for everything.

I would also like to acknowledge the friends who have supported me throughout this journey. Connor, thank you for your endless support of 'Science' and for all the good times, you're a great friend! A huge thanks to the men of Div 11 Abbotsford for the best final season one could ask for. This year has been undoubtedly the toughest year of my thesis. A large part of getting through it has been thanks to Hio—thank you for your love, patience, support, and company during the long nights and weekends.

Eddie Holmes—Thank you for being the best supervisor one could ask for. Learning from you has been a real pleasure and the way you approach science is truly inspiring. You were so welcoming from the beginning, even when I showed up unannounced a couple months before I was meant to start. I am grateful for the trust and freedom you gave me to purse the projects that interested me as well as your encouragement when these didn't work out as expect. Even during the hectic covid periods I have always felt your support. This thesis wouldn't be what it is without your guidance and the occasional push. I am forever grateful for all the opportunities you have given me.

Erin Harvey—thank you for taking me on as a student at the beginning of my PhD. I owe you for the countless edits of just about everything. Thank you for all your support and continuous attempts to get me out of my shell. I won't forget the Western Australia fieldtrip—sorry again on behalf of that alpaca —and the conference adventures. Finally, thank you for introducing me to Hunter who is curled up next to me as I write this. I am truly grateful for all the time, effort, and support you have so generously given me.

I would like to thank my Master's supervisors, Rachael Gallagher and Jemma Geoghegan, as well as my high school science teacher, Nicholas Briffa. Without their support and encouragement, I likely wouldn't be on this path.

I am appreciative of the Holmies lab which welcomed me so warmly. They are an incredible group of researchers and genuinely wonderful people all around. Erin, Jackie, Justine, Julien, Kosuke, Mary, Michelle, Wei-Shan, Bec, Bethan, Callum, Kate, Krispin, Lauren, Liv, Sabrina, Susana, and Vince—thank you all. I will hold onto the Holmies shirt and tote bag with pride forever (thanks Sabrina!). I am indebted to all of you for your support, advice and friendship over the years. I wish you all the best and hope we have the chance to work together again someday!

Vince, Mary, Erin, Sabrina, Jackie—thank you for trusting me and allowing me to contribute to your wonderful projects.

I have had the pleasure of working with amazing collaborators over the past three and a half years. Your contributions have undoubtedly improved this thesis. Joe Grove, thank you for taking me under your wing and guiding me so patiently through the world of glycoproteins. Working with you has been a joy. Your approach and love of science are truly inspiring. I really appreciate the hospitality you showed me in Glasgow. Jane Hall and Karrie Rose at Taronga, thank you for all your advice and for providing access to fantastic samples across numerous projects—best of luck with your thesis, Jane! Brendan Alting, our Myall Lakes trip, despite the terrible weather, was an unforgettable experience. I must also thank my Glasgow tour guide, Spyros Lytras and to Annabelle Olsson and Rhys Parry for their contributions. Thank you as well to the many other collaborators, farm owners and university staff whose contributions and acts of kindness have helped along the way, even if I cannot name them all individually here.

Jon Mifsud

# Table of Contents


***Authorship Attribution Statement*** ..... ***ii***

***Statement of Originality*** ..... ***vii***

***Acknowledgements*** ..... ***viii***

***List of Figures*** ..... ***xv***

***List of Tables*** ..... ***xvii***

***Abstract*** ..... ***1***

***Chapter 1: General Introduction*** ..... ***2***

1.1 Impact of viruses in history ..... 2

1.2 Identification of the first virus and the beginnings of virus discovery ..... 2

1.3 Virus culture ..... 3

1.4 Molecular identification of viruses ..... 4

1.5 The beginnings of metagenomic sequencing ..... 6

1.6 Metatranscriptomic virus discovery workflow ..... 9

1.7 Transcriptome mining ..... 13

1.8 Changing understanding of viruses from metagenomics ..... 15

1.8.1 Metagenomics, virus diversity and taxonomy ..... 16

1.8.2 Metagenomics and disease ..... 18

1.8.3 Metagenomics, agricultural and wildlife viruses ..... 19

1.8.4 Metagenomics and companion animals ..... 20

1.8.5 Virus-host evolutionary relationships ..... 21

1.9 Sequence saturation and distant evolutionary relationships ..... 24

1.10 Protein evolution and structure ..... 26

1.11 Protein structure prediction ..... 27

1.11.1 AlphaFold ..... 27

1.11.2 Protein language models ..... 29

1.12 Protein structure comparison ..... 29

1.13 Protein structure phylogenetics ..... 31

1.14 Thesis rationale ..... 34
1.15 References ..... 38
***Chapter 2. A novel papillomavirus in a New Zealand fur seal (Arctocephalus forsteri) with oral lesions ..... 50***
2.1 Abstract ..... 51
2.2 Introduction ..... 51
2.3 Methods ..... 53
2.3.1 Sample collection and processing ..... 53
2.3.2 RNA extraction and metatranscriptomic sequencing ..... 53
2.3.3 Identification of novel virus sequences ..... 54
2.3.4 A. forsteri papillomavirus 1 RT-PCR ..... 54
2.3.5 Genome analysis and annotation ..... 54
2.3.6 Assessment of sequencing library composition ..... 55
2.3.7 Phylogenetic analysis ..... 55
2.4 Results ..... 56
2.4.1 Overview of metatranscriptomic data ..... 56
2.4.2 Identification of a novel papillomavirus in a seal with oral papilloma ..... 56
2.4.3 Papillomavirus genome construction and confirmation ..... 58
2.4.4 Evolutionary relationship and genomic properties of AforPV1 ..... 59
2.4.5 Gene expression of AforPV1 ..... 62
2.4.6 SL16 library composition ..... 63
2.5 Discussion ..... 63
2.6 Data availability ..... 66
2.7 Acknowledgements ..... 66
2.8 References ..... 67
2.9 Supplementary information ..... 71
***Chapter 3. Diverse Viral Pathogens in Australian Canines: Limited Geographic Structure and the First Detection of an RNA Virus in Dingoes ..... 75***
3.1 Abstract ..... 76

3.2 Introduction ..... 77

3.3. Methods ..... 79

3.3.1 Sample collection and processing ..... 79

3.3.2 Sample collection and processing ..... 79

3.3.3 RNA extraction and metatranscriptomic sequencing ..... 80

3.3.4 Identification of novel virus sequences ..... 80

3.3.5 Genome extension, annotation and abundance ..... 81

3.3.6 Assessment of sequencing library composition ..... 82

3.3.7 Phylogenetic analysis ..... 82

3.3.8 Statistical comparisons of virus count, diversity, and abundance ..... 83

3.3.9 Statistical comparisons of genetic and geographical distances ..... 83

3.4 Results ..... 83

3.4.1 Identification of potential dog and human pathogens in Australia ..... 84

3.4.2 Co-occurrence of viruses in faecal libraries ..... 93

3.4.3 Genetic structure and evolution of Australian sequences ..... 93

3.5 Discussion ..... 99

3.6 Data availability ..... 103

3.7 Acknowledgements ..... 103

3.8 References ..... 104

3.9 Supplementary Information ..... 110

***Chapter 4. Transcriptome Mining Expands Knowledge of RNA Viruses Across the Plant Kingdom ..... 112***

4.1 Abstract ..... 113

4.2 Introduction ..... 114

4.3 Methods ..... 115

4.3.1 Transcriptome data generation ..... 115

4.3.2 Surveying for viruses in the 1KP ..... 116

4.3.3 Virus filtering and abundance calculations ..... 117

4.3.4 Genome extension and annotation ..... 117

4.3.5 Detection of endogenous virus elements ..... 118

4.3.6 Assessing library contamination by eukaryotes, bacteria, and protozoa ..... 119

4.3.7 Phylogenetic analysis of plant viruses ..... 119

4.3.8 Assigning plant host clades ..... 120

4.4 Results ..... 120
4.4.1 Diversity and abundance of plant viruses ..... 121
4.4.2 Presence of contaminants in sequencing libraries ..... 124
4.4.3 Phylogenetic analysis of identified viruses ..... 127
4.4.4 Long-term virus-host evolutionary relationships ..... 154
4.5 Discussion ..... 156
4.6 Data availability ..... 161
4.7 Acknowledgements ..... 161
4.8 References ..... 162
4.9 Supplementary Information ..... 170
***Chapter 5. Transcriptome mining extends the host range of the Flaviviridae to non-bilaterians ..... 187***
5.1 Abstract ..... 188
5.2 Introduction ..... 189
5.3 Methods ..... 191
5.3.1 Screening of SRAs for flavivirid-like sequences ..... 191
5.3.2 Tunicate collection, RNA extraction, and metagenomic next-generation sequencing ..... 191
5.3.3 Identification of novel flavivirid genomes ..... 192
5.3.4 Genome extension and annotation ..... 192
5.3.5 Detection of endogenous virus elements ..... 193
5.3.6 Assessment of library composition ..... 193
5.3.7 Phylogenetic analysis ..... 194
5.3.8 Assessment of cross-species virus transmission ..... 194
5.4 Results ..... 195
5.4.1 Genus Flavivirus ..... 203
5.4.2 Genus *Pestivirus* ..... 205
5.4.3 Genus *Hepacivirus* ..... 208
5.4.4 An unclassified flavivirid-like virus ..... 210
5.4.5 Genetic composition of sequencing libraries ..... 210
5.4.6 Long-term virus-host evolutionary relationships ..... 212
5.5 Discussion ..... 214
5.6 Data availability ..... 217
5.7 Acknowledgements ..... 217

5.8 References ..... 219
5.9 Supplementary Figures ..... 226
***Chapter 6. Mapping glycoprotein structure reveals Flaviviridae evolutionary history ..... 232***
6.1 Abstract ..... 233
6.2 Introduction ..... 233
6.3 Methods ..... 235
6.3.1 Compilation of *Flaviviridae* sequence set ..... 235
6.3.2 Structure prediction and homology search ..... 236
6.3.3 Phylogenetic analysis ..... 238
6.4 Results ..... 243
6.4.1 *Flaviviridae* contains three major clades ..... 243
6.4.2 Building a *Flaviviridae* protein foldome ..... 245
6.4.3 *Flaviviridae* glycoprotein discovery ..... 246
6.4.4 Glycoproteins follow ecological niche ..... 249
6.4.5 LGFs harbour novel and acquired proteins ..... 250
6.4.6 Evolutionary history of the *Flaviviridae* ..... 253
6.5 Discussion ..... 257
6.6 Data availability ..... 258
6.7 Acknowledgements ..... 258
6.8 References ..... 259
6.9 Supplementary Information ..... 267
***Chapter 7. General discussion ..... 301***

# List of Figures

*Figure 1.1 Metagenomic era of virus discovery.* ..... 6
*Figure 1.2 Virus discovery protocol* ..... 8
*Figure 1.3 Expansion of the Sequence Read Archive* ..... 14
*Figure 1.4 Increasing rate of metagenomic virus studies and novel species submissions over time* ..... 16
*Figure 1.5 Coevolutionary analysis of the Coronaviridae* ..... 23
*Figure 1.6. Effect of saturation on estimating divergence and the increase in levels of conservation when comparing nucleotide, amino acid, and measures of structure* ..... 25
*Figure 1.7. Workflow of evolutionary information extraction in AlphaFold2* ..... 28
*Figure 1.8 Overview of FoldSeek's classification of tertiary interactions into a 20-state 3Di alphabet* ..... 31
*Figure 1.9. Schematic illustrating the phylogenetic signal "sweet spot" of nucleotides, amino acids, and structural data across varying levels of sequence divergence.* ..... 33
*Figure 2.1 Papillary proliferations, cellular atypia, and the presence of koilocytes in oral lesions.* ..... 57
*Figure 2.2 Genome organization of Arctocephalus forsteri papillomavirus 1 (AforPV1).* ..... 59
*Figure 2.3 Placement of Arctocephalus forsteri papillomavirus 1 (AforPV1) within the taupapillomaviruses* ..... 61
*Figure 3.1 Map showing the composition and number of the dog and dingo samples across mainland Australia* ..... 84
*Figure 3.2. Boxplot showing the differences in virus number, percentage of total library reads and virus diversity between sample type (faeces, nasal swab, and rectal swab), host sex (female and male) and host age (adult, young adult, and puppy)* ..... 85
*Figure 3.3 Maximum likelihood phylogenetic trees of four viral pathogens* ..... 87
*Figure 3.4 Maximum likelihood phylogenetic trees of canine rotaviruses* ..... 92
*Figure 3.5 Maximum likelihood phylogenies of kobuviruses, vesivirus, dicipivirus and mamastroviruses* ..... 95
*Figure 3.6 Maximum likelihood phylogenies of sapoviruses, noroviruses, picornaviruses and copiparvoviruses* ..... 97
*Figure 3.7 Virus co-occurrence heatmap matrix* ..... 98
*Figure 4.1 Virus abundance and diversity across the plant kingdom* ..... 123
*Figure 4.2 Taxonomic assignments of reads in select One Thousand Plant Transcriptomes Initiative (1KP) libraries* ..... 126

*Figure 4.3 Phylogenetic relationships and genome organization of the beny-like viruses...129*
*Figure 4.4 Phylogenetic relationships and genome organization of the viruses within the order Tymovirales .....132*
*Figure 4.5 Phylogenetic relationships, genome organization and movement protein motids of the viruses identified within the virus family Secoviridae.....135*
*Figure 4.6 Phylogenetic relationships and genome organization of the viruses within the virus families Narnaviridae and Mitoviridae.....138*
*Figure 4.7 Phylogenetic relationships and genome organization of the endorna-like and unclassified (+)ssRNA virus.....141*
*Figure 4.8 Phylogenetic relationships and genome organization of the phenui-like and bunya-like viruses .....143*
*Figure 4.9 Phylogenetic relationships and genome organization of select negative-strand RNA virus families.....146*
*Figure 4.10 Phylogenetic relationships of the viruses within the order Durnavirales.....149*
*Figure 4.11 Phylogenetic relationships of the viruses within the order Ghabrivirales.....152*
*Figure 4.12 Tanglegram and reconciliation analysis of select plant virus .....155*
*Figure 5.1 Phylogeny of the Flaviviridae. .....196*
*Figure 5.2 Phylogenetic relationships of the flavi-like viruses identified in this study.....204*
*Figure 5.3 Phylogenetic relationships of the pesti-like viruses identified in this study .....207*
*Figure 5.4 Phylogenetic relationships of the hepaciviruses viruses identified in this study. 209*
*Figure 5.5 Taxonomic assignments of contigs in sequencing libraries.....211*
*Figure 5.6 Tanglegram of rooted phylogenetic trees for representative virus groups and their hosts .....213*
*Figure 6.1 Generation of a protein foldome for the Flaviviridae.....244*
*Figure 6.2 Discovery of glycoproteins across the Flaviviridae .....248*
*Figure 6.3 Novel and acquired proteins in a large genome flavivirus. .....251*
*Figure 6.4 Structurally informed phylogenetics. .....254*
*Figure 6.5 Proposed evolutionary history of the Flaviviridae .....256*

# List of Tables

Table 1. Comparison of the International Committee on Taxonomy of Viruses (ICTV) description of the *Flaviviridae* compared to the known (unclassified) diversity .....18

# Abstract

Metagenomic sequencing, particularly total RNA sequencing (i.e., meta-transcriptomics), has emerged as a powerful tool for virus discovery. Although the known virosphere has expanded dramatically in recent years, the viruses associated with much of the tree of life remain largely undiscovered, limiting our ability to infer the evolutionary and ecological processes that shape viral diversity. The central aim of my thesis was to explore viral diversity in under-sampled species critical to answering fundamental questions about virus evolution and ecology. To achieve this, I combined metatranscriptomic sequence data with state-of-the-art phylogenetic and structural analysis. I determined a novel papillomavirus, a likely aetiological agent of oral papilloma-like lesions in a fur seal, and characterised the virome of Australian dogs, including the first RNA virus in dingoes. I also conducted broad virus discovery surveys of publicly available transcriptome data sets. By analysing species from across the breadth of the plant kingdom I identified 104 potentially novel viruses, including sequences divergent enough to represent novel virus families, as well as the first virus genomes in bryophytes and lycophytes. Through an analysis of flaviviruses across the Metazoa I discovered 32 novel flavivirus sequences. These included an orthoflavi-like virus in a coral, marking the earliest diverging metazoan host, and a substantial expansion of the host range of pestiviruses. I next used the expanded *Flaviviridae* sequence data set to investigate glycoprotein distribution within the family. Through structural homology searches I uncovered a complex evolutionary history for the *Flaviviridae*, characterised by bacterial gene capture and inter-genus recombination. In sum, the research in this thesis expands the known virosphere and highlights how novel and highly divergent viruses can be key to uncovering the fundamental mechanisms, patterns, and processes of virus evolution.

# Chapter 1: General Introduction

## 1.1 Impact of viruses in history

Viruses have had enormous health and economic impacts on humans, other animals, and plants. In the case of humans this is evident from the ongoing COVID-19 global pandemic, in which more than 770 million confirmed cases and 7 million deaths have been reported as of November 10, 2024 (https://covid19.who.int). At the same time, outbreaks of such diseases as Crimean-Congo haemorrhagic fever, dengue fever, Middle East respiratory syndrome (MERS), mpox, and avian influenza A (H5N1), are occurring worldwide (https://www.emro.who.int/pandemic-epidemic-diseases/outbreaks/index.html) (Neumann and Kawaoka 2024). However, the phenomenon of disease emergence is not recent, and evidence of the impact caused by viral infections spans millennia. For example, descriptions of a smallpox-like disease in Chinese medical writings date back to the 4th century CE (Geddes 2006). Although the causative pathogen can sometimes be difficult to determine from historical records alone, metagenomic sequencing of human archaeological remains (i.e., “ancient DNA”) from over 31,000 to 150 years ago uncovered variola virus sequences—the causative agent of smallpox—across Europe, indicating that this virus has spanned this region since the late 6th century (Mühlemann et al. 2020). The effects of viruses are not confined to humans; they also impact many agricultural species and wildlife. For example, foot-and-mouth disease (FMD) is a highly contagious disease of cloven-hoofed animals such as in cattle, pigs, sheep, and goats, disrupting meat and milk production, and the causative virus (FMDV) was one of the first isolated (see below). A possible incursion of this virus into Australia is estimated to have a direct economic impact of at least $80 billion over a 10 year period (Buetre 2013; Hone et al. 2022). Similarly, in plants, Tomato yellow leaf curl virus (TYLCV) has spread globally, and infection causes up to 100% yield loss, resulting in enormous economic damage worldwide (Prasad et al. 2020). Indeed, in regions of heavy TYLCV disease pressure tomato cultivation is restricted to periods of the year where TYLCV incidence is low (Moriones and Navas-Castillo 2000).

## 1.2 Identification of the first virus and the beginnings of virus discovery

Investigation of the infectious mottling and stunting of tobacco plants (now known as Tobacco mosaic disease), by Dmitri Ivanovsky and Adolf Mayer, led to the discovery of viruses as discreet disease-causing entities distinct from other pathogens. Mayer and Ivanovsky independently passed extracts from infected plants through filters. Ivanovsky used the newly developed Pasteur–Chamberland filter to trap bacteria and found that some

extracts retained their infectivity. This led Mayer and Ivanovsky to conclude that the agent was likely bacterial, or in Ivanovsky's case, a bacterial toxin. Martinus Beijerinck later replicated this work, but unlike his predecessors, he proposed a different hypothesis for the nature of the disease-causing agent. Beijerinck found that this agent could propagate in plants but not in broth, unlike bacteria, suggesting a reliance on the host. Beijerinck termed this agent "contagium vivum fluidum", Latin for a 'living soluble contagion'. The ability of these agents to pass through filters also led to their being called "filterable viruses". Similar techniques were later used by Friedrich Loeffler and Paul Frosch in 1898 to identify FMDV (Brown 2003; Löffler and Frosch 1898), marking the discovery of the first known animal virus. It was not until 25 years later, with the advent of plaque assays (d'Herelle 1961) and viral particle electron microscopy, that viruses were first visualised (Kausche et al. 1939). These methods confirmed that viruses were particulate, not fluid as Beijerinck had initially hypothesised. This led to the term "filterable" being dropped, leaving the name "virus".

### 1.3 Virus culture

In the early 1900s, the filtration of diseased animal tissues led to the discovery of numerous viral agents. During this era, virus discovery in animals was laborious and required live animals, which led to the development of laboratory models and, later, the use of embryonated chicken eggs. The modern era of molecular virology arguably began with the advent of single-cell culturing and the discovery that viruses such as variola virus (the causative agent of smallpox) and poliovirus could be replicated in cultured cells (Enders et al. 1949). Since these cells could be grown in vitro in flasks, they were less expensive than live animals and easier to examine microscopically. This technique was not only essential for virus discovery but also made it more convenient to manipulate and grow viruses to achieve high titres. Numerous attempts to produce poliovirus between the 1930s and 1950s largely failed due to challenges such as insufficient animal models and incomplete viral inactivation (Hovern 2018). The development of a method to grow poliovirus in non-nervous human tissue cultures by John Enders, Thomas Weller, and Frederick Robbins for which they were jointly awarded the 1954 Nobel Prize in Physiology or Medicine, enabled more controlled and scalable vaccine development. Using these methods, two key poliovirus vaccines, the Salk inactivated polio vaccine (IPV) and the Sabin oral polio vaccine (OPV). Vaccination programs using IPV and OPV led to the eradication of polio in much of the world (Nathanson and Kew 2010). It is estimated that these vaccination programs prevented 29 million cases of paralytic polio between 1960 and 2021 (Badizadegan et al.

2022). Such large-scale vaccination programs would not have been possible without the industrial-scale vaccine production enabled by cell culture techniques.

The controlled conditions in which viruses could be grown and propagated enabled by cell cultural were vital for techniques such as electron microscopy, which led to the discovery and visualisation of hepatitis B virus amongst others (Dane et al. 1970). The main downside of this cell culture and electron microscopy-based approach to virus discovery was that the vast majority of viruses are not easily cultivatable. This is due to factors such as low virus abundance, the lack of a suitable tissue culture system, or incomplete knowledge of their true hosts.

Through the use of virus culture and electron microscopy, researchers began to uncover the remarkable diversity of viruses in terms of genome composition, replication mechanisms, and genetic material. Unlike the other domains of life, which primarily use double-stranded DNA as their genetic material, viruses comprise either RNA or DNA, which can be present as single- or double-stranded, monocistronic or polycistronic, and linear or circular in topology. Based upon these features and the genome expression mechanisms, viruses have been classified into seven groups, known as the seven “Baltimore classes”, named after their proposer, David Baltimore (Baltimore 1971), and which generally reflect the broader phylogenetic relationships of viruses. The Baltimore classification includes seven groups:

Class I: Double-stranded DNA genome ((+/–) dsDNA)
Class II: Single-stranded DNA genome ((+) ssDNA)
Class III: Double-stranded RNA genome ((+/–) dsRNA)
Class IV: Positive-sense single-stranded RNA genome ((+) ssRNA)
Class V: Negative-sense single-stranded RNA genome ((–) ssRNA)
Class VI: Reverse-transcribing single-stranded RNA genome ((+) ssRNA-RT)
Class VII: Double-stranded DNA genome with an RNA intermediate in its replication cycle ((+/–) dsDNA-RT), which was later added (Baltimore 1971; Koonin et al. 2021).

### 1.4 Molecular identification of viruses

The field of virus discovery was revolutionised with the development of methods for the analysis of nucleic acid, including chain-terminating approaches for DNA sequencing (Sanger sequencing) in 1977 (Sanger et al. 1977) and polymerase chain reaction (PCR) for DNA amplification in 1985 (Mullis et al. 1986). This allowed researchers to easily amplify and

sequence known viruses that were difficult to culture or of disease interest, provided that the appropriate target sequence was available. By combining this with degenerative primers targeting conserved parts of the genome in a technique coined “consensus PCR”, it was possible to detect novel viruses that are related to the target sequence rather than being an exact match. Consensus PCR methods typically target two viral proteins that characterise the vast majority of viruses: the capsid and the polymerase. The viral capsid, an outer protein shell, protects the viral genome and facilitates cell attachment and entry. The polymerase, depending on the genome type and replication strategy is either an RNA-dependent RNA polymerase (RdRp), a DNA-dependent RNA polymerase, or an RNA-dependent DNA polymerase, and is responsible for virus replication. Although of great utility, targeting these proteins using consensus PCR also has key drawbacks: it is largely constrained by the limited diversity of available virus sequences and requires researchers to design primers specific to each virus group of interest. Unlike cellular organisms, which have the 16S small subunit of ribosomal RNA (rRNA) as a universal marker gene, viruses lack an equivalent. Even conserved genes such as RdRp within RNA viruses (realm Riboviria) exhibit high diversity, with some highly divergent viruses exhibiting less than the 5% similarity in amino acid sequence expected by chance alone. As a result, the view of all virus diversity, termed the virosphere, was constrained to closely related viruses of those previously sequenced or cultured (Figure 1.1).

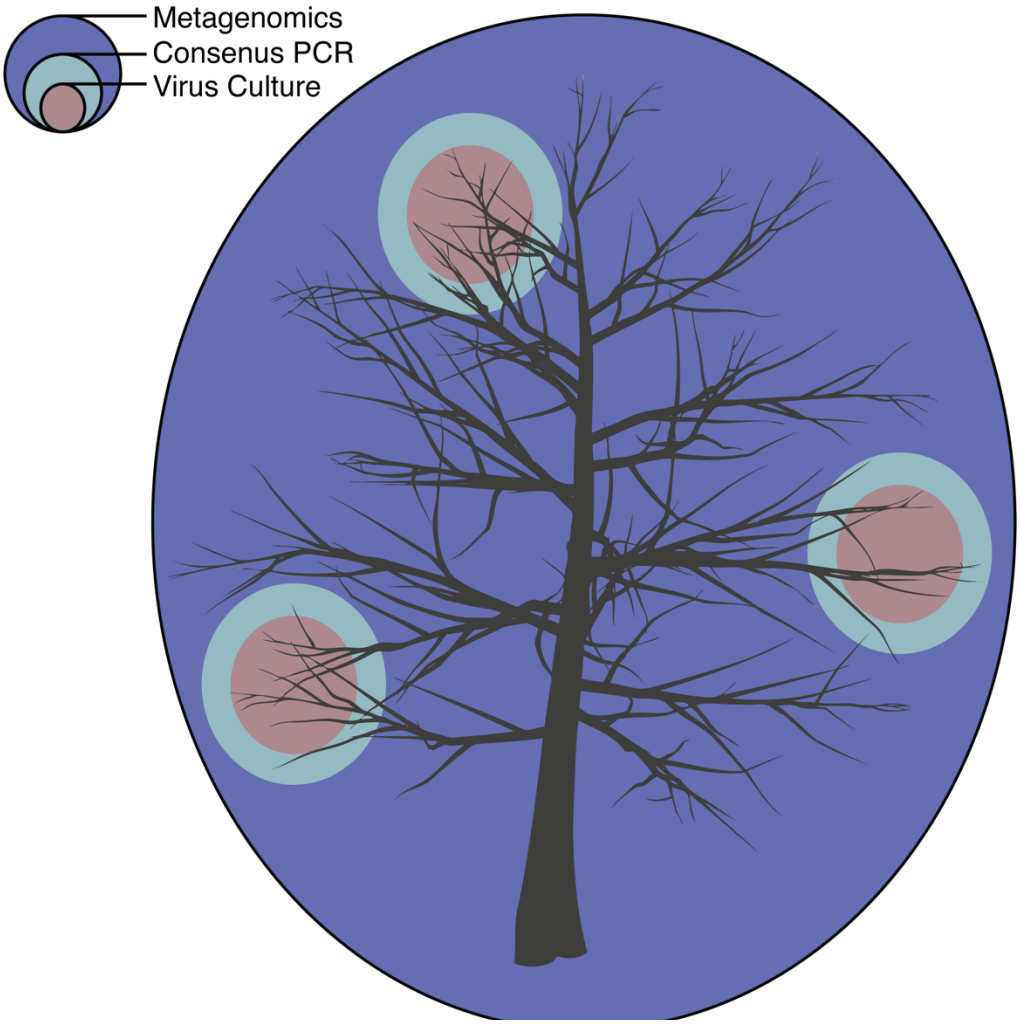



***Figure 1.1 Metagenomic era of virus discovery.*** *Tree schematic used to represent the total phylogenetic diversity of viruses. The coloured circles represent the extent of the viral diversity that can be captured with virus culture, consensus PCR, and metagenomic methods. This tree depicts a single tree for all viruses for visualisation purposes only. As there is no conserved gene across all viruses it is unlikely that viruses share a single common ancestor. Instead, evidence suggests that viruses have arisen independently on multiple occasions (Holmes 2011).*

### 1.5 The beginnings of metagenomic sequencing

A notable advancement in viral discovery came with the development of sequence-independent amplification methods that allow unbiased amplification of genetic material without requiring prior knowledge of target sequences. Early methods included sequence-independent single-primer amplification (SISPA) (Reyes and Kim 1991) and suppression subtractive hybridisation (Ambrose and Clewley 2006). These early methods laid the foundation for what is now recognised as metagenomic sequencing, a term coined by Jo

Handelsman in 1998 to describe the “cloning and functional analysis of the collective genomes of soil microflora” (Handelsman et al. 1998), which has since expanded outside of soil to include samples of all types. These breakthroughs became feasible with the advent of faster and more affordable sequencing technologies such as Roche 454 pyrosequencing (2004). Solexa (later commercialised as Illumina in 2007), PacBio (2011) and Oxford Nanopore Technology (ONT) sequencers (2015). With these advancements, the cost and error rate of genome sequencing have decreased dramatically, reducing from a median error rate of 0.473% for the Illumina MiSeq, released in 2011, compared to 0.109% for the Illumina NovaSeq 6000 released in 2017 (Schirmer et al. 2016). This has made metagenomic methods more accessible, accurate and appealing for a broader range of applications (Van Dijk et al. 2014).

Metagenomics is typically divided into two categories: targeted metagenomics and shotgun sequencing. Targeted metagenomics typically involves amplicon-based sequencing focusing on a specific marker such as the highly conserved 16S rRNA gene in bacteria, widely used for phylogenetic analysis (Sanschagrin and Yergeau 2014). While a targeted metagenomics approach such as the 16S rRNA gene is of considerable use for profiling bacterial communities, these approaches have obvious limitations for virus discovery. These include focus on a small, amplified region rather than entire genomes and reliance on universal marker genes, which are absent in viruses. As a result, amplicon-based metagenomics has seen limited adoption for profiling viral communities. In shotgun metagenomic sequencing, all genomic information contained within the sample, derived from various hosts is fragmented and randomly sequenced without prior knowledge of their sequences. Genome assembly is then performed *in silic*o. Shotgun sequencing of the entire RNA content (i.e., the transcriptome) – although various steps for virus enrichment and/or host depletion are commonly employed (see below) – is herein referred to as “metatranscriptomics”. The procedure and its application to virus discovery are detailed in Figure 1.2. Other metagenomic approaches, such as long-read metagenomics, are also gaining popularity, particularly for sequencing large and complex phage genomes. However, it has seen limited adoption for RNA viruses, as these genomes are relatively manageable to assemble using short-read sequencing.

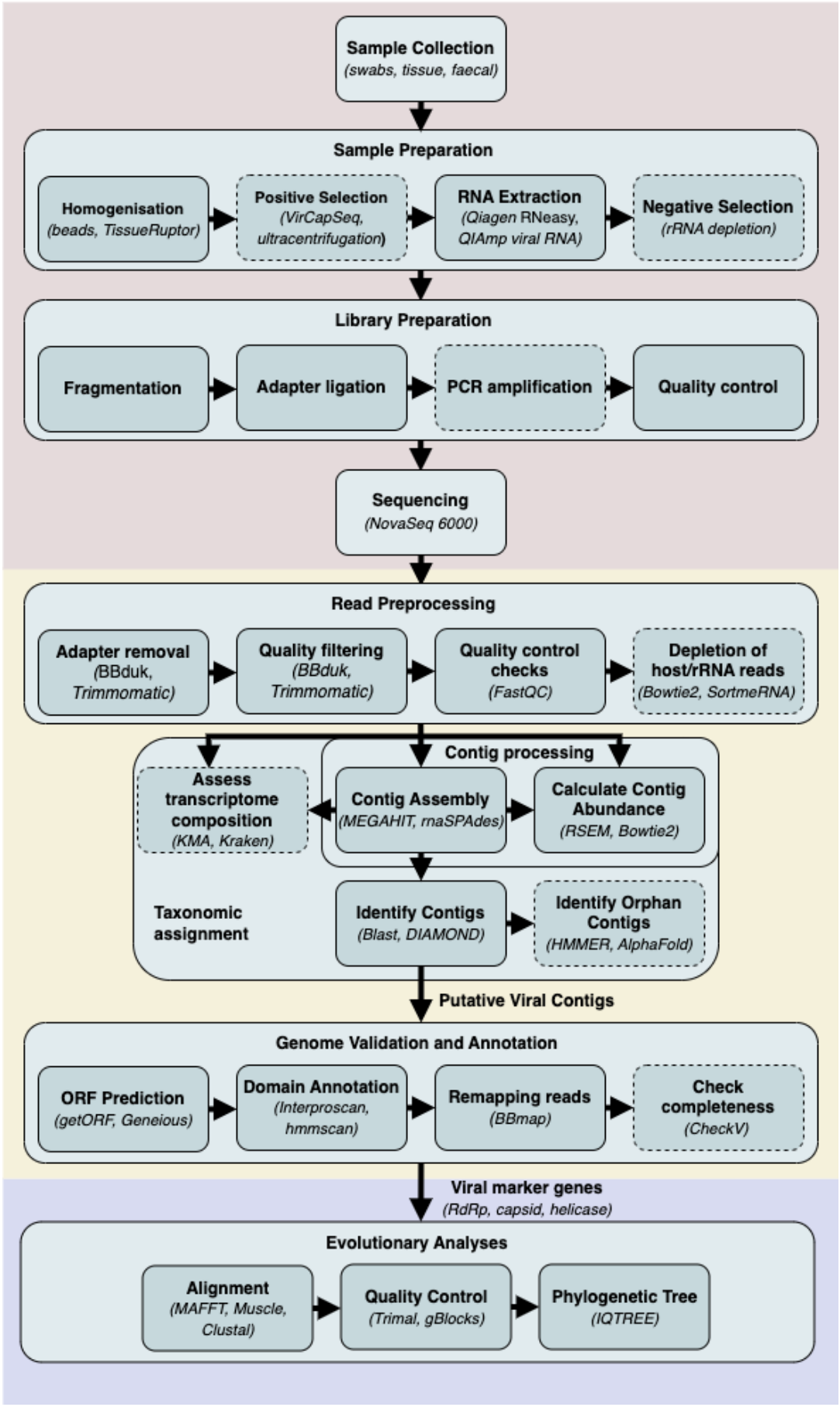

Sample Collection
(swabs, tissue, faecal)
Sample Preparation
Homogenisation
(beads, TissueRuptor)
Positive Selection
(VirCapSeq, ultracentrifugation)
RNA Extraction
(Qiagen RNeasy, QIAmp viral RNA)
Negative Selection
(rRNA depletion)
Library Preparation
Fragmentation
Adapter ligation
PCR amplification
Quality control
Sequencing
(NovaSeq 6000)
Read Preprocessing
Adapter removal
(BBduk, Trimmomatic)
Quality filtering
(BBduk, Trimmomatic)
Quality control checks
(FastQC)
Depletion of host/rRNA reads
(Bowtie2, SortmeRNA)
Contig processing
Assess transcriptome composition
(KMA, Kraken)
Contig Assembly
(MEGAHIT, rnaSPAdes)
Calculate Contig Abundance
(RSEM, Bowtie2)
Taxonomic assignment
Identify Contigs
(Blast, DIAMOND)
Identify Orphan Contigs
(HMMER, AlphaFold)
Putative Viral Contigs
Genome Validation and Annotation
ORF Prediction
(getORF, Geneious)
Domain Annotation
(Interproscan, hmmscan)
Remapping reads
(BBmap)
Check completeness
(CheckV)
Viral marker genes
(RdRp, capsid, helicase)
Evolutionary Analyses
Alignment
(MAFFT, Muscle, Clustal)
Quality Control
(Trimal, gBlocks)
Phylogenetic Tree
(IQTREE)

***Figure 1.2 Virus discovery protocol.*** *An example metagenomic virus discovery protocol comprised of the following sections: (i) Sample preparation: the sample is homogenised, and RNA is extracted. Situationally, positive and negative selection methods may be applied to increase the viral signal by concentrating or enriching viruses or by depleting background signal. (ii) Library preparation: The extracted RNA is transcribed into complementary DNA (cDNA), fragmented, and ligated with adapters at each end before sequencing. While not used in this thesis, targeted amplification can be used during library preparation. (iii) Bioinformatic analysis: Sequence reads are pre-processed and assembled into contiguous sequences (contigs) which are then classified using sequence similarity-based methods. Reads are mapped back onto contigs to estimate abundance and coverage of the putative viral sequences. These sequences are annotated and validated to assess for completeness and genome composition. (iv) Evolutionary analyses: Conserved domains such as the RNA-dependent-RNA polymerase (RdRp) are extracted and combined with reference sequences to generate multiple sequence alignments. These alignments are then used to infer maximum likelihood phylogenies. Examples of specific methods, instrumentation, or software are listed in italics under each step of the pipeline. Dotted borders around boxes indicate an optional or situationally useful step.*

### 1.6 Metatranscriptomic virus discovery workflow

Metatranscriptomics begins with sample processing (Figure 1.2) and has the advantage that it is highly flexible in regard to sample type and hence applicable to a range of tissues from vertebrates to plants, water, and air samples. As RNA is readily degraded by RNases and environmental conditions, samples often require cold storage in the field or stabilisers like RNAlater (Ambion) or DNA/RNA Shield (Zymo). Next, the RNA extraction procedure involves homogenising the sample and placing it in a lysis buffer. DNase is then used to reduce genomic DNA. In metagenomic libraries, viruses are often of relatively low abundance compared to the vast genetic material from cellular organisms (Parris et al. 2022). This low abundance poses a risk of underrepresentation and affects the quality of assembled contigs. This issue persists in metatranscriptomics, where host and bacterial rRNA commonly represent over 95% of RNA libraries, potentially overpowering the viral signal (Benes et al. 2011; O'Neil et al. 2013). As a consequence, both positive and negative selection techniques have been used to address this by enriching the viral RNA or depleting background RNA, respectively. At this stage of the extraction positive enrichment can be performed and typically involves separating the virions from the host material. This can be achieved through centrifugation (low-speed or ultracentrifugation, depending on the sample

type) or targeted enrichment such as VirCapSeq (Briese et al. 2015) as well as poly(A) selection, although targeted enrichment is performed after RNA extraction and purification.

Despite considerable utility, each of these positive selection (i.e., enrichment) approaches has limitations that have prevented them from seeing widespread adoption in virus discovery (although they can be of considerable utility for the diagnostic detection of know pathogens). For example, ultracentrifugation may bias virome composition through the exclusion of non-enveloped viruses (Tan et al. 2019) and potentially those that have been lysed or degraded. Furthermore, ultracentrifugation can introduce contaminating nucleic acids from the density gradient solution. Capture-based methods, such as VirCapSeq, allow for the 100- to 10,000 fold increase in viral reads compared to traditional high-throughout sequencing (Anderson et al. 2018). However, the viral reads are limited to those with sequence similarity to the diversity of the viruses from which the probes were constructed, which in this case is 207 viral taxa known to infect vertebrates (Briese et al. 2015). In the context of this thesis, metatranscriptomics refers specifically to the total sequencing of RNA using shotgun sequencing without methods of positive selection and targeted enrichment unless otherwise specified.

With the total RNA extraction extracted, negative selection (i.e., depletion) methods are typically applied. These commonly involve the depletion of mitochondrial and cytoplasmic ribosomal RNAs (rRNA) using kits such as the Ribo-Zero Gold Kit (Illumina). This requires the use of probes that bind rRNA which are subsequently captured and removed. The key advantage of rRNA depletion approaches is that it is essentially unbiased as viruses do not possess rRNA genes. Unlike methods that selectively enrich for viruses—potentially introducing biases—it strengthens the viral signal by depleting the majority of the rRNA which comprises a large component of the host reads. RNA is then reverse-transcribed to complementary DNA (cDNA) and subjected to library preparation and sequencing on the desired platform.

The raw sequencing data generated from total RNA sequencing is processed bioinformatically (Figure 1.2). In this thesis, this typically involves quality trimming and the removal of Illumina HiSeq adapters and other technical sequences from the raw reads using Trimmomatic (Bolger et al. 2014). Trimming is performed as low-quality read ends or adapter contamination can introduce artefacts leading to incorrect assemblies downstream. FastQC is then used to diagnose common issues with sequencing data, such as problems

with base sequence quality, per-base GC content, read length distribution, sequence duplication, overrepresented sequences, adapter contamination, and k-mer content (Andrews 2017). In certain cases in this thesis, the *in silico* depletion of host sequence has been conducted by mapping the reads to a reference genome of the host using Bowtie2 (Langmead and Salzberg 2012). Since host reads often comprise a large proportion of the total sequencing library, pre-assembly host-mapping subtraction can drastically reduce the number of sequences for downstream assembly and homology searches. This reduction decreases computational resource requirements and potentially increases the accuracy of analyses with reduced false positive hits to host genes. However, this approach is only situationally used as it requires a high-quality genome closely related to the sequenced host, which is not always available. Additionally, remnants of viral sequences integrated into host genomes, known as endogenous viral elements (EVEs) are commonly found. Exogenous virus reads can be lost when mapping against a host genome containing EVEs sharing homology to these reads (Brait et al. 2024). At this stage, residual rRNA from the host or other organisms can be bioinformatically removed using programs such as SortmeRNA (Kopylova et al. 2012).

Following preprocessing described above, reads are assembled *de novo* into long contiguous pieces of sequence (contigs) without the use of a reference genome. Two common types of commonly used *de novo* assemblers are greedy algorithm assemblers and De Bruijn graph assemblers. In this thesis, the latter is employed using the program MEGAHIT (Li et al. 2015b). Briefly, the process of *de novo* assembly using De Bruijn graphs involves splitting each read into overlapping sequences called k-mers. Each k-mer forms a node in the De Bruijn graph. Nodes are connected by edges when one k-mer overlaps with another. Once all k-mers within all reads are graphed, the graph is simplified by merging linear paths into single sequences. The assembler tries to find the path through the graph that visits the most edges. MEGAHIT implements this method and uses iterative k-mer sizes, which allows it to capture broad and finer read overlaps (Li et al. 2015b). Reads can be mapped back onto the constructed contigs using programs such as Bowtie2, BBMap, and RSEM (Bushnell 2014; Langmead and Salzberg 2012; Li and Dewey 2011) to provide estimates of genome read coverage and abundance.

The taxonomic classification of sequences is typically conducted using sequence homology-based search methods (but see Section 1.12: Protein Evolution and Structure). In this thesis, this predominantly involves the use of the NCBI BLAST suite (Camacho et al.

2009) and DIAMOND BLAST tools (Buchfink et al. 2021). This process typically involves comparing the nucleotide contig sequences to protein databases using the rapid and highly sensitive protein DIAMOND BLASTx tools. BLASTx, also known as a translated BLAST, is used as it translates nucleotide sequences in all six reading frames and compares them to the reference database. The key benefit of this is that it eliminates the need to predict open reading frames (ORFs), which can be difficult to accurately do with diverse metagenomic sequence sets and challenging with partial coding sequences. In this thesis, BLASTx is frequently used to compare contigs to virus-specific databases, such as the RdRp-scan core protein sequence database (Charon et al. 2022) and the protein version of the Reference Viral Databases (RVDB) (Bigot et al. 2019; Goodacre et al. 2018). It is common practice to retain those contigs with positive hits to viral sequences which substantially reduces the sequence set. To exclude potential false positives, these contigs are queried against the NCBI nucleotide database and the NCBI non-redundant protein (nr) database. Putative virus contigs are then analysed in Geneious (https://www.geneious.com) (Kearse et al. 2012), where ORF prediction occurs. Functional domains are then annotated using InterProScan, which compares the virus amino acid sequences against the InterPro, a resource that integrates multiple domain and functional site databases (Blum et al. 2024; Jones et al. 2014).

To infer phylogenetic trees, nucleotide or amino acid sequences with similarity to virus proteins, typically the polymerase or capsid, are extracted and aligned with related reference sequences to form multiple sequence alignments (MSA) using programs such as MAFFT, MUSCLE and CLUSTALO (Edgar 2022; Katoh and Standley 2013; Sievers et al. 2011). Ambiguously aligned regions such as those with a high proportion of gaps, are removed using TrimAl (Capella-Gutiérrez et al. 2009) from the alignment, as this has been shown to increase the accuracy of MSA-based applications (Talavera and Castresana 2007). From these alignments, phylogenetic trees can be estimated using various methodologies. Although distance-matrix methods such as the unweighted pair group method with arithmetic mean (UPGMA) or neighbour-joining (NJ) were once popular, their use has subsided as it is evident that evolutionary information is lost in such a process. As a consequence, character-based methods such as maximum likelihood (ML), maximum parsimony, and Bayesian inference are now predominant. In this thesis, I use maximum likelihood inference through IQ-TREE (Minh et al. 2020). ML is a process of statistical inference that calculates the probability of a given tree based on the observed sequences and the likelihood of their evolution under that tree. This is achieved by estimating the

probabilities of different types of nucleotide substitution. Trees that require more substitutions at internal nodes to explain the observed phylogeny are generally assigned lower likelihoods. ML methods account for varying probabilities of amino acid or nucleotide substitutions by employing models of sequence evolution, known as (nucleotide or amino acid) substitution models. These models estimate the substitution probability for nucleotides or amino acids at a given site. In this thesis, a variety of nucleotide and amino acid substitution models were used. The substitution models were generally variations of the General Matrix (LG) model (Le and Gascuel 2008) combined with Gamma models to account for variable substitution rates across sites (Yang 1994). The appropriate substitution model for a given data set is often unknown *a priori*. Hence, model choice can be performed using ModelFinder (Kalyaanamoorthy et al. 2017), which computes the log-likelihoods of maximum parsimony phylogenies for a wide variety of substitution models. For each model, a Bayesian information criterion (BIC) score is calculated and the model with the lowest score is selected (Kalyaanamoorthy et al. 2017).

### 1.7 Transcriptome mining

Researchers do not always need to collect and sequence their own samples. Upon manuscript acceptance, authors are routinely requested to upload metagenomic data in the form of raw reads to publicly available databases. The largest of these repositories is the Sequence Read Archive (SRA) (Leinonen et al. 2010). Coinciding with the development of low-cost, low-error and high-throughput sequencing technologies, the amount of freely available metagenomic data has grown at a remarkable rate (Katz et al. 2022). This increase is reflected in the expansion from roughly two quadrillion bases in 2013 to 86 quadrillion bases by 2023, encompassing over 5 million individual sequencing runs (Figure 1.3) (Charon et al. 2024). Not only has the amount of sequencing data increased, but the diversity of hosts represented has also expanded, growing from 1,843 unique organisms in 2010 to 48,461 in 2023. Similarly, the proportion of the database associated with human sequencing projects has decreased from 58% in 2020 to 39% in 2023 (Figure 1.3). Taken together, the SRA now provides a valuable snapshot of the diversity of life on Earth, particularly in terms of vertebrates. The vast repository of sequencing data can be bioinformatically analysed (Figure 1.2) and serves as an invaluable resource for virus

discovery.

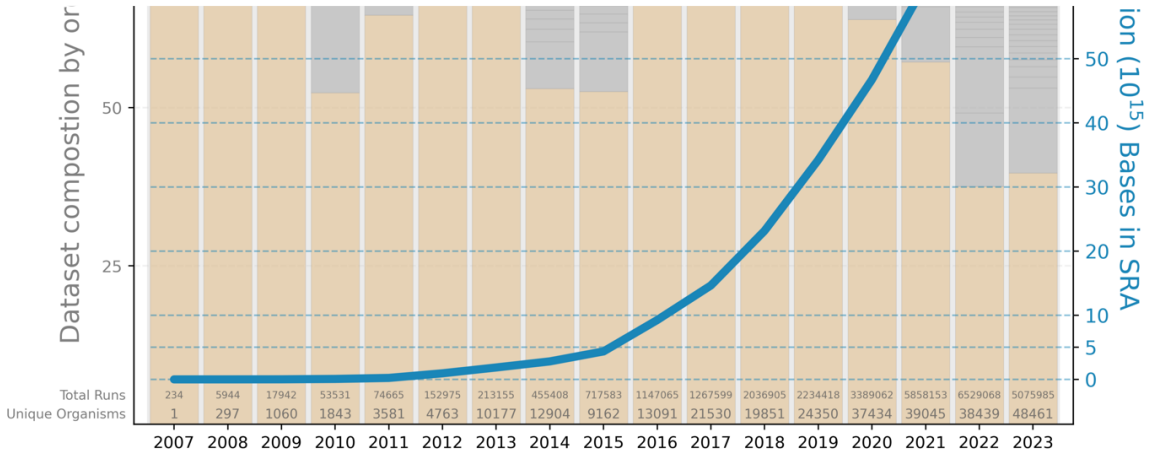



***Figure 1.3 Expansion of the Sequence Read Archive****. Left axis, stacked bars show the species composition proportion in the Sequence Read Archive (SRA) across time. Brown represents human and grey other. Right axis, blue lines, the cumulative number of bases in the SRA across time. The number of total sequencing runs and unique organisms are displayed above the x-axis. Data taken from SRA metadata available via BigQuery (nih-sra-datastore.sra.metadata). Figure reproduced from Charon, J., Olendraite, I., Forgia, M., Chong, L. C., Hillary, L. S., Roux, S., ... & Neri, U. (2024). Consensus statement from the first RdRp Summit: advancing RNA virus discovery at scale across communities. Frontiers in Virology, 4, 1371958.*

Transcriptome mining not only leverages prior monetary investments in sampling and sequencing, but it enables a broader diversity of host species to be targeted. Virus discovery efforts are often directed towards diseased hosts, particularly those of economic importance (Harvey and Holmes 2022). While this anthropocentric bias still extends beyond virology, the diverse research objectives outside of virology result in access to sequencing data that would be difficult to justify in a virology-focused grant. For example, the 1000 Plant Transcriptomes Initiative (1KP) (Leebens-Mack et al. 2019) sampled plants across all major taxonomic groups and broad geographical regions, most of which are of little economic importance. Given the large biochemical diversity of these samples, and the fact that they are non-model organisms (e.g., ferns, mosses, and green algae) for which standard RNA extraction kits were not developed, sampling and extraction of RNA can be

challenging. In such cases, custom RNA extraction protocols are required (Carpenter et al. 2019). The secondary use of this SRA data enables us to leverage this expertise. In Chapter 4 I use this resource to identify viruses across the breadth of the plant kingdom.

The SRA has been leveraged by initiatives such as the Serratus project, which analysed raw reads from over 5 million runs using a translated-nucleotide search against a database of viral RdRps. (Edgar et al. 2022) These results have been developed into a public interface that has facilitated numerous virus discovery projects (Bergner et al. 2021; Debat et al. 2023; Debat and Bejerman 2023; Forgia et al. 2023; Petrone et al. 2023). Similarly, genome assemblies were recently constructed for the entire SRA (Chikhi et al. 2024) and are now searchable through a kmer-BLAST interface (https://logan-search.org) which will undoubtedly enhance the efficiency of virus discovery. In Chapter 5 I use this resource to narrow down the number of SRA libraries to search for identifying novel members of the *Flaviviridae*.

There are several limitations associated with the use of the SRA that must be considered during study design. Within a given species of interest, sequencing libraries are likely to be prepared using different methods, with or without positive or negative selection, and sequenced on different platforms at varying depths. While these variations are unlikely to affect the conclusions of a pure virus discovery study, they introduce substantial biases in cross-study statistical comparisons of virus diversity and abundance (Visser et al. 2016). Additionally, the use of public data limits control over sampling methodology and contamination control, which poses challenges to data reliability and, more specifically, the accurate assignment of virus host species. As a result, the host information provided in the SRA should not be assumed to correspond to the actual host of the virus without further investigation. This could include using read taxonomic classification tools such as KMA or CCMetagen (Clausen et al. 2018; Marcelino et al. 2020) to examine library composition, employing phylogenetic analyses to examine host patterns, and understanding the sampling and sequencing methodologies associated with the libraries.

### 1.8 Changing understanding of viruses from metagenomics

Metatranscriptomics has proven useful for many areas of virology and underpins much of the work in this thesis. Below I discuss the impact of metagenomics on our understanding of viral disease, taxonomy, diversity, and evolution.

### 1.8.1 Metagenomics, virus diversity and taxonomy

The onset of low-cost, low-error, and high-throughput sequencing platforms has resulted in the widespread adoption of metagenomics and the rapid increase in virus genomes uploaded to the NCBI database (Figure 1.4). However, the increase in the number of virus genomes generated has introduced challenges for the classification and incorporation of these sequences into taxonomy.

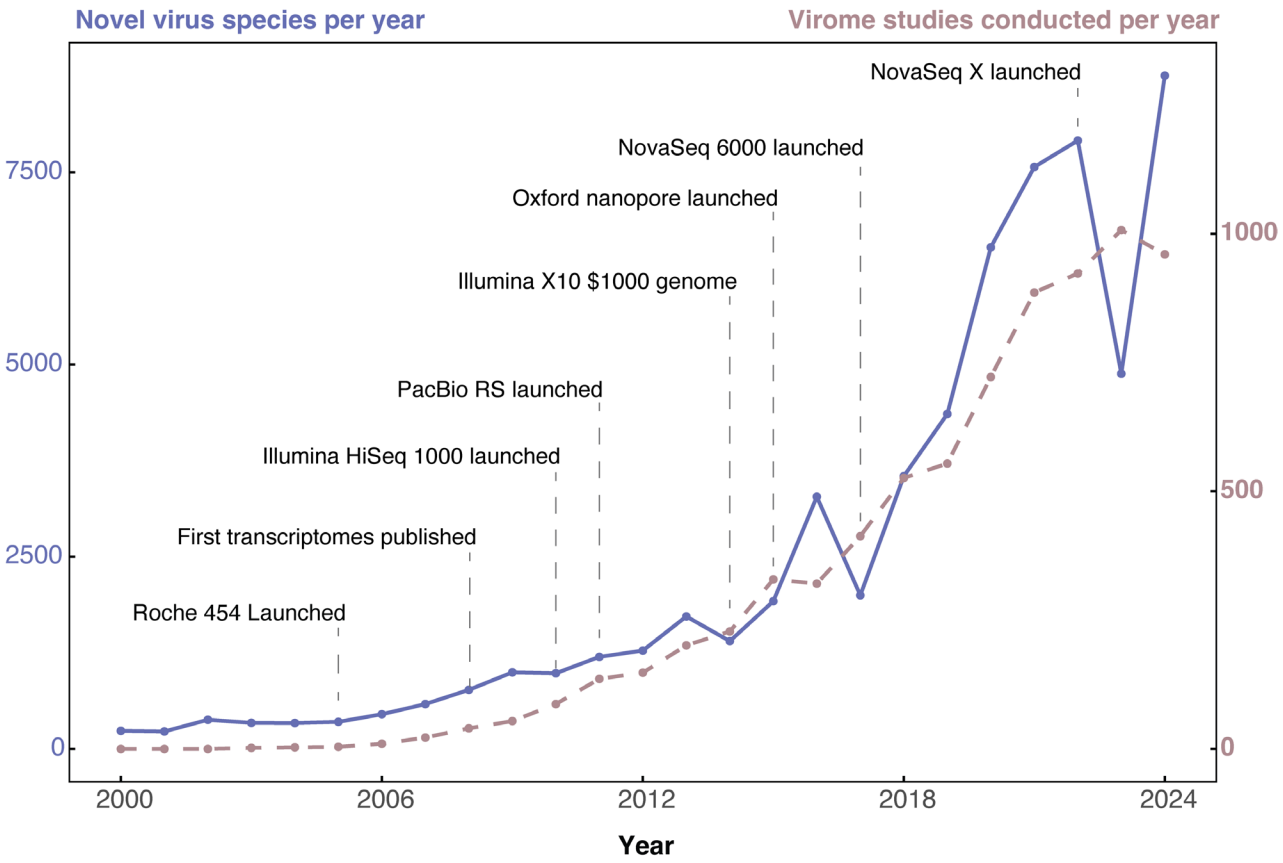



***Figure 1.4 Increasing rate of metagenomic virus studies and novel species submissions over time.*** *Left axis, in blue, the number of novel virus organisms added to the NCBI nucleotide database from the years 2000 to 2024. Right axis, in purple, the number of viral metagenomic studies published in NCBI's PubMed databases conducted from the years 2000 to 2024. Study data was obtained using the search query [(metagenomic OR metatranscriptomic) AND (virus OR virome)]. Key events in the history of metagenomics are shown above the line graph. Figure adapted from Chang, W.E., Harvey, E., Mahar, J., Geoghegan, J., Firth, C., Shi, M., Simon-Loriere, E., Holmes, E., & Wille, M. (2024). Improving the reporting of metagenomic virome-scale data. Zenodo.* https://doi.org/10.5281/zenodo.10846429

Virus taxonomy is governed by the International Committee on Taxonomy of Viruses (ICTV) which was first established in 1966 under the name International Committee on Nomenclature of Viruses (ICNV). The ICTV publishes taxonomic reports and nomenclature for virus groups and is in charge of reviewing and accepting taxonomic proposals to create or modify existing taxa. In the pre-metagenomics era, both the naming of viruses (nomenclature) and their hierarchical biological or evolutionary classification (taxonomy)—such as family, genus, and species—were primarily based on observable phenotypic features. These features included the symmetry and structure of virus particles, type of nucleic acid, and other biochemical properties. The formal ratification of new species required substantial information on the virus's host range, morphology, and replication cycle. This process typically necessitated culturing the virus—an approach that is not always feasible—and significant wet lab investment. Over the first 40 years since its inception (1966-2004), the ICTV successfully classified nearly all known viruses (n = 1899) at the family, genus, and species levels.

As more virus genomic sequences became available, the phylogenetic relationships among virus sequences was considered alongside host range, morphology, and replication cycle as part of the criteria for virus classification, although never used as the sole basis (Simmonds 2015). This has fundamentally changed with next-generation sequencing, particularly metagenomic virus discovery. For example, a recent study of environmental metatranscriptomes using an artificial intelligence-based protein structure homology searches identified 161,979 potentially novel RNA virus species as well as likely new phyla of RNA viruses (Hou et al. 2024). This finding represents an 8.6-fold expansion of RNA virus diversity compared to the number of RNA viruses classified by the ICTV at that time (Hou et al. 2024; Simmonds et al. 2024). Many of these newly identified viruses lack information about their hosts and morphology. It is unrealistic to expect that these viruses—often of limited anthropocentric importance—would receive the same level of experimental classification as disease-causing agents (Simmonds et al. 2017).

Given the breadth of virus diversity now available and the increasing information that can be extracted from genomic sequences—including evolutionary relationships, genome composition and features, motifs, secondary and tertiary structures, and expression and replication strategies—some ICTV study groups have begun accepting taxonomic proposals for novel species based solely on metagenomic data. For example, *Rotavirus I* was ratified based entirely on a genome sequence obtained from metagenomic sequencing

(Mihalov-Kovács et al. 2015). Despite these advances, species demarcation criteria for many virus families have yet to formally accept sequence-based comparisons in this manner, resulting in the current virus classification lagging behind the known diversity and considerable confusion in the field. For example, the current taxonomic report for the *Flaviviridae* does not yet include segmented and large genome flaviviruses or the wide diversity of metazoan and plant hosts this group infects (Table 1).

***Table 1. Comparison of the International Committee on Taxonomy of Viruses (ICTV) description of the Flaviviridae compared to the known (unclassified) diversity***

| **Characteristic** | **ICTV Description (2019)** | **Current Diversity** |
|---|---|---|
| Genome | 9.0–13 kb<br>+sense<br>non-segmented RNA | 8.2-40 kb<br>+sense<br>Segmented RNA (jingmenviruses)<br>Non-segmented RNA (all other viruses) |
| Host range | Mammals<br>most *Orthoflavivirus* are arthropod-borne | Majority of metazoan groups<br>Potentially diatoms, plants and fungi |
| Taxonomy | four genera and 97 species | ~300-400 species |

### 1.8.2 Metagenomics and disease

With the advent of cell culture, electron microscopy, and early molecular identification techniques, researchers were able to identify pathogens for diseases with no attributable cause. Metagenomics, specifically metatranscriptomics, has greatly expanded these capabilities (Hoffmann et al. 2012; Palacios et al. 2008; Xu et al. 2011). A key advantage of these techniques is their ability to recover novel and divergent pathogens, including those that may represent recent zoonotic events. A notable example of this was the rapid identification of SARS-CoV-2 from a worker at the Wuhan seafood market at the onset of the COVID-19 pandemic (Wu et al. 2020), which in turn led to the design of vaccines for COVID-19 within days. In other cases, the pathogen may be known but not previously associated with clinical disease, making it unlikely to have been targeted using traditional methods of virus discovery such as PCR. For example, a recent investigation into unexplained paediatric hepatitis and the association with human adenovirus (HAdV) a virus not commonly associated with hepatitis. Interestingly, in Scottish patients, only low levels of HAdV DNA were detected, and no inclusions, viral particles, or proteins were observed in

the liver tissue (Ho et al. 2023). Metagenomics followed by PCR analysis identified adeno-associated virus 2 (AAV2), a parvovirus, in all but one case and suggested that, in combination with other factors, AAV2 likely triggered immune-mediated hepatic disease (Ho et al. 2023). Importantly, metatranscriptomics captures more than viral information: it captures the entire spectrum of gene expression, including that of the host in question. This allows researchers to investigate host responses to viral infection (e.g., cytokine activity) through the comparison of sick and healthy individuals with and without the virus (Morfopoulou et al. 2023; Rajagopala et al. 2021). Additionally, metatranscriptomics has also proven useful in solving complex disease associations such as those observed in human clinical cases of immunocompromised individuals who are often co-infected with a range of common and uncommon pathogens including viruses, bacteria, fungi, and parasites (Han et al. 2019; Parize et al. 2017).

Metagenomic virus discovery has also resulted in a changing view of virus virulence. In particular, it is now evident that viruses aren't strictly disease-causing for their hosts and do not exist in isolation, but rather are part of a broader holobiont made up of bacteria, archaea, eukaryotes, and other viruses. Indeed, seemingly healthy hosts can harbour multiple viruses and even known pathogens (He et al. 2022). As a result, it is likely that the one-pathogen-one-disease model that characterised early studies of disease association is overly simplistic, and instead that observed virulence likely results from a combination of factors including virus, other microbes, host, and environment (Casadevall and Pirofski 2001; Geoghegan and Holmes 2018).

### 1.8.3 Metagenomics, agricultural and wildlife viruses

As a result of the enormous economic impact that viral disease can have on agriculture, many studies have used metagenomics to characterise the virome of agricultural animals and crops (Kwok et al. 2020; Rivarez et al. 2023). Indeed, outside of humans, pigs, chickens, and cows are the most frequently reported hosts for viruses in the NCBI virus database (Brister et al. 2015), collectively accounting for 33% of sequences. As infectious diseases in humans and agricultural animals often originate from wildlife, it has long been recognised that the health of humans, animals, and ecosystems are interconnected (Kruse et al. 2004). Approaches considering this view is commonly referred to as the "One Health" framework which has been adopted by various organisations including the World Organisation for Animal Health (WOAH). The overarching goal of this research is to improve

the health of humans, animals, and ecosystems (https://www.woah.org/en/what-we-do/global-initiatives/one-health/)

A key component of the One Health approach involves studying viruses in sentinel animals (Bartlett and Uhart 2024). Sentinel animals are believed to provide early warnings of broader issues that could affect human health or ecosystem stability (Bossart 2011) and are identified based on factors such as their ecological niche, biology, or interactions with other species. Sentinel species often inhabit areas near humans and may share physiological characteristics such as susceptibility to certain infectious agents, making them valuable indicators of disease emergence. As a case in point, seals are recognised as sentinel species for ocean and human health (Bossart 2011). In the Baltic Sea, seals are seen as key indicator species and are monitored for viruses, including Phocine Distemper Virus and influenza A (Sonne et al. 2020). These viruses are not only of importance due to their association with mass seal mortalities (Puryear et al. 2023), but also because of their potential for zoonotic transmission to humans. Indeed, outbreaks of Influenza A and sealpox viruses have been documented as zoonotic, with cases of transmission to humans (Abdelwhab and Mettenleiter 2023). In Chapter 2 I examine the virome of the marine sentinel, the New Zealand fur seal (*Arctocephalus forsteri*), which is the most abundant seal species in the Australian-New Zealand region.

### 1.8.4 Metagenomics and companion animals

Companion animals, particularly domestic dogs (*Canis lupus familiaris*) have also been used as sentinel species (Bowser and Anderson 2018). For example, dogs have been used to monitor tick- and mosquito-borne viruses that infect humans, such as tick-borne encephalitis virus and West Nile virus, as well as bacterial pathogens like *Borrelia burgdorferi*, the causative agent of Lyme disease (Duncan et al. 2005; Könenkamp et al. 2022; Rocheleau et al. 2017; Roelandt et al. 2011). Due to their close proximity to humans, viruses affecting these animals theoretically present a risk of becoming zoonotic or spilling over into wildlife (Ghasemzadeh and Namazi 2015; He et al. 2022). For example, dogs are a major vector of rabies, causing ~99% of an estimated 59,000 human deaths each year (Hampson et al. 2015). In addition, reverse zoonosis (zooanthroponosis) events are also of considerable concern. These occur when viruses are transmitted from humans to animals. For example, in a large-scale survey of companion animals (n = 919) in northern Italy, 3.3% of dogs and 5.8% of cats had measurable SARS-CoV-2 neutralising antibody titres (Patterson et al. 2020). In particular, dogs from COVID-19 positive households were

significantly more likely to test positive than those from COVID-19 negative households. In addition, zooanthroponotic viruses could gain mutations in the intermediate host and spill back into humans. An investigation of a SARS-CoV-2 outbreak in mink farms and mink farm employees in the Netherlands revealed that human sequences were nearly identical to mink sequences from the same farm and often clustered deeply within the diversity of mink-associated sequences. This suggests onwards human transmission of SARS-CoV-2 from minks to humans (Oude Munnink et al. 2021; Tan et al. 2022). Within the mink sequences there was a 3.5 fold increase in C->U compared with U->C substitutions, suggestive of adaption to minks (Oude Munnink et al. 2021). In Chapter 3, I conduct a metatranscriptomic survey of the viruses in Australian domestic dogs as well as the endemic dingoes (*Canis lupus dingo*) and identify potential zoonotic agents. Australia's geographic isolation, strict biosecurity measures, and native dingo populations provide a unique opportunity to study viral evolution and transmission in companion animals.

### 1.8.5 Virus-host evolutionary relationships

The integration of sensitive protein sequence search algorithms such as BLASTP (Altschul et al. 1990), DIAMOND (Buchfink et al. 2021) with assembled metagenomic sequencing libraries has enabled the extension of virus families and the identification of new virus groups at the edge of known viral diversity. These new groups often fall between established families and orders, revealing greater evolutionary continuity across virus groups and prompting the amalgamation of smaller groups into larger orders and even phyla (Hou et al. 2024; Neri et al. 2022). The linkage between different virus groups has allowed for a more stable reconstruction of virus evolutionary history of which several aspects are discussed below.

Coinciding with the increased sampling coverage of vertebrate and invertebrate organisms (Harding et al. 2022; Li et al. 2015a; Paraskevopoulou et al. 2021; Shi et al. 2016; Shi et al. 2018), phylogenetic analyses of the newly discovered viruses have fundamentally changed our view of how viruses have evolved alongside their hosts. In particular, many of the virus genera and families that were once thought to be strictly associated with mammals have now been identified across all vertebrate classes and have distant relatives identified in invertebrates, in turn suggesting that these viruses have co-diverged with their hosts over millions of years (Petrone et al. 2023; Shi et al. 2016; Shi et al. 2018). An informative example is provided by the *Orthomyxoviridae* a family of enveloped, segmented, negative-sense single-stranded RNA viruses. Prior to the influx of metagenomic-derived virus

sequences this family comprised five genera: *Influenza virus A*, *B*, and *C*, along with *Isavirus* and *Thogotovirus* which were associated with mammals, birds, fish, and ticks (Allison et al. 2015). Over the past decade, large-scale metagenomic virus discovery projects in invertebrates and vertebrates (Harvey et al. 2019; Li et al. 2015a; Parry et al. 2020; Petrone et al. 2023; Shi et al. 2016; Shi et al. 2018; Xu et al. 2022a; Xu et al. 2022b) have identified ten times more novel orthomyxoviruses than were discovered in the 79 years following the identification of the first orthomyxovirus (Influenza A virus) (Dudas and Batson 2023). Notably, the finding of influenza B-like viruses in amphibians and fish suggests that influenza viruses have been associated with vertebrates throughout their entire evolutionary history (Parry et al. 2020; Petrone et al. 2023).

Long-term evidence of virus-host associations can also be found within host genomes. Infection of germ cell lines can result in the integration of viral genetic material, known as endogenous viral elements (EVEs), into host chromosomes allowing vertical transmission to future generations. Although the integration and fixation of EVEs derived from non-retroviral RNA viruses are rare compared to retroviruses, many examples have been documented (Crochu et al. 2004; Katzourakis and Gifford 2010; Li et al. 2022). For example, the recent finding of an orthoflavivirus-like derived EVEs in the peach blossom jellyfish genome (*Craspedacusta sowerbii*, phylum *Cnidaria*) suggests that an ancient origin of the flaviviruses, likely close to the emergence of the metazoans 750–800 million years ago (Bamford et al. 2022). Due to the sparsity of basal metazoans viruses identified and experimentally characterised the ancestral features of RNA virus lineages are difficult to infer. Specifically, the lack of traditional circulatory or organs system raises the question as to how transmission and cell-to-cell spread of viruses in basal metazoans occurs.

Virus evolution is punctuated by countless (horizontal) cross-species transmission events, a process that is key to disease emergence. By comparing the phylogenetic topologies of host and virus, estimations of the frequencies of these events have been made, in turn demonstrating that cross-species transmission is a near universal feature among viruses (Geoghegan et al. 2017). Interestingly, the proportion of cross-species transmission events to co-divergence is not consistent across Baltimore classes. In particular, double-stranded DNA genomes (dsRNA) were found to exhibit more frequent co-divergence while RNA viruses show greater topological incongruence, potentially reflecting more frequent host switching (Geoghegan et al. 2017). Metagenomically identified viruses have helped reveal this complex interaction between cross-species transmission and virus-host co-divergence.

For example, the coronavirus subfamily *Letovirinae* has been identified in diverse vertebrates, including amphibians, bony fish, and jawless vertebrates like the pouched lamprey (*Geotria australis*) (*Geotria australis*) (Edgar et al. 2022; Lauber et al. 2024; Miller et al. 2021). Phylogenetic trees of the *Letovirinae* generally reflect that of their hosts with mammals clustering together followed by birds. However, key violations of virus-host co-divergence are evident, indicative of cross-species transmission events (Figure 1.5). For example, instead of falling basal to bony fish *Letovirinae*, lamprey-associated letoviruses are phylogenetically positioned within the bony fish clade. Additionally, two amphibian-associated letoviruses occur both within and basal to the fish clades (Figure 1.5). These patterns suggest a complex history of cross-species transmission occurring on a backbone of virus-host co-divergence over long evolutionary timescales. As more viromes from diverse organisms across the tree of life are examined, our understanding of these dynamics will undoubtedly improve.

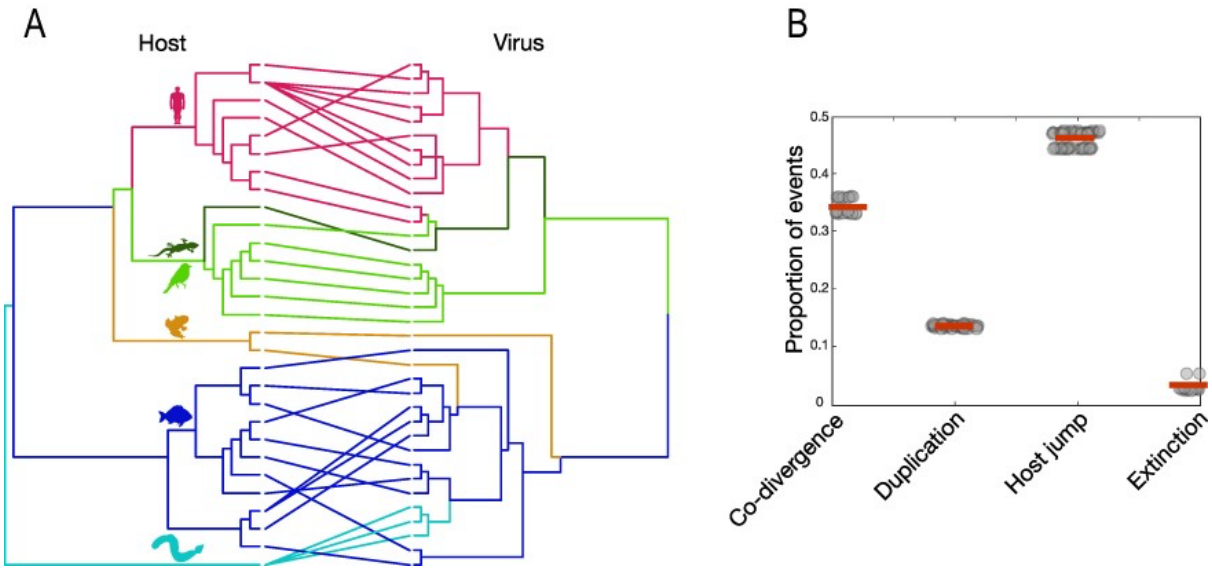



***Figure 1.5 Coevolutionary analysis of the Coronaviridae**. (A) Tanglegram comprising the phylogenetic trees for the Coronaviridae (right) and their hosts (left). Branches are coloured by host clade, human = pink, reptile = dark green, light green = avian, yellow = amphibian, dark blue = bony fish, light blue = cartilaginous fish. (B) Reconciliation analysis of the Coronaviridae shows the proportion of specific predicted evolutionary events for each maximum parsimony reconciliation, with the mean indicated by a red horizontal line. The 'event costs' associated with incongruences between trees were conservative towards co-divergence and defined here as follows: 0 for co-divergence, 1 for duplication, 1 for host-jumping and 1 for extinction. Figure reproduced from Miller, A. K., Mifsud, J. C., Costa, V. A., Grimwood, R. M., Kitson, J., Baker, C., ... & Geoghegan, J. L. (2021). Slippery when wet:*

*cross-species transmission of divergent coronaviruses in bony and jawless fish and the evolutionary history of the Coronaviridae. Virus Evolution, 7(2), veab050.*

### 1.9 Sequence saturation and distant evolutionary relationships

The identification and characterisation of viruses generally rely on sequence similarity-based comparisons to known viruses. Specifically, information about novel sequences is inferred from their homology to known sequences. However, the rapid evolutionary rates that characterise RNA viruses, that are approximately $\sim 10^4$ to $10^6$ times higher than those of cellular organisms (Duffy et al. 2008), make it challenging to meaningfully compare evolutionarily distant sequences, even for the most conserved regions of the RNA virus genome, such as the RdRp (Holmes and Duchêne 2019). This is largely down to a combination of their rapid rate at which they evolve, as amino acid sequence search and sequence alignment methods perform poorly in detecting evolutionary-related proteins sharing less than 25% identical residues (Venclovas 2012). Specifically, at this level of divergence, multiple substitutions are likely to accumulate at the same site leading to a substantial underestimation of divergence, known as "site saturation" (Wilke et al. 2009). This would result in uncertainty in downstream analyses, such as MSA, the estimation of genetic distances (which can be considerable), and phylogenetic inference (Figure 1.6a).

Efforts to have been made to estimate a phylogenetic tree of all RNA viruses. Using an iterative clustering and alignment procedure, Wolf et al. (2018) constructed a single RdRp alignment comprising 4,627 taxa and 1,028 reverse transcriptase (RT) sequences. These results classify the RNA viruses into five major branches and imply that dsRNA viruses evolved from positive-sense RNA viruses on at least two occasions, while negative-sense RNA viruses evolved from dsRNA viruses. This has provided the foundation for the mega-taxonomy classification of RNA viruses, which now includes six recognised realms (Koonin et al. 2024). Despite revealing several putative sequence motifs indicative of a common ancestry among RdRp sequences, these results—and a subsequent comment and response (Holmes and Duchêne 2019; Wolf Yuri et al. 2019)—highlight challenges in performing such an analysis. The alignment presented several key issues. Notably, all sites within the alignment including the highly conserved GDD motif contained gaps. Additionally, the pairwise identity between sequences was often extremely low with a mean value of 7.7% and a minimum of 1%. This approaches the level of pairwise identity expected by chance alone (5%). In this alignment, the mean pairwise distance neared the maximum possible value of substitutions per site (0.93/1). When the alignment was trimmed

only a maximum of five sites were retained, with some trimming methods retaining none. Overall, this highlights the difficulties in inferring meaningful phylogenies from highly divergent viral sequences.

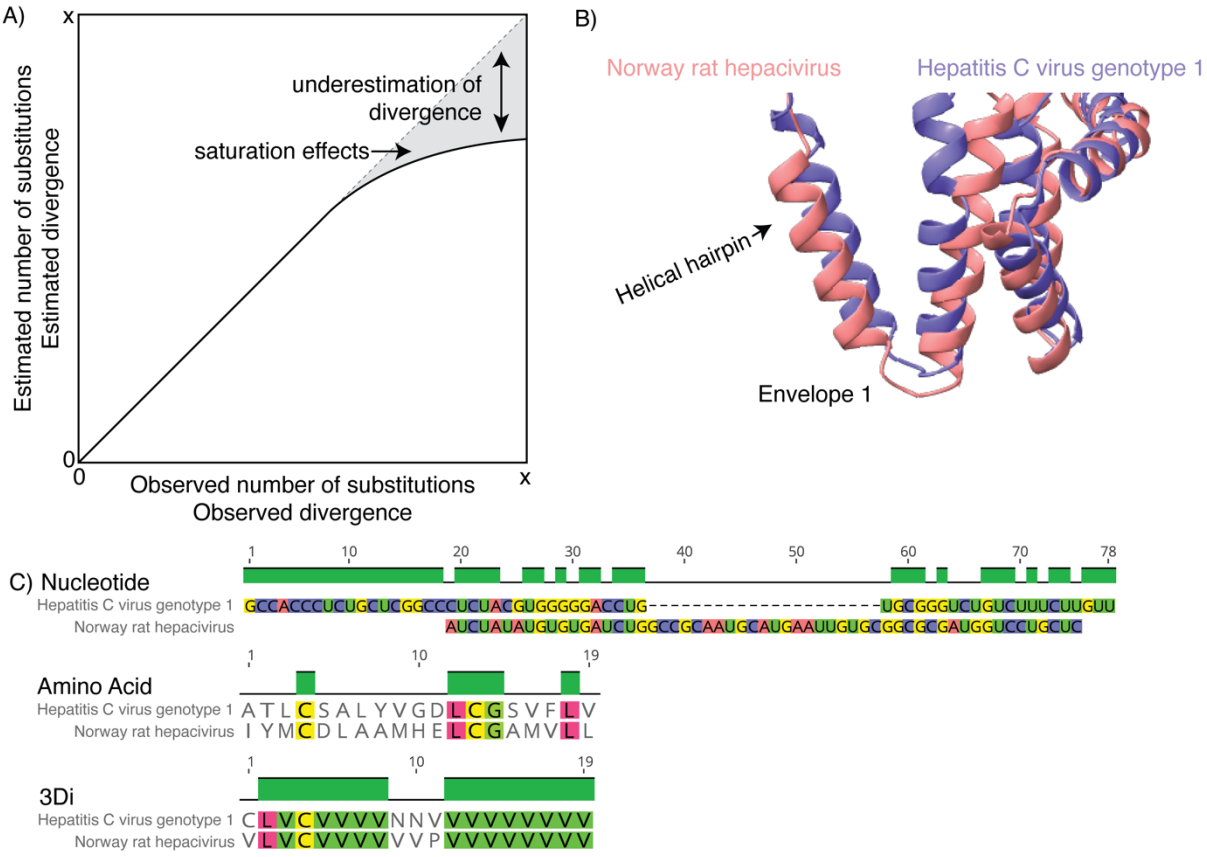


***Figure 1.6. Effect of saturation on estimating divergence and the increase in levels of conservation when comparing nucleotide, amino acid, and measures of structure.*** *(a) When comparing genetically divergent sequences, genetic saturation (i.e., multiple substitutions at the same site in a sequence, or identical substitutions in different sequences) can cause the loss of phylogenetic signal and underestimation of divergence with the number of the observed mutations (solid line) lower than the actual number of differences (dotted line). Figure reproduced from Wilke, T., Schultheiß, R., & Albrecht, C. (2009) (b) FATCAT (Li et al. 2020) pairwise structural alignment of ColabFold model of E1 structures for Hepatitis C virus genotype 1 (purple) and Norway rat hepacivirus 1 (pink) taken from Mifsud et al. (2024), structures have been cropped to focus on the helical hairpin. (c) Pairwise nucleotide, amino acid, and 3Di alignments of the helical hairpin region for Hepatitis C virus genotype 1 (NC_004102) and Norway rat hepacivirus 1 (NC_025672). Alignments were constructed with MAFFT E-INS-I (Katoh and Standley 2013), 3Di was obtained from the structures generated by Mifsud et al. (2024) using FoldMason (Gilchrist et al. 2024).*

### 1.10 Protein evolution and structure

As protein structures appear three to ten times more conserved than amino acid sequences (Illergård et al. 2009), structural information is the logical next step for identifying and characterising highly divergent viruses. Protein function is closely tied to protein structure and while structure is ultimately derived from sequence, the relationship between the two is complex (Krissinel 2007). For example, structures are highly likely to be homologous when sequence similarity is high, but the inverse is not always true. Sequences in the “twilight zone” (20-35% amino acid sequence identity) and below can still exhibit structural homology, particularly in the inner core of monomeric units (Krissinel 2007). Additionally, given the close tie to function, evolutionary pressure favours the maintenance of protein structure over sequence (i.e., selective constraints). In some cases, multiple amino acid mutations may occur without significantly altering structural integrity (Griffiths 2002). For example, homologous folds are observed across the E1 glycoproteins of pestiviruses, hepaciviruses, and pegiviruses despite these proteins sharing little to no similarity at the sequence level (Oliver et al. 2023) (Figure 1.6b). Given this, using comparisons of structural similarity represents a powerful way of inferring homology between highly divergent sequences and potentially leading to more accurate alignments (Hasegawa and Holm 2009)

and phylogenies (Lundin et al. 2012; Moi et al. 2023; Puente-Lelievre et al. 2024) (Figure 1.6c).

## 1.11 Protein structure prediction

Until recently, the determination of virus protein structure required labour-intensive, and expensive techniques, such as X-ray crystallography, nuclear magnetic resonance, and cryogenic electron microscopy. Consequently, the number of experimentally determined structures was relatively small—200,988 entries in the Protein Data Bank (PDB)—compared to the vast number of protein sequences in UniProtKB (229,580,745 as of January 2023), with only a small fraction of these structures belonging to viruses (Bertoline et al. 2023). The lack of available structures has limited large-scale evolutionary structural comparisons for viruses. In his acceptance speech for the 1972 Nobel Prize in Chemistry, Christian Anfinsen stated that in theory a protein's amino acid sequence should fully determine its structure (Anfinsen 1992). This marks the beginning of a five-decade-long effort to address this problem.

### 1.11.1 AlphaFold

The problem of predicting structure from amino acid has been solved in some part with protein structural prediction software using artificial intelligence (AI), namely AlphaFold (Abramson et al. 2024; Jumper et al. 2021; Senior et al. 2020). Indeed, the research leading to the development of AlphaFold was recently recognised with the Noble Prize in Chemistry in 2024 (Abriata 2024). The underlying methodology behind AlphaFold (specifically AlphaFold2) has been extensively documented in Jumper et al. (2021), but in brief it is comprised of three core parts. The first, takes an amino acid sequence as input and conducts similarity searches against a clustered amino acid database containing sequences from reference databases, metagenomes, and metatranscriptomes. The query and target sequences obtained from the database are then used to construct an MSA. In practice, this is now commonly done using ColabFold, which replaces the AlphaFold2 homology search with the 40–60-fold faster MMseqs2, among other optimisations (Mirdita et al. 2022; Steinegger and Söding 2017). The MSA is central to AlphaFold, providing key evolutionary information such as which residues are conserved, which residues are likely to vary, and the correlation between different columns are extracted (Figure 1.7). Using amino acid charge and hydrophobicity it can explore how amino acids may interact with each other and the effect of mutations observed in the MSA. In the second part known as the transformer module a numerical representation of the MSA and the amino acid residue

pairs (i.e., pairs within a sequence that are predicted to be in contact) are generated. This module is a type of neural network designed to learn the context and relationships between sequences (Vaswani 2017). Finally, the representation of the MSA and residue pairs is input into a structural module that predicts a 3D structure of the protein. This process is iterative, as the resulting information, including the representations and predicted structure are passed back through the three stages. The structures generated alongside the software's release outcompeted all other prediction methods (Pereira et al. 2021) and, in many cases compared equally to experimentally determined structures (Akdel et al. 2022), all while being produced at a fraction of the time and cost.

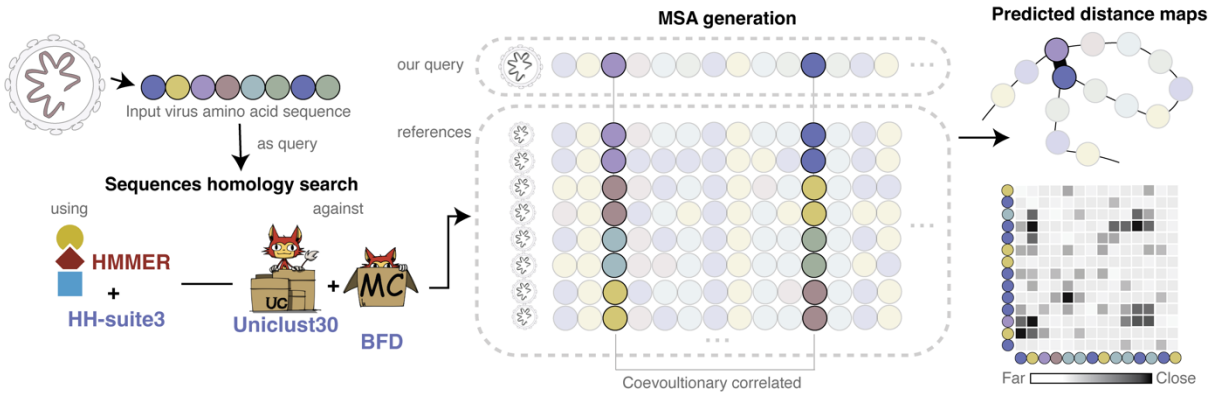



***Figure 1.7. Workflow of evolutionary information extraction in AlphaFold2.*** *This schematic illustrates the process of generating a multiple sequence alignment (MSA) from an input amino acid sequence, followed by the prediction of contact distance. The input query undergoes a homology search against reference databases (UniClust30 and BFD) using HMMER and HH-suite3. The resulting MSA captures evolutionary patterns, which are used to predict inter-residue distances, represented here as a distance map. The predicted distance maps panel was adapted from Marks et al. (2011) and "How does AlphaFold2 work?" Jarnestad, 2024, https://www.nobelprize.org/uploads/2024/10/fig2_ke_en_24.pdf.*

The speed and ease at which structures can now be generated is highlighted by the size of the AlphaFold Database (AlphaFold DB). First released in July 2021 by DeepMind, the team behind AlphaFold, it consisted of an initial 300,000 structures which has now expanded to over 214 million (Varadi et al. 2024), dwarfing the number of experimentally determined structures. However, viruses are currently excluded from the AlphaFold DB due to the need for improved support for polyproteins (https://alphafold.ebi.ac.uk/faq). This gap is being addressed by external databases (Nomburg et al. 2024; Steinegger et al. 2024).

### 1.11.2 Protein language models

Large language models (LLMs) similar to widely used systems like ChatGPT represent an alternative approach to understanding protein evolution (Achiam et al. 2023). LLMs are neural network AI systems that are trained on large masses of text data from which they are learning patterns and contextual usage of words. These models can used with any sequential data including biological data. In protein language models (pLMs) each residue is treated as a token akin to a word in a sentence. These models are unsupervised, meaning the training data doesn't contain manual labels. For example, in the context of protein structure, a specific amino acid sequence might be secondary structure tags, "alpha helix", or information about contact pairs. Instead, these models learn by masking and generating input training data, which, in this case, are amino acid sequences.

pLMs such as ESM-2 have been used in combination with a protein folding module, together called ESMFold, to generate structure predictions directly from an amino acid sequence (Lin et al. 2023). While ESMFold appears less accurate than AlphaFold in comparison to experimental benchmarks, it has advantages in speed and in its ability to infer structure directly from the sequence without requiring an MSA (Lin et al. 2023). This approach is particularly useful in "dark" regions of viral diversity where the MSA depth (i.e., the number of sequences with detectable homology) is low. This approach was recently incorporated into the deep-learning virus detection software LucaProt (Hou et al. 2024). Here, the incorporation of structural information predicted using ESMFold increased the model's specificity from 41.8% to 98.28% and accuracy from 94.9% to 99.98% and allowed it to find homologous sequence matches deeper into viral space (Hou et al. 2024).

## 1.12 Protein structure comparison

The abundance of available protein structures and the ease at which structures can be generated have opened countless research avenues. Central to these is the comparison of structures, which in turn enables evolutionary inferences to be made. One approach to detecting similarity between structures involves comparing geometric distances, typically done using three-dimensional (3D) superposition. This approach aims to align two or more 3D structures by minimising the distance (e.g., Root Mean Square Deviation (RMSD)) between their equivalent atoms. Here, lower RMSD scores refer to a higher degree of similarity between structures. A superposition approach is well-established through tools such as TM-align (Zhang and Skolnick 2005) or DALI (Holm and Rosenstrï 2010). Other structural comparison metrics have been extensively used such as template modelling (TM)

Score, which unlike RMSD is not dependent on inherent protein size and focuses on global structural topology (Zhang and Skolnick 2005) as well as Local Distance Difference Test (LDDT) which is a superposition-free score that examines the differences in local distances between atoms in the query and reference structure (Mariani et al. 2013). A key limitation of these methods is their speed and scalability. For example, it requires a month of computing time on a single CPU core to compare a single protein structure against a protein structure database of 100 million structures using TM-align (Zhang and Skolnick 2005). This makes all-versus-all comparisons unfeasible, even with large computational resources available (van Kempen et al. 2024).

One approach to improving the speed of structural alignments is to represent the amino acid backbone of proteins as characters and align these using traditional sequence alignment methods. Early methods, such as SA-search and 3D-BLAST focused on secondary structures (Guyon et al. 2004; Yang and Tung 2006). More recently, tools like FoldSeek and Reseek have incorporated tertiary structural information (Edgar 2024; van Kempen et al. 2024). FoldSeek addresses some of the limitations of geometric-based alignment (see Hasegawa and Holm (2009)) by encoding the interactions between each residue and its nearest neighbours by Euclidean distance. These interactions are summarised into one of 20 distinct character states for each residue, termed 3Di (Figure 1.8). This encoding reduces computation times by four to five orders of magnitude with only minor decreases in sensitivity compared to state-of-the-art geometric-based methods like Dali and TM-align (van Kempen et al. 2024). Consequently, large-scale structural searches and alignments are much more feasible.

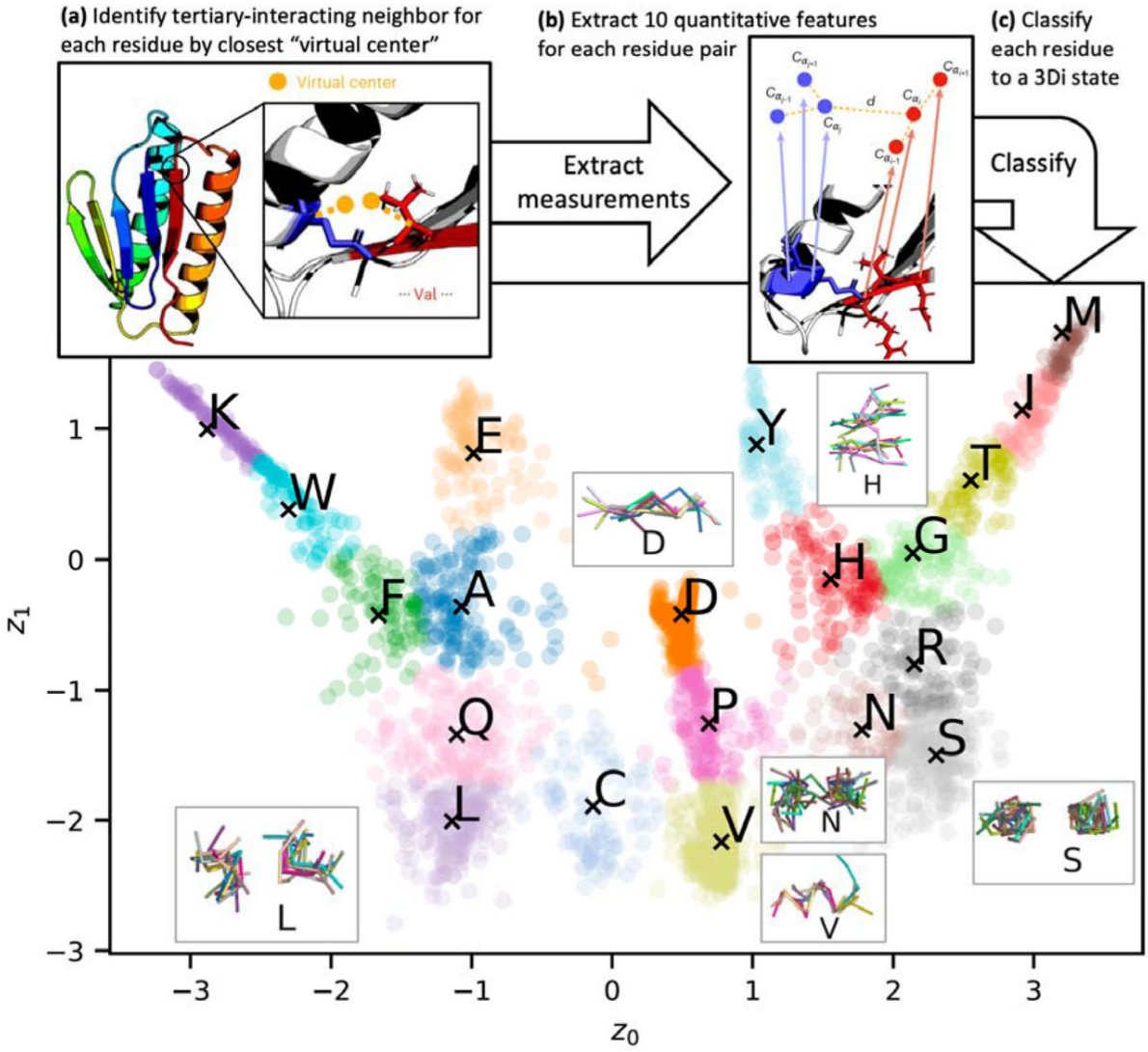



***Figure 1.8 Overview of FoldSeek's classification of tertiary interactions into a 20-state 3Di alphabet.*** *For each residue in the protein structure, (a) a virtual centre is defined, and the nearest* neighbour *based on the closest virtual centre is identified (b) 10 quantitative features containing neighbourhood descriptors are extracted for that residue and its pair (c) discretised into one of 20-character states using a quantised variational autoencoder. Figure reproduced from Puente-Lelievre et al. (2024)*

### 1.13 Protein structure phylogenetics

Given the resilience of structural features to saturation over deep evolutionary timescales, phylogeneticists and structural biologists have long sought to infer phylogenetic trees from protein structure (Eventoff et al. 1975). Various structural comparison measures, such as RMSD, LDDT, RMSD, Z-score, or $Q_{score}$, have been successfully used to estimate phylogenetic trees using methods like NJ (Bujnicki 2000; Garau et al. 2005; Lundin et al. 2012) and Bayesian phylogenetic analysis (Challis and Schmidler 2012; Herman et al. 2014).

Virus structural phylogenetics is not a new field, yet the number of studies and analysed structures remains limited (Butkovic et al. 2023; Laanto et al. 2017; Ravantti et al. 2020). This is in part due to the limitations inherited from structural distance methodologies in that they are computationally intensive and up until recently hampered by the lack of available structures. Where virus structural phylogenetic studies have been conducted, they typically infer NJ phylogenies from structural distances, such that evolutionary information is invariably lost in the estimation of pairwise distances. Further, unlike the well-defined distance matrices used for nucleotide and amino acids, those used for structural comparison matrices such as LDDT or TM-score remain largely unvalidated. Additional limitations include the lack of a probabilistic model describing the underlying generative processes, as discussed further in Puente-Lelievre et al. (2024), and the susceptibility of these matrices to distortion caused by conformational changes (Malik et al. 2020).

As a scalable solution to protein structure alignment, structural alphabets are increasingly being explored in structural phylogenetics (Garg and Hochberg 2024; Moi et al. 2023; Mutz et al. 2024; Puente-Lelievre et al. 2024). The 3Di alphabet of FoldSeek, which appears to retain signals of core conserved folds and exhibits high sensitivity in homology searches could represent a highly conserved and phylogenetically informative character (Puente-Lelievre et al. 2024). Recently, Moi et al. (2023) developed the software FoldTree and demonstrated the power of FoldSeek-based structural phylogenetics in their comparison of RRNPPA quorum sensing receptor structures. RRNPPA proteins allow gram-positive bacteria, and their associated plasmids and viruses to sense population density and regulate biological processes in response. Using outputs from FoldSeek they obtained LDDT, TM-score, and a custom distance metric based on the fraction of identical amino acids from 3Di and amino acid-based alignments (Fident). A key observation was that NJ phylogenies derived from the combination of sequence and structural alignments and the associated Fident distance, outperformed those based solely on LDDT or TM or sequence alignment on measures of taxonomic congruence with known taxonomy (Moi et al. 2023).

Another approach was proposed by Puente-Lelievre et al. (2024), who suggested that since 3Di is characterised by 20 distinct character states—the same number as amino acids—it can be directly incorporated into traditional bioinformatic and phylogenetic programs as characters. This integration opens the possibility of applying ML, Bayesian inference, and modern ultrafast statistical measures of phylogenetic uncertainty. As 3Di is highly conserved, and hence likely slow evolving, they may be particularly beneficial for resolving

deep phylogenetic nodes where amino acid signal approaches saturation (Figure 1.7). Using this approach, Puente-Lelievre et al. (2024) compared and combined 3Di states with amino acids, partition models and a 3DI-specific rate matrix to examine 53 structures from the ferritin-like superfamily. Overall, they found that the partition model phylogeny derived from both amino acids and 3Di avoided long-branch attraction errors observed in non-structural phylogenies and was orders of magnitude faster to compute. Although the validity of 3Di states as phylogenetically informative characters remains largely untested (see Puente-Lelievre et al. (2024) and Garg and Hochberg (2024)), this approach offers an exciting avenue for rapid and sensitive structural phylogenetics (Mutz et al. 2024). In Chapter 6, I follow this approach to infer phylogenies using 3Di character states representing *Flaviviridae* envelope proteins and demonstrate the effectiveness of this method for reconstruction deep evolutionary history.

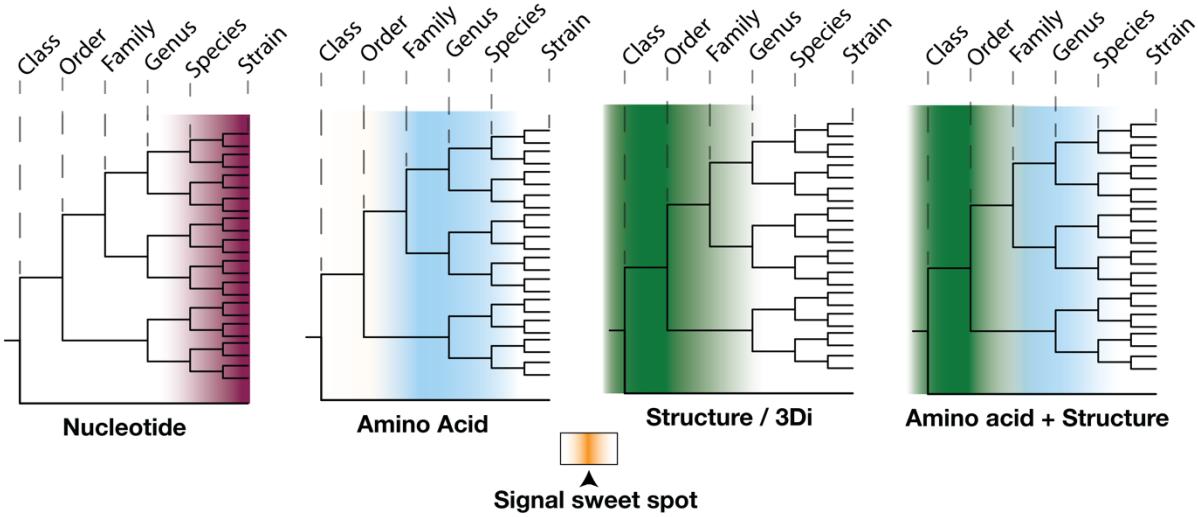



***Figure 1.9. Schematic illustrating the phylogenetic signal "sweet spot" of nucleotides, amino acids, and structural data across varying levels of sequence divergence.*** *The intensity of the gradient represents the hypothetical strength of evolutionary information provided by each data type, where darker areas indicate a stronger phylogenetic signal. The sweet spot is where the gradient intensity is the strongest. The combination of structural and amino acid information is shown as a potential approach to simultaneously resolve both deep evolutionary and finer-scale relationships (e.g., species level).*

### 1.14 Thesis rationale

The COVID-19 pandemic has highlighted the urgent need to better understand virus emergence, ecology and evolution. Unfortunately, our knowledge of viral diversity is largely restricted to those viruses that cause diseases in humans and agriculturally important plants and animals. Even less is known about the ecological and evolutionary processes that shape virus diversity despite the importance of understanding pathogen evolution. Thus, the central aim of my thesis was to examine the viral diversity in species critical to answering these questions. These were often under-sampled as a species or in a specific context (i.e., Australia) and included important sentinels for viral disease. Expanding the scope and timescale further, I investigated key evolutionary transitions in plants and metazoans. To achieve these aims, I combined metatranscriptomic sequencing and the secondary analysis of transcriptome data with state-of-the-art phylogenetic and structural analysis. Below, I provide summaries of each chapter, outlining the aims, approaches, and key findings.

In Chapter 2, I combine traditional veterinary diagnostic techniques and metatranscriptomic sequencing of tissue samples to investigate the virome of New Zealand fur seals. These animals are the predominant seal species in the Australian-New Zealand region and serve as important indicators of marine environmental health. Despite this, little is known about the infectious diseases affecting these animals. I identified a novel *Taupapillomavirus* in an individual exhibiting clinically and histologically confirmed oral papilloma-like lesions, marking the first known occurrence of this disease and virus in fur seals. This study highlights the value of combining metagenomic sequencing with conventional gross and histological examination to characterise potential aetiological agents in wildlife, ultimately contributing to disease detection, control, and conservation efforts.

In Chapter 3, I examine the virome of Australia's dogs and dingoes. Dogs are infected by numerous pathogens that cause a substantial disease burden. Due to their proximity to humans, these viruses may pose zoonotic risk. Australia's geographic isolation, strict biosecurity measures, and endemic dingo populations present a unique model for studying the spread and evolution of canine viruses. However, the genomic data on canine pathogens in Australia remains limited, which hampers preparedness for potential future outbreaks. Using a metatranscriptomic approach, I analysed dogs and dingoes from across various geographical locations and sample types, revealing the presence of 81 vertebrate-associated viruses belonging to 16 distinct species, including the first dingo-associated

RNA virus and important dog pathogens associated with canine infectious respiratory disease syndrome and gastroenteritis. Through this work I expand the knowledge of viral diversity in Australian canines which in turns improves our understanding of viral movement into and within Australia and the potential zoonotic risks associated with dogs and dingoes.

In Chapter 4, I mine publicly available metagenomic data from the 1000 Plant Transcriptomes Initiative (1KP) to describe virus diversity across the breadth of the plant kingdom. I identified 104 novel viruses, greatly expanding our understanding of RNA virus diversity in lower plants and algae. These results suggest that several RNA plant virus groups are associated with older evolutionary plant lineages than previously thought and that the apparent scarcity of RNA viruses in lower plants to date likely reflects a lack of investigation rather than their absence.

In Chapter 5, I adopt a similar approach by screening publicly available metazoan transcriptome data to identify novel viruses related to the *Flaviviridae.* The recent discovery of a flavivirus-derived endogenous viral element in a jellyfish (phylum Cnidaria) implies that the flaviviruses arose early in the evolution of the Metazoa. However, at the time of this study, no exogenous flavivirus sequences had been identified in these hosts. I identified 32 novel viral sequences, including a flavivirus-like sequence in an octocoral (phylum Cnidaria), and greatly expanded the known host range of the pestiviruses. Together, these data help resolve the timescale of flavivirus evolution, suggesting that the origin of orthoflaviviruses likely coincided with the emergence of metazoans approximately 750–800 million years ago.

Building on this data, in Chapter 6, I combined phylogenetic analyses with protein structure prediction to survey glycoproteins across the entire *Flaviviridae.* Canonical orthoflaviruses possess class II envelope proteins, while the molecular mechanisms of membrane fusion for the hepaciviruses, pegiviruses, and pestiviruses remain largely uncharacterised. Using phylogenetics combined with protein structure prediction, I identified previously unknown glycoproteins in over 100 species, revealing that the hepaciviruses, pegiviruses, and pestiviruses share a common membrane fusion syßstem that was strictly associated with vertebrate infection. In contrast, the class II system was found in a broader range of hosts and genomic compositions. These results further our understanding of viral fusion mechanisms and reveal the events that have shaped the evolution and ecology of the *Flaviviridae.*

Finally, in Chapter 7, I place the findings of my thesis in a broader context, discussing the implications, limitations and future directions of this work.

I also contributed to twelve papers during my candidature, which are featured in the Appendices. Appendix I is a metagenomic investigation of the viruses in the invasive common carp and native Australian fishes in the Murray–Darling Basin. Appendix II is an analysis of the evolutionary history of the *Coronaviridae*, extending the host phylogenetic range of this family to include a diverse array of fish. Appendix III provides an ecological analysis of viruses present in a small reef fish community from the Great Barrier Reef (GBR). Appendix IV provides evidence for an ancient aquatic origin for the *Articulavirales* using a combination of total RNA sequencing and transcriptome mining virus discovery methods. Appendix V is an analysis of publicly available transcriptome data to identify viruses in endangered Australian marsupial species. Appendix VI extends upon Appendix III, sampling an additional 65 fish species from another GBR island and characterising the overlap in viruses, bacteria, and single-celled eukaryotes between these islands. Appendix VII reports the discovery of a large-genome flavi-like virus that encodes a ~40kb genome with no known proofreading mechanism, suggesting that RNA viruses have evolved multiple solutions to overcome the mutational error threshold. Appendix VIII investigates the virome of farmland and sediment across Australia's east and west coast, identifying over 2000 novel viruses. Appendix IX examines how the rapid adaptive radiation of the cichlid fishes over the last 10 million years has shaped the diversity and evolution of the viruses they carry. Appendix X is a consensus statement that provides insights and recommendations for advancing RNA virus discovery. Appendix XI is a surveillance project of viral biosecurity threats to Australia's native and agricultural plants using gut metatranscriptomes of vertebrate animals and ticks. Appendix XII is a metatranscriptomic survey of the primary and publicly available RNA libraries of tunicates to identify RNA viruses.

Metatranscriptomic sequencing for virus discovery has proven a powerful tool for discovering and characterising viral diversity. Throughout my PhD research, I use this approach to identify aetiological agents, characterise highly divergent viruses, and expand the known host ranges of several virus families. Notably, through the focus on the *Flaviviridae*, I demonstrated how novel and highly divergent viruses—often discovered in hosts of limited economic importance—can be key to reconstructing virus origins and

evolutionary histories. These findings provide insights into the evolution of viral proteins that influence host range, tissue tropism, and pathogenesis.

### 1.15 References


Abdelwhab, E. M. and Mettenleiter, T. C. (2023), 'Zoonotic Animal Influenza Virus and Potential Mixing Vessel Hosts', *Viruses,* 15 (4), 980.

Abramson, J., et al. (2024), 'Accurate structure prediction of biomolecular interactions with AlphaFold 3', *Nature,* 630 (8016), 493-500.

Abriata, L. A. (2024), 'The Nobel Prize in Chemistry: past, present, and future of AI in biology', *Communications Biology,* 7 (1), 1409.

Achiam, J., et al. (2023), 'Gpt-4 technical report', *arXiv preprint arXiv:2303.08774*.

Akdel, M., et al. (2022), 'A structural biology community assessment of AlphaFold2 applications', *Nature Structural & Molecular Biology,* 29 (11), 1056-67.

Allison, A. B., et al. (2015), 'Cyclic Avian Mass Mortality in the Northeastern United States Is Associated with a Novel Orthomyxovirus', *Journal of Virology,* 89 (2), 1389-403.

Altschul, S. F., et al. (1990), 'Basic local alignment search tool', *Journal of Molecular Biology,* 215 (3), 403-10.

Ambrose, H. E. and Clewley, J. P. (2006), 'Virus discovery by sequence-independent genome amplification', *Reviews in Medical Virology,* 16 (6), 365-83.

Anderson, M. E., et al. (2018), 'Highly sensitive virome capture sequencing technique VirCapSeq-VERT identifies partial noncoding sequences but no active viral infection in cutaneous T-cell lymphoma', *Journal of Investigative Dermatology,* 138 (7), 1671-73.

Andrews, S. (2017), 'FastQC: a quality control tool for high throughput sequence data. 2010'.

Anfinsen, C. B. (1992), 'Studies on the principles that govern the folding of protein chains', *Chemistry, 1971-1980,* 5, 55.

Badizadegan, K., Kalkowska, D. A., and Thompson, K. M. (2022), 'Polio by the Numbers—A Global Perspective ', *The Journal of Infectious Diseases,* 226 (8), 1309-18.

Baltimore, D. (1971), 'Expression of animal virus genomes', *Bacteriological Reviews,* 35 (3), 235-41.

Bamford, C. G., et al. (2022), 'Comparative analysis of genome-encoded viral sequences reveals the evolutionary history of flavivirids (family Flaviviridae)', *Virus Evolution,* 8 (2).

Bartlett, M. L. and Uhart, M. (2024), 'Leveraging one health as a sentinel approach for pandemic resilience', *Virology Journal,* 21 (1), 269.

Benes, V., Blake, J., and Doyle, K. (2011), 'Ribo-Zero Gold Kit: improved RNA-seq results after removal of cytoplasmic and mitochondrial ribosomal RNA', *Nature Methods*.

Bergner, L. M., et al. (2021), 'Diversification of mammalian deltaviruses by host shifting', *Proceedings of the National Academy of Sciences,* 118 (3), e2019907118.

Bertoline, L. M., et al. (2023), 'Before and after AlphaFold2: An overview of protein structure prediction', *Frontiers in Bioinformatics,* 3, 1120370.

Bigot, T., et al. (2019), 'RVDB-prot, a reference viral protein database and its HMM profiles', *F1000Research,* 8.

Blum, M., et al. (2024), 'InterPro: the protein sequence classification resource in 2025', *Nucleic Acids Research*.

Bolger, A. M., Lohse, M., and Usadel, B. (2014), 'Trimmomatic: a flexible trimmer for Illumina sequence data', *Bioinformatics,* 30 (15), 2114-20.

Bossart, G. D. (2011), 'Marine mammals as sentinel species for oceans and human health', *Veterinary Pathology,* 48 (3), 676-90.

Bowser, N. H. and Anderson, N. E. (2018), 'Dogs (Canis familiaris) as Sentinels for Human Infectious Disease and Application to Canadian Populations: A Systematic Review', *Veterinary Sciences,* 5 (4), 83.

Brait, N., et al. (2024), 'A tale of caution: How endogenous viral elements affect virus discovery in transcriptomic data', *Virus Evolution,* 10 (1), vead088.

Briese, T., et al. (2015), 'Virome capture sequencing enables sensitive viral diagnosis and comprehensive virome analysis', *MBio,* 6 (5), 01491-15.

Brister, J. R., et al. (2015), 'NCBI viral genomes resource', *Nucleic Acids Research,* 43, 571-7.

Brown, F. (2003), 'The history of research in foot-and-mouth disease', *Virus Research,* 91 (1), 3-7.

Buchfink, B., Reuter, K., and Drost, H.-G. (2021), 'Sensitive protein alignments at tree-of-life scale using DIAMOND', *Nature Methods,* 18 (4), 366-68.

Buetre, B. (2013), 'Potential socio-economic impacts of an outbreak of foot-and-mouth disease in Australia'.

Bujnicki, J. M. (2000), 'Phylogeny of the restriction endonuclease-like superfamily inferred from comparison of protein structures', *Journal of Molecular Evolution,* 50 (1), 39-44.

Bushnell, B. (2014), 'BBMap: a fast, accurate, splice-aware aligner', (Lawrence Berkeley National Lab (LBNL), Berkeley, CA.).

Butkovic, A., et al. (2023), 'Plant virus movement proteins originated from jelly-roll capsid proteins', *PLoS Biology,* 21 (6), e3002157.

Camacho, C., et al. (2009), 'BLAST+: architecture and applications', *BMC Bioinformatics,* 10 (1), 1-9.

Capella-Gutiérrez, S., Silla-Martínez, J. M., and Gabaldón, T. (2009), 'trimAl: a tool for automated alignment trimming in large-scale phylogenetic analyses', *Bioinformatics,* 25 (15), 1972-73.

Carpenter, E. J., et al. (2019), 'Access to RNA-sequencing data from 1,173 plant species: The 1000 Plant transcriptomes initiative (1KP)', *GigaScience,* 8 (10).

Casadevall, A. and Pirofski, L. a. (2001), 'Host-Pathogen Interactions: The Attributes of Virulence', *The Journal of Infectious Diseases,* 184 (3), 337-44.

Challis, C. J. and Schmidler, S. C. (2012), 'A stochastic evolutionary model for protein structure alignment and phylogeny', *Molecular Biology and Evolution,* 29 (11), 3575-87.

Chang, W. S., et al. (2024), 'Improving the reporting of metagenomic virome-scale data', *Zenodo*.

Charon, J., et al. (2022), 'RdRp-scan: A bioinformatic resource to identify and annotate divergent RNA viruses in metagenomic sequence data', *Virus Evolution,* 8 (2), veac082.

Charon, J., et al. (2024), 'Consensus statement from the first RdRp Summit: advancing RNA virus discovery at scale across communities', *Frontiers in Virology,* 4, 1371958.

Chikhi, R., et al. (2024), 'Logan: Planetary-Scale Genome Assembly Surveys Life's Diversity', *bioRxiv*, 2024.07.30.605881.

Clausen, P. T., Aarestrup, F. M., and Lund, O. (2018), 'Rapid and precise alignment of raw reads against redundant databases with KMA', *BMC Bioinformatics,* 19 (1), 1-8.

Crochu, S., et al. (2004), 'Sequences of flavivirus-related RNA viruses persist in DNA form integrated in the genome of Aedes spp. mosquitoes', *Journal of General Virology,* 85 (7), 1971-80.

d'Herelle, M. (1961), 'Sur un microbe invisible antagoniste des bacilles dysentériques', *Acta Kravsi*.

Dane, D., Cameron, C., and Briggs, M. (1970), 'Virus-like particles in serum of patients with Australia-antigen-associated hepatitis', *The Lancet,* 295 (7649), 695-98.

Debat, H. and Bejerman, N. (2023), 'Two novel flavi-like viruses shed light on the plant-infecting koshoviruses', *Archives of Virology,* 168 (7), 184.

Debat, H., Garcia, M. L., and Bejerman, N. (2023), 'Expanding the Repertoire of the Plant-Infecting Ophioviruses through Metatranscriptomics Data', *Viruses,* 15 (4), 840.

Dudas, G. and Batson, J. (2023), 'Accumulated metagenomic studies reveal recent migration, whole genome evolution, and undiscovered diversity of orthomyxoviruses', *Journal of Virology,* 97 (10), e01056-23.

Duffy, S., Shackelton, L. A., and Holmes, E. C. (2008), 'Rates of evolutionary change in viruses: patterns and determinants', *Nature Reviews Genetics,* 9 (4), 267-76.

Duncan, A. W., et al. (2005), 'The Dog as a Sentinel for Human Infection: Prevalence of Borrelia burgdorferi C6 Antibodies in Dogs from Southeastern and Mid-Atlantic States', *Vector-Borne and zoonotic diseases,* 5 (2), 101-09.

Edgar, R. C. (2022), 'Muscle5: High-accuracy alignment ensembles enable unbiased assessments of sequence homology and phylogeny', *Nature Communications,* 13 (1), 6968.

Edgar, R. C. (2024), 'Protein structure alignment by reseek improves sensitivity to remote homologs', *Bioinformatics*.

Edgar, R. C., et al. (2022), 'Petabase-scale sequence alignment catalyses viral discovery', *Nature,* 602 (7895), 142-47.

Enders, J. F., Weller, T. H., and Robbins, F. C. (1949), 'Cultivation of the Lansing strain of poliomyelitis virus in cultures of various human embryonic tissues', *Science,* 109 (2822), 85-87.

Eventoff, W., Rossmann, M. G., and Brändén, C.-I. (1975), 'The Evolution of Dehydrogenases and Kinase', *CRC Critical Reviews in Biochemistry,* 3 (2), 111-40.

Forgia, M., et al. (2023), 'Hybrids of RNA viruses and viroid-like elements replicate in fungi', *Nature Communications,* 14 (1), 2591.

Garau, G., Di Guilmi, A. M., and Hall, B. G. (2005), 'Structure-based phylogeny of the metallo- $\beta$ -lactamases', *Antimicrobial Agents and Chemotherapy,* 49 (7), 2778-84.

Garg, S. G. and Hochberg, G. K. (2024), 'A general substitution matrix for structural phylogenetics', *bioRxiv*, 2024.09.19.613819.

Geddes, A. M. (2006), 'The history of smallpox', *Clinics in Dermatology,* 24 (3), 152-57.

Geoghegan, J. L. and Holmes, E. C. (2018), 'The phylogenomics of evolving virus virulence', *Nature Reviews Genetics,* 19 (12), 756-69.

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 (2), e1006215.

Ghasemzadeh, I. and Namazi, S. (2015), 'Review of bacterial and viral zoonotic infections transmitted by dogs', *Journal of Medicine and Life,* 8, 1.

Gilchrist, C. L., Mirdita, M., and Steinegger, M. (2024), 'Multiple Protein Structure Alignment at Scale with FoldMason', *bioRxiv*, 2024.08. 01.606130.

Goodacre, N., et al. (2018), 'A reference viral database (RVDB) to enhance bioinformatics analysis of high-throughput sequencing for novel virus detection', *mSphere,* 3 (2), e00069-18.

Griffiths, A. J. (2002), *Modern genetic analysis: integrating genes and genomes* (1: Macmillan).

Guyon, F., et al. (2004), 'SA-Search: a web tool for protein structure mining based on a Structural Alphabet', *Nucleic Acids Research,* 32 (suppl_2), W545-W48.

Hampson, K., et al. (2015), 'Estimating the global burden of endemic canine rabies', *PLoS Neglected Tropical Diseases,* 9 (4), e0003709.

Han, D., et al. (2019), 'mNGS in clinical microbiology laboratories: on the road to maturity', *Critical Reviews in Microbiology,* 45 (5-6), 668-85.

Handelsman, J., et al. (1998), 'Molecular biological access to the chemistry of unknown soil microbes: a new frontier for natural products', *Chemistry & Biology,* 5 (10), R245-R49.

Harding, E. F., et al. (2022), 'Revealing the uncharacterised diversity of amphibian and reptile viruses', *ISME Communications,* 2 (1), 95.

Harvey, E. and Holmes, E. C. (2022), 'Diversity and evolution of the animal virome', *Nature Reviews Microbiology,* 20 (6), 321-34.

Harvey, E., et al. (2019), 'Extensive Diversity of RNA Viruses in Australian Ticks', *Journal of Virology,* 93 (3).

Hasegawa, H. and Holm, L. (2009), 'Advances and pitfalls of protein structural alignment', *Current Opinion in Structural Biology,* 19 (3), 341-48.

He, W.-T., et al. (2022), 'Virome characterization of game animals in China reveals a spectrum of emerging pathogens', *Cell,* 185 (7), 1117-29.e8.

Herman, J. L., et al. (2014), 'Simultaneous Bayesian estimation of alignment and phylogeny under a joint model of protein sequence and structure', *Molecular Biology and Evolution,* 31 (9), 2251-66.

Ho, A., et al. (2023), 'Adeno-associated virus 2 infection in children with non-A–E hepatitis', *Nature,* 617 (7961), 555-63.

Hoffmann, B., et al. (2012), 'Novel orthobunyavirus in cattle, Europe, 2011', *Emerging Infectious Diseases,* 18 (3), 469.

Holm, L. and Rosenstrï, P. (2010), 'Dali server: conservation mapping in 3D', *Nucleic Acids Research,* 38 (suppl_2), W545-W49.

Holmes, E. C. (2011), 'What does virus evolution tell us about virus origins?', *Journal of Virology,* 85 (11), 5247-51.

Holmes, E. C. and Duchêne, S. (2019), 'Can sequence phylogenies safely infer the origin of the global virome?', *MBio,* 10 (2), 00289-19.

Hone, S., et al. (2022), 'Discount Rates and Risk in the Economic Analysis of Agricultural Projects', *ABARES: Canberra, Australia*.

Hou, X., et al. (2024), 'Using artificial intelligence to document the hidden RNA virosphere', *Cell,* 187 (24), 6929-42.e16.

Hovern, D. (2018), 'The Trials and Triumphs of the American Polio Vaccine'.

Illergård, K., Ardell, D. H., and 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.

Jones, P., et al. (2014), 'InterProScan 5: genome-scale protein function classification', *Bioinformatics,* 30 (9), 1236-40.

Jumper, J., et al. (2021), 'Highly accurate protein structure prediction with AlphaFold', *Nature,* 596 (7873), 583-89.

Kalyaanamoorthy, S., et al. (2017), 'ModelFinder: fast model selection for accurate phylogenetic estimates', *Nature Methods,* 14 (6), 587-89.

Katoh, K. and Standley, D. M. (2013), 'MAFFT multiple sequence alignment software version 7: improvements in performance and usability', *Molecular Biology and Evolution,* 30 (4), 772-80.

Katz, K., et al. (2022), 'The Sequence Read Archive: a decade more of explosive growth', *Nucleic Acids Research,* 50 (D1), D387-D90.

Katzourakis, A. and Gifford, R. J. (2010), 'Endogenous viral elements in animal genomes', *PLoS Genetics,* 6 (11), e1001191.

Kausche, G. A., Pfankuch, E., and Ruska, H. (1939), 'Die Sichtbarmachung von pflanzlichem virus im Übermikroskop', *Naturwissenschaften,* 27, 292-99.

Kearse, M., et al. (2012), 'Geneious Basic: an integrated and extendable desktop software platform for the organization and analysis of sequence data', *Bioinformatics,* 28 (12), 1647-49.

Könenkamp, L., et al. (2022), 'Antibody ratios against NS1 antigens of tick-borne encephalitis and West Nile viruses support differential flavivirus serology in dogs', *Transboundary and Emerging Diseases,* 69 (5), e2789-e99.

Koonin, E. V., Krupovic, M., and Agol, V. I. (2021), 'The Baltimore classification of viruses 50 years later: how does it stand in the light of virus evolution?', *Microbiology and Molecular Biology Reviews,* 85 (3), 00053-21.

Koonin, E. V., et al. (2024), 'Megataxonomy and global ecology of the virosphere', *The ISME Journal,* 18 (1).

Kopylova, E., Noé, L., and Touzet, H. (2012), 'SortMeRNA: fast and accurate filtering of ribosomal RNAs in metatranscriptomic data', *Bioinformatics,* 28 (24), 3211-17.

Krissinel, E. (2007), 'On the relationship between sequence and structure similarities in proteomics', *Bioinformatics,* 23 (6), 717-23.

Kruse, H., kirkemo, A.-M., and Handeland, K. (2004), 'Wildlife as source of zoonotic infections', *Emerging Infectious Diseases,* 10 (12), 2067-72.

Kwok, K. T., et al. (2020), 'Virus metagenomics in farm animals: a systematic review', *Viruses,* 12 (1), 107.

Laanto, E., et al. (2017), 'Virus found in a boreal lake links ssDNA and dsDNA viruses', *Proceedings of the National Academy of Sciences,* 114 (31), 8378-83.

Langmead, B. and Salzberg, S. L. (2012), 'Fast gapped-read alignment with Bowtie 2', *Nature Methods,* 9 (4), 357-59.

Lauber, C., et al. (2024), 'Deep mining of the Sequence Read Archive reveals major genetic innovations in coronaviruses and other nidoviruses of aquatic vertebrates', *Plos Pathogens,* 20 (4), e1012163.

Le, S. Q. and Gascuel, O. (2008), 'An improved general amino acid replacement matrix', *Molecular Biology and Evolution,* 25 (7), 1307-20.

Leebens-Mack, J. H., et al. (2019), 'One thousand plant transcriptomes and the phylogenomics of green plants', *Nature,* 574 (7780), 679-85.

Leinonen, R., et al. (2010), 'The sequence read archive', *Nucleic Acids Research,* 39, 19-21.

Li, B. and Dewey, C. N. (2011), 'RSEM: accurate transcript quantification from RNA-Seq data with or without a reference genome', *BMC Bioinformatics,* 12 (1), 323.

Li, C.-X., et al. (2015a), 'Unprecedented genomic diversity of RNA viruses in arthropods reveals the ancestry of negative-sense RNA viruses', *Elife,* 4, e05378.

Li, D., et al. (2015b), 'MEGAHIT: an ultra-fast single-node solution for large and complex metagenomics assembly via succinct de Bruijn graph', *Bioinformatics,* 31 (10), 1674-76.

Li, Y., et al. (2022), 'Endogenous Viral Elements in Shrew Genomes Provide Insights into Pestivirus Ancient History', *Molecular Biology and Evolution,* 39 (10), msac190.

Li, Z., et al. (2020), 'FATCAT 2.0: towards a better understanding of the structural diversity of proteins', *Nucleic Acids Research,* 48 (W1), W60-W64.

Lin, Z., et al. (2023), 'Evolutionary-scale prediction of atomic-level protein structure with a language model', *Science,* 379 (6637), 1123-30.

Löffler, F. and Frosch, P. (1898), 'Report of the commission for research on foot-and-mouth disease', *Zent. Bakt. Parasitkde. Abt. I,* 23, 371-91.

Lundin, D., et al. (2012), 'Use of structural phylogenetic networks for classification of the ferritin-like superfamily', *Journal of Biological Chemistry,* 287 (24), 20565-75.

Malik, A. J., Poole, A. M., and Allison, J. R. (2020), 'Structural Phylogenetics with Confidence', *Molecular Biology and Evolution,* 37 (9), 2711-26.

Marcelino, V. R., et al. (2020), 'CCMetagen: comprehensive and accurate identification of eukaryotes and prokaryotes in metagenomic data', *Genome Biology,* 21 (1), 103.

Mariani, V., et al. (2013), 'lDDT: a local superposition-free score for comparing protein structures and models using distance difference tests', *Bioinformatics,* 29 (21), 2722-28.

Marks, D. S., et al. (2011), 'Protein 3D Structure Computed from Evolutionary Sequence Variation', *PLOS One,* 6 (12), e28766.

Mifsud, J. C. O., et al. (2024), 'Mapping glycoprotein structure reveals Flaviviridae evolutionary history', *Nature*, 1-9.

Mihalov-Kovács, E., et al. (2015), 'Candidate new rotavirus species in sheltered dogs, Hungary', *Emerging Infectious Diseases,* 21 (4), 660.

Miller, A. K., et al. (2021), 'Slippery when wet: cross-species transmission of divergent coronaviruses in bony and jawless fish and the evolutionary history of the Coronaviridae', *Virus Evolution,* 7 (2).

Minh, B. Q., et al. (2020), 'IQ-TREE 2: New Models and Efficient Methods for Phylogenetic Inference in the Genomic Era', *Molecular Biology and Evolution,* 37 (5), 1530-34.

Mirdita, M., et al. (2022), 'ColabFold: making protein folding accessible to all', *Nature Methods,* 19 (6), 679-+.

Moi, D., et al. (2023), 'Structural phylogenetics unravels the evolutionary diversification of communication systems in gram-positive bacteria and their viruses', *bioRxiv*.

Morfopoulou, S., et al. (2023), 'Genomic investigations of unexplained acute hepatitis in children', *Nature,* 617 (7961), 564-73.

Moriones, E. and Navas-Castillo, J. (2000), 'Tomato yellow leaf curl virus, an emerging virus complex causing epidemics worldwide', *Virus Research,* 71 (1-2), 123-34.

Mühlemann, B., et al. (2020), 'Diverse variola virus (smallpox) strains were widespread in northern Europe in the Viking Age', *Science,* 369 (6502), eaaw8977.

Mullis, K., et al. (1986), 'Specific enzymatic amplification of DNA in vitro: the polymerase chain reaction', *Cold Spring Harbor symposia on quantitative biology* (51: Cold Spring Harbor Laboratory Press), 263-73.

Mutz, P., et al. (2024), 'The protein structurome of Orthornavirae and its dark matter', *bioRxiv*, 2024.09.25.615016.

Nathanson, N. and Kew, O. M. (2010), 'From emergence to eradication: the epidemiology of poliomyelitis deconstructed', *American Journal of Epidemiology,* 172 (11), 1213-29.

Neri, U., et al. (2022), 'Expansion of the global RNA virome reveals diverse clades of bacteriophages', *Cell,* 185 (21), 4023-37. e18.

Neumann, G. and Kawaoka, Y. (2024), 'Highly pathogenic H5N1 avian influenza virus outbreak in cattle: the knowns and unknowns', *Nature Reviews Microbiology,* 22 (9), 525-26.

Nomburg, J., et al. (2024), 'Birth of protein folds and functions in the virome', *Nature*, 1-8.

O'Neil, D., Glowatz, H., and Schlumpberger, M. (2013), 'Ribosomal RNA depletion for efficient use of RNA-seq capacity', *Current Protocols in Molecular Biology,* 103 (1), 4.19. 1-4.19. 8.

Oliver, M. R., et al. (2023), 'Structures of the Hepaci-, Pegi-, and Pestiviruses envelope proteins suggest a novel membrane fusion mechanism', *PLoS Biology,* 21 (7).

Oude Munnink, B. B., et al. (2021), 'Transmission of SARS-CoV-2 on mink farms between humans and mink and back to humans', *Science,* 371 (6525), 172-77.

Palacios, G., et al. (2008), 'A new arenavirus in a cluster of fatal transplant-associated diseases', *New England Journal of Medicine,* 358 (10), 991-98.

Paraskevopoulou, S., et al. (2021), 'Viromics of extant insect orders unveil the evolution of the flavi-like superfamily', *Virus Evolution,* 7 (1).

Parize, P., et al. (2017), 'Untargeted next-generation sequencing-based first-line diagnosis of infection in immunocompromised adults: a multicentre, blinded, prospective study', *Clinical Microbiology and Infection,* 23 (8), 574. e1-74. e6.

Parris, D. J., Kariithi, H., and Suarez, D. L. (2022), 'Non-target RNA depletion strategy to improve sensitivity of next-generation sequencing for the detection of RNA viruses in poultry', *Journal of Veterinary Diagnostic Investigation,* 34 (4), 638-45.

Parry, R., et al. (2020), 'Divergent Influenza-Like Viruses of Amphibians and Fish Support an Ancient Evolutionary Association', *Viruses,* 12 (9), 1042.

Patterson, E. I., et al. (2020), 'Evidence of exposure to SARS-CoV-2 in cats and dogs from households in Italy', *Nature Communications,* 11 (1), 6231.

Pereira, J., et al. (2021), 'High-accuracy protein structure prediction in CASP14', *Proteins: Structure, Function, and Bioinformatics,* 89 (12), 1687-99.

Petrone, M. E., et al. (2023), 'Evidence for an ancient aquatic origin of the RNA viral order Articulavirales', *Proceedings of the National Academy of Sciences,* 120 (45), e2310529120.

Prasad, A., et al. (2020), 'Tomato Yellow Leaf Curl Virus: Impact, Challenges, and Management', *Trends in Plant Science,* 25 (9), 897-911.

Puente-Lelievre, C., et al. (2024), 'Tertiary-interaction characters enable fast, model-based structural phylogenetics beyond the twilight zone', *bioRxiv*, 2023.12.12.571181.

Puryear, W., et al. (2023), 'Highly Pathogenic Avian Influenza A(H5N1) Virus Outbreak in New England Seals, United States', *Emerging Infectious Diseases,* 29 (4), 786-91.

Rajagopala, S. V., et al. (2021), 'Metatranscriptomics to characterize respiratory virome, microbiome, and host response directly from clinical samples', *Cell reports methods,* 1 (6).

Ravantti, J. J., Martinez-Castillo, A., and Abrescia, N. G. A. (2020), 'Superimposition of Viral Protein Structures: A Means to Decipher the Phylogenies of Viruses', *Viruses-Basel,* 12 (10).

Reyes, G. R. and Kim, J. P. (1991), 'Sequence-independent, single-primer amplification (SISPA) of complex DNA populations', *Molecular and cellular probes,* 5 (6), 473-81.

Rivarez, M. P. S., et al. (2023), 'In-depth study of tomato and weed viromes reveals undiscovered plant virus diversity in an agroecosystem', *Microbiome,* 11 (1), 60.

Rocheleau, J. P., et al. (2017), 'Emerging arboviruses in Quebec, Canada: assessing public health risk by serology in humans, horses and pet dogs', *Epidemiology and Infection,* 145 (14), 2940-48.

Roelandt, S., et al. (2011), 'Tick-Borne Encephalitis Virus Seropositive Dog Detected in Belgium: Screening of the Canine Population as Sentinels for Public Health', *Vector-Borne and zoonotic diseases,* 11 (10), 1371-76.

Sanger, F., Nicklen, S., and Coulson, A. R. (1977), 'DNA sequencing with chain-terminating inhibitors', *Proceedings of the National Academy of Sciences,* 74 (12), 5463-67.

Sanschagrin, S. and Yergeau, E. (2014), 'Next-generation sequencing of 16S ribosomal RNA gene amplicons', *Journal of visualized experiments: JoVE,* (90), 51709.

Schirmer, M., et al. (2016), 'Illumina error profiles: resolving fine-scale variation in metagenomic sequencing data', *BMC Bioinformatics,* 17 (1), 125.

Senior, A. W., et al. (2020), 'Improved protein structure prediction using potentials from deep learning', *Nature,* 577 (7792), 706-10.

Shi, M., et al. (2018), 'The evolutionary history of vertebrate RNA viruses', *Nature,* 556 (7700), 197-202.

Shi, M., et al. (2016), 'Redefining the invertebrate RNA virosphere', *Nature,* 540 (7634), 539.

Sievers, F., et al. (2011), 'Fast, scalable generation of high-quality protein multiple sequence alignments using Clustal Omega', *Molecular Systems Biology,* 7 (1), 539.

Simmonds, P. (2015), 'Methods for virus classification and the challenge of incorporating metagenomic sequence data', *Journal of General Virology,* 96 (6), 1193-206.

Simmonds, P., et al. (2024), 'Changes to virus taxonomy and the ICTV Statutes ratified by the International Committee on Taxonomy of Viruses (2024)', *Archives of Virology,* 169 (11), 236.

Simmonds, P., et al. (2017), 'Virus taxonomy in the age of metagenomics', *Nature Reviews Microbiology,* 15 (3), 161-68.

Sonne, C., et al. (2020), 'A review of pathogens in selected Baltic Sea indicator species', *Environment International,* 137, 105565.

Steinegger, M. and Söding, J. (2017), 'MMseqs2 enables sensitive protein sequence searching for the analysis of massive data sets', *Nature Biotechnology,* 35 (11), 1026-28.

Steinegger, M., Levy Karin, E., and Kim, R. S. (2024), 'BFVD-a large repository of predicted viral protein structures', *bioRxiv*, 2024.09. 08.611582.

Talavera, G. and Castresana, J. (2007), 'Improvement of phylogenies after removing divergent and ambiguously aligned blocks from protein sequence alignments', *Systematic Biology,* 56 (4), 564-77.

Tan, C. C. S., et al. (2019), 'A novel method for the capture-based purification of whole viral native RNA genomes', *AMB Express,* 9 (1), 45.

Tan, C. C. S., et al. (2022), 'Transmission of SARS-CoV-2 from humans to animals and potential host adaptation', *Nature Communications,* 13 (1), 2988.

Van Dijk, E. L., et al. (2014), 'Ten years of next-generation sequencing technology', *Trends in Genetics,* 30 (9), 418-26.

van Kempen, M., et al. (2024), 'Fast and accurate protein structure search with Foldseek', *Nature Biotechnology,* 42 (2).

Varadi, M., et al. (2024), 'AlphaFold Protein Structure Database in 2024: providing structure coverage for over 214 million protein sequences', *Nucleic Acids Research,* 52 (D1), D368-D75.

Vaswani, A. (2017), 'Attention is all you need', *Advances in Neural Information Processing Systems*.

Venclovas, Č. (2012), 'Methods for sequence–structure alignment', *Homology Modeling: Methods and Protocols*, 55-82.

Visser, M., et al. (2016), 'Next-generation sequencing for virus detection: covering all the bases', *Virology Journal,* 13 (1), 85.

Wilke, T., Schultheiß, R., and Albrecht, C. (2009), 'As time goes by: a simple fool's guide to molecular clock approaches in invertebrates', *American Malacological Bulletin,* 27 (1/2), 25-45.

Wolf, Y. I., et al. (2018), 'Origins and evolution of the global RNA cirome', *MBio,* 9 (6).

Wolf Yuri, I., et al. (2019), 'Reply to Holmes and Duchêne, “Can Sequence Phylogenies Safely Infer the Origin of the Global Virome?”: Deep Phylogenetic Analysis of RNA Viruses Is Highly Challenging but Not Meaningless', *MBio,* 10 (2), 10.1128/mbio.00542-19.

Wu, F., et al. (2020), 'A new coronavirus associated with human respiratory disease in China', *Nature,* 579 (7798), 265-69.

Xu, B., et al. (2011), 'Metagenomic analysis of fever, thrombocytopenia and leukopenia syndrome (FTLS) in Henan Province, China: discovery of a new bunyavirus', *PLoS pathogens,* 7 (11), e1002369.

Xu, Y., et al. (2022a), 'Identification of diverse viruses associated with grasshoppers unveils the parallel relationship between host phylogeny and virome composition', *Virus Evolution,* 8 (2).

Xu, Z., et al. (2022b), 'Virome of Bat-Infesting Arthropods: Highly Divergent Viruses in Different Vectors', *Journal of Virology,* 96 (4), e01464-21.

Yang, J.-M. and Tung, C.-H. (2006), 'Protein structure database search and evolutionary classification', *Nucleic Acids Research,* 34 (13), 3646-59.

Yang, Z. (1994), 'Maximum likelihood phylogenetic estimation from DNA sequences with variable rates over sites: approximate methods', *Journal of Molecular Evolution,* 39, 306-14.

Zhang, Y. and Skolnick, J. (2005), 'TM-align: a protein structure alignment algorithm based on the TM-score', *Nucleic Acids Research,* 33 (7), 2302-09.

# Chapter 2. A novel papillomavirus in a New Zealand fur seal (Arctocephalus forsteri) with oral lesions

Jonathon C.O. Mifsud[1], Jane Hall[2,3], Kate Van Brussel[1], Karrie Rose[3], Rhys H. Parry[4], Edward C. Holmes[1], Erin Harvey[1]

1. Sydney Institute for Infectious Diseases, School of Medical Sciences, The University of Sydney, Sydney, NSW 2006, Australia.
2. Centre for Planetary Health and Food Security, School of Environment and Science, Griffith University, Nathan, Queensland, Australia
3. Australian Registry of Wildlife Health, Taronga Conservation Society Australia, Mosman, NSW 2088, Australia
4. School of Chemistry and Molecular Biosciences, The University of Queensland, Brisbane, QLD 4067, Australia

## 2.1 Abstract

Despite being the predominant seal species in the Australian-New Zealand region and serving as a key indicator of marine environmental health, little is known about infectious diseases in New Zealand fur seals (Long-nosed fur seal; *Arctocephalus forsteri*). Several papillomaviruses have been identified in earless seals and sea lions, with the latter linked to cutaneous plaques and invasive squamous cell carcinoma. To date, no papillomaviruses have been reported in fur seals. We used traditional veterinary diagnostic techniques and metatranscriptomic sequencing of tissue samples to investigate the virome of New Zealand fur seals. We identified a novel papillomavirus, provisionally termed *A. forsteri* papillomavirus 1 (AforPV1) in an animal with clinically and histologically identified oral papilloma-like lesions. RT-PCR confirmed the presence of AforPV1 only in oral papilloma samples from the affected individual. Phylogenetic analysis of the complete 7926bp genome of AforPV1 revealed that it grouped with taupapillomaviruses found in related Carnivora species. These findings highlight the need for further research into the disease associations and impact of undiagnosed and novel viruses on New Zealand fur seals.

## 2.2 Introduction

New Zealand fur seals (Long-nosed fur seal; *Arctocephalus forsteri*) are the most abundant seal species in the Australian-New Zealand region (Goldsworthy et al. 2003). Their breeding colonies span the southern coast of Australia, ranging from Western Australia to Tasmania, as well as the coastlines and offshore islands of New Zealand and its subantarctic islands (Shaughnessy 1999). As long-lived marine mammals that feed at high trophic levels, New Zealand fur seals serve as important marine sentinels and offer insights into the health of aquatic ecosystems (Bossart 2011).

Despite their importance, our understanding of infectious diseases in New Zealand fur seals, particularly those due to viruses, remains limited. A single circovirus has been identified from a New Zealand fur seal fecal sample (Sikorski et al. 2013), although no details on the health status of the seal or the potential for pathogenicity were provided. The impact viruses can have on other seal species is well established, with numerous viral diseases described, including urogenital carcinoma associated with Otarine herpesvirus 1 in California sea lions (*Zalophus californianus*) and a South American fur seal (*A. australis*) (Dagleish et al. 2013; Deming et al. 2021) and vesicular exanthema associated with San Miguel sea lion virus reported in sea lions, fur seals, and elephant seals along the western

coast of the United States (Bossart and Duignan 2019). Viral diseases in seals can have far-reaching consequences (Colegrove et al. 2005). For instance, outbreaks of seal influenza A (H10N7) across Europe in 2014 (Bodewes et al. 2015) and highly pathogenic avian influenza A (H5N1) in New England, USA, in 2023, both caused mass mortalities of harbour seals (*Phoca vitulina*) (Puryear et al. 2023). Additionally, seals may have the potential to affect human health through zoonotic virus transmission, particularly during close interactions such as stranding and mass mortality events (Abdelwhab and Mettenleiter 2023).

The *Papillomaviridae* are a family of non-enveloped, double-stranded DNA viruses, typically 7500bp in length, that exhibit high host and tissue specificity, primarily infecting skin and mucosal surfaces. Infection can manifest in a range of ways, from asymptomatic, to the formation of self-resolving papillary masses, to malignant epithelial cancers (Syrjänen 2018). This family of viruses has a broad host range, spanning diverse species and environments. Apart from the genus *Alphapapillomavirus* within the subfamily *Secondpapillomavirinae*, which is known to infect fish, all other papillomaviruses belong to the subfamily *Firstpapillomavirinae* and are associated with reptiles, birds, and mammals, from both terrestrial and aquatic environments (Van Doorslaer et al. 2018).

Few papillomaviruses have been identified in pinnipeds. *Z. californianus* papillomavirus (genus *Dyonupapillomavirus*) was reported from California sea lions (*Z. californianus*) presenting with axillary and preputial papillomatous lesions (Rivera et al. 2012). This virus has been associated with sporadic cases of both *in situ* and invasive squamous cell carcinoma (Luff et al. 2018). In addition, various genera of papillomaviruses have been identified in the faeces of apparently healthy Weddell seals (*Leptonychotes weddellii*) (Smeele et al. 2018).

Given the potential for pinniped viruses to cause population-level effects, and the risk posed to veterinarians and rehabilitators from zoonotic diseases, it is crucial to expand our knowledge of seal viruses. Accordingly, we employed traditional veterinary diagnostic techniques in conjunction with metatranscriptomic sequencing of a variety of tissues from seals, with or without clinical or pathological signs of viral infection, to explore the virome of New Zealand fur seals.

## 2.3 Methods

### 2.3.1 Sample collection and processing

Samples from live seals were collected by a veterinarian between 2003 and 2021 for diagnostic purposes under a License to Rehabilitate Injured, Sick or Orphaned Protected Wildlife (no. MWL000100542) issued by the New South Wales (NSW) Department of the Environment. Samples from deceased, beach cast, or euthanased seals were collected by the Australian Registry of Wildlife Health (Registry), a conservation science program of Taronga Conservation Society Australia, during routine necropsy for disease surveillance in accordance with NSW National Parks and Wildlife Act 1974, section 132c, Scientific Licence number SL100104. These samples were collected under the auspices of the Taronga Conservation Society Australia's Opportunistic Sample Policy (approval no. R22D34), and pursuant to NSW Office of Environment and Heritage-issued scientific licenses SL10469 and SL100104. Tissue samples comprising various skin and mucosa samples, liver, brain, lung, and kidney were variably collected from 18 New Zealand fur seals and stored at −80°C until RNA extraction. A set of tissues representing each organ system was also collected into 10% neutral buffered formalin, embedded in paraffin wax, sectioned and mounted on a glass slide, stained with hematoxylin and eosin, and examined by light microscopy at ×200, ×400, and ×1000 magnification.

### 2.3.2 RNA extraction and metatranscriptomic sequencing

Individual tissue aliquots were placed into 600µl of lysis buffer containing 0.5% foaming reagent (Reagent DX, Qiagen) and 1% β-mercaptoethanol (Sigma-Aldrich), and tissue was homogenized using a TissueRuptor (Qiagen) at a speed of 5000rpm for up to one minute. The homogenate was centrifuged at maximum speed (15,200rpm) for three minutes to eliminate any remaining tissue residue. RNA was then extracted from the supernatant using the RNeasy Plus Mini Kit (Qiagen), following the manufacturer's protocol. Extracted RNA was combined by tissue type into 14 pools with a median of three samples per pool (minimum=1, maximum=5) (Supplementary Table 1). Sequencing libraries were constructed using the TruSeq Total RNA Library Preparation Protocol (Illumina). Host ribosomal RNA (rRNA) was depleted using the Ribo-Zero Plus Kit (Illumina), and paired-end sequencing (150bp) was performed on the NovaSeq 6000 platform (Illumina). Library construction and sequencing were performed by the Australian Genome Research Facility (AGRF).

### 2.3.3 Identification of novel virus sequences

Virus identification followed the BatchArtemisSRAMiner pipeline(Mifsud 2023). Briefly, sequencing reads underwent quality trimming and adapter removal using Trimmomatic (v0.38) with parameters SLIDINGWINDOW:4:5, LEADING:5, TRAILING:5, and MINLEN:25, prior to assembly (Bolger et al. 2014). *De novo* assembly was conducted using MEGAHIT (v1.2.9) (Li et al. 2015). Assembled contigs were compared to the RdRp-scan RdRp core protein sequence database (v0.90) (Charon et al. 2022) and the protein version of the Reference Viral Databases (v23.0) (Goodacre et al. 2018) using DIAMOND BLASTx (v2.0.9) with an E-value cut-off of $1\times10^{-5}$ (Buchfink et al. 2021). To exclude potential false positives, contigs with hits to virus sequences were used as a query against the NCBI nucleotide database (as of March 2022) using BLASTn (Camacho et al. 2009) and the NCBI non-redundant protein (nr) database (as of March 2022) using DIAMOND BLASTx. Using BLASTx and BLASTn matches, virus-like contigs associated with non-vertebrate hosts were excluded.

### 2.3.4 A. forsteri papillomavirus 1 RT-PCR

RT-PCR was performed on total RNA from the five individual samples that made up library SL16 using primers designed based upon the novel *A. forsteri* papillomavirus 1 (AforPV1) fragments obtained by metatranscriptomic sequencing. SuperScript IV One Step RT-PCR (Invitrogen) and the forward primer 5′-TGGAACGTTGACCTGAGAGA 3’ and reverse primer 5′-AAGGATACGGTCCGTTCTGA-3′ were used to amplify a missing 689bp section between the L1 and E6 genes. A second set of primers, forward primer 5′-ATACACTCCGTCTTGGGACG-3′ and reverse primer 5′-CAGTTACAAAGCTTCGAGGGT-3′, was used to check the region surrounding the stop codon of E2. The resulting amplicon product was subsequently used for Sanger sequencing at the AGRF.

### 2.3.5 Genome analysis and annotation

To examine the genome coverage of each virus, sequence reads were mapped onto virus-like contigs using BBMap (v37.98) (Bushnell 2014), and areas of heterogeneous coverage were manually checked using Geneious (v11.0.9). Where possible, the extremities of contigs were manually extended and re-submitted to read mapping until the contig appeared complete or no overhanging extremities were observed. Sequences of vector origin were detected using VecScreen (https://www.ncbi.nlm.nih.gov/tools/vecscreen/) and removed. GetORF from EMBOSS (v6.6.0) was used to predict open reading frames (ORFs)

(Rice et al. 2000). To annotate protein functional domains, the InterProScan software package (v5.56) was used with the TIGRFAMs (v15.0), SFLD (v4.0), PANTHER (v15.0), SuperFamily (v1.75), PROSITE (v2022_01), CDD (v3.18), Pfam (v34.0), Hamap (v 2023_01), SMART (v7.1), PRINTS (v42.0), PIRSF (v3.10) and CATH-Gene3D databases (v4.3.0) (Jones et al. 2014). The completeness and quality of viral sequences were assessed by visual inspection and the CheckV pipeline (Nayfach et al. 2021). AforPV1 gene expression levels were calculated using htseq-count (v2.0.3) with non-default parameters "-s reverse --nonunique fraction". AforPV1 binding sites were predicted by manual sequence comparisons, while GC content was calculated in Geneious with a sliding window of 40 nucleotides. AforPV1 spliced ORFs E1^E4 and E8^E2 were manually predicted by aligning AforPV1 to Canis familiaris papillomavirus 19 isolate tvmb1 (KX599536). Support for the splice junctions was assessed using ViReMa (v0.25) (Sotcheff et al. 2023). Circos (v0.69-6) was used to produce the circular genome graphs for AforPV1 (Krzywinski et al. 2009). The marker gene cytochrome c oxidase subunit I (COX1) was identified by querying the contig set against the nr database using DIAMOND BLASTx. The abundances of COX1 and viral transcripts were determined by individually mapping the SL16 reads to each using RNA-Seq by Expectation Maximization (RSEM) software (v1.3.0) (Li and Dewey 2011).

### 2.3.6 Assessment of sequencing library composition

To identify any possible contaminant sequences or coinfecting bacteria or fungi, contigs from the SL16 library were aligned to the custom NCBI nt database using the KMA aligner and the CCMetagen program (Clausen et al. 2018; Marcelino et al. 2020). Species related to known pathogens of mammals were manually confirmed through BLASTn and read mapping against reference genes using BBMap.

### 2.3.7 Phylogenetic analysis

Phylogenetic trees of the putative papillomavirus identified here were inferred using a maximum likelihood approach. Representative genomes ($n$=117) from each of the papillomavirus genera were downloaded from The Papillomavirus Episteme (PaVE) (https://pave.niaid.nih.gov/) (Van Doorslaer et al. 2017). The amino acid sequences of four genes (L1, L2, E1, and E2) were obtained for these sequences along with the novel papillomavirus identified here and individually aligned using MAFFT (v7.402), quality trimmed using trimAl (v1.2), and concatenated to form a single alignment. All phylogenetic trees were estimated using IQ-TREE2 (Anisimova et al. 2011). Branch support was calculated using 1000 bootstrap replicates with the UFBoot2 algorithm and an

implementation of the SH-like approximate likelihood ratio test within IQ-TREE2 (Anisimova et al. 2011). The best-fit model of amino acid substitution was determined using the Akaike information criterion (AIC), the corrected AIC, and the Bayesian information criterion with the ModelFinder function in IQ-TREE2 (Anisimova et al. 2011). This process was then repeated with a subset of papillomaviruses, namely the taupapillomaviruses and the related gamma- and pipapillomaviruses.

## 2.4 Results

### 2.4.1 Overview of metatranscriptomic data

In total, 14 metatranscriptomic libraries were constructed from pools of tissue from 18 individual New Zealand fur seals (library statistics are summarised in Supplementary Table 1). The extracted RNA was pooled according to tissue type. No mammalian-associated viruses were identified from the brain, liver, lung, or kidney. However, a novel papillomavirus was recovered from oral tissue library SL16 and is discussed in further detail below.

### 2.4.2 Identification of a novel papillomavirus in a seal with oral papilloma

In October 2002, an immature male New Zealand fur seal (Registry #3254) was taken into rehabilitation care after being found hauled out on a beach near Narooma, New South Wales (NSW), Australia, in an emaciated body condition and exhibiting dehydration, severe anemia, and several deep skin wounds over the right hip and right hind flipper, presumably associated with a failed predation attempt by a shark. Upon examination, the seal was noted to have several small raised, sometimes pedunculated and coalescing papillary masses on the roof of the mouth with similar but smaller lesions, evident on the caudoventral right and ventral left aspects of the tongue (Figure 2.10a,b). There was also a circumferential zone of mucosal pallor around the right mandibular canine tooth (Figure 2.10c). Biopsies of the papillary masses were taken from the roof of the mouth and tongue region in February 2003.

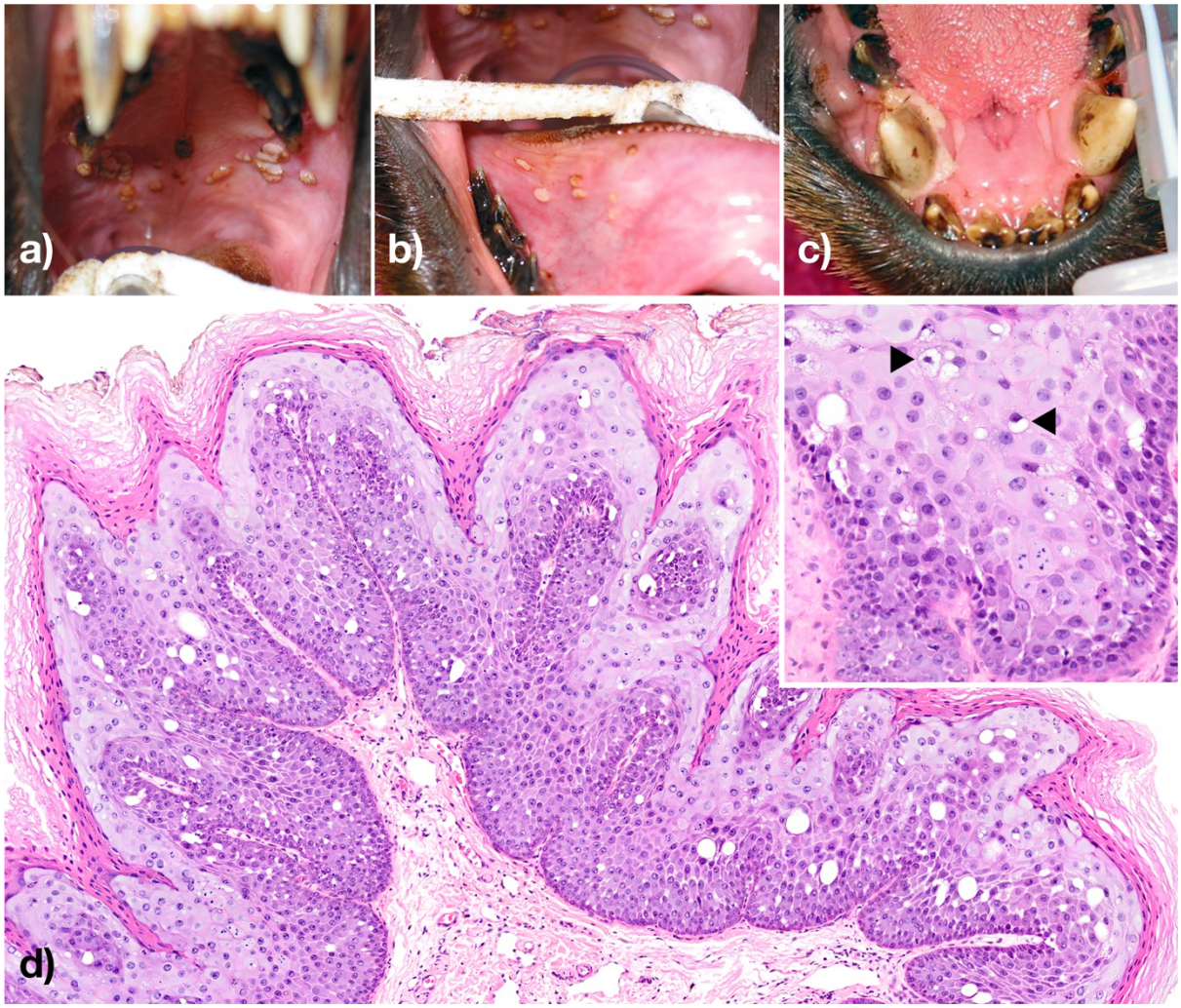



***Figure 2.10 Papillary proliferations, cellular atypia, and the presence of koilocytes in oral lesions.*** *Multifocally coalescing sessile to papillary proliferations of the palatine and lingual epithelium of a New Zealand fur seal Registry #3254 (a, b). A zone of mucosal pallor surrounding the mandibular canine tooth (c). Demarcated lingual epithelial proliferation, with basophilia, and expansion of the stratum spinosum and stratum corneum (d—inset higher magnification of the lesion demonstrating cellular atypia and the presence of koilocytes—black arrowheads).*

Microscopic examination of multiple formalin-fixed oral lesion biopsies revealed sharply demarcated zones of lingual epithelial proliferation characterised by basophilia, binucleation of cells throughout the stratum basale, and an epithelium irregularly thickened by expansion of the stratum spinosum and stratum corneum (Figure 2.10d). Cells in the stratum spinosum were large and often atycal, with large nuclei and abundant pale to basophilic cytoplasm. Epithelial cells with pyknotic or karyorrhectic nuclei, and intercellular oedema were multifocally evident throughout the affected stratum spinosum. Multifocal koilocytes,

epithelial cells exhibiting nuclear enlargement and hyperchromasia in conjunction with perinuclear cytoplasmic vacuolation, were also evident within the stratum spinosum. The lingual lamina propria beneath the epithelial lesions appeared normal.

### 2.4.3 Papillomavirus genome construction and confirmation

Several papillomavirus-like contigs were assembled from library SL16. Using a combination of read mapping and RT-PCR, a complete circular genome of 7926bp here termed *A. forsteri* papillomavirus 1 (AforPV1) was recovered, which was confirmed by CheckV (100% genome completion, high confidence) (Figure 11). AforPV1 was present in high abundance in SL16 with 0.18% (*n*=130,197 reads) of total library reads mapping to the genome, slightly less than the abundance of the mitochondrial COX1 gene in this library (0.27%, *n*=195,218 reads). RT-PCR confirmed that AforPV1 was limited to seal Registry #3254, with both tissue samples taken from under the tongue testing positive. No other tissues or blood samples were available from this individual.

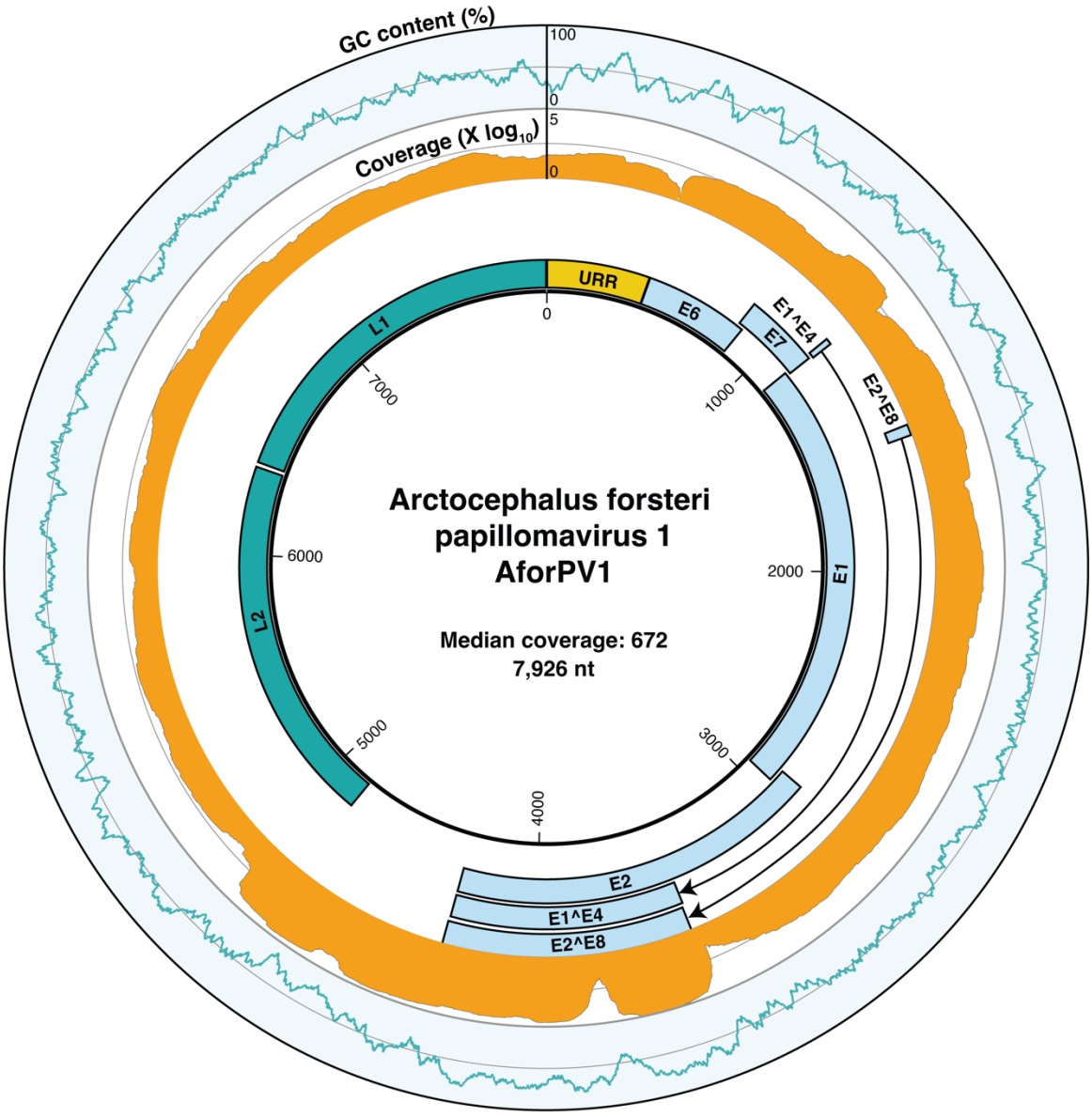



***Figure 11.2 Genome organization of Arctocephalus forsteri papillomavirus 1 (AforPV1).*** *The outer ring indicates the percentage GC content (blue) over a 40bp sliding window, the middle ring represents read coverage across the genome, log10 transformed (orange), and the inner ring illustrates the predicted ORFs and the upstream regulatory region of AforPV1. The "A" in the first start codon of the ORF E6 is assigned as position one.*

### 2.4.4 Evolutionary relationship and genomic properties of AforPV1

AforPV1 falls as a distinct lineage within the papillomaviruses, with its closest relative Mustela putorius papillomavirus 1 (MpPV1) (NC_022253) isolated from a European polecat (*Mustela putorius*, family *Mustelidae*) with which it shares 66% nucleotide identity across the genome (Figure 2.12a). However, the degree of similarity between AforPV1 and MpPV1 varied between the six genes tested (E1, E2, E6, E7, L1 and L2), ranging from 50% (E2) to

74% (L1) nucleotide identity. To determine the evolutionary history of AforPV1 a phylogenetic analysis was conducted using a concatenated alignment of four genes (L1, L2, E1, and E2) as per the ICTV guidelines (Van Doorslaer et al. 2018). This analysis placed AforPV1 within the genus *Taupapillomavirus*, forming a clade with MpPV1 and a papillomavirus associated with the Weddell seal, *L. weddellii* papillomavirus 5 (LwPV5), albeit with poor bootstrap support (SH-aLRT=43% and UFboot=64%) (Figure 2.12b). Given that the ICTV species demarcation (Van Doorslaer et al. 2018) for the taupapillomaviruses is <70% nucleotide identity across the genome, we suggest that AforPV1 represents a novel species within the genus *Taupapillomavirus* (family *Papillomaviridae*).

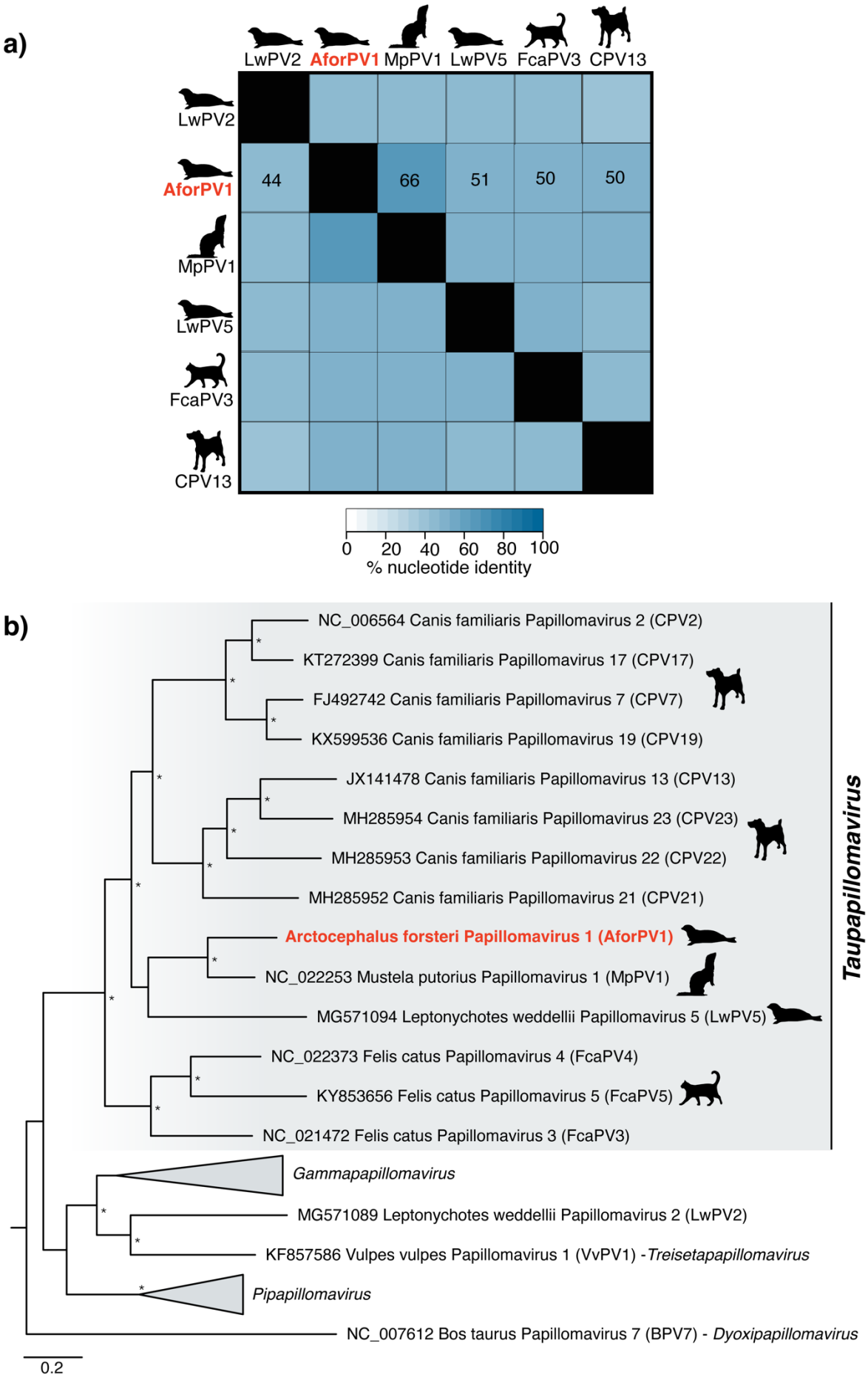

a)
LwPV2
AforPV1
MpPV1
LwPV5
FcaPV3
CPV13
44
66
51
50
50
0
20
40
60
80
100
% nucleotide identity
b)
NC_006564 Canis familiaris Papillomavirus 2 (CPV2)
KT272399 Canis familiaris Papillomavirus 17 (CPV17)
FJ492742 Canis familiaris Papillomavirus 7 (CPV7)
KX599536 Canis familiaris Papillomavirus 19 (CPV19)
JX141478 Canis familiaris Papillomavirus 13 (CPV13)
MH285954 Canis familiaris Papillomavirus 23 (CPV23)
MH285953 Canis familiaris Papillomavirus 22 (CPV22)
MH285952 Canis familiaris Papillomavirus 21 (CPV21)
Arctocephalus forsteri Papillomavirus 1 (AforPV1)
NC_022253 Mustela putorius Papillomavirus 1 (MpPV1)
MG571094 Leptonychotes weddellii Papillomavirus 5 (LwPV5)
NC_022373 Felis catus Papillomavirus 4 (FcaPV4)
KY853656 Felis catus Papillomavirus 5 (FcaPV5)
NC_021472 Felis catus Papillomavirus 3 (FcaPV3)
Taupapillomavirus
Gammapapillomavirus
MG571089 Leptonychotes weddellii Papillomavirus 2 (LwPV2)
KF857586 Vulpes vulpes Papillomavirus 1 (VvPV1) - Treisetapapillomavirus
Pipapillomavirus
NC_007612 Bos taurus Papillomavirus 7 (BPV7) - Dyoxipapillomavirus
0.2

***Figure 2.12 Placement of Arctocephalus forsteri papillomavirus 1 (AforPV1) within the taupapillomaviruses.*** *(a) Percentage identity matrix of representative taupapillomaviruses for each host species compared to AforPV1. (b) Phylogenetic relationship of the taupapillomaviruses and other closely related genera. An ML phylogenetic tree based on the conserved amino acid sequences of the L1, L2, E1, E2 genes shows, in red, the topological position of AforPV1, in the context of its closest relatives. Animal silhouettes depict the virus-host associations for the taupapillomaviruses. All branches are scaled to the number of amino acid substitutions (model LG+F+I+G4) per site, and the tree is midpoint rooted for clarity only. An asterisk indicates node support where SH-aLRT≥80% and UFboot≥95%.*

The genome organization of AforPV1 is consistent with other taupapillomaviruses, comprising eight open reading frames (ORFs) (E1, E2, E4, E6, E7, E8, L1, and L2) as well as the spliced ORFs of E1^E4 and E8^E2. The presence of the E8^E2 and E1^E4 splice junctions was supported by 89 and 13,700 reads, respectively. No E5 ORF was detected. Six of the eight ORFs had matches to domains associated with the E1, E2, E6, E7, L1, and L2 ORFs (Figure 11.2). No domains were detected for the E4 and E8 ORFs. A premature stop codon, which resulted in E2 being 238bp (79 aa) shorter than its closest relatives, was present. This stop codon was in a region of high coverage and was confirmed by Sanger sequencing. Two zinc-binding sites (CXXC-X29-CXXC) were found in E6, although E6 lacked a PDZ-binding motif (ETQL) in its C-terminus. An alternative pRB-binding site (retinoblastoma tumor suppressor-binding domain) (LXSXE) was detected in E7, consistent with other taupapillomaviruses (Smeele et al. 2018; Wang et al. 2010). A cyclin RXL motif (KRRLF) and ATP-binding site (GXXXXGK[T/S]) were detected in E1. The upstream regulatory region (URR), defined as the region between the L1 stop codon and E6 start codon, was 430bp in length. The URR contained four E2-binding sites (ACCN2-11GGT), one Nf1-binding site (TTGGC), one palindromic E1-binding site (ATTGTTXXXAACAAT) and a TATA box. The predicted ORFs, protein products, and binding sites are shown in Supplementary Table 2.

### 2.4.5 Gene expression of AforPV1

Gene expression analysis of AforPV1 revealed variation among its genes (Supplementary Table 3). Notably, E2 displayed relatively high expression levels (194,773 reads per kilobase per million mapped reads [RPKM]). Expression was concentrated in specific regions, encompassing nucleotides 3048–3331 and 3397–3872, which corresponded to the predicted alternatively spliced isoforms, E8^E2 and E1^E4, each recording RPKM values of

315,227 and 315,461, respectively. In contrast, the remaining early genes E1, E6, and E7 exhibited comparatively lower RPKM values of 39,158, 12,894, and 10,751, respectively. The late gene L2 exhibited higher expression (63,078 RPKM) than L1 (4693 RPKM), although this was predominantly observed at the beginning of the L2 gene (nucleotides 4402–4510).

### 2.4.6 SL16 library composition

To examine whether other pathogens were present in SL16 (the library in which AforPV1 was found), contigs assembled from rRNA-depleted reads were assessed for taxonomic associations using CCMetagen. SL16 was predominately (85%) comprised of contigs belonging to eared seals (family *Otariidae*), while 5% was attributed to other pinnipeds (e.g., walruses), and non-chordates including bacteria (<0.1%) (Supplementary Figure 2.1). Among the non-chordate abundance, 52% were associated with bacteria from various families, and 48% were linked to eukaryotes, specifically fungi (43%), arthropods (2%), algae (2%), and diatoms (1%) (Supplementary Table 4, Supplementary Figure 2.1). The fungi were primarily assigned to *Candida albicans* (32%).

Further analysis of the CCMetagen results employing BLASTn revealed that a contig with hits to the *Tannerellaceae* (reference sequence *Tannerella forsythia*, HG784150.1) shared 81% nucleotide similarity (evalue=7e-93) with the tetratricopeptide repeat protein of *T. forsythia*. The contigs associated with this species were highly fragmented, so we were unable to make a definitive judgment regarding the taxonomy of this bacterium. Additionally, a fragment associated with the *Enterobacteriaceae* (reference sequence *Shigella dysenteriae* EU855235.1) was found to be identical to a hypothetical gene found in *Escherichia coli* (e-value=0.0, EFO55302.1). The remaining contigs associated with *Enterobacteriaceae* share the greatest sequence similarity with *E. coli* (e-value=0.0, 95% nucleotide identity, CP017061).

In addition, we identified a partial genome that may represent a novel *Gammaherpesvirus* in a different individual without pathological evidence of viral infection. Information on this putative virus can be found in Supplementary Note 2.1 and Supplementary Figure 2.2

## 2.5 Discussion

We used a metatranscriptomic approach to identify potential viral etiological agents in New Zealand fur seals with and without gross and microscopic pathology. Using this approach

on a variety of tissues, we were able to identify novel papillomavirus and a likely etiology for the oral papilloma-like lesions described.

Papillomaviruses have been found in a wide range of mammalian species, with 85 hosts identified to date (Van Doorslaer et al. 2017). Despite this, our understanding of papillomaviruses in seals and their associated pathology remains limited. We identified a novel papillomavirus, the first described in fur seals, and the first report of oral papilloma-like lesions in this host group. Microscopic examination of the oral lesions revealed substantial epithelial cellular atypia and the presence of koilocytes consistent with the cytopathic effects associated with papillomaviruses in humans and other animals (Betz 2019; Sundberg et al. 2000). Although such lesions are often of limited clinical significance—as observed in this case study where the lesions self-resolved—there is potential for these plaques to progress into invasive squamous cell carcinomas (Munday et al. 2016).

A critical factor in papillomavirus-induced oncogenesis is the integration of the virus into the host genome and its impact on the expression of the viral oncogenes E6/E7. These genes are negatively regulated by E2, and integration, which is frequently observed in the HPV (Arias-Pulido et al. 2006) can disrupt the E2 gene, leading to upregulation of the E6/E7 oncogenes (Münger et al. 2004). We present preliminary evidence that AforPV1 may exist in both episomal and integrated forms within this individual. The gene expression profile of AforPV1 infection, particularly the high E2/E6 ratio, is typically indicative of episomal infection in HPV16 infection, although exceptions exist (Qiu et al. 2021). Conversely, the AforPV1 E2 gene appears truncated, suggesting that AforPV1 may also be present in an integrated form. The integration of HPV has been shown to trigger substantial host genome alterations resulting in the complete loss of function (Schmitz et al. 2012) or increased expression in target genes surrounding virus integration sites (Ojesina et al. 2014). Further investigation is required to confirm whether AforPV1 is present in an integrated form and its effect on host gene expression.

While establishing causality is challenging, as papillomaviruses often asymptomatically infect the skin, there are several key pieces of evidence indicating that AforPV1 was infecting the New Zealand fur seal and may be associated with disease in these animals, although further investigation is warranted. AforPV1 appears to be highly abundant within the metatranscriptome data, with SL16 transcripts representing 0.18% of reads in a pool of

three individuals, two of which were PCR-negative for this virus. AforPV1 was limited to an individual with papilloma-like lesions, as determined by gross and histological examination prior to the molecular discovery of this virus. This, together with the construction of the entire genome of AforPV1 from metatranscriptomic data, makes it unlikely that this sequence represents an endogenous viral element (i.e., a viral sequence that was passed vertically in the germline of the host) but rather the presence of both integration and episomal AforPV1 replication.

There is no evidence to suggest that another pathogen was responsible for the oral lesions of seal Registry #3254. *C. albicans* was the only microorganism detected in relatively high abundance (0.05% of SL16 abundance in CCMetagen) and is a common oral fungus, acting most commonly as a commensal organism. Although *C. albicans* can be an opportunistic pathogen, it is not known to cause papillary lesions, and these organisms were not evident within multiple histological sections of the lesions (Dunn et al. 1984). Bacteria were also present in the SL16 library, such as *E. coli* and a sequence with homology to *Tannerella sp*., although in the latter, the highly fragmented nature of the sequences assembled prevented our ability to taxonomically classify this species further. While *E. coli* and certain *Tannerella* species (e.g., *T. forsythia*) are known to be pathogenic, they are not known to cause oral lesions consistent with those described (Gomes et al. 2016; Holt and Ebersole 2005).

A phylogenetic analysis provides additional evidence that AforPV1 is indeed associated with a seal host: it groups with the Carnivora-associated taupapillomaviruses in a clade with seal and polecat papillomaviruses, LwPV5 (MG571094) (Smeele et al. 2018) and MpPV1 (NC_022253) (Smits et al. 2013). Although there is some evidence of recombination and host switching within the *Papillomaviridae*, the known taupapillomaviruses appear to have generally co-diverged with their hosts (Van Doorslaer 2013).

The absence of viruses across many of our samples could be attributed to numerous causes. If low abundance viruses are present, it is possible that we did not have the sequencing depth to recover these over the host signal, that viruses are too divergent to detect using similarity-based methods, or that this simply could reflect the absence of infection entirely at the point of sampling. This is supported by the absence of evidence of viral disease in the histopathology reported from these seals.

Through a metatranscriptomic survey of tissue samples from New Zealand fur seals, we identified a novel papillomavirus. The identification of AforPV1 in a seal exhibiting oral papilloma-like lesions emphasises the value of combining metagenomic sequencing with conventional gross and histological examination to discover novel wildlife pathogens. Such work may ultimately contribute to improved disease detection, control, and conservation efforts. Additional research is necessary to ascertain the clinical relevance of these viruses in New Zealand fur seals.

### 2.6 Data availability

All *A. forsteri* sequence reads are available on the NCBI Sequence Read Archive (SRA) under BioProject PRJNA1013207. The viral genomes assembled in this study have been deposited in the GenBank and assigned the accession number OR531434, OR590706 and OR590707. The sequences, alignments, and phylogenetic trees generated in this study are available at https://github.com/JonathonMifsud/Identification-of-a-novel-papillomavirus-in-a-New-Zealand-Fur-seal-with-oral-papilloma. The underlying code for this study is not publicly available but may be made available to qualified researchers on reasonable request from the corresponding author.

### 2.7 Acknowledgements

We thank Dr. Wei Shan Chang for their assistance with the genome organization graph. We acknowledge the University of Sydney's high-performance computing cluster Artemis for providing the computing resources used for this study. We acknowledge the Taronga Conservation Society Australia and the New South Wales Department of Environment for their ongoing funding of the Australian Registry of Wildlife Health, and we thank the NSW National Parks and Wildlife Service, Organization for the Rescue and Research of Cetaceans in Australia, and Taronga Wildlife Hospital for their ongoing efforts to rescue, rehabilitate, and care for injured and debilitated seals across NSW. We thank PhyloPic (https://www.phylopic.org) for the animal silhouettes used in Figure 3 under Public Domain Dedication. E.C.H. is supported by an NHMRC Investigator Fellowship (GNT2017197). J.C.O.M. is supported by an Australian Government Research Training Program (RTP) Scholarship. The funder played no role in study design, data collection, analysis, and interpretation of data, or the writing of this manuscript.

## 2.8 References


Abdelwhab, E. M. and Mettenleiter, T. C. (2023), 'Zoonotic Animal Influenza Virus and Potential Mixing Vessel Hosts', *Viruses,* 15 (4), 980.

Anisimova, M., et al. (2011), 'Survey of Branch Support Methods Demonstrates Accuracy, Power, and Robustness of Fast Likelihood-based Approximation Schemes', *Systematic Biology,* 60 (5), 685-99.

Arias-Pulido, H., et al. (2006), 'Human papillomavirus type 16 integration in cervical carcinoma in situ and in invasive cervical cancer', *Journal of Clinical Microbiology,* 44 (5), 1755-62.

Betz, S. J. (2019), 'HPV-related papillary lesions of the oral mucosa: a review', *Head and Neck Pathology,* 13 (1), 80-90.

Bodewes, R., et al. (2015), 'Avian Influenza A(H10N7) virus-associated mass deaths among harbor seals', *Emerging Infectious Diseases,* 21 (4), 720-2.

Bolger, A. M., Lohse, M., and Usadel, B. (2014), 'Trimmomatic: a flexible trimmer for Illumina sequence data', *Bioinformatics,* 30 (15), 2114-20.

Bossart, G. and Duignan, P. (2019), 'Emerging viruses in marine mammals', *CABI Reviews,* (2018), 1-17.

Bossart, G. D. (2011), 'Marine mammals as sentinel species for oceans and human health', *Veterinary Pathology,* 48 (3), 676-90.

Buchfink, B., Reuter, K., and Drost, H.-G. (2021), 'Sensitive protein alignments at tree-of-life scale using DIAMOND', *Nature Methods,* 18 (4), 366-68.

Bushnell, B. (2014), 'BBMap: a fast, accurate, splice-aware aligner', (Lawrence Berkeley National Lab (LBNL), Berkeley, CA.).

Camacho, C., et al. (2009), 'BLAST+: architecture and applications', *BMC Bioinformatics,* 10 (1), 1-9.

Charon, J., et al. (2022), 'RdRp-scan: A bioinformatic resource to identify and annotate divergent RNA viruses in metagenomic sequence data', *Virus Evolution,* 8 (2), veac082.

Clausen, P. T., Aarestrup, F. M., and Lund, O. (2018), 'Rapid and precise alignment of raw reads against redundant databases with KMA', *BMC Bioinformatics,* 19 (1), 1-8.

Colegrove, K. M., Greig, D. J., and Gulland, F. M. (2005), 'Causes of live strandings of northern elephant seals (Mirounga angustirostris) and Pacific harbor seals (Phoca vitulina) along the central California coast, 1992-2001', *Aquatic Mammals,* 31 (1), 1.

Dagleish, M., et al. (2013), 'The first report of otarine herpesvirus-1-associated urogenital carcinoma in a South American fur seal (Arctocephalus australis)', *Journal of Comparative Pathology,* 149 (1), 119-25.

Deming, A. C., et al. (2021), 'Unlocking the role of a genital herpesvirus, otarine herpesvirus 1, in California Sea lion cervical cancer', *Animals,* 11 (2), 491.

Dunn, J. L., Buck, J. D., and Spotte, S. (1984), 'Candidiasis in captive pinnipeds', *Journal of the American Veterinary Medical Association,* 185 (11), 1328-30.

Goldsworthy, S. D., et al. (2003), 'Trophic interactions between marine mammals and Australian fisheries: an ecosystem approach', *Marine mammals: Fisheries, tourism and management issues* (2006: CSIRO Publishing Melbourne), 62-99.

Gomes, T. A., et al. (2016), 'Diarrheagenic escherichia coli', *Brazilian Journal of Microbiology,* 47 (Suppl 1), 3-30.

Goodacre, N., et al. (2018), 'A reference viral database (RVDB) to enhance bioinformatics analysis of high-throughput sequencing for novel virus detection', *mSphere,* 3 (2), e00069-18.

Holt, S. C. and Ebersole, J. L. (2005), 'Porphyromonas gingivalis, Treponema denticola, and Tannerella forsythia: the ‘red complex’, a prototype polybacterial pathogenic consortium in periodontitis', *Periodontology 2000,* 38 (1), 72-122.

Jones, P., et al. (2014), 'InterProScan 5: genome-scale protein function classification', *Bioinformatics,* 30 (9), 1236-40.

Krzywinski, M., et al. (2009), 'Circos: an information aesthetic for comparative genomics', *Genome Research,* 19 (9), 1639-45.

Li, B. and Dewey, C. N. (2011), 'RSEM: accurate transcript quantification from RNA-Seq data with or without a reference genome', *BMC Bioinformatics,* 12 (1), 323.

Li, D., et al. (2015), 'MEGAHIT: an ultra-fast single-node solution for large and complex metagenomics assembly via succinct de Bruijn graph', *Bioinformatics,* 31 (10), 1674-76.

Luff, J. A., et al. (2018), 'Cutaneous squamous cell carcinoma associated with Zalophus californianus papillomavirus 1 in a California sea lion', *Journal of Veterinary Diagnostic Investigation,* 30 (4), 572-75.

Marcelino, V. R., et al. (2020), 'CCMetagen: comprehensive and accurate identification of eukaryotes and prokaryotes in metagenomic data', *Genome Biology,* 21 (1), 103.

Mifsud, J. C. O. (2023), 'BatchArtemisSRAMiner: v1.0.0', (Zenodo).

Munday, J. S., et al. (2016), 'Genomic characterisation of canine papillomavirus type 17, a possible rare cause of canine oral squamous cell carcinoma', *Veterinary Microbiology,* 182, 135-40.

Münger, K., et al. (2004), 'Mechanisms of human papillomavirus-induced oncogenesis', *Journal of Virology,* 78 (21), 11451-60.

Nayfach, S., et al. (2021), 'CheckV assesses the quality and completeness of metagenome-assembled viral genomes', *Nature Biotechnology,* 39 (5), 578-85.

Ojesina, A. I., et al. (2014), 'Landscape of genomic alterations in cervical carcinomas', *Nature,* 506 (7488), 371-75.

Puryear, W., et al. (2023), 'Highly Pathogenic Avian Influenza A(H5N1) Virus Outbreak in New England Seals, United States', *Emerging Infectious Diseases,* 29 (4), 786-91.

Qiu, Q., et al. (2021), 'Integrated analysis of virus and host transcriptomes in cervical cancer in Asian and Western populations', *Genomics,* 113 (3), 1554-64.

Rice, P., Longden, I., and Bleasby, A. (2000), 'EMBOSS: the European molecular biology open software suite', *Trends in Genetics,* 16 (6), 276-77.

Rivera, R., et al. (2012), 'Characterization of a novel papillomavirus species (ZcPV1) from two California sea lions (Zalophus californianus)', *Veterinary Microbiology,* 155 (2-4), 257-66.

Schmitz, M., et al. (2012), 'Loss of gene function as a consequence of human papillomavirus DNA integration', *International Journal of Cancer,* 131 (5), E593-E602.

Shaughnessy, P. D. (1999), *The action plan for Australian seals* (Environment Australia).

Sikorski, A., Dayaram, A., and Varsani, A. (2013), 'Identification of a Novel Circular DNA Virus in New Zealand Fur Seal (Arctocephalus forsteri) Fecal Matter', *Genome Announcements,* 1 (4), e00558-13.

Smeele, Z. E., et al. (2018), 'Diverse papillomaviruses identified in Weddell seals', *The Journal of General Virology,* 99 (4), 549.

Smits, S. L., et al. (2013), 'Metagenomic analysis of the ferret fecal viral flora', *PLOS One,* 8 (8), e71595.

Sotcheff, S., et al. (2023), 'ViReMa: a virus recombination mapper of next-generation sequencing data characterizes diverse recombinant viral nucleic acids', *GigaScience,* 12, giad009.

Sundberg, J., et al. (2000), 'Feline papillomas and papillomaviruses', *Veterinary Pathology,* 37 (1), 1-10.

Syrjänen, S. (2018), 'Oral manifestations of human papillomavirus infections', *European Journal of Oral Sciences,* 126 (S1), 49-66.

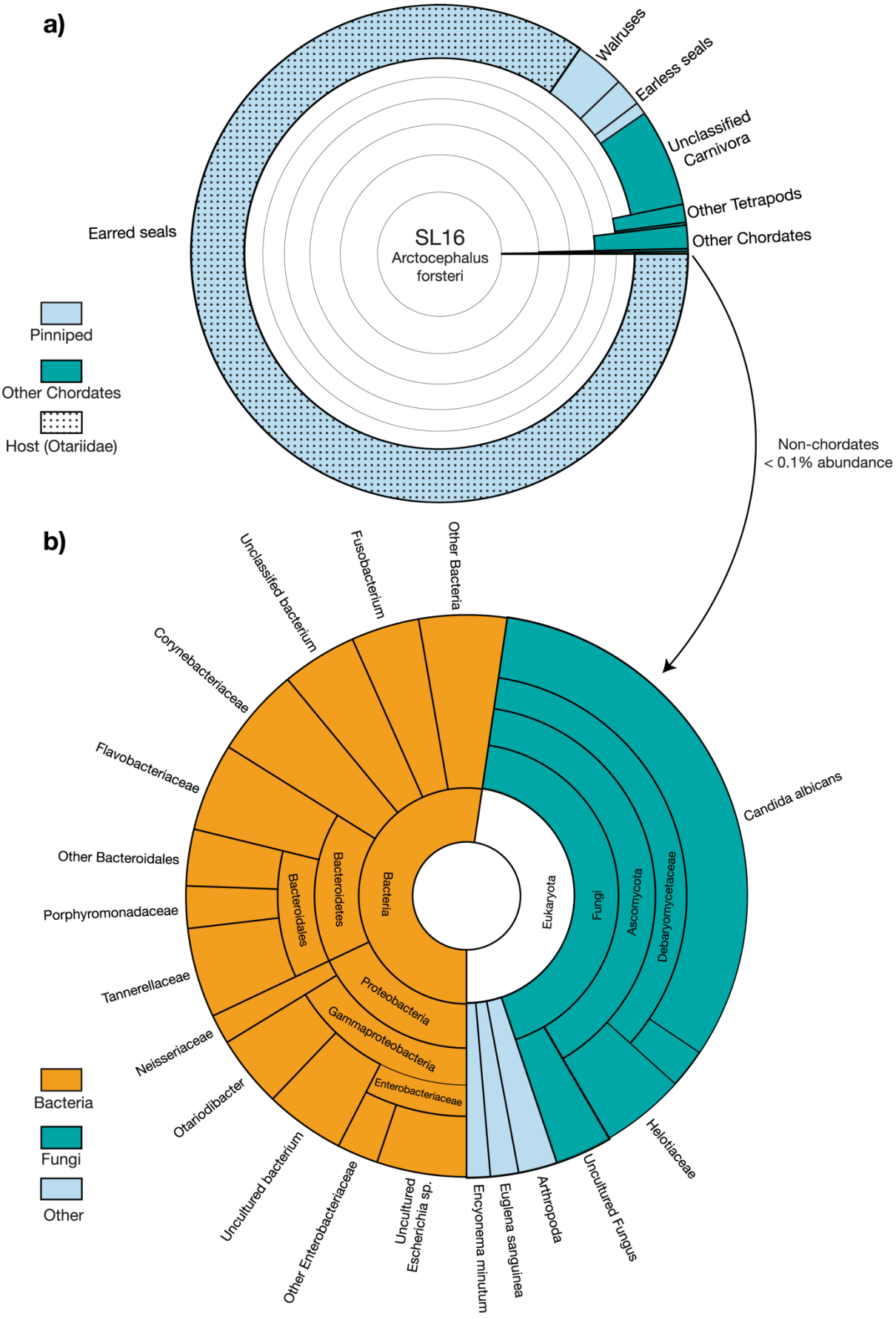
Van Doorslaer, K. (2013), 'Evolution of the Papillomaviridae', *Virology,* 445 (1), 11-20.

Van Doorslaer, K., et al. (2017), 'The Papillomavirus Episteme: a major update to the papillomavirus sequence database', *Nucleic Acids Research,* 45 (D1), D499-D506.

Van Doorslaer, K., et al. (2018), 'ICTV virus taxonomy profile: Papillomaviridae', *Journal of General Virology,* 99 (8), 989-90.

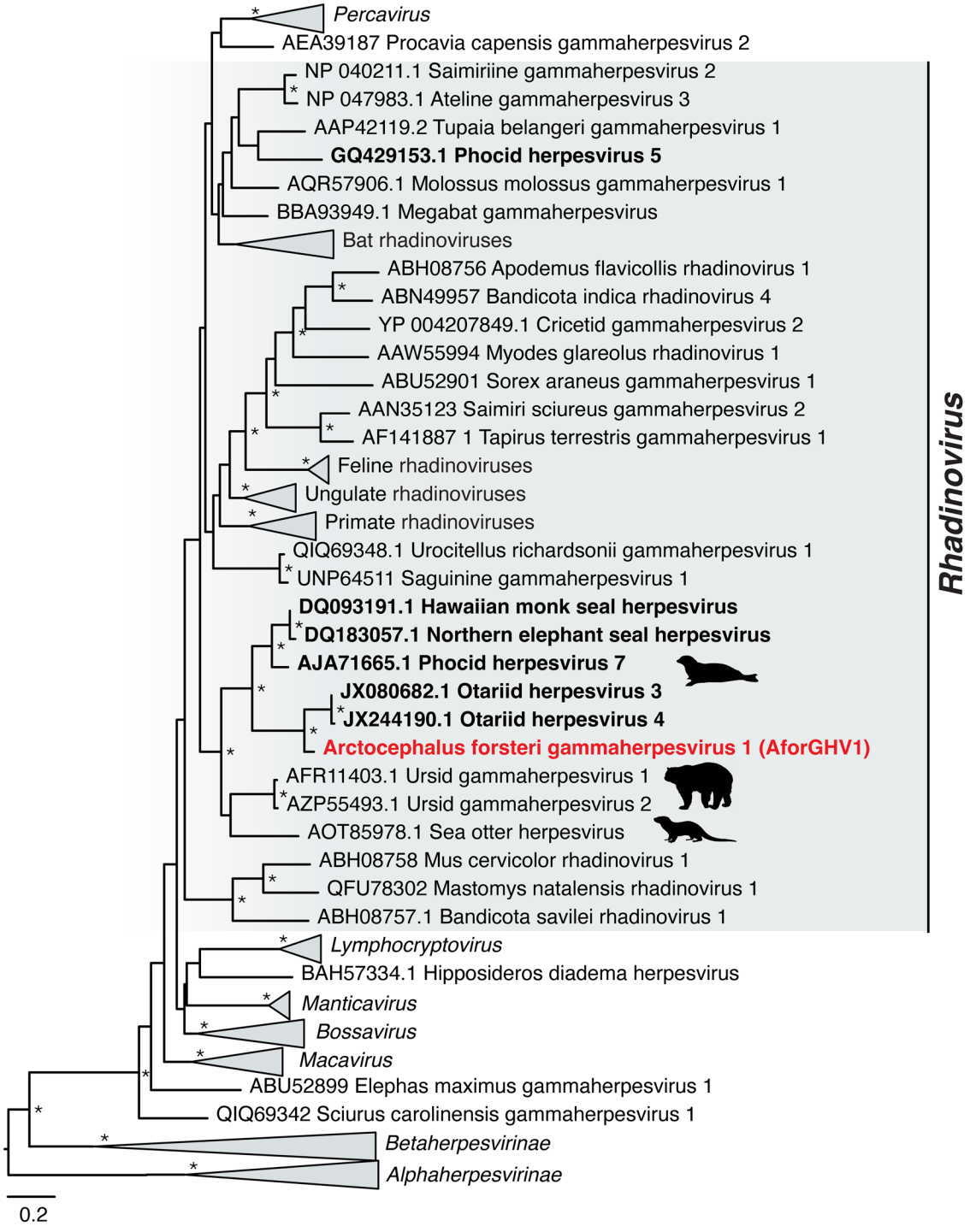
Wang, J., et al. (2010), 'The Canine Papillomavirus and Gamma HPV E7 Proteins Use an Alternative Domain to Bind and Destabilize the Retinoblastoma Protein', *PLoS Pathogens,* 6 (9), e1001089.

## 2.9 Supplementary information



***Supplementary Figure 2.1. Taxonomic assignments of contigs in sequencing libraries***

*(A) Krona graphs illustrating the relative abundance of taxa in the SL16 metatranscriptome at varying taxonomic levels. For clarity, a maximum depth of six taxonomic levels was chosen. (B) Krona graph of the non-chordate subset of taxa. Across both panels segments are highlighted based on the species' taxonomic grouping. Dots have been used to signify where contigs have been taxonomically assigned within the same family (Otariidae) as the New Zealand fur seal (Arctocephalus forsteri). Contigs without any matches in the database are not shown.*



***Supplementary Figure 2.1. Investigation of a partial gammaherpesvirus.*** *Phylogenetic relationships of the Herpesviridae. An ML phylogenetic tree based on the conserved amino acid sequences of the DNA polymerase gene with Arctocephalus forsteri gammaherpesvirus 1 (AforGHV1) shown in red and in the context of its closest relatives. The tip labels of pinniped associated herpesviruses are bolded. Animal silhouettes depict the virus-host associations for the rhadinoviruses. All branches are scaled to the number of amino acid substitutions (model LG+F+R9) per site, and the tree is midpoint rooted for clarity only. An asterisk indicates node support where SH-aLRT >= 80% and UFboot >= 95%.*

**Supplementary Note 2.1. Investigation of a partial gammaherpesvirus.** In the SL16 library, several fragmented contigs exhibited similarities to gammaherpesviruses, comprising 0.005% of the total reads (n = 3,503). RT-PCR analysis confirmed that this sequence was exclusive to the oral tissue of seal Registry #10741, with no presence in the genital tissue of this individual. This seal was an immature male rescued at Cronulla, NSW, Australia, in September 2015 presenting in an emaciated body condition with various traumatic injuries and was euthanased due to a poor prognosis for recovery. Gross and histopathological examinations confirmed the animal's emaciated state, the presence of intestinal helminths (within the expected range for a free-ranging pinniped), severe traumatic injuries, and a neutrophilia consistent with a systemic inflammatory response secondary to infected wounds. Genital and oral mucosal tissue samples were taken from this individual for RNA extraction and pooled to form library SL16, the same library in which the papillomavirus contigs were identified.

Two contigs containing a partial polymerase gene (3,460 bp) and the major capsid protein (12,493 bp) were assembled from the RNA libraries. In an effort to recover more of the herpesvirus genome we used a viral particle enrichment protocol 1 designed to isolate and randomly amplify both DNA and RNA from the oral tissue of a juvenile male seal TARZ-10741. Briefly, tissue was homogenised, filtered, centrifuged, and nuclease treated prior to viral nucleic acid extraction with the QIAamp viral RNA mini kit (Qiagen) 1,2. Following this, nucleic acids were randomly amplified using the Whole Transcriptome Amplification kit (WTA2, Sigma Aldrich) with modifications 1, and purified using the GenElute PCR cleanup kit (Sigma Aldrich). The DNA library was prepared using the Illumina DNA M preparation kit and sequenced on the Illumina NovaSeq 6000 platform at the AGRF. As the standard clean-up protocol depletes amplicons < 500 bp, these amplicons underwent Illumina purification using beads at a ratio of 1.8x volume to supernatant.

DNA viral particle enrichment yielded related reads (6,310 reads when mapped to NC_035117, the closest relative with a 60% nucleotide identity cut-off), but we were unable to substantially extend the contigs recovered through RNA-seq or recover other core gammaherpesvirus genes. The partial polymerase sequence shared 72% amino acid identity with phocid herpesvirus 7 (AJA71665.1). Phylogenetic analysis revealed that this sequence grouped within a clade of pinniped-associated gammaherpesviruses within the genus *Rhadinovirus* (Supplementary Figure 2.2). We tentatively assign this virus as Arctocephalus forsteri gammaherpesvirus 1 (AforGHV1). Both AforGHV1 fragments have been deposited in GenBank and assigned accession numbers OR590706 and OR590707. Further work is undoubtedly needed to characterise this virus and to understand the presence and impact of gammaherpesviruses in New Zealand fur seals.

The following Supplementary tables are available as .xlsx files:

**Supplementary Table 1. Sample collection and library details**

*https://static-content.springer.com/esm/art%3A10.1038%2Fs44298-024-00020-w/MediaObjects/44298_2024_20_MOESM2_ESM.xlsx*

**Supplementary Table 2. Predicted ORFs, protein products and binding sites**

*https://static-content.springer.com/esm/art%3A10.1038%2Fs44298-024-00020-w/MediaObjects/44298_2024_20_MOESM3_ESM.xlsx*

**Supplementary Table 3. AforPV1 gene expression analysis**

*https://static-content.springer.com/esm/art%3A10.1038%2Fs44298-024-00020-w/MediaObjects/44298_2024_20_MOESM4_ESM.xlsx*

**Supplementary Table 4. SL16 metatranscriptome composition**

*https://static-content.springer.com/esm/art%3A10.1038%2Fs44298-024-00020-w/MediaObjects/44298_2024_20_MOESM5_ESM.xlsx*

# Chapter 3. Diverse Viral Pathogens in Australian Canines: Limited Geographic Structure and the First Detection of an RNA Virus in Dingoes

Jonathon C.O. Mifsud[1], Erin Harvey[1,] Kate Van Brussel[1], Annabelle Olsson[2], Benjamin J. Pitcher[3,4], Jane Hall[5], Heather Fenton[5], Brendan F. Alting[6], Sabrina Sadiq[1], Edward C. Holmes[1]

1. School of Medical Sciences, The University of Sydney, Sydney, New South Wales, 2006, Australia.
2. Wildlife Health and Conservation Hospital, Sydney School of Veterinary Science, The University of Sydney, Sydney, New South Wales, 2006, Australia.
3. Taronga Institute of Science and Learning, Taronga Conservation Society, Dubbo and Sydney, New South Wales, Australia
4. School of Natural Sciences, Faculty of Science and Engineering, Macquarie University, Sydney, New South Wales, Australia
5. Australian Registry of Wildlife Health, Taronga Conservation Society Australia, Mosman, New South Wales, 2088, Australia
6. Centre for Ecosystem Science, School of Biological, Earth and Environmental Sciences, University of New South Wales, Sydney, New South Wales, Australia

### 3.1 Abstract

Viruses impose a substantial disease burden on dogs that may pose zoonotic risks through close contacts with humans. Australia's geographic isolation, strict biosecurity measures, and native dingo populations present a unique model for understanding the spread and evolution of canine viruses. However, aside from a few well-characterised pathogens like canine parvovirus and canine herpesvirus, genomic data are limited for many common dog viruses, limiting our understanding of their evolution and disease ecology. Using a metatranscriptomic approach we identified the viruses in dogs and dingoes from various geographical locations across mainland Australia and sample types, revealing the presence of 86 vertebrate-associated viruses belonging to 16 distinct species including a new vesivirus-like species. Many of these viruses identified here have not previously been sequenced in Australia. We identified important dog pathogens associated with canine infectious respiratory disease syndrome—such as canine pneumovirus, canine herpesvirus, and canine respiratory coronavirus—and gastroenteritis, including canine parvovirus, canine coronavirus, and rotavirus A. The sequences of Australian canine viruses often occupied multiple distinct clades phylogenetically and had little geographic structure, suggesting multiple virus introductions and subsequent spread across the country. Notably, we identified the first RNA virus – rotavirus A – in a dingo. This virus was phylogenetically distinct from dog-associated rotavirus A sequences and more closely related to viruses found in humans and bats, indicative of the past cross-species transmission of a reassortant virus into dingoes, and shows dingoes and domestic dogs may have distinct viromes. Our findings greatly expand the knowledge of viral diversity in Australian canines, improving our understanding of viral movement into and within Australia, as well as the potential zoonotic risks associated with dogs and dingoes.

### 3.2 Introduction

Australia is host to two species of true dogs (tribe Canini)—the domestic dog (*Canis lupus familiaris*) and the dingo (*Canis lupus dingo*)—that arrived separately in Australia. The first species to arrive was the dingo, which appears most closely related to New Guinea singing dogs (*Canis lupus hallstromi*) (Fillios and Taçon 2016). The arrival of dingoes likely occurred at least 3,000 years ago, pre-dating the introduction of domestic dog breeds through Western colonisation from the 18th century onwards (Fillios and Taçon 2016). Dingoes can hybridise with domestic dogs, although recent evidence suggests that interbreeding is relatively rare and that behavioural barriers or assortative mating preferences exist (Cairns et al. 2023). Dingoes remain free-living across much of mainland Australia, often coming into contact with humans, wildlife, and domestic dogs (Bombara et al. 2017; Souilmi et al. 2024).

Viruses including canine parvovirus (CPV), canine distemper virus (CDV), and canine respiratory coronavirus (CRCoV) impact dog health in Australia (Kelman et al. 2019; Smith et al. 2021; Wyllie et al. 2016). Given the close proximity dogs share with humans and wildlife, viruses affecting these animals theoretically present a risk of becoming zoonotic or spilling over into wildlife (Ghasemzadeh and Namazi 2015; He et al. 2022). For example, the transmission of CPV between domestic and wild carnivores has been widely reported (Canuti et al. 2022; Kelman et al. 2020b; Nandi and Kumar 2010; Ndiana et al. 2021) and has been associated with clinical symptoms of parvovirus disease (Steinel et al. 2000; Steinel et al. 2001). Reverse zoonoses events – in which viruses are transmitted from humans to animals – are also of concern, with the transmission of human SARS-CoV-2 to companion animals, including dogs, an important example (Patterson et al. 2020). Moreover, efforts to manage canine viruses such as those that cause rabies have substantial economic consequences for pet owners and government agencies (Animal Health Australia 2011). Australia's aquatic borders and strict biosecurity laws regarding importation present a unique, but currently understudied model, for understanding virus spread and evolution in a relatively isolated locality (Harvey and Foster 2024; Polak 2019).

To date, much of our knowledge of viruses in Australian dogs is restricted to DNA viruses such as CPV (Kelman et al. 2020a). In instances where RNA viruses have been examined, this has been largely limited to clinical and epidemiological investigations without the generation of gene sequence data (Marshall et al. 1984; Wyllie et al. 2016), although a few exceptions exist (Bhatta et al. 2019; Crispe et al. 2011; Moreno et al. 2017; Moreno et al.

2018; Naylor et al. 2001; Naylor et al. 2002; Smith et al. 2021). This knowledge gap is even more apparent in dingoes, with only four cases of CPV reported in dingo puppies and a dingo-dog hybrid (Van Arkel et al. 2019). No metagenomic virus discovery studies of dingoes have been conducted, and it is not currently known if dogs and dingoes share any RNA viruses.

Two metagenomic studies have focused on the faecal virome of Australian shelter dogs with acute diarrhoea (Moreno et al. 2017) and chronic enteropathy (Moreno et al. 2018), identifying numerous DNA and RNA virus families, including the *Astroviridae*, *Coronaviridae* and *Parvoviridae*. However, both studies employed PCR-based sequence-independent, single-primer amplification (SISPA) of viral and genomic DNA. Given the inherent limitations (e.g., uneven and sometimes primer-specific coverage profiles) of this method and the increased potential of artefacts (e.g., repetitive regions) during *de novo* assembly, in most cases viral sequences were not classified at the species level, or made publicly available, nor were phylogenetic analyses conducted (Moreno et al. 2017; Moreno et al. 2018). Exceptions to this were canine kobu- and astroviruses, where nucleic acids from a single sample identified to have these viruses by SISPA metagenomic sequencing where used to determine the complete genome sequence using Sanger sequencing. These sequences were highly similar to known canine kobu- and astroviruses, sharing 99.41% and 99.14% nucleotide identity with the closest relatives, respectively. In addition, the limited genomic data for many dog viruses globally has hampered our understanding of the patterns of canine viral movement into and within Australia.

Given our limited understanding of canine virus evolution and disease ecology, combined with Australia's geographic isolation, strict biosecurity measures, and native dingo populations there is clear merit in describing the viruses in domestic dog and dingo populations in Australia. Accordingly, we employed a metatranscriptomic (i.e., total RNA-sequencing) to describe the viruses in domestic dog and dingo populations in Australia. We examined multiple tissue types (faecal, tissue, nasal and rectal swabs) across three states—Queensland (QLD), New South Wales (NSW), and Western Australia (WA)—and from different contexts, including dog shelters and rescues, community veterinary clinics, and wild populations. With this data in hand, we investigated virus diversity and evolution within the unique Australian context. In particular, we aimed to: (i) conduct the first metagenomic survey of viruses in dingoes, (ii) compare the viromes of domestic dogs and dingoes to determine the extent of cross-species virus transmission between them, (iii) identify any

potential zoonotic pathogens present, and (iv) examine the phylogenetic structure, geographical distribution, and connectivity of Australian canine viruses to infer patterns of virus introduction and spread across the country.

## 3.3. Methods

### 3.3.1 Sample collection and processing

All methods were approved by Taronga Conservation Society Australia's Animal Ethics Committee (approval number 4c/10/21) and the University of Sydney Animal Ethics Committee (approval number 2022/2096). Carcasses of Myall Lake dingoes were transferred to the Australian Registry of Wildlife Health at Taronga Zoo, Sydney, Australia, for gross and histologic examination, as well as sample collection in accordance with NSW National Parks and Wildlife Act 1974, section 132c, Scientific License number SL100104. Where possible (i.e., not chemically euthanised) carcasses were returned to country for repatriation by Traditional Owners.

### 3.3.2 Sample collection and processing

Faeces, rectal and nasal swabs, tissues, and ectoparasites were collected from canines across Australia. Specifically, faecal samples were collected from four dog shelters in NSW: Central West, Illawarra, Coffs Harbour, and Hunter Valley, and one shelter in Perth, WA. Matching sets of nasal and rectal swabs were collected from dogs during three community veterinary clinics in Hope Vale, QLD. Tissue samples were collected from dingoes found deceased or euthanised on welfare grounds from Hawks Nest and Myall Lakes National Park, NSW. Carcasses were stored in a commercial-style chest freezer before being transported to the registry. The carcasses were allowed to thaw and necropsied the following day. Tissues were stored at -80 °C. All other sample types were placed in RNA*later*. A faecal sample was also opportunistically collected from Myall Lakes National Park (-32.626E, 152.201S). Trail cameras (browning strike force HD) were deployed at dingo scent-marking sites in Myall Lakes National Park as part of a separate study. This faecal sample was collected at one of these sites 21 hours after a dingo was seen on a trail camera displaying behaviours consistent with having just scent marked (slightly off camera), in the exact location the sample was collected and stored in a portable -80°C freezer. Additional dingo faecal samples were collected from a dingo rescue centre in WA and transported on ice to laboratory facilities. All samples were then stored at -80°C until RNA extraction. Dog age, gender, and general health status information were recorded by technicians for the shelter samples and by a veterinarian for the Hope Vale samples. Dog

ages were classified into three categories: puppies (less than 1 year old), young adults (1–2 years old), and adults (over 2 years old). Most animals showed no obvious signs of pathology. Those with signs of pathology are recorded in Supplementary Table 1.

### 3.3.3 RNA extraction and metatranscriptomic sequencing

Individual aliquots of each sample were placed into 600 µl of lysis buffer containing 1% β-mercaptoethanol (Sigma-Aldrich). Faecal samples and swabs were homogenised in a QIAshredder column (Qiagen). Dingo tissue was homogenised using a TissueRuptor (Qiagen) at a speed of 5,000 rpm for up to one minute with 0.5% antifoaming reagent (Reagent DX, Qiagen) added to the lysis buffer. RNA was then extracted from the supernatant using the RNeasy Plus Mini Kit (Qiagen), following the manufacturer's protocol. Extracted RNA was pooled in roughly equimolar ratios, grouping by location, sample type, gender, condition, and age. This resulted in 70 pools with a median of 2.5 samples per pool (minimum = 1, maximum = 6) (Supplementary Table 1). Sequencing libraries were constructed using the Truseq Total RNA Library Preparation Protocol (Illumina). Host ribosomal RNA (rRNA) was depleted using the Ribo-Zero Plus Kit (Illumina) and paired-end sequencing (150 bp) was performed on the NovaSeq 6000 platform (Illumina). Library construction and sequencing were performed by the Australian Genome Research Facility (AGRF) in Melbourne, Australia. Seventeen of the Hope Vale libraries were sequenced over two lanes. Each library generated from a lane was processed through the pipeline and analysed individually. When assembling final genomes and estimating abundance reads from both libraries, they were used as a single input.

### 3.3.4 Identification of novel virus sequences

The sequencing libraries were processed following the BatchArtemisSRAMiner pipeline (v1.0.4) (Mifsud 2023). Reads underwent quality trimming and adapter removal using Trimmomatic (v0.38) with parameters SLIDINGWINDOW:4:5, LEADING:5, TRAILING:5, and MINLEN:25, prior to assembly (Bolger et al. 2014). *De novo* assembly was conducted using MEGAHIT (v1.2.9) (Li et al. 2015). Assembled contigs were compared to the RdRp-scan (RNA-dependent RNA polymerase) RdRp core protein sequence database (v0.90) (Charon et al. 2022) and the protein version of the Reference Viral Databases (v26.0) (Bigot et al. 2019; Goodacre et al. 2018) using Diamond BLASTx (v2.1.6) (Buchfink et al. 2021) with an e-value cut-off of $1 \times 10^{-4}$ for the RdRp-scan database and $1 \times 10^{-10}$ for the RVDB database. To exclude potential false positives, contigs with hits to virus sequences were used as a query against the NCBI nucleotide database (as of December, 2023) using

BLASTN (Camacho et al. 2009) and the NCBI non-redundant protein (nr) database (as of December, 2023) using Diamond BLASTx. We used sequence similarity to known vertebrate viruses (e.g., % amino acid similarity), and in some cases phylogenetic grouping, to determine if a virus was potentially vertebrate infecting. Viruses identified as likely of non-vertebrate origin were excluded from further analysis.

To remove viruses present as a result of index-hopping from another library, we excluded those that meet the following three criteria: (1) sequenced on the same lane, (2) the total read count was <0.1% of the read count in the other library, and (3) were >99% identical at the nucleotide level. To screen for endogenous viral element (EVEs) within the viral-like contigs, all putative virus-like nucleotide sequences were compared to the corresponding host genome (*Canis lupus familiaris* or *Canis lupus dingo*) using the BLASTN algorithm with an E-value cutoff of $1 \times 10^{-20}$. In addition, all virus-like sequences were checked for host gene contamination using the contamination function implemented in CheckV (v0.8.1) (Nayfach et al. 2021). No EVEs were detected.

To infer whether a sequence represents a novel virus we used the demarcations set by the International Committee of Viral Taxonomy (ICTV) for each virus family. Where no demarcations were provided, we classified a virus as novel if it shared <95% amino acid identity across the RdRp and represented a distinct clade.

### 3.3.5 Genome extension, annotation and abundance

To potentially extend contigs and check for areas of heterogeneous coverage, sequence reads were mapped onto virus-like contigs with BBMap (v37.98) (Bushnell 2014). For known dog viruses, the closest full-genome representative sequence as determined by web BLASTN was used as a reference for mapping. The resulting library consensus was then remapped with 98% identity. ORFs were predicted using the Geneious Prime Find ORFs tool (v2022.0) (https://www.geneious.com/ ) (Kearse et al. 2012). Protein functional domains were annotated using the InterProScan software package (v5.56-89.0) (Jones et al. 2014) with the TIGRFAMs (v15.0), SFLD (v4.0), PANTHER (v15.0), SuperFamily (v1.75), PROSITE (v2022_01), CDD (v3.18), Pfam (v34.0), SMART (v7.1), PRINTS (v42.0), and CATH-Gene3D databases (v4.3.0). Classification of Rotavirus A (RVA) genomes was performed using the Rotavirus A Genotyping Tool Version 0.1 (https://www.rivm.nl/mpf/typingtool/rotavirusa/) (check Supplementary Table 2 for detailed results).

The abundance of each virus contig was calculated by using Bowtie2 (v2.2.5) (Langmead and Salzberg 2012). Trimmed reads were mapped against each virus contig at 99% identity. Samtools (v1.6) (Li et al. 2009) was used to calculate read coverage. Abundance was reported as a percentage of the total number of reads mapped divided by the total number of trimmed reads in that library.

### 3.3.6 Assessment of sequencing library composition

Library composition was assessed by Kraken2 (v2.1.2) (Breitwieser et al. 2018; Wood et al. 2019) for the purpose of confirming host species and investigating pathogen co-occurrence in libraries in which we identified canine infectious respiratory disease syndrome associated viruses. Kraken2 was ran on the trimmed reads using non-standard options --minimum-hit-groups 3 and --report-minimizer-data against the nt Kraken data set 5/30/2024 obtained from https://benlangmead.github.io/aws-indexes/k2. Bracken (v2.9) (Lu et al. 2017) was then used to estimate the abundance of each species, KrackenTools (v1.2) (Lu et al. 2022) and Krona (v2.8.1) (Ondov et al. 2011) were used to process and visualise the output.

### 3.3.7 Phylogenetic analysis

To determine the evolutionary relationships of the viral sequences obtained to known viruses, we inferred maximum likelihood phylogenetic trees derived from nucleotide and amino acid multiple sequence alignments. The sequence type (nucleotide or amino acid) and region (RdRp or capsid) used for the alignment were dependent on standard practice for the virus group in question, such as species demarcation criteria established by the ICTV and the degree of sequence similarity between the identified and reference sequences. For example, when analysing the phylogenetic relationships within a single virus species (e.g., canine associated *Sapporo virus* sequences), nucleotide sequences from the capsid region were typically used, as the amino acid sequences of RdRp tend to be nearly identical. Sequences determined in this study were aligned with reference sequences using MAFFT (v7.402) (Katoh and Standley 2013) using the E-INS-I algorithm. Ambiguously aligned regions were removed using trimAl (v1.4) (Capella-Gutiérrez et al. 2009), employing variable conservation thresholds (i.e., minimum percentage of alignment columns to retain) and gap thresholds (i.e., the minimum fraction of sequences without a gap needed to keep a column), as well as the automated parameter selection method gappyout.

Phylogenetic trees were estimated using IQ-TREE 2 (v2.1.7) (Minh et al. 2020) with the selection of the best-fit substitution model determined using ModelFinder (Kalyaanamoorthy et al. 2017). The sequence type, trimAl parameters and IQ-TREE 2 substitution model selected are presented in Supplementary Table 3. Branch support was calculated using 1,000 bootstrap replicates with the UFBoot2 algorithm and an implementation of the SH-like approximate likelihood ratio test within IQ-TREE 2. Phylogenetic trees were annotated using the R packages ape (5.7.1) (Paradis and Schliep 2019). Phytools (2.1.1) (Revell 2023) and ggtree (v3.8) (Yu et al. 2017), and further edited in Adobe Illustrator.

### 3.3.8 Statistical comparisons of virus count, diversity, and abundance

To examine how virus count, diversity, and abundance varied across sample type, host sex, and host age, we conducted Kruskal-Wallis tests using the stats package (v4.2.3) in R. Virus count was defined as the number of dog-associated viruses per library, while diversity was virus count summarised at a species level as to count only one virus per species. Abundance was measured as the percentage of total library reads assigned to that virus. For visualisation, boxplots were generated using the ggplot2 package (v3.5.1) in R.

### 3.3.9 Statistical comparisons of genetic and geographical distances

To determine the relationship between virus sequence relatedness and geographical distance, we performed Mantel tests for each virus species. As they are so closely related, canine vesivirus (CaVV) and the canine vesi-like viruses identified in our study were treated as a single species. Pairwise sequence identity percentages across the genome were obtained using BLASTN. In the case of segmented viruses only the RdRp-containing segment was used. Genetic distance was calculated by subtracting the pairwise sequence identity (%) from 100. Geographical distances between sampling locations were calculated using the Haversine formula using the distHaversine function from the geosphere package (v1.5-18) in R. Mantel tests were conducted using Spearman's rank correlation coefficient with 999 permutations through the mantel function from the vegan package (v2.6-4) in R.

## 3.4 Results

We characterised the virome of domestic dogs and dingoes sampled from locations across Australia. Total RNA-sequencing was performed on 185 samples pooled into 70 libraries (87 when considering the Hope Vale libraries sequenced on two lanes), grouping samples by location, sample type, sex, condition, and age (Figure 3.1). A total of 8,292,278,589

sequence reads were generated with a median of 155,755,544 (range 63,732,230 – 237,451,130). *De novo* assembly of the sequencing reads resulted in a median of 144,407 (range 251 – 738,858) contigs with a range of per library, for a total of 19,399,943 contigs. Library statistics are provided in Supplementary Table 4.

***Figure 3.13 Map showing the composition and number of the dog and dingo samples across mainland Australia.*** *The pie chart reflects sample type, and the size reflects the number of individual samples collected at each location. Sample sites where dingoes were collected are labelled in italics.*

### 3.4.1 Identification of potential dog and human pathogens in Australia

We focused our analysis on vertebrate infecting viruses, thereby excluding those that are likely associated with diet or gut flora such as the picobirnaviruses that are likely associated with bacteria (Sadiq et al. 2024). Overall, we detected 86 vertebrate-associated viruses, 82 of which were RNA viruses and 4 DNA viruses. These viruses were classified into 16 virus

species, 15 of which are known dog viruses as well as one novel species in the *Caliciviridae* family of RNA viruses. Members of the RNA virus order *Picornavirales* were the most frequent both in terms of the number of distinct species (n = 6) and the number of libraries (n = 34) (Supplementary Table 5). Members of the family *Picornaviridae* comprised a range of genera: *Kobuvirus*, *Dicipivirus*, *Mischivirus*, and *Sapovirus*. The remaining viruses were members of the *Caliciviridae, Coronaviridae*, *Pneumoviridae*, *Astroviridae*, and DNA viruses from the families *Parvoviridae* and *Orthoherpesviridae*. In contrast to many viral metagenomic studies in other mammals, nearly all the viruses discovered here are known rather than novel.

The number of canine viruses, their diversity in terms of the number of species, and their abundance varied across sample type. In particular, faeces contained significantly higher canine virus numbers (H (2) = 17, P = 0.0002), diversity (H (2) = 19.7, P = 0.00005) and abundance (H (2) = 19.7, P = 0.00001) than other sample types. There was also a significant difference in canine virus abundance between sexes (H (1) = 5.353, P = 0.02) with males having a higher overall abundance. No significant difference was observed in the number or diversity of canine viruses. There was no detectable relationship between these variables and dog age (Figure 3.2, Supplementary Table 6).

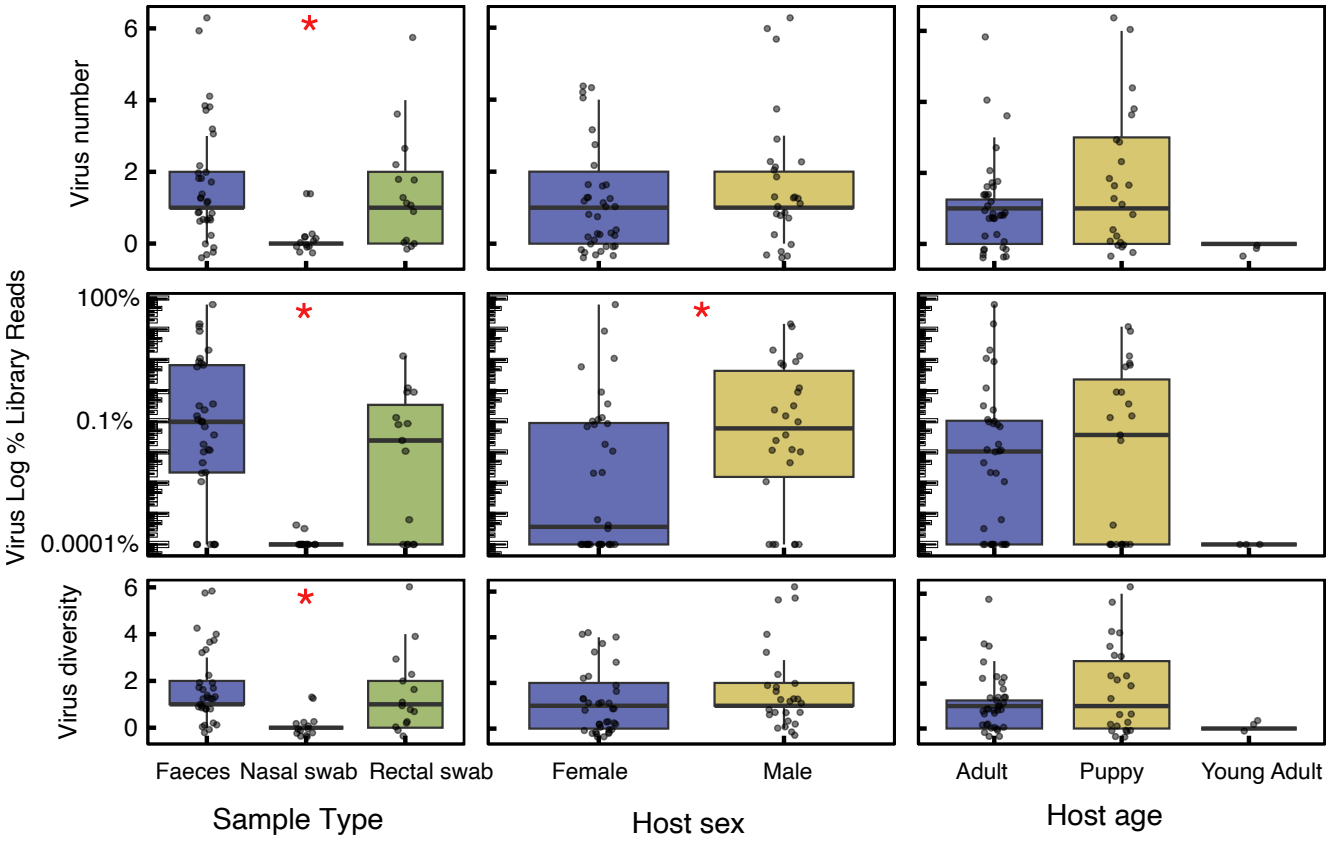


***Figure 3.14. Boxplot showing the differences in virus number, percentage of total library reads and virus diversity between sample type (faeces, nasal swab, and rectal swab), host sex (female and male) and host age (adult, young adult, and puppy).*** *Significant differences between the groups as determined by a Kruskal-Wallis rank sum test are shown with a red asterisk. In the boxplots, the middle lines show the median, and upper and lower hinges show the first and third quartiles.*

We now describe specific virus groups in turn.

***Canine infectious respiratory disease syndrome (CIRDS).*** This is a syndrome associated with respiratory symptoms and collectively described as “kennel cough”. It is associated with multiple bacterial and viral pathogens including *Bordetella bronchiseptica*, CRCoV, Canine pneumovirus (CnPnV), CDV, and Canine herpesvirus (CHV). We identified fragments of CnPnV, CRCoV and CHV in three different sequencing libraries (Supplementary 4). A 369 bp fragment of CnPnV was assembled from hopevale_40, a library comprising a single nasal swab from a female puppy, at a relatively low abundance of 0.36 fragments per kilobase of transcript per million mapped reads (FPKM) and representing $4.27 \times 10^{-6}$ % of total library reads. The associated veterinary report described the dog as “mangey, thin, kennel cough”. This fragment was most closely related to a virus associated with a 2012 CIRDC outbreak in Italy (NC_025344, 98.9% nucleotide identity) (Decaro et al. 2014). Phylogenetic trees inferred from the partial RdRp fragment placed this virus with the European Bari/100-12 sequence rather than those identified from outbreaks in Thailand or the US (Piewbang and Techangamsuwan 2019; Thieulent et al. 2024) (Figure 3.3a).

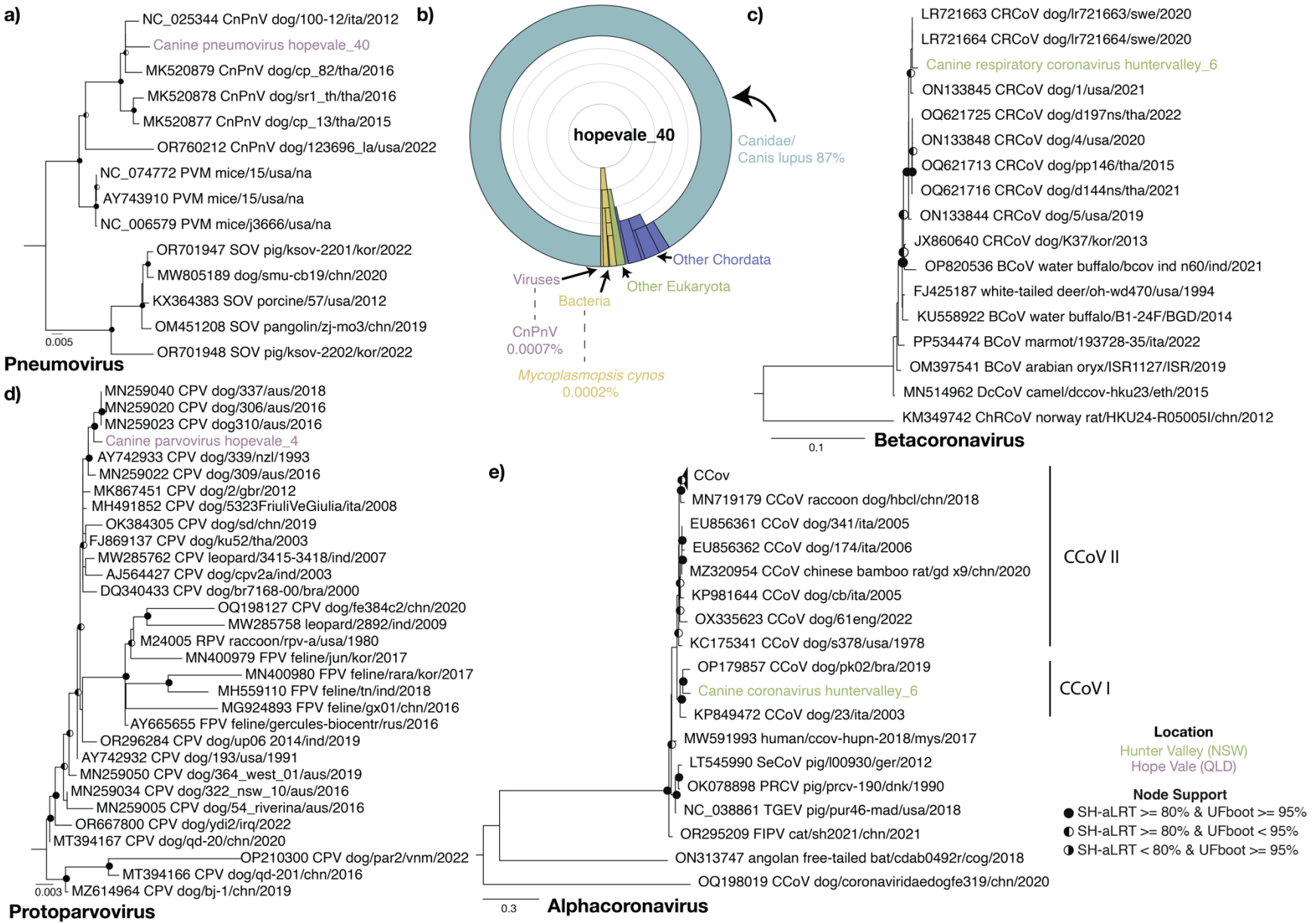

a)
NC_025344 CnPnV dog/100-12/ita/2012
Canine pneumovirus hopevale_40
MK520879 CnPnV dog/cp_82/tha/2016
MK520878 CnPnV dog/sr1_th/tha/2016
MK520877 CnPnV dog/cp_13/tha/2015
OR760212 CnPnV dog/123696_la/usa/2022
NC_074772 PVM mice/15/usa/na
AY743910 PVM mice/15/usa/na
NC_006579 PVM mice/j3666/usa/na
OR701947 SOV pig/ksov-2201/kor/2022
MW805189 dog/smu-cb19/chn/2020
KX364383 SOV porcine/57/usa/2012
OM451208 SOV pangolin/zj-mo3/chn/2019
OR701948 SOV pig/ksov-2202/kor/2022
0.005
Pneumovirus
b)
hopevale_40
Canidae/
Canis lupus 87%
Other Chordata
Other Eukaryota
Bacteria
Viruses
CnPnV
0.0007%
Mycoplasmopsis cynos
0.0002%
c)
LR721663 CRCoV dog/lr721663/swe/2020
LR721664 CRCoV dog/lr721664/swe/2020
Canine respiratory coronavirus huntervalley_6
ON133845 CRCoV dog/1/usa/2021
OQ621725 CRCoV dog/d197ns/tha/2022
ON133848 CRCoV dog/4/usa/2020
OQ621713 CRCoV dog/pp146/tha/2015
OQ621716 CRCoV dog/d144ns/tha/2021
ON133844 CRCoV dog/5/usa/2019
JX860640 CRCoV dog/K37/kor/2013
OP820536 BCoV water buffalo/bcov ind n60/ind/2021
FJ425187 white-tailed deer/oh-wd470/usa/1994
KU558922 BCoV water buffalo/B1-24F/BGD/2014
PP534474 BCoV marmot/193728-35/ita/2022
OM397541 BCoV arabian oryx/ISR1127/ISR/2019
MN514962 DcCoV camel/dccov-hku23/eth/2015
KM349742 ChRCoV norway rat/HKU24-R05005I/chn/2012
0.1
Betacoronavirus
d)
MN259040 CPV dog/337/aus/2018
MN259020 CPV dog/306/aus/2016
MN259023 CPV dog310/aus/2016
Canine parvovirus hopevale_4
AY742933 CPV dog/339/nzl/1993
MN259022 CPV dog/309/aus/2016
MK867451 CPV dog/2/gbr/2012
MH491852 CPV dog/5323FriuliVeGiulia/ita/2008
OK384305 CPV dog/sd/chn/2019
FJ869137 CPV dog/ku52/tha/2003
MW285762 CPV leopard/3415-3418/ind/2007
AJ564427 CPV dog/cpv2a/ind/2003
DQ340433 CPV dog/br7168-00/bra/2000
OQ198127 CPV dog/fe384c2/chn/2020
MW285758 leopard/2892/ind/2009
M24005 RPV raccoon/rpv-a/usa/1980
MN400979 FPV feline/jun/kor/2017
MN400980 FPV feline/rara/kor/2017
MH559110 FPV feline/tn/ind/2018
MG924893 FPV feline/gx01/chn/2016
AY665655 FPV feline/gercules-biocentr/rus/2016
OR296284 CPV dog/up06 2014/ind/2019
AY742932 CPV dog/193/usa/1991
MN259050 CPV dog/364_west_01/aus/2019
MN259034 CPV dog/322_nsw_10/aus/2016
MN259005 CPV dog/54_riverina/aus/2016
OR667800 CPV dog/ydi2/irq/2022
MT394167 CPV dog/qd-20/chn/2020
OP210300 CPV dog/par2/vnm/2022
MT394166 CPV dog/qd-201/chn/2016
MZ614964 CPV dog/bj-1/chn/2019
0.003
Protoparvovirus
e)
CCov
MN719179 CCoV raccoon dog/hbcl/chn/2018
EU856361 CCoV dog/341/ita/2005
EU856362 CCoV dog/174/ita/2006
MZ320954 CCoV chinese bamboo rat/gd x9/chn/2020
KP981644 CCoV dog/cb/ita/2005
OX335623 CCoV dog/61eng/2022
KC175341 CCoV dog/s378/usa/1978
OP179857 CCoV dog/pk02/bra/2019
Canine coronavirus huntervalley_6
KP849472 CCoV dog/23/ita/2003
MW591993 human/ccov-hupn-2018/mys/2017
LT545990 SeCoV pig/l00930/ger/2012
OK078898 PRCV pig/prcv-190/dnk/1990
NC_038861 TGEV pig/pur46-mad/usa/2018
OR295209 FIPV cat/sh2021/chn/2021
ON313747 angolan free-tailed bat/cdab0492r/cog/2018
OQ198019 CCoV dog/coronaviridaedogfe319/chn/2020
0.3
Alphacoronavirus
CCoV II
CCoV I
Location
Hunter Valley (NSW)
Hope Vale (QLD)
Node Support
SH-aLRT >= 80% & UFboot >= 95%
SH-aLRT >= 80% & UFboot < 95%
SH-aLRT < 80% & UFboot >= 95%

***Figure 3.15 Maximum likelihood phylogenetic trees of four viral pathogens.***
*Phylogenies were inferred using nucleotide sequences from the (a) Pneumovirus RdRp gene, (c) Betacoronavirus RdRp, (d) Protoparvovirus VP1, and (e) Alphacoronavirus ORF1b. The trees are midpoint rooted for clarity with the exception of the protoparvoviruses which are rooted on the node leading to OP210300, MT394166, MZ614964. Branch lengths reflect the number of substitutions per site. Strong node support (SH-aLRT >= 80% and UFboot >= 95%) is indicated by a black circle while intermediate support (SH-aLRT <80% and UFboot >95% or SH-aLRT >80% and UFboot <95%) is indicated by a half-filled circle. The tip labels of the sequences identified in this study are coloured by sampling location. (b). Krona chart showing the classification of reads associated with dogs and other eukaryotes as well as the potential agents of canine infectious respiratory disease syndrome.*

To examine whether any other pathogens were present in the hopevale_40 library, rRNA-depleted reads were assessed for taxonomic associations using Kraken. This revealed that most of the bacterial reads identified were associated with Mycoplasmatota (83% of bacteria, 2% of library), and a small percentage (0.01%) of these were associated with *Mycoplasmopsis cynos,* part of the CIRDS complex. *Mycoplasmopsis cynos* percentage library abundance was slightly less (0.0001%) than CnPnV (0.0007%). However, the majority of the Mycoplasmatota reads assigned at the species level were from *Mycoplasma cavipharyngis* (38%) and *Mycoplasmoides fastidiosum* (12%), which are not associated with disease (Figure 3.3b). No other dog pathogens were detected, and no CnPnV was found in the associated rectal swab from this individual.

Fragments of CRCoV, a *Betacoronavirus* associated with CIRDS, were assembled from the huntervalley_6 library that comprised five adult male dog faecal samples. There were no reports of dogs with respiratory symptoms or general illness. Both fragments were at low abundance, representing 0.24 and 0.32 FPKM and $3.14 \times 10^{-6}$ and $5.23 \times 10^{-6}$ percentage reads, respectively. Interestingly, there was only a single nucleotide change between these two fragments, 339 and 388 bp in length, and a sequence (LR721664.1) associated with a *Betacoronavirus* associated CIRDS outbreak in Swedish dogs, that was thought to have entered Sweden in approximately 2010 and has since become endemic (Wille et al. 2020). Phylogenies inferred from the RdRp region support this close association with the Swedish sequences (Figure 3.3c).

Fragments of CHV, an alphaherpesvirus, were identified in the hopevale_21 library which comprised nasal swabs from three adult female dogs without respiratory symptoms. The fragments shared 99-100% nucleotide identity with CHV sequence V1154 (KT819631.1). Phylogenetic analysis of the polymerase fragment placed this sequence within the larger diversity of CHV, although there was insufficient sequence to reliably infer evolutionary history and whether it was related to other Australian sequences.

***Gastrointestinal viruses.*** Gastroenteritis in dogs has been linked to a number of viruses including CPV, Canine coronavirus (CCoV), Rotavirus A (RVA), and potentially Canine vesivirus (CaVV) and Canine astrovirus (CaAstV) (Martella et al. 2012; Renshaw Randall et al. 2018). RVA is a leading cause of acute diarrhoea in humans, and dog-associated sequences have been associated with zoonotic transmission to humans. Notably, we identified three species of these known pathogens in seemingly healthy dogs.

The complete genome of CPV was assembled in the hopevale_4 library that comprised rectal swabs from three male puppies with no signs of disease. The sequence represents 0.003% of total library reads. Across the VP1 and NS1 genes this sequence shared the highest nucleotide identity with those from New Zealand in 1993 (AY742933). Indeed, phylogenetic analysis of the VP1 placed this sequence with sequences sampled in 2016 and 2018 from NSW in what is termed as the CPV-2a clade 5 (Kwan et al. 2021) (Figure 3.3d).

CCoV was assembled from the huntervalley_6 library which comprised a pool of 5 adult male dog faecal samples who were seemingly healthy, and represented 0.34% of total library read abundance. This sequence is distinct from the previously identified CCoVs in Australia, sharing 70%, 60%, and 50% amino acid identity in the spike to the QLD CCoV-IIb sequence (MW383487) (Smith et al. 2021) and two divergent NSW sequences [AF516907 (Naylor et al. 2002) and AF327928 (Naylor et al. 2001)], respectively. Phylogenetic analysis of huntervalley_6 ORF1b placed huntervalley_6 in CCoV I (Figure 3.3e). As with other CCoV I sequences, an additional accessory protein gene ORF3 was predicted between nucleotides 25,518 and 26,141.

Coding complete genomes of RVA were recovered in four libraries from the Hunter Valley (n = 4) and Central West shelter (n = 1), in both puppies and adults. Of particular note, partial VP1 (407 bp), NSP3 (395 bp) and NSP1 (317 bp) sequences were also recovered in the

library of a dingo from Myall Lakes National Park that comprised a single faecal sample. With the exception of the fragments in the myalllakesdingo_4 library (7.55 x $10^{-6}$%), these viruses were very highly abundant, on average representing 22% of library reads (min = 0.015%, max = 64.45%). Viral abundance was also higher in adult dogs than puppies. Despite this, both the dogs and dingoes were reported to be healthy.

The RVA sequences were classified using an approach used by the Rotavirus Classification Working Group (RCWG) based on the 11 genome segments (Matthijnssens et al. 2011). Accordingly, our sequences were classified as 'G3-P[3]-I3-R3-C3-M3-A9-N2-T3-E3-H6'. The Myall Lake dingo sequences were also classified as R3 (VP1) and T3 (NSP3), but typing could not be performed with the NSP1 fragment. This classification was supported by phylogenetic analysis of the VP4 gene, which placed the sequences in a monophyletic group with other members of the same genotype identified in dogs from Thailand (Figure 3.4a). As no VP4 or VP7 sequences were recovered for the myalllakesdingo_4 RVA, a phylogenetic analysis was performed using VP1 (Figure 3.4b). This revealed that the Myall Lakes dingo RVA was distinct from the those identified in dogs, sharing ~90% nucleotide identity across the fragments (Figure 3.4a). Instead, notably, the dingo RVA VP1 sequence grouped with a human associated RVA sequence (KU597747) identified from a child with diarrhoea from China in 2014 which is thought to have originated from a reassortment from bat MYAS33 or simian rotaviruses (Dong et al. 2016). This similarity to non-canine associated sequences is also apparent in the other two fragments recovered in this library, with all three fragments having closest nucleotide matches to *Myotis horsfieldii* (OR868019, OR868019) and *Rhinolophus affinis* (OR868034) RVA sequences, identified in bats from China in 2016 and 2017, respectively.

The taxonomic composition of the myalllakedingo_4 library was further analysed to check for evidence of another possible source of RVA origin (dietary or contamination). The myalllakedingo_4 library was predominately comprised of bacteria (97%). Of the eukaryotic reads detected in this library, 57% were associated with Arthropoda, 21% of which were associated with the dung flies (*Sepsis cynipsea*) (Figure 3.4c). Mammalian reads, which represent just 0.3% of all reads, were 64% *Panthera onca* and 16% *Canis lupus* (Figure 3.4d). To examine this further, we used BLASTN to compare the myalllakedingo_4 contigs to the nucleotide database and found that 74% of the vertebrate associated contigs with an abundance of >= 50 reads were associated with *Canis*. Notably, contigs with top hits to *Panthera onca* were associated with the reference *Panthera onca* voucher GTJAG006 16S

ribosomal RNA gene (KX419654) which, when further examined, was found to be a misannotated *Fusobacteriaceae* sequence. No human associated contigs were found to have >40 reads. A similar library composition was observed in the myalllakedingo_5 library, with 92% of sequence reads assigned to bacteria and 0.6% assigned to the genus *Canis*. In this case, the faecal sample was collected directly from the individual during dissection under sterile conditions.

A rarely detected rotavirus species, Rotavirus I (RVI), was recovered from five libraries from the Hunter Valley (n = 4) and Central West shelter (n = 1). Fragments were also detected in centralwest_4 library, but no VP1 or VP4 sequences were recovered. There was no clear link between the presence of this virus and pathology in the dogs (Supplementary Table 1). RVI sequences formed a single clade as a sister group to the known RVI sequences (Figure 3.4e), and there was very little diversity between our sequences (99.4% average nucleotide identity based on a multiple sequence alignment of the four sequences). RVI abundance varied between 0.0001% to 0.77% of total library reads, with an average of 0.19%.

Sequences related to CaVV were detected in faeces and rectal swab samples from 12 libraries. There was no association between the presence of CaVV and pathology in the dogs (Supplementary Table 1). Phylogenetic analysis of the VP1 demonstrates that Australian CaVV sequences formed multiple independent lineages within the diversity of this virus, including a distinct clade that was associated with Illawarra and Central West shelter dogs (Figure 3.5b). This clade appears divergent from known CaVV sequences, sharing ~72% amino acid identity with its closest relative from a canine nasal swab taken in China in 2024 (XHB18322). Consequently, this clade likely represents a novel species within the genus *Vesivirus*.

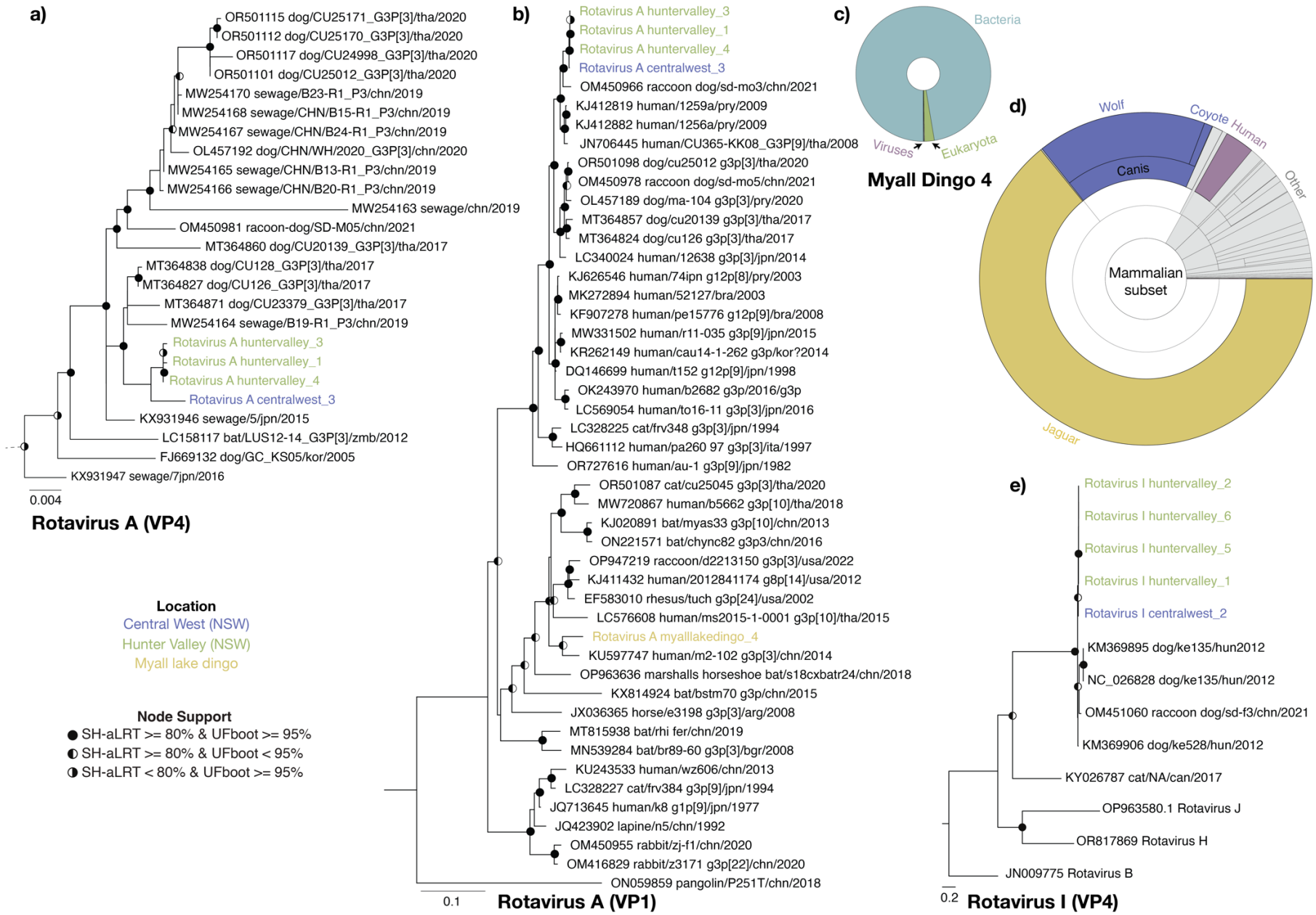

a)
OR501115 dog/CU25171_G3P[3]/tha/2020
OR501112 dog/CU25170_G3P[3]/tha/2020
OR501117 dog/CU24998_G3P[3]/tha/2020
OR501101 dog/CU25012_G3P[3]/tha/2020
MW254170 sewage/B23-R1_P3/chn/2019
MW254168 sewage/CHN/B15-R1_P3/chn/2019
MW254167 sewage/CHN/B24-R1_P3/chn/2019
OL457192 dog/CHN/WH/2020_G3P[3]/chn/2020
MW254165 sewage/CHN/B13-R1_P3/chn/2019
MW254166 sewage/CHN/B20-R1_P3/chn/2019
MW254163 sewage/chn/2019
OM450981 racoon-dog/SD-M05/chn/2021
MT364860 dog/CU20139_G3P[3]/tha/2017
MT364838 dog/CU128_G3P[3]/tha/2017
MT364827 dog/CU126_G3P[3]/tha/2017
MT364871 dog/CU23379_G3P[3]/tha/2017
MW254164 sewage/B19-R1_P3/chn/2019
Rotavirus A huntervalley_3
Rotavirus A huntervalley_1
Rotavirus A huntervalley_4
Rotavirus A centralwest_3
KX931946 sewage/5/jpn/2015
LC158117 bat/LUS12-14_G3P[3]/zmb/2012
FJ669132 dog/GC_KS05/kor/2005
KX931947 sewage/7jpn/2016
0.004
Rotavirus A (VP4)
Location
Central West (NSW)
Hunter Valley (NSW)
Myall lake dingo
Node Support
● SH-aLRT >= 80% & UFboot >= 95%
◐ SH-aLRT >= 80% & UFboot < 95%
◑ SH-aLRT < 80% & UFboot >= 95%
b)
Rotavirus A huntervalley_3
Rotavirus A huntervalley_1
Rotavirus A huntervalley_4
Rotavirus A centralwest_3
OM450966 raccoon dog/sd-mo3/chn/2021
KJ412819 human/1259a/pry/2009
KJ412882 human/1256a/pry/2009
JN706445 human/CU365-KK08_G3P[9]/tha/2008
OR501098 dog/cu25012 g3p[3]/tha/2020
OM450978 raccoon dog/sd-mo5/chn/2021
OL457189 dog/ma-104 g3p[3]/pry/2020
MT364857 dog/cu20139 g3p[3]/tha/2017
MT364824 dog/cu126 g3p[3]/tha/2017
LC340024 human/12638 g3p[3]/jpn/2014
KJ626546 human/74ipn g12p[8]/pry/2003
MK272894 human/52127/bra/2003
KF907278 human/pe15776 g12p[9]/bra/2008
MW331502 human/r11-035 g3p[9]/jpn/2015
KR262149 human/cau14-1-262 g3p/kor?2014
DQ146699 human/t152 g12p[9]/jpn/1998
OK243970 human/b2682 g3p/2016/g3p
LC569054 human/to16-11 g3p[3]/jpn/2016
LC328225 cat/frv348 g3p[3]/jpn/1994
HQ661112 human/pa260 97 g3p[3]/ita/1997
OR727616 human/au-1 g3p[9]/jpn/1982
OR501087 cat/cu25045 g3p[3]/tha/2020
MW720867 human/b5662 g3p[10]/tha/2018
KJ020891 bat/myas33 g3p[10]/chn/2013
ON221571 bat/chync82 g3p3/chn/2016
OP947219 raccoon/d2213150 g3p[3]/usa/2022
KJ411432 human/2012841174 g8p[14]/usa/2012
EF583010 rhesus/tuch g3p[24]/usa/2002
LC576608 human/ms2015-1-0001 g3p[10]/tha/2015
Rotavirus A myalllakedingo_4
KU597747 human/m2-102 g3p[3]/chn/2014
OP963636 marshalls horseshoe bat/s18cxbatr24/chn/2018
KX814924 bat/bstm70 g3p/chn/2015
JX036365 horse/e3198 g3p[3]/arg/2008
MT815938 bat/rhi fer/chn/2019
MN539284 bat/br89-60 g3p[3]/bgr/2008
KU243533 human/wz606/chn/2013
LC328227 cat/frv384 g3p[9]/jpn/1994
JQ713645 human/k8 g1p[9]/jpn/1977
JQ423902 lapine/n5/chn/1992
OM450955 rabbit/zj-f1/chn/2020
OM416829 rabbit/z3171 g3p[22]/chn/2020
ON059859 pangolin/P251T/chn/2018
0.1
Rotavirus A (VP1)
c)
Bacteria
Viruses
Eukaryota
Myall Dingo 4
d)
Wolf
Coyote
Human
Canis
Other
Mammalian subset
Jaguar
e)
Rotavirus I huntervalley_2
Rotavirus I huntervalley_6
Rotavirus I huntervalley_5
Rotavirus I huntervalley_1
Rotavirus I centralwest_2
KM369895 dog/ke135/hun2012
NC_026828 dog/ke135/hun/2012
OM451060 raccoon dog/sd-f3/chn/2021
KM369906 dog/ke528/hun/2012
KY026787 cat/NA/can/2017
OP963580.1 Rotavirus J
OR817869 Rotavirus H
JN009775 Rotavirus B
0.2
Rotavirus I (VP4)

***Figure 3.16 Maximum likelihood phylogenetic trees of canine rotaviruses.*** *Phylogenies were inferred using (a) Rotavirus A (RVA) VP1, (b) RVA VP4, and (c) Rotavirus I VP4. The trees are midpoint rooted for clarity, with branch lengths reflecting the number of substitutions per site. Strong node support (SH-aLRT >= 80% and UFboot >= 95%) is indicated by a black circle, while intermediate support (SH-aLRT <80% and UFboot >95% or SH-aLRT >80% and UFboot <95%) is indicated by a half-filled circle. (b) RVA VP4 is a clade extracted from a larger G3P3 RVA phylogeny available in the GitHub repository (https://github.com/JonathonMifsud/Australian-Canine-Virome/). The tip labels of sequences identified in this study are coloured by sampling location. (d) Krona chart showing the classification of reads associated in the myalllakedingo_4 library, Super kingdom level classification of Bacteria, Viruses and Eukaryota are labelled. (e) mammalian subset of the myalllakedingo_4 library composition as shown in d.*

### 3.4.2 Co-occurrence of viruses in faecal libraries

Cases of gastroenteritis are commonly characterised by coinfections between multiple viruses, with seemingly weak or unknown associations with disease. In addition to the viruses described above, we detected the following viruses in faeces or rectal swab libraries: Canine kobuvirus (CaKoV) (n = 17 libraries), CaVV/CaVV-like (n = 12), Canine picodicistrovirus (CPDV) (n = 10), CaAstV (n = 8), sapovirus (SaV) GXIII (n = 7), Norovirus (NoV) (n = 6), Canine picornavirus (CanPV) (n = 5) and ucalufa virus (n = 1) (Figure 3.5, 6). In these libraries, the presence of a second dog virus was detected ~50% of the time (19/38 libraries). The average number of co-occurring viruses per library was 2.02 (range 1 - 6). HU6 and CW3 had six, the former containing CaAstV, CaCV, CCoV, CRCoV, and RVI. There was no relationship between this and the number of individual samples per library and likewise between sampling location and the number of co-occurring viruses. CaKoV frequently co-occurred with other dog viruses (14/17) and was rarely the only virus found in that library. Notably, where noroviruses and SaVs were identified, CaKoV was also detected 83% and 71% of the time (Figure 3.7).

### 3.4.3 Genetic structure and evolution of Australian sequences

Understanding the evolution of dog viruses is hampered by the lack of sequences available in public databases. Indeed, eight of the 15 dog virus species detected in this study represent the first sequences available from Australia. For many of these species we found multiple distinct sequences within and between sampling locations and, more broadly,

Australian states and territories. In some cases, near identical (>99% identity) viruses are detected across states that are separated by over 2,000 km.

Of note, the same viral species were often found in multiple sampling locations and states. For example, CaAstV was present across all three states sampled (Figure 3.5d). In some cases, these viruses were near identical in sequence. Virus sharing, defined here as sequences sharing >99% pairwise nucleotide identity across the genome, was observed on 13 occasions, 11 of which occurred within sampling locations (11/13). On two occasions near identical virus NoV genotype GVI sequences were detected in Central West and Hope Vale, sharing 99.2% nucleotide identity across the genome (Figure 3.6b, c). The second was a CanPV sequence that shared 99.8% identity between Central West and Coffs Harbour shelters (Figure 3.6d). Conversely, we also observed extensive sequence diversity for a given virus species within a sampling location. For example, CaAstV identified in the huntervalley_6 and huntervalley_7 libraries shared only 89% nucleotide identity. On three occasions what appeared to be multiple distinct viruses belonging to the same species were detected within a given library. These included three CaAstVs in hopevale_4, three CaCV in perth_5 and two CPDVs in perth_2 libraries. These libraries were pools comprised of three or more faecal samples.

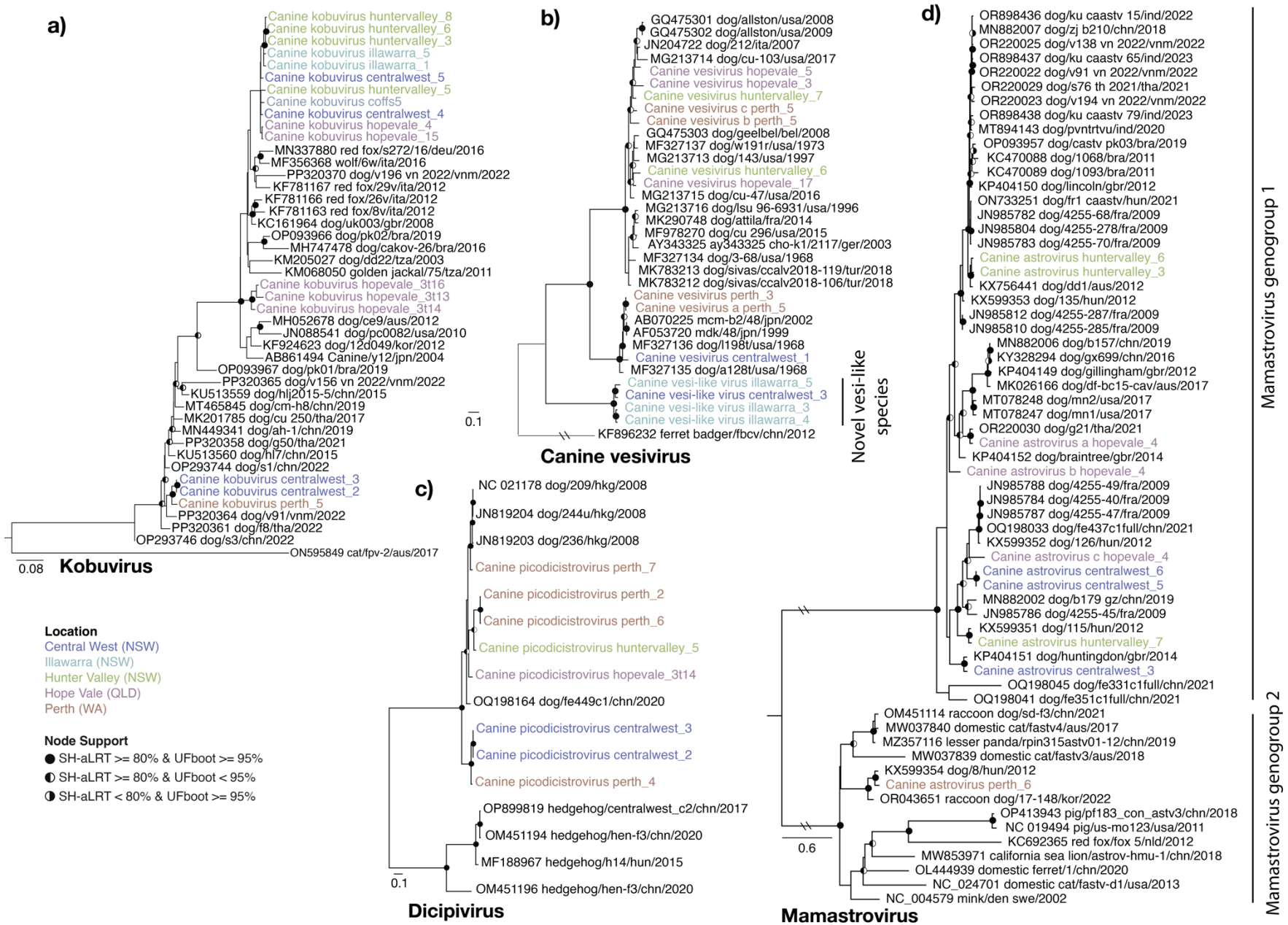

a)
Canine kobuvirus huntervalley_8
Canine kobuvirus huntervalley_6
Canine kobuvirus huntervalley_3
Canine kobuvirus illawarra_5
Canine kobuvirus illawarra_1
Canine kobuvirus centralwest_5
Canine kobuvirus huntervalley_5
Canine kobuvirus coffs5
Canine kobuvirus centralwest_4
Canine kobuvirus hopevale_4
Canine kobuvirus hopevale_15
MN337880 red fox/s272/16/deu/2016
MF356368 wolf/6w/ita/2016
PP320370 dog/v196 vn 2022/vnm/2022
KF781167 red fox/29v/ita/2012
KF781166 red fox/26v/ita/2012
KF781163 red fox/8v/ita/2012
KC161964 dog/uk003/gbr/2008
OP093966 dog/pk02/bra/2019
MH747478 dog/cakov-26/bra/2016
KM205027 dog/dd22/tza/2003
KM068050 golden jackal/75/tza/2011
Canine kobuvirus hopevale_3t16
Canine kobuvirus hopevale_3t13
Canine kobuvirus hopevale_3t14
MH052678 dog/ce9/aus/2012
JN088541 dog/pc0082/usa/2010
KF924623 dog/12d049/kor/2012
AB861494 Canine/y12/jpn/2004
OP093967 dog/pk01/bra/2019
PP320365 dog/v156 vn 2022/vnm/2022
KU513559 dog/hlj2015-5/chn/2015
MT465845 dog/cm-h8/chn/2019
MK201785 dog/cu 250/tha/2017
MN449341 dog/ah-1/chn/2019
PP320358 dog/g50/tha/2021
KU513560 dog/hl7/chn/2015
OP293744 dog/s1/chn/2022
Canine kobuvirus centralwest_3
Canine kobuvirus centralwest_2
Canine kobuvirus perth_5
PP320364 dog/v91/vnm/2022
PP320361 dog/f8/tha/2022
OP293746 dog/s3/chn/2022
ON595849 cat/fpv-2/aus/2017
0.08
Kobuvirus
b)
GQ475301 dog/allston/usa/2008
GQ475302 dog/allston/usa/2009
JN204722 dog/212/ita/2007
MG213714 dog/cu-103/usa/2017
Canine vesivirus hopevale_5
Canine vesivirus hopevale_3
Canine vesivirus huntervalley_7
Canine vesivirus c perth_5
Canine vesivirus b perth_5
GQ475303 dog/geelbel/bel/2008
MF327137 dog/w191r/usa/1973
MG213713 dog/143/usa/1997
Canine vesivirus huntervalley_6
Canine vesivirus hopevale_17
MG213715 dog/cu-47/usa/2016
MG213716 dog/lsu 96-6931/usa/1996
MK290748 dog/attila/fra/2014
MF978270 dog/cu 296/usa/2015
AY343325 ay343325 cho-k1/2117/ger/2003
MF327134 dog/3-68/usa/1968
MK783213 dog/sivas/ccalv2018-119/tur/2018
MK783212 dog/sivas/ccalv2018-106/tur/2018
Canine vesivirus perth_3
Canine vesivirus a perth_5
AB070225 mcm-b2/48/jpn/2002
AF053720 mdk/48/jpn/1999
MF327136 dog/l198t/usa/1968
Canine vesivirus centralwest_1
MF327135 dog/a128t/usa/1968
Canine vesi-like virus illawarra_5
Canine vesi-like virus centralwest_3
Canine vesi-like virus illawarra_3
Canine vesi-like virus illawarra_4
KF896232 ferret badger/fbcv/chn/2012
0.1
Canine vesivirus
Novel vesi-like species
c)
NC 021178 dog/209/hkg/2008
JN819204 dog/244u/hkg/2008
JN819203 dog/236/hkg/2008
Canine picodicistrovirus perth_7
Canine picodicistrovirus perth_2
Canine picodicistrovirus perth_6
Canine picodicistrovirus huntervalley_5
Canine picodicistrovirus hopevale_3t14
OQ198164 dog/fe449c1/chn/2020
Canine picodicistrovirus centralwest_3
Canine picodicistrovirus centralwest_2
Canine picodicistrovirus perth_4
OP899819 hedgehog/centralwest_c2/chn/2017
OM451194 hedgehog/hen-f3/chn/2020
MF188967 hedgehog/h14/hun/2015
OM451196 hedgehog/hen-f3/chn/2020
0.1
Dicipivirus
d)
OR898436 dog/ku caastv 15/ind/2022
MN882007 dog/zj b210/chn/2018
OR220025 dog/v138 vn 2022/vnm/2022
OR898437 dog/ku caastv 65/ind/2023
OR220022 dog/v91 vn 2022/vnm/2022
OR220029 dog/s76 th 2021/tha/2021
OR220023 dog/v194 vn 2022/vnm/2022
OR898438 dog/ku caastv 79/ind/2023
MT894143 dog/pvntrtvu/ind/2020
OP093957 dog/castv pk03/bra/2019
KC470088 dog/1068/bra/2011
KC470089 dog/1093/bra/2011
KP404150 dog/lincoln/gbr/2012
ON733251 dog/fr1 caastv/hun/2021
JN985782 dog/4255-68/fra/2009
JN985804 dog/4255-278/fra/2009
JN985783 dog/4255-70/fra/2009
Canine astrovirus huntervalley_6
Canine astrovirus huntervalley_3
KX756441 dog/dd1/aus/2012
KX599353 dog/135/hun/2012
JN985812 dog/4255-287/fra/2009
JN985810 dog/4255-285/fra/2009
MN882006 dog/b157/chn/2019
KY328294 dog/gx699/chn/2016
KP404149 dog/gillingham/gbr/2012
MK026166 dog/df-bc15-cav/aus/2017
MT078248 dog/mn2/usa/2017
MT078247 dog/mn1/usa/2017
OR220030 dog/g21/tha/2021
Canine astrovirus a hopevale_4
KP404152 dog/braintree/gbr/2014
Canine astrovirus b hopevale_4
JN985788 dog/4255-49/fra/2009
JN985784 dog/4255-40/fra/2009
JN985787 dog/4255-47/fra/2009
OQ198033 dog/fe437c1full/chn/2021
KX599352 dog/126/hun/2012
Canine astrovirus c hopevale_4
Canine astrovirus centralwest_6
Canine astrovirus centralwest_5
MN882002 dog/b179 gz/chn/2019
JN985786 dog/4255-45/fra/2009
KX599351 dog/115/hun/2012
Canine astrovirus huntervalley_7
KP404151 dog/huntingdon/gbr/2014
Canine astrovirus centralwest_3
OQ198045 dog/fe331c1full/chn/2021
OQ198041 dog/fe351c1full/chn/2021
OM451114 raccoon dog/sd-f3/chn/2021
MW037840 domestic cat/fastv4/aus/2017
MZ357116 lesser panda/rpin315astv01-12/chn/2019
MW037839 domestic cat/fastv3/aus/2018
KX599354 dog/8/hun/2012
Canine astrovirus perth_6
OR043651 raccoon dog/17-148/kor/2022
OP413943 pig/pf183_con_astv3/chn/2018
NC 019494 pig/us-mo123/usa/2011
KC692365 red fox/fox 5/nld/2012
MW853971 california sea lion/astrov-hmu-1/chn/2018
OL444939 domestic ferret/1/chn/2020
NC_024701 domestic cat/fastv-d1/usa/2013
NC_004579 mink/den swe/2002
0.6
Mamastrovirus
Mamastrovirus genogroup 1
Mamastrovirus genogroup 2
Location
Central West (NSW)
Illawarra (NSW)
Hunter Valley (NSW)
Hope Vale (QLD)
Perth (WA)
Node Support
● SH-aLRT >= 80% & UFboot >= 95%
◐ SH-aLRT >= 80% & UFboot < 95%
◑ SH-aLRT < 80% & UFboot >= 95%

***Figure 3.17 Maximum likelihood phylogenies of kobuviruses, vesivirus, dicipivirus and mamastroviruses.*** *Phylogenies were inferred using (a) Kobuvirus VP1 nucleotide, (b) Canine vesivirus VP1 amino acid, (c) Dicipivirus VP1 amino acid, and (d) Mamastrovirus capsid nucleotide sequences. The trees are midpoint rooted with the exception of Canine vesivirus which is rooted on the Ferret badger calicivirus (KF896232) outgroup, the sequence of which has for been removed for visual clarity. Branch lengths reflect the number of substitutions per site. Strong node support (SH-aLRT >= 80% and UFboot >= 95%) is indicated by a black circle while intermediate support (SH-aLRT <80% and UFboot >95% or SH-aLRT >80% and UFboot <95%) is indicated by a half-filled circle. The tip labels of sequences identified in this study are coloured by sampling location. The two lines on the branches of the Mamastrovirus tree indicate that these branches were shortened for visualisation purposes.*

Finally, we tested whether there was an association between pairwise nucleotide identity and the geographical distance between sampling locations for each virus species (i.e., whether sequences from shelters within the same region, such as NSW, share greater similarity compared to sequences from shelters in different regions, such as NSW and WA). Overall, we observed little correlation between geographical distance and genetic distance (Supplementary Table 7, Supplementary Figure 3.1). The exceptions were the small positive correlations observed for CanPV ($R = 0.78$, $p = 0.050$) and CaAstV ($R = 0.67$, $p = 0.023$), indicating that as geographic distance increased, so too did genetic distance.

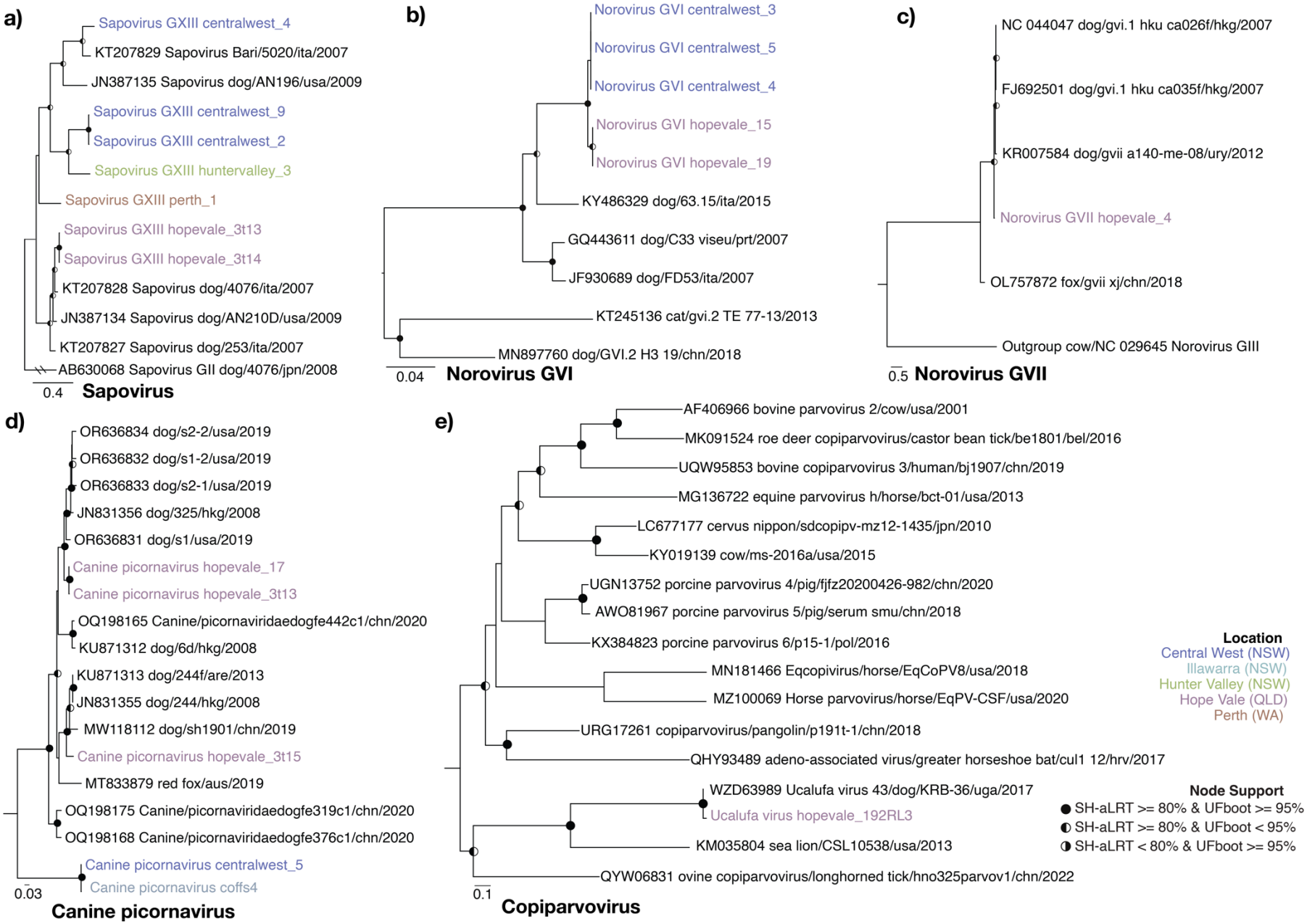

a)
Sapovirus GXIII centralwest_4
KT207829 Sapovirus Bari/5020/ita/2007
JN387135 Sapovirus dog/AN196/usa/2009
Sapovirus GXIII centralwest_9
Sapovirus GXIII centralwest_2
Sapovirus GXIII huntervalley_3
Sapovirus GXIII perth_1
Sapovirus GXIII hopevale_3t13
Sapovirus GXIII hopevale_3t14
KT207828 Sapovirus dog/4076/ita/2007
JN387134 Sapovirus dog/AN210D/usa/2009
KT207827 Sapovirus dog/253/ita/2007
AB630068 Sapovirus GII dog/4076/jpn/2008
0.4
Sapovirus
b)
Norovirus GVI centralwest_3
Norovirus GVI centralwest_5
Norovirus GVI centralwest_4
Norovirus GVI hopevale_15
Norovirus GVI hopevale_19
KY486329 dog/63.15/ita/2015
GQ443611 dog/C33 viseu/prt/2007
JF930689 dog/FD53/ita/2007
KT245136 cat/gvi.2 TE 77-13/2013
MN897760 dog/GVI.2 H3 19/chn/2018
0.04
Norovirus GVI
c)
NC 044047 dog/gvi.1 hku ca026f/hkg/2007
FJ692501 dog/gvi.1 hku ca035f/hkg/2007
KR007584 dog/gvii a140-me-08/ury/2012
Norovirus GVII hopevale_4
OL757872 fox/gvii xj/chn/2018
Outgroup cow/NC 029645 Norovirus GIII
0.5
Norovirus GVII
d)
OR636834 dog/s2-2/usa/2019
OR636832 dog/s1-2/usa/2019
OR636833 dog/s2-1/usa/2019
JN831356 dog/325/hkg/2008
OR636831 dog/s1/usa/2019
Canine picornavirus hopevale_17
Canine picornavirus hopevale_3t13
OQ198165 Canine/picornaviridaedogfe442c1/chn/2020
KU871312 dog/6d/hkg/2008
KU871313 dog/244f/are/2013
JN831355 dog/244/hkg/2008
MW118112 dog/sh1901/chn/2019
Canine picornavirus hopevale_3t15
MT833879 red fox/aus/2019
OQ198175 Canine/picornaviridaedogfe319c1/chn/2020
OQ198168 Canine/picornaviridaedogfe376c1/chn/2020
Canine picornavirus centralwest_5
Canine picornavirus coffs4
0.03
Canine picornavirus
e)
AF406966 bovine parvovirus 2/cow/usa/2001
MK091524 roe deer copiparvovirus/castor bean tick/be1801/bel/2016
UQW95853 bovine copiparvovirus 3/human/bj1907/chn/2019
MG136722 equine parvovirus h/horse/bct-01/usa/2013
LC677177 cervus nippon/sdcopipv-mz12-1435/jpn/2010
KY019139 cow/ms-2016a/usa/2015
UGN13752 porcine parvovirus 4/pig/fjfz20200426-982/chn/2020
AWO81967 porcine parvovirus 5/pig/serum smu/chn/2018
KX384823 porcine parvovirus 6/p15-1/pol/2016
MN181466 Eqcopivirus/horse/EqCoPV8/usa/2018
MZ100069 Horse parvovirus/horse/EqPV-CSF/usa/2020
URG17261 copiparvovirus/pangolin/p191t-1/chn/2018
QHY93489 adeno-associated virus/greater horseshoe bat/cul1 12/hrv/2017
WZD63989 Ucalufa virus 43/dog/KRB-36/uga/2017
Ucalufa virus hopevale_192RL3
KM035804 sea lion/CSL10538/usa/2013
QYW06831 ovine copiparvovirus/longhorned tick/hno325parvov1/chn/2022
0.1
Copiparvovirus
Location
Central West (NSW)
Illawarra (NSW)
Hunter Valley (NSW)
Hope Vale (QLD)
Perth (WA)
Node Support
SH-aLRT >= 80% & UFboot >= 95%
SH-aLRT >= 80% & UFboot < 95%
SH-aLRT < 80% & UFboot >= 95%

***Figure 3.18 Maximum likelihood phylogenies of sapoviruses, noroviruses, picornaviruses and copiparvoviruses.*** *Phylogenies were inferred using (a) Sapovirus VP1 nucleotide, (b) Norovirus GVI VP1 nucleotide, (c) Norovirus GVII VP1 nucleotide, (d) Canine picornavirus RdRp nucleotide, and (e) Copiparvovirus RdRp amino acid. The trees are midpoint rooted for clarity, with branch lengths reflecting the number of substitutions per site. Strong node support (SH-aLRT >= 80% and UFboot >= 95%) is indicated by a black circle while intermediate support (SH-aLRT <80% and UFboot >= 95% or SH-aLRT >80% and UFboot <95%) is indicated by a half-filled circle. The tip labels of sequences identified in this study are coloured by sampling location. The two lines on the branches of the Sapovirus tree indicate that these branches were shortened for visualisation purposes.*

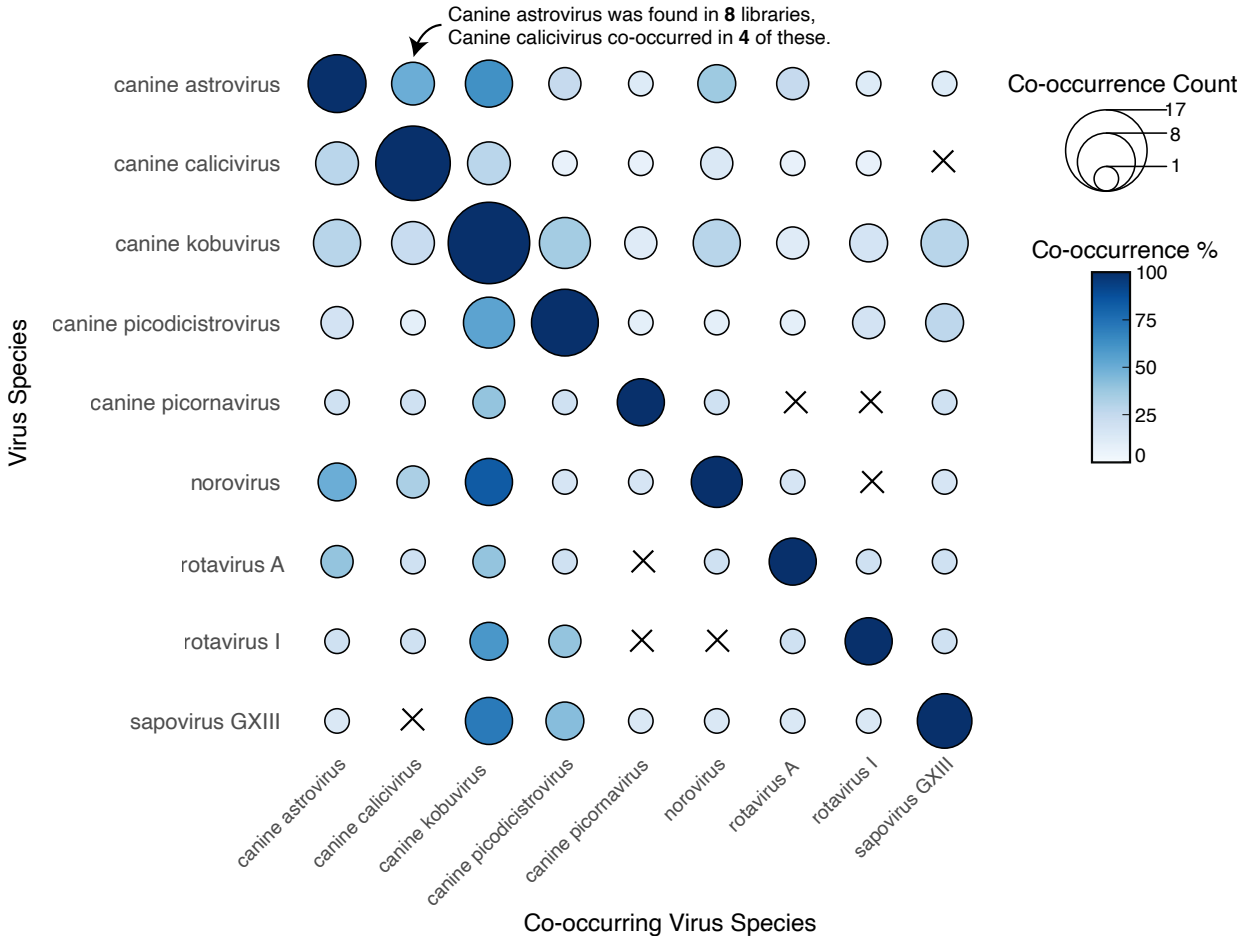



***Figure 3.19 Virus co-occurrence heatmap matrix.*** *For virus species that occur more than once, the frequency and percentage at which they are found with other viruses within a given library is shown. The size of the circle indicates the number of occasions while the colour*

*represents the percentage of time the "Co-occurring Virus Species" was identified in the libraries in which the "Virus Species" was found. An "X" indicates that no co-occurrences between these two virus species were found.*

### 3.5 Discussion

Viruses cause disease in both domestic and wild dogs (including dingoes) in Australia, resulting in health and financial burdens due to treatment, control, and biosecurity efforts. Given the close contact between dogs, humans, and wildlife, these viruses can have complex ecology with multiple host ranges and spill over events (Day et al. 2012; He et al. 2022; Wu et al. 2024). However, aside from a few well-characterised pathogens like CPV and CHV, genomic data are limited for many of the dog viruses prevalent overseas, restricting understanding of potential population-level impacts and disease ecology that may be relevant to wild and domestic animal management as well as public health. Australia's geographic isolation, strict biosecurity measures, and native dingo populations provide a unique opportunity to study viral evolution and transmission. Through a metatranscriptomic approach we characterised the viruses in dogs and dingoes from a variety of geographical locations and sample types, revealing the presence of 86 vertebrate associated viruses belonging to 16 distinct species. These included important canine and potentially zoonotic pathogens, many of which have not previously been sequenced in Australia.

Of note, we identified the first RNA virus – rotavirus A – in dingoes. In contrast to the only previously known dingo virus—a parvovirus in a dog-dingo hybrid that was closely related to domestic dog sequences (Van Arkel et al. 2019)—these RVA fragments are most closely related to human, and bat-associated viruses. This strongly suggests that the dingo RVA resulted from the cross-species transmission, although the exact source species is uncertain. It is notable that the RVA infected dingo lived in proximity to an urban centre and occupied a den site close to a bat colony. While not detected in the RNA of this dingo, preliminary diet analysis of dingo faeces in Myall Lakes National Park have identified some human derived DNA, most likely from the consumption of faeces (Myall Lakes Dingo/Dapin Project Unpublished data). Although domestic dog associated DNA viruses have previously been detected in dingoes (Van Arkel et al. 2019), this represents the first dingo virus that does not appear derived from those found in domestic dogs, and further shows that dogs and dingoes do not necessarily share the

same viruses despite their close genetic relationship. Importantly, there was no clear evidence that the RVA fragments were the result of contamination or misidentification of faeces during sample. The faecal sample was collected in the exact location where a dingo had been displaying behaviours consistent with having just scent marked 21 hours prior. No domestic dogs have previously been detected at that camera site, and dingoes in this area have high (>99%) dingo genetic ancestry (Cairns et al. 2020). Likewise, there was no strong evidence of contamination during extraction or dietary, such as a dingo consuming human or bat faeces, when the reads from the library in question were classified taxonomically. Most mammalian-associated reads from non-dingo organisms were misannotated bacterial reads, with canine-associated contigs comprising the majority of the mammalian-associated subset in this library. Despite this, because only fragments of VP1, NSP1, and NSP3 were identified, further work is needed to recover the complete genome of this virus and confirm its association with dingoes.

We detected six known pathogens associated with canine infectious respiratory disease syndrome (CnPnV, CHV, and CRCoV) and gastroenteritis (CPV, CCoV, and RVA). We also frequently detected CaKoV, CaVV, CaAstV, SaV, NoV, RVI in faecal samples, although they have only limited associations with gastroenteric disease and are largely uncharacterised outside their detection in symptomatic cases (Caddy and Goodfellow 2015; Li et al. 2011; Mihalov-Kovács et al. 2015; Ntafis et al. 2010; Renshaw Randall et al. 2018; Ribeiro et al. 2017). Despite the presence of six known canine pathogen—some at high abundance—there was strikingly little association between virus detection and clinical disease. It is important to note that there were no reports of diarrhoeal or gastrointestinal disease among the dogs sampled, and only one report of respiratory symptoms (kennel cough), from which CnPnV was recovered. These results suggest that these viruses—including zoonotic viruses like RVA—may cause transient or subclinical infections in dogs yet still be present in high abundance. It is possible that there as unreported diarrhoeal or gastrointestinal disease in these dogs or that disease occurred outside the observational window. This finding highlights the importance of including both healthy and sick dogs in pathogen surveillance and control efforts, as well as understanding that the disease manifestation may be multifactorial and associated with additional stressors.

Rotavirus A can cause severe, sometimes fatal, gastroenteritis in both humans and animals. The G3P[3] genotype is common in dogs globally and has occasionally been identified in acute

gastroenteritis cases in humans, including in Australia (Roczo-Farkas et al. 2017). These instances are thought to have been zoonotic transmission events originating in dogs, cats, or ruminants (Khamrin et al. 2006; Theamboonlers et al. 2013). While non-human RVAs such as G3P[3] are present in a small percentage of human gastroenteritis cases, through interspecies transmissions these viruses may contribute to reassortment and vaccine escape (Gentsch et al. 2005; Roczo-Farkas et al. 2017). Recently, RVAs with a unique genetic constellation, denoted 'G3-P[3]-I3-R3-C3-M3-A9-N2-T3-E3-H6', were discovered in Thailand (Theamboonlers et al. 2013), and the RVAs identified in shelter dogs in this study share the same constellation. Notably, this genotype contains segments that appear to be reassortments from human and feline RVA. Combined with the high viral abundance the detected in dog faeces, these findings highlight the potential risk of zoonotic transmission of these viruses.

For many viruses detected in this study, this is the first description of an Australian variant. These new sequences allow for initial evolutionary analyses, and although the number of sequences from Australian dogs was limited, they revealed that Australian viral lineages form multiple distinct clades within several virus groups (e.g., CanPV, SaV, CaCV, CaKoV, and mamastroviruses). In addition, the number of distinct clusters among Australian sequences suggests that there may have been multiple introductions of these viruses into Australia. However, there was often sequence divergence (>1% across the genome) between Australian sequences and the closest related sequences from overseas, complicating the reconstruction of potential transmission routes into Australia. This is likely due to a combination of older introductions and limited global sampling.

It is tempting to hypothesise how these viruses may have entered Australia. One possible explanation is the importation of domestic dogs from overseas, although introduced feral red foxes (V*ulpes vulpes*) are also common on the mainland. Around 6000 dogs are imported into Australia annually under strict importation protocols. With the exception of a select few countries (New Zealand, Norfolk Island, and Cocos Island), all are required to stay a minimum of 10 days in quarantine which can extend up to 180 days in some circumstances (https://www.agriculture.gov.au/about/news/on-the-record/importing-dogs-aus). Dogs are required to be vaccinated against rabies and recommended canine distemper, hepatitis, parvovirus, and parainfluenza vaccination. In combination with additional measures such as veterinary checkups, individual rooms, and strict protocols involving bedding and waste

management, this system likely reduces the risk of exotic viruses entering Australia, particularly for known pathogens. Nevertheless, as previously shown for CDV, viral shedding can exceed the current minimum quarantine period, potentially resulting in virus transmission after release (Allen et al. 2023). Several of the enteric viruses detected in this study are shown to be environmentally stable for extended periods of time and capable of transmission by fomites (Boone and Gerba 2007). Given this, it may be possible for these viruses to be transmitted through contaminated clothing or shoes. Other routes, such as the illegal entry of dogs or other carnivores on boats and the importation of dog semen, could also introduce viruses (Charles 2022).

Conversely, some virus clades consisted solely of Australian sequences that are relatively divergent from their international counterparts; for example, we detected a novel lineage of CaVV-like viruses which likely represent a novel species. In some cases (e.g., CaKoVs) one clade comprises members from five sampling locations across two states, spanning much of Australia's east coast (over 2,000 km), and was distinct from non-Australian CaKoVs (95% nucleotide identity with the nearest relatives). This suggests that these lineages may have persisted and spread across Australia for some time. Notably, there was also little geographical pattern observed in the Australian dog viruses described. This lack of pattern extends to genome identity, as geographically closer locations did not consistently contain more similar viruses, with the exceptions of CaKoV and CanPV. This is clearly apparent within the CaAstVs, in which some sequences sampled in different states share greater sequence identity than those within a single sampling location. The lack of geographical structure indicates virus connectivity/transmission between states, perhaps via the human- mediated interstate movement of dogs. It is also interesting to note that all imported dogs are sent to the same quarantine facility in Mickleham, Victoria. Indeed, despite separation by over 2000 km, the norovirus GVI sequences identified here shared 99.2% nucleotide identity across the genome.

In sum, through our characterisation of viruses in dogs and dingoes across Australia, we sequenced known dog pathogens in the country for the first time and identified the first RNA virus in dingoes which was distinct from viruses associated with domestic dogs. These findings provide insights into the viral diversity in Australian canines, with implications for understanding patterns in viral movement into and within Australia and the potential zoonotic risks associated with canines and dingoes.

### 3.6 Data availability

All raw data (fastq files) generated for this study are available in the NCBI SRA database under BioProject XX, BioSample accessions XX, and SRA accessions XX. The viral genomes assembled in this study have been deposited in the GenBank and assigned the accession numbers XX:XX. The sequences, alignments, and phylogenetic trees generated in this study are available at: https://github.com/JonathonMifsud/Australian-Canine-Virome

### 3.7 Acknowledgements

We acknowledge the University of Sydney's high-performance computing cluster Artemis for providing the computing resources used for this study. We thank Gemma Ma, Nicole Balzer, Deanna Douglas, Hannah Dreaver and Kasey Bridge for assistance with sample collection. Thanks to NSW NPWS staff for assistance with sample collection and storage. The Myall Lakes Dingo/Dapin Project acknowledges the Worimi Traditional Owners on who's land their research takes place. E.C.H. is supported by an National Health and Medical Research Council (NHMRC) Investigator Fellowship (GNT2017197) and by an Australian Research Council (ARC) Discovery Project grant (DP240101313). J.C.O.M. is supported by an Australian Government Research Training Program (RTP) Scholarship.

### 3.8 References


Allen, C., et al. (2023), 'Prolonged persistence of canine distemper virus RNA, and virus isolation in naturally infected shelter dogs', *PLOS One,* 18 (1), e0280186.

Animal Health Australia (2011), 'Disease Strategy Rabies: (Version 3.0)', *Australian Veterinary Emergency Plan (AUSVETPLAN)* (Canberra, ACT).

Bhatta, T. R., et al. (2019), 'Detection and characterisation of canine astrovirus, canine parvovirus and canine papillomavirus in puppies using next generation sequencing', *Scientific Reports,* 9 (1), 4602.

Bigot, T., et al. (2019), 'RVDB-prot, a reference viral protein database and its HMM profiles', *F1000Research,* 8.

Bolger, A. M., Lohse, M., and Usadel, B. (2014), 'Trimmomatic: a flexible trimmer for Illumina sequence data', *Bioinformatics,* 30 (15), 2114-20.

Bombara, C., et al. (2017), 'Roaming of dogs in remote Indigenous communities in northern Australia and potential interaction between community and wild dogs', *Australian Veterinary Journal,* 95 (6), 182-88.

Boone, S. A. and Gerba, C. P. (2007), 'Significance of fomites in the spread of respiratory and enteric viral disease', *Appl Environ Microbiol,* 73 (6), 1687-96.

Breitwieser, F. P., Baker, D. N., and Salzberg, S. L. (2018), 'KrakenUniq: confident and fast metagenomics classification using unique k-mer counts', *Genome Biology,* 19 (1), 198.

Buchfink, B., Reuter, K., and Drost, H.-G. (2021), 'Sensitive protein alignments at tree-of-life scale using DIAMOND', *Nature Methods,* 18 (4), 366-68.

Bushnell, B. (2014), 'BBMap: a fast, accurate, splice-aware aligner', (Lawrence Berkeley National Lab (LBNL), Berkeley, CA.).

Caddy, S. L. and Goodfellow, I. (2015), 'Complete genome sequence of canine astrovirus with molecular and epidemiological characterisation of UK strains', *Veterinary Microbiology,* 177 (1-2), 206-13.

Cairns, K. M., et al. (2020), 'Geographic hot spots of dingo genetic ancestry in southeastern Australia despite hybridisation with domestic dogs', *Conservation Genetics,* 21 (1), 77-90.

Cairns, K. M., et al. (2023), 'Genome-wide variant analyses reveal new patterns of admixture and population structure in Australian dingoes', *Molecular Ecology,* 32 (15), 4133-50.

Camacho, C., et al. (2009), 'BLAST+: architecture and applications', *BMC Bioinformatics,* 10 (1), 1-9.

Canuti, M., et al. (2022), 'Co-circulation of five species of dog parvoviruses and canine adenovirus type 1 among gray wolves (Canis lupus) in northern Canada', *Transboundary and Emerging Diseases,* 69 (5), e1417-e33.

Capella-Gutiérrez, S., Silla-Martínez, J. M., and Gabaldón, T. (2009), 'trimAl: a tool for automated alignment trimming in large-scale phylogenetic analyses', *Bioinformatics,* 25 (15), 1972-73.

Charles, C. (2022), 'First detection of canine herpesvirus (CHV) in dog semen: a contribution to the study of venereal transmission risk by real-time PCR and complementary serological analysis', *Sciences du Vivant*, 4021592.

Charon, J., et al. (2022), 'RdRp-scan: A bioinformatic resource to identify and annotate divergent RNA viruses in metagenomic sequence data', *Virus Evolution,* 8 (2), veac082.

Crispe, E., et al. (2011), 'Infection of dogs with equine influenza virus: evidence for transmission from horses during the Australian outbreak', *Australian Veterinary Journal,* 89, 27-28.

Day, M. J., et al. (2012), 'Surveillance of Zoonotic Infectious Disease Transmitted by Small Companion Animals', *Emerging Infectious Diseases,* 18 (12), e1-e1.

Decaro, N., et al. (2014), 'Full-Genome Analysis of a Canine Pneumovirus Causing Acute Respiratory Disease in Dogs, Italy', *PLOS One,* 9 (1).

Dong, H., et al. (2016), 'Genomic Characterization of an Unusual Human G3P [3] Rotavirus with Multiple Cross-species Reassortment', *Chinese Journal of Virology,* 32 (2), 129-40.

Fillios, M. A. and Taçon, P. S. (2016), 'Who let the dogs in? A review of the recent genetic evidence for the introduction of the dingo to Australia and implications for the movement of people', *Journal of Archaeological Science: Reports,* 7, 782-92.

Gentsch, J. R., et al. (2005), 'Serotype Diversity and Reassortment between Human and Animal Rotavirus Strains: Implications for Rotavirus Vaccine Programs', *The Journal of Infectious Diseases,* 192, S146-S59.

Ghasemzadeh, I. and Namazi, S. (2015), 'Review of bacterial and viral zoonotic infections transmitted by dogs', *Journal of Medicine and Life,* 8, 1.

Goodacre, N., et al. (2018), 'A reference viral database (RVDB) to enhance bioinformatics analysis of high-throughput sequencing for novel virus detection', *mSphere,* 3 (2), e00069-18.

Harvey, E. and Foster, C. S. (2024), 'Emerging viral threats in Australia', *Microbiology Australia,* 45 (1), 32-37.

He, W.-T., et al. (2022), 'Virome characterization of game animals in China reveals a spectrum of emerging pathogens', *Cell,* 185 (7), 1117-29.e8.

Jones, P., et al. (2014), 'InterProScan 5: genome-scale protein function classification', *Bioinformatics,* 30 (9), 1236-40.

Kalyaanamoorthy, S., et al. (2017), 'ModelFinder: fast model selection for accurate phylogenetic estimates', *Nature Methods,* 14 (6), 587-89.

Katoh, K. and Standley, D. M. (2013), 'MAFFT multiple sequence alignment software version 7: improvements in performance and usability', *Molecular Biology and Evolution,* 30 (4), 772-80.

Kearse, M., et al. (2012), 'Geneious Basic: an integrated and extendable desktop software platform for the organization and analysis of sequence data', *Bioinformatics,* 28 (12), 1647-49.

Kelman, M., et al. (2019), 'The geographic distribution and financial impact of canine parvovirus in Australia', *Transboundary and Emerging Diseases,* 66 (1), 299-311.

Kelman, M., et al. (2020a), 'A history of canine parvovirus in Australia: what can we learn?', *Australian Veterinary Journal,* 98 (10), 504-10.

Kelman, M., et al. (2020b), 'Phylogenetic and Geospatial Evidence of Canine Parvovirus Transmission between Wild Dogs and Domestic Dogs at the Urban Fringe in Australia', *Viruses-Basel,* 12 (6).

Khamrin, P., et al. (2006), 'Molecular characterization of a rare G3P [3] human rotavirus reassortant strain reveals evidence for multiple human-animal interspecies transmissions', *Journal of Medical Virology,* 78 (7), 986-94.

Kwan, E., et al. (2021), 'Analysis of canine parvoviruses circulating in Australia reveals predominance of variant 2b and identifies feline parvovirus-like mutations in the capsid proteins', *Transboundary and Emerging Diseases,* 68 (2), 656-66.

Langmead, B. and Salzberg, S. L. (2012), 'Fast gapped-read alignment with Bowtie 2', *Nature Methods,* 9 (4), 357-59.

Li, D., et al. (2015), 'MEGAHIT: an ultra-fast single-node solution for large and complex metagenomics assembly via succinct de Bruijn graph', *Bioinformatics,* 31 (10), 1674-76.

Li, H., et al. (2009), 'The sequence alignment/map format and SAMtools', *Bioinformatics,* 25 (16), 2078-79.

Li, L., et al. (2011), 'Viruses in diarrhoeic dogs include novel kobuviruses and sapoviruses', *Journal of General Virology,* 92, 2534-41.

Lu, J., et al. (2017), 'Bracken: estimating species abundance in metagenomics data', *PeerJ Computer Science,* 3, e104.

Lu, J., et al. (2022), 'Metagenome analysis using the Kraken software suite', *Nature Protocols,* 17 (12), 2815-39.

Marshall, J., et al. (1984), 'Viruses and virus-like particles in the faeces of dogs with and without diarrhoea', *Australian Veterinary Journal,* 61 (2), 33-38.

Martella, V., et al. (2012), 'Enteric disease in dogs naturally infected by a novel canine astrovirus', *Journal of Clinical Microbiology,* 50 (3), 1066-69.

Matthijnssens, J., et al. (2011), 'Uniformity of rotavirus strain nomenclature proposed by the Rotavirus Classification Working Group (RCWG)', *Archives of Virology,* 156, 1397-413.

Mifsud, J. C. O. (2023), 'BatchArtemisSRAMiner: v1.0.0', (Zenodo).

Mihalov-Kovács, E., et al. (2015), 'Candidate new rotavirus species in sheltered dogs, Hungary', *Emerging Infectious Diseases,* 21 (4), 660.

Minh, B. Q., et al. (2020), 'IQ-TREE 2: new models and efficient methods for phylogenetic inference in the genomic era', *Molecular Biology and Evolution,* 37 (5), 1530-34.

Moreno, P. S., et al. (2018), 'Characterization of the fecal virome in dogs with chronic enteropathy', *Veterinary Microbiology,* 221, 38-43.

Moreno, P. S., et al. (2017), 'Characterisation of the canine faecal virome in healthy dogs and dogs with acute diarrhoea using shotgun metagenomics', *PLOS One,* 12 (6), e0178433.

Nandi, S. and Kumar, M. (2010), 'Canine parvovirus: current perspective', *Indian Journal of Virology,* 21, 31-44.

Nayfach, S., et al. (2021), 'CheckV assesses the quality and completeness of metagenome-assembled viral genomes', *Nature Biotechnology,* 39 (5), 578-85.

Naylor, M. J., et al. (2001), 'Identification of canine coronavirus strains from feces by S gene nested PCR and molecular characterization of a new Australian isolate', *Journal of Clinical Microbiology,* 39 (3), 1036-41.

Naylor, M. J., et al. (2002), 'Molecular characterization confirms the presence of a divergent strain of canine coronavirus (UWSMN-1) in Australia', *Journal of Clinical Microbiology,* 40 (9), 3518-22.

Ndiana, L. A., et al. (2021), 'Circulation of diverse protoparvoviruses in wild carnivores, Italy', *Transboundary and Emerging Diseases,* 68 (4), 2489-502.

Ntafis, V., et al. (2010), 'Outbreak of Canine Norovirus Infection in Young Dogs', *Journal of Clinical Microbiology,* 48 (7), 2605-08.

Ondov, B. D., Bergman, N. H., and Phillippy, A. M. (2011), 'Interactive metagenomic visualization in a Web browser', *BMC Bioinformatics,* 12 (1), 385.

Paradis, E. and Schliep, K. (2019), 'ape 5.0: an environment for modern phylogenetics and evolutionary analyses in R', *Bioinformatics,* 35 (3), 526-28.

Patterson, E. I., et al. (2020), 'Evidence of exposure to SARS-CoV-2 in cats and dogs from households in Italy', *Nature Communications,* 11 (1), 6231.

Piewbang, C. and Techangamsuwan, S. (2019), 'Phylogenetic evidence of a novel lineage of canine pneumovirus and a naturally recombinant strain isolated from dogs with respiratory illness in Thailand', *BMC Veterinary Research,* 15 (1).

Polak, K. (2019), 'Dog transport and infectious disease risk: an international perspective', *Veterinary Clinics: Small Animal Practice,* 49 (4), 599-613.

Renshaw Randall, W., et al. (2018), 'Characterization of a Vesivirus Associated with an Outbreak of Acute Hemorrhagic Gastroenteritis in Domestic Dogs', *Journal of Clinical Microbiology,* 56 (5), 10.1128/jcm.01951-17.

Revell, L. J. (2023), 'phytools 2.0: An updated R ecosystem for phylogenetic comparative methods (and other things)', *bioRxiv*, 2023.03. 08.531791.

Ribeiro, J., et al. (2017), 'Extra-intestinal detection of canine kobuvirus in a puppy from Southern Brazil', *Archives of Virology,* 162, 867-72.

Roczo-Farkas, S., Kirkwood, C. D., and Bines, J. E. (2017), 'Australian rotavirus surveillance program: annual report, 2016', *Communicable Diseases Intelligence,* 41 (4).

Sadiq, S., Holmes, E. C., and Mahar, J. E. (2024), 'Genomic and phylogenetic features of the Picobirnaviridae suggest microbial rather than animal hosts', *Virus Evolution,* 10 (1), veae033.

Smith, C. S., et al. (2021), 'Identification of a canine coronavirus in Australian racing Greyhounds', *Journal of Veterinary Diagnostic Investigation,* 34 (1), 77-81.

Souilmi, Y., et al. (2024), 'Ancient genomes reveal over two thousand years of dingo population structure', *Proceedings of the National Academy of Sciences,* 121 (30), e2407584121.

Steinel, A., et al. (2000), 'Genetic characterization of feline parvovirus sequences from various carnivores', *Journal of General Virology,* 81 (2), 345-50.

Steinel, A., et al. (2001), 'Parvovirus infections in wild carnivores', *Journal of Wildlife Diseases,* 37 (3), 594-607.

Theamboonlers, A., et al. (2013), 'Complete genome analysis of a rare human G3P [9] rotavirus posing as an AU-1 like strain', *SpringerPlus,* 2, 1-5.

Thieulent, C. J., et al. (2024), 'Coding-complete genome sequences of two strains of canine pneumovirus derived from dogs with upper respiratory disease in the United States', *Microbiology Resource Announcements,* 13 (2), e0105723.

Van Arkel, A., et al. (2019), 'The relationship between reported domestic canine parvovirus cases and wild canid distribution', *Heliyon,* 5 (9).

Wille, M., et al. (2020), 'Evolutionary genetics of canine respiratory coronavirus and recent introduction into Swedish dogs', *Infection, Genetics and Evolution,* 82, 104290.

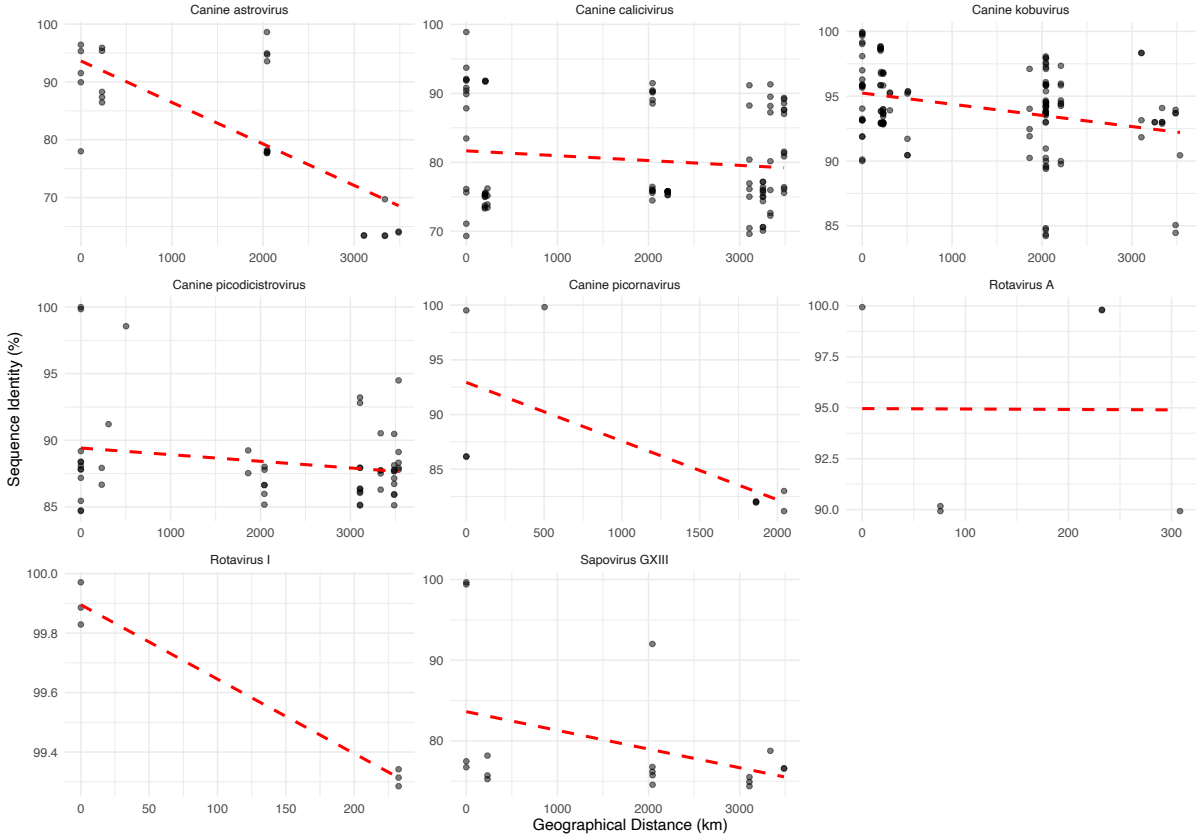
Wood, D. E., Lu, J., and Langmead, B. (2019), 'Improved metagenomic analysis with Kraken 2', *Genome Biology,* 20 (1), 257.

Wu, W.-C., et al. (2024), 'Meta-transcriptomic analysis of companion animal infectomes reveals their diversity and potential roles in animal and human disease', *mSphere,* 9 (8), e00439-24.

Wyllie, S., Kelman, M., and Ward, M. (2016), 'Epidemiology and clinical presentation of canine distemper disease in dogs and ferrets in Australia, 2006–2014', *Australian Veterinary Journal,* 94 (7), 215-22.

Yu, G., et al. (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.

## 3.9 Supplementary Information




***Supplementary Figure 3.1.*** *Association between geographical distance and sequence identity across viral species. Each panel represents a distinct virus species as labelled, with points indicating sequence pairs plotted by geographical distance between sampling locations (km) and pairwise nucleotide sequence identity (%). The red dashed line represents the linear regression fit.*

The following Supplementary tables are available as .xlsx files: https://drive.google.com/drive/folders/1S7XKKlecDYz3TaOgFvu991KA62mPZBTF?usp=share_link

**Supplementary Table 1.** Sample metadata for each library

**Supplementary Table 2.** Rotavirus A typing results

**Supplementary Table 3.** Sequence type, trimming methods, and phylogenetic models

**Supplementary Table 4.** Contig statistics for each library

**Supplementary Table 5.** Summary information for the viruses identified in this study

**Supplementary Table 6.** Data underlying the boxplot in Figure 3.2

**Supplementary Table 7.** Data underlying the genetic and geographical distances mantel test

# Chapter 4. Transcriptome Mining Expands Knowledge of RNA Viruses Across the Plant Kingdom

Jonathon C.O. Mifsud[1,2], Rachael V. Gallagher[2,3], Edward C. Holmes[1], Jemma L. Geoghegan[4,5]

1. Sydney Institute for Infectious Diseases, School of Life and Environmental Sciences and School of Medical Sciences, The University of Sydney, Sydney, NSW 2006, Australia.
2. Department of Biological Sciences, Macquarie University, North Ryde, NSW 2109 Australia
3. Hawkesbury Institute for the Environment, Western Sydney University, Locked Bag 1797, Penrith, NSW 2751, Australia
4. Department of Microbiology and Immunology, University of Otago, Dunedin 9016, New Zealand.
5. Institute of Environmental Science and Research, Wellington 5018, New Zealand.

### 4.1 Abstract

Our current understanding of plant viruses stems largely from those affecting economically important plants. Yet, plant species in cultivation represent a small and bias subset of the plant kingdom. Here, we describe virus diversity and abundance in 1079 transcriptomes from species across the breadth of the plant kingdom (Archaeplastida) by analysing open-source data from the One Thousand Plant Transcriptomes Initiative (1KP). We identified 104 potentially novel viruses, of which 40% comprised single-stranded positive-sense RNA viruses across eight orders, including members of the *Hepelivirales, Tymovirales*, *Cryppavirales*, *Martellivirales* and *Picornavirales*. One-third of the newly described viruses comprised double-stranded RNA viruses from the orders *Durnavirales* and *Ghabrivirales*. The remaining were negative-sense RNA viruses from the *Rhabdoviridae*, *Aspiviridae*, *Yueviridae*, *Phenuiviridae* and the newly proposed *Viridisbunyaviridae*. Our analysis considerably expands the known host range of 13 virus families to include lower plants (e.g., *Benyviridae* and *Secoviridae*) and four virus families to include algae hosts (e.g., *Tymoviridae* and *Chrysoviridae*). More broadly, however, a co-phylogeny analysis revealed that the evolutionary history of these families is largely driven by cross-species transmission events. The discovery of the first 30 kDa movement protein in a non-vascular plant, suggests that the acquisition of plant virus movement proteins occurred prior to the emergence of the plant vascular system. Together, these data highlight that numerous RNA virus families are associated with older evolutionary plant lineages than previously thought and that the apparent scarcity of RNA viruses found in lower plants likely reflects a lack of investigation rather than their absence.

## 4.2 Introduction

Viruses are responsible for almost 50% of all emerging plant disease (Anderson 2016). Historically, virus characterisation has focused on pathogenic viruses that infect species of economic importance with 69% of the current phytovirosphere — the total assemblage of plant viruses—identified in cultivated plant species even though they represent less than 0.17% of known plant diversity (Mifsud 2020; Wren et al. 2006). Importantly, the advent of metagenomic sequencing technology enables the comprehensive screening of plant tissues for novel and known viruses (Roossinck et al. 2015). Despite this, virus diversity in the vast majority of plants remains unquantified (Dolja et al. 2020).

Our ability to infer the origins and diversification of the phytovirosphere requires adequate sampling of the viruses across the plant kingdom. Several key plant groups are severely underrepresented or absent in previous studies, including charophytes, lower plants, gymnosperms and several angiosperm orders (Dolja et al. 2020; Shates et al. 2019). Improving knowledge across these groups will undoubtedly help uncover the evolutionary history of plant virus lineages. For instance, an analysis of the evolutionary history of viruses from rhodophytes and charophytes might reveal how key evolutionary transitions of plants – such as terrestrialisation – have shaped the contemporary land plant virome (Dolja et al. 2020). Similarly, through broad sampling across the plant kingdom, we can acquire a better understanding of the acquisition of viruses through cross-species transmission from plant-associated organisms such as invertebrates, fungi, or protists (Dolja et al. 2020).

The majority (68%) of the currently documented genera of plant viruses have positive-sense single-stranded RNA (+ssRNA) genomes and the majority of virus diversity is known only from angiosperms (Dolja and Koonin 2018) (Figure 4.1). Currently, 16 viruses belonging to 12 virus families have been found in gymnosperms (Han et al. 2002; Nibert et al. 2016; Sidharthan et al. 2021; Veliceasa et al. 2006; Yang et al. 2020). Outside of several viruses found in ferns, we know little of the diversity of viruses in the lycophytes, bryophytes and charophytes that together encompass ~27,000 species (Christenhusz and Byng 2016; Dawes 2016; Nibert et al. 2018; Valverde and Sabanadzovic 2009) (Figure 4.1). A partial analysis of published transcriptome data detected homologs of the canonical RNA virus RNA-dependent RNA polymerase (RdRp) in algae, several lower plants and gymnosperms (Mushegian et al. 2016).

However, it is uncertain whether viruses that infect freshwater algae—which include the *Zygnematophyceae* ancestors of land plants—resemble those infecting angiosperms or those infecting the chlorophytes, which are dominated by double stranded DNA (dsDNA) viruses, particularly from the *Phycodnaviridae* (Short et al. 2020). To date, two +ssRNA viruses related to the benyvirids have been identified in freshwater algae (Gibbs et al. 2011; Vlok et al. 2019). Unlike the Chlorophyta, the Charophyta characteristically contain plasmodesmata and homologs of the key components of the land plant innate immune system, both of which have been speculated to explain the apparent absence of double-strand (ds) DNA viruses in land plants (Brunkard and Zambryski 2017; Dolja et al. 2020; Han 2019). In particular, the plasmodesmata is thought to provide a physical barrier preventing the entry of medium to large virions which lack movement proteins (i.e., dsDNA viruses). An understanding of the viruses infecting the Charophyta and other lower plants is required to effectively test these ideas.

Transcriptome mining has become an efficient method of virus discovery that leverages previous investment (Bejerman et al. 2021; Gilbert et al. 2019; Greninger 2018; Grimwood et al. 2021; Lauber et al. 2017; Miller et al. 2021; Parry et al. 2020). To this end, we mined the transcriptome data generated by the One Thousand Plant Transcriptomes Initiative (1KP) using sequence homology searches of known plant viruses. The 1KP project provides a major untapped source of polyA-selected transcriptome data for virus discovery drawn from species across the breadth of the plants in a broad sense including the *Viridiplantae* (green plants), Glaucophyta, and Rhodophyta (red algae) (Carpenter et al. 2019; Leebens-Mack et al. 2019). Our broad aim was to revise our understanding of the phytovirosphere using data across the plant kingdom as well as providing insights into the origins and diversification of plant viruses.

## 4.3 Methods

### 4.3.1 Transcriptome data generation

The 1KP generated RNA sequencing libraries from 1,143 species across the breadth of the plant kingdom (Leebens-Mack et al. 2019). Thirty chromista and rhodophyte species were also included. Due to the diversity of species examined, samples were obtained from multiple sources including field collections, greenhouses, culture collections and laboratory specimens (Johnson et al. 2012). For the majority of species, young leaves or shoots were collected, although occasionally a mix of vegetative and reproductive tissues was used. To avoid RNA

degradation, RNA extraction was performed immediately after tissue collection or tissue was frozen in liquid nitrogen and stored in a -80˚C until extraction (Johnson et al. 2012). Several extraction protocols were used including CTAB and TRIzol (see (Johnson et al. 2012) for complete details). All sequencing was conducted at BGI-Shenzhen, China, using a combination of in-house protocols or TruSeq chemistry (Johnson et al. 2012). All libraries were prepared from polyA RNA. Paired-end sequencing was initially completed using Illumina GAII machines (11% of libraries) with a ~72bp read length but later the HiSeq platform was used (89% of libraries) with a 90 bp read length (Johnson et al. 2012).

### 4.3.2 Surveying for viruses in the 1KP

Raw transcriptomes (n = 1079, belonging to 960 plants species) from the 1KP major release were downloaded from the NCBI Short Read Archive (SRA) database (BioProject accession PRJEB21674) and converted to FASTQ format using the SRA Toolkit program fastq-dump in combination with the parallel-fastq-dump wrapper (https://github.com/rvalieris/parallel-fastq-dump) (Leinonen et al. 2010). One hundred transcriptomes within the BioProject were not publicly available (released 22/08/2019) at the commencement of this study and thus not analysed. Transcriptomes from the 1KP pilot study (BioProject accession PRJEB4921) and secondary project (BioProject accession PRJEB8056) were similarly not analysed. To reduce the downstream computing resources, raw sequences were mapped to their respective host genome scaffold using bowtie2 (Langmead and Salzberg 2012). Genome scaffolds were assembled as part of a previous study (Leebens-Mack et al. 2019). Where genome scaffolds were not available (n = 2) reads were assembled *de novo*. Trinity RNA-seq (v2.1.1) was used to quality trim and assemble *de novo* the unaligned reads captured from mapping (Haas et al. 2013). The assembled contigs were then assigned to known virus families and annotated through similarity searches against the NCBI nucleotide database (nt), the non-redundant protein database (nr) and a custom viral RdRp database using BLASTN and Diamond (BLASTX) (Altschul et al. 1990; Buchfink et al. 2015). To filter out weak BLAST sequence matches an e-value cut-off of $1 \times 10^{-10}$ was employed. To identify potential false positives, putative viral contigs were manually compared across the three BLAST searches (nt, nr and RdRp) to ensure matches to virus-associated sequences were consistent.

### 4.3.3 Virus filtering and abundance calculations

For all analyses, we focused on virus families known to infect plants or algae. As our analyses rely on sequence-based similarity searches for virus detection it is necessarily biased towards viruses that exhibit similarity to existing virus families. Together, the Virus-Host database (Mihara et al. 2016) and the International Committee on Taxonomy of Viruses (http://www.ictvonline.org/) were used to develop a list of plant virus families and genera to filter out virus-like contigs associated with vertebrate, invertebrate or fungi hosts based upon their top BLASTx and BLASTn matches. Packages within the Tidyverse collection (v1.3.0) in RStudio were used to complete these tasks (Wickham et al. 2019). Where the host was ambiguous (e.g., belonged to a family or genera known to infect both plant and fungal species) the contig was inspected manually.

The relative abundance of each transcript within the host transcriptome was calculated using RNA-Seq by Expectation-Maximization (v1.2.28) (Li and Dewey 2011). To account for variation in the number of unaligned reads between libraries after mapping, contig abundance was standardised by the total number of unaligned paired reads. Contigs under 200 nucleotides in length were excluded from further analysis.

### 4.3.4 Genome extension and annotation

Where a novel virus-like contig was discovered, we re-assembled the complete library – without removing host reads – in an attempt to recover a complete virus genome. For all re-assembled libraries, we recalculated abundance measurements to account for both host and non-host reads (Supplementary Table 4). We further re-assembled all libraries belonging to non-flowering plants (n = 402). Reads were mapped onto virus-like contigs using Bbmap and heterogeneous coverage and potential misassemblies were manually resolved using Geneious (v11.0.9) (Bushnell 2014; Kearse et al. 2012).

To determine whether a virus was novel, we followed the criteria as specified by The International Committee on Taxonomy of Viruses (http://www.ictvonline.org/). Novel viruses were named using a combination of the host common name - if documented – and the associated virus taxonomic group (e.g., *Interrupted club-moss deltapartitivirus*). In cases where host assignment proved difficult the suffix “associated” was added to the host name to signify this (e.g., *Calypogeia fissa associated deltaflexivirus*). Where the taxonomic position of a virus

was ambiguous the suffix "-like" was used (e.g., *Goldenrod fern qin-like virus*). Virus acronyms were created using a combination of the first and/or second letters of the host common name - if documented – and virus taxonomic group (e.g., *Leucodon julaceus beny-like virus* (LjBV)). Where multiple related viruses were found in the same host, we assigned each a number (e.g., *Aulacomnium heterostichum bunyavirus 1* (AuheBV1)).

The percentage identity among virus sequences was calculated via multiple sequence alignments using Clustal Omega (v1.2.3) (Sievers et al. 2011). The RdRp protein coding domain was used for all sequence alignments. Percentage identity matrices were converted to heat map plots using a custom R script provided by (Gilbert et al. 2019).

To characterise functional domains, predicted protein sequences along with their closest viral relatives were subjected to a domain-based search using the Conserved Domain Database (v3.18) (https://www.ncbi.nlm.nih.gov/Structure/cdd/cdd.shtml) and cross-referenced with the PFAM (v34.0) and Uniclust30 (v2018_08) databases available within the MMseqs2 webserver (Mirdita et al. 2019). To recover additional annotations, we used HHpred within the MPI Bioinformatics Toolkit webserver to query the PDB_mmCIF70 (v12_Oct), SCOPe70 (v2.08), UniProt-SwissProt-viral70 (v3_Nov_2021) and TIGRFAMs (v15.0) databases (Zimmermann et al. 2018). Virus genome diagrams were produced using the program littlegenomes (Lay 2021). Where available NCBI/GenBank CDS information was used to annotate reference virus sequences (Sayers et al. 2021).

#### 4.3.5 Detection of endogenous virus elements

All genome scaffolds produced by the 1KP were used as a database which we queried using the protein translations of the viruses discovered in this study. Endogenous viral elements (i.e., EVEs) were detected using the tblastn algorithm (Gertz et al. 2006). The search threshold was limited to 100 amino acids in length with an e-value cut off of $1\times10^{-20}$. Where multiple hits across several plant scaffolds were observed we manually examined the sequence. Suspected EVEs were queried against a subset whole-genome shotgun contig database which included green plants (taxid: 33090) and rhodophytes (taxid: 2763). In addition, the virus-like sequences discovered in this study were checked for host gene contamination using the contamination function implemented in CheckV (v0.8.1) (Nayfach et al. 2021). All EVEs were removed.

### 4.3.6 Assessing library contamination by eukaryotes, bacteria, and protozoa

For libraries in which a novel virus was discovered we investigated whether reads belonging to other eukaryotes were also present in the sequencing libraries. To achieve this, we obtained taxonomic identification for raw reads in each library – without the removal of host reads – by aligning them to the NCBI nt database using the KMA aligner and the CCMetagen program (Clausen et al. 2018; Marcelino et al. 2020). Sequence abundance was calculated by counting the number of nucleotides matching the reference sequence with an additional correction for template length (the default parameter in KMA). Krona charts generated by CCMetagen were edited were further edited in Adobe Illustrator (https://www.adobe.com) (Ondov et al. 2011). Library contamination was also assessed by the 1KP and used to inform our host-virus assignments (Carpenter et al. 2019).

### 4.3.7 Phylogenetic analysis of plant viruses

Phylogenetic trees of the plant-associated viruses discovered here were inferred using a maximum likelihood approach. We combined our translated virus contigs with known virus protein sequences from each respective virus family taken from NCBI/GenBank (Sayers et al. 2021). Sequences were then aligned with the program Clustal Omega (v1.2.3) with default parameters (Sievers et al. 2011). Sites of ambiguity were removed using trimAl (v1.2) with a gap threshold of 0.9 and a variable conserve value (Capella-Gutiérrez et al. 2009). To estimate phylogenetic trees, selection of the best-fit model of amino acid substitution was determined using the Akaike information criterion, corrected AIC, and the Bayesian information criterion with the ModelFinder function (-m MFP) in IQ-TREE (Kalyaanamoorthy et al. 2017; Nguyen et al. 2015). All phylogenetic trees were created using IQ-TREE with 1000 bootstrap replicates. Phylogenetic trees were annotated with FigTree (v1.4.4) (Rambaut and Drummond 2012) and further edited in Adobe Illustrator (https://www.adobe.com).

To visualise the occurrence of cross-species transmission and virus-host co-divergence across plant virus families, we reconciled the co-phylogenetic relationship between viruses and their hosts. In each case a vascular plant host cladogram was constructed using trees from (Zanne et al. 2014) and (Smith and Brown 2018), using the R package V.PhyloMaker (v0.1.0) (Jin and Qian 2019). As lower plants and non-plant species are not present in the V.PhyloMaker megatree, these hosts were added to the cladogram using the phyloT software, a phylogenetic tree generator based on NCBI taxonomy (http://phylot.biobyte.de/) as well as topologies

available in the appropriate literature. Host information was obtained from the NCBI Virus database (accessed 14/12/2021) and available literature (Hatcher et al. 2017) A tanglegram that graphically represents the correspondence between host and virus trees was created using the R packages phytools (v0.7-80) and APE (v5.5) (Paradis and Schliep 2019; Revell 2012). Virus sequences from each family were obtained through a broad survey of all virus genomic data available on GenBank. The virus phylogenies used in the co-phylogenies were constructed as detailed above. To quantify the relative frequencies of cross-species transmission versus virus-host co-divergence we reconciled the co-phylogenetic relationship between viruses and their hosts using the Jane package (Conow et al. 2010). This employs a maximum parsimony approach to determine the best 'map' of the virus phylogeny onto the host phylogeny. The cost of duplication, host-jump and extinction event types were set to 1.0, while host-virus co-divergence was set to zero as it was considered the null event. The reconciliation proceeds by minimising the total event cost. The number of generations and the population size was both set to 100. Jane was chosen over its successor eMPRess as it allows for a virus to be associated with multiple host species and handle polytomies (Santichaivekin et al. 2020). For a multi-host virus, we represented each association as a polytomy on the virus phylogeny.

#### 4.3.8 Assigning plant host clades

Each plant host was assigned to each clade in a previous study based upon their phylogenetic positioning and lineage information (Leebens-Mack et al. 2019). To improve clarity when colouring the phylogenies (although not the tanglegrams) we reduced the number of clades from 25 to ten (core eudicots, basal eudicots, monocots, basalmost angiosperms, gymnosperms, fern and fern allies, non-vascular, chlorophyte/charophyte, rhodophyte and lastly Chromista) by combining those that were closely related or potentially overlapping to increase the number of species in each group (Supplementary Table 1).

### 4.4 Results

We characterised the viruses found in the transcriptomes of 960 plant species within the 1KP major release. The transcriptomes represented a broad taxonomic sampling across the Archaeplastida (green plants, glaucophytes and rhodophytes). Sequencing libraries had a median of 25,187,714 paired reads (range 10,156,464–46,650,336). A median of 82% of reads (range 1%-96%) in these libraries mapped to host genome scaffolds and were subsequently

removed. *De novo* assembly of the sequencing reads resulted in a median of 36,015 contigs (range 1,396–146,217) per library, with a total of 41,256,176 contigs generated (Supplementary Table 2).

### 4.4.1 Diversity and abundance of plant viruses

In total, virus-like transcripts were found for 603 plant species; 69% of these were plant-associated while numerous identified sequences shared high similarity to non-plant associated viruses including those known to infect fungi, invertebrate and vertebrate hosts. Among the non-plant-associated virus transcripts, 34% were unclassified (10% of total virus-like transcripts) such that they were most closely related to a virus sequence with little to no taxonomic information (i.e., a virus sequence classified as only belonging to the *Riboviria*). If an RdRp-like region was detected in an unclassified virus-like transcript we further assessed whether it could be plant-associated (see Phylogenetic analysis of identified viruses). The remaining non-plant-associated virus transcripts were largely classified within the *Orthomyxoviridae* (vertebrate associated) (25%), *Rhabdoviridae* (invertebrate associated) (17%), *Partitiviridae* (fungus associated) (10%), *Mimiviridae* (amoeboid associated) (10%) and *Adenoviridae* (vertebrate associated) (7%) and excluded from the remainder of this study. These sequences are discussed in more detail in the section on “Presence of contaminants in sequencing libraries”. Although some of these viruses could represent plant infection, we made the conservative decision to remove them from the analysis.

We detected transcripts closely associated with viruses containing single and double-stranded DNA and RNA genomes. The majority of virus-like sequences belonged to families with +ssRNA genomes (61%) or reverse-transcribing dsDNA viruses (22%) (Figure 4.1). The +ssRNA virus transcripts were predominately classified within the *Betaflexiviridae* (30%), *Potyviridae* (19%), *Secoviridae* (16%) and *Alphaflexiviridae* (10%) (Supplementary Table 3). Negative-sense single-stranded RNA (-ssRNA) virus transcripts were classified within the *Aspiviridae* (0.04%), *Rhabdoviridae* (6%) and *Tospoviridae* (3%) (*Phenuiviridae* and *Yueviridae* transcripts were later detected in the unclassified virus-like transcripts) (Supplementary Table 3). dsDNA virus transcripts with sequence similarities to the *Phycodnaviridae* were detected in the algae samples. These phycodna-like virus transcripts frequently encoded the chitinase and DNA ligase genes that are homologous to those in distantly related host organisms including fungi and bacteria. As it is difficult to determine whether these transcripts represent

*Phycodnaviridae* sequences or contamination, we excluded all phycodnavirus-related sequences. All remaining dsDNA viruses were exclusively reverse-transcribing viruses from the *Caulimoviridae*. We failed to detect any sequences that shared homology with several plant virus families including *Reoviridae*, *Nanoviridae* and *Fimoviridae* (although see the Discussion for caveats).

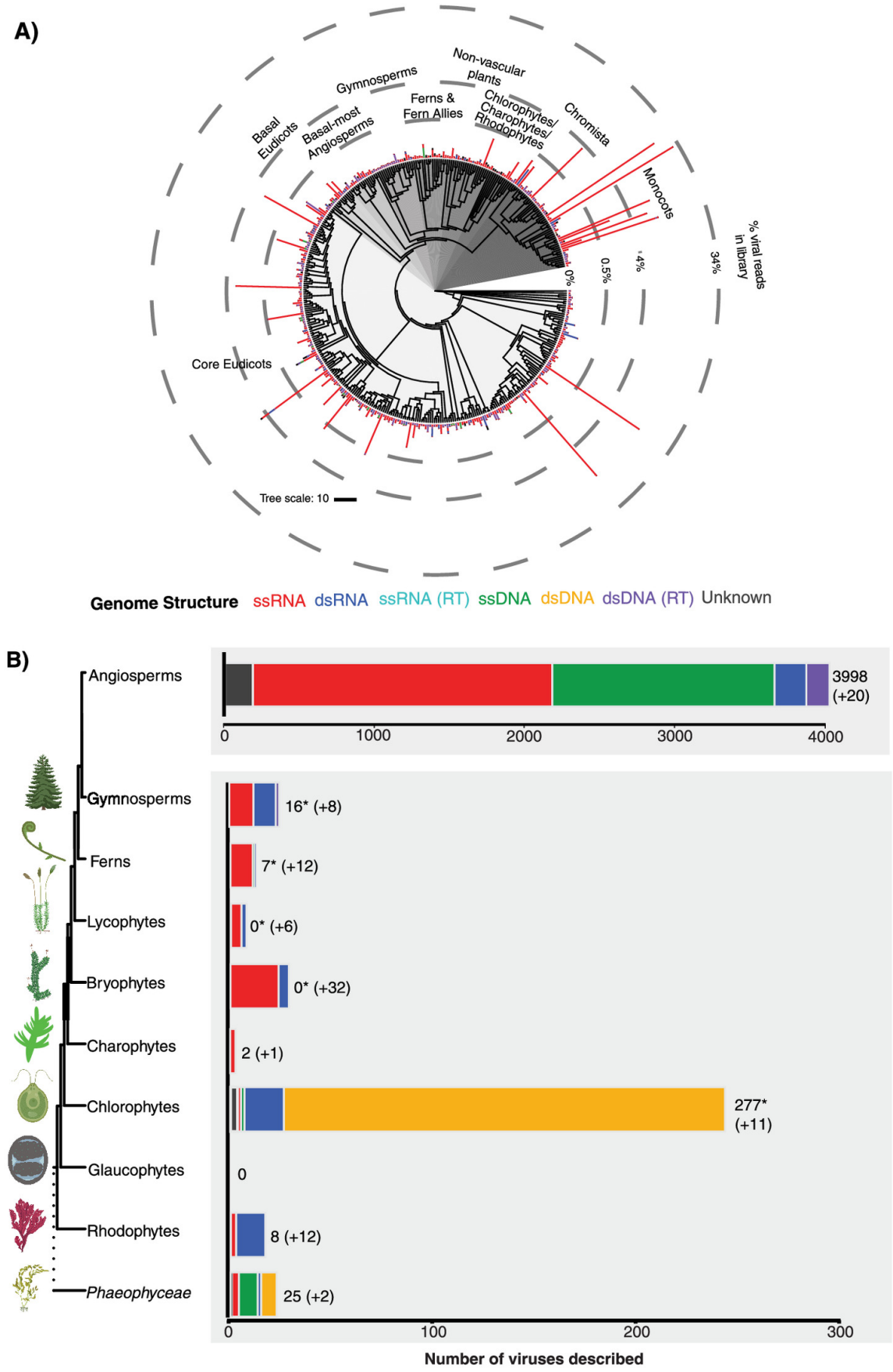

A)
Gymnosperms
Non-vascular plants
Ferns & Fern Allies
Chlorophytes/ Charophytes/ Rhodophytes
Chromista
Monocots
Basal Eudicots
Basal-most Angiosperms
Core Eudicots
% viral reads in library
34%
4%
0.5%
0%
Tree scale: 10
Genome Structure ssRNA dsRNA ssRNA (RT) ssDNA dsDNA dsDNA (RT) Unknown
B)
Angiosperms
Gymnosperms
Ferns
Lycophytes
Bryophytes
Charophytes
Chlorophytes
Glaucophytes
Rhodophytes
Phaeophyceae
3998 (+20)
0
1000
2000
3000
4000
16* (+8)
7* (+12)
0* (+6)
0* (+32)
2 (+1)
277* (+11)
0
8 (+12)
25 (+2)
0
100
200
300
Number of viruses described

***Figure 4.20 Virus abundance and diversity across the plant kingdom.*** (*a) Phylogram of virus composition across the One Thousand Plant Transcriptomes Initiative (1KP) samples. Plant-associated virus abundance was summarised for each plant species and normalised using a Box-Cox transformation. The height of each bar represents the percentage of virus reads detected in each plant species (after the removal of host reads). Plant clades are labelled and differentiated by shades of grey. The 1KP ASTRAL tree was used as the basis for this tree (Leebens-Mack et al. 2019). Clade and abundance annotations were added using the Interactive Tree of Life (iTOL) web-based tool (Letunic and Bork 2019). (b) The phytovirosphere across the Plantae and Phaeophyceae. A schematic tree of the evolution of major plant groups. Each bar represents the number of total viruses formally or likely associated with each host group and is coloured by virus genome composition. The total number of viruses for each plant group plus those found in this study is also shown at the end of each bar. The Virus-Host (Mihara et al. 2016) and NCBI virus databases (Brister et al. 2015) combined with literature searches were used to obtain virus counts. Lineage branches are not drawn to scale. To our knowledge, no viruses have been found in the Glaucophytes. Plant and algae images were obtained from BioRender.com or drawn in Adobe Illustrator (https://www.adobe.com). *Transcriptome scaffolds from libraries belonging to these host groups shared homology to virus RdRps and were partially analysed but not assembled or deposited to GenBank (Mushegian et al. 2016).*

There was a large range of viral abundance in each library ($5\times10^{-6}$% – 31% reads after host-associated reads were removed). Viruses with +ssRNA genomes accounted for the vast majority (99.8%) of virus abundance (Figure 4.1, Supplementary Table 3). As expected, virus discovery was concentrated in the flowering plants (angiosperms). For instance, plant virus-like sequences were frequently discovered in the core eudicots and monocots (i.e., 73% of libraries in which plant virus transcripts were found). The detection rate of plant viruses was highest in the most basal angiosperms (57%) and monocots (50%). No significant difference in virus abundance was observed between sequencing platforms (Genome Analyzer II and Illumina HiSeq 2000; $p = 0.327$).

### 4.4.2 Presence of contaminants in sequencing libraries

The bacterial, fungal and insect species that live in or on plant tissues are commonly sampled within plant sequencing libraries (Carpenter et al. 2019), although contamination from other

plants is also a possibility during sample preparation or sequencing. To quantify the extent of library contamination we used the KMA and CCMetagen tools (Figure 4.2). Among the libraries analysed (n = 95), bacteria were consistently detected representing a median of 1.5% of total abundance (range 0.01%-33%). A median of 2% of library abundance was associated with fungal sequences (range 0%-53%). Arthropods and chordates were also commonly detected across libraries (found in 87 and 89 libraries, respectively) but at lower abundance (median 0.15%, range 0%-11.4%). The presence of chordate associated reads is likely attributed to various routes of sample contamination (e.g., faeces) or during sample processing and sequencing.

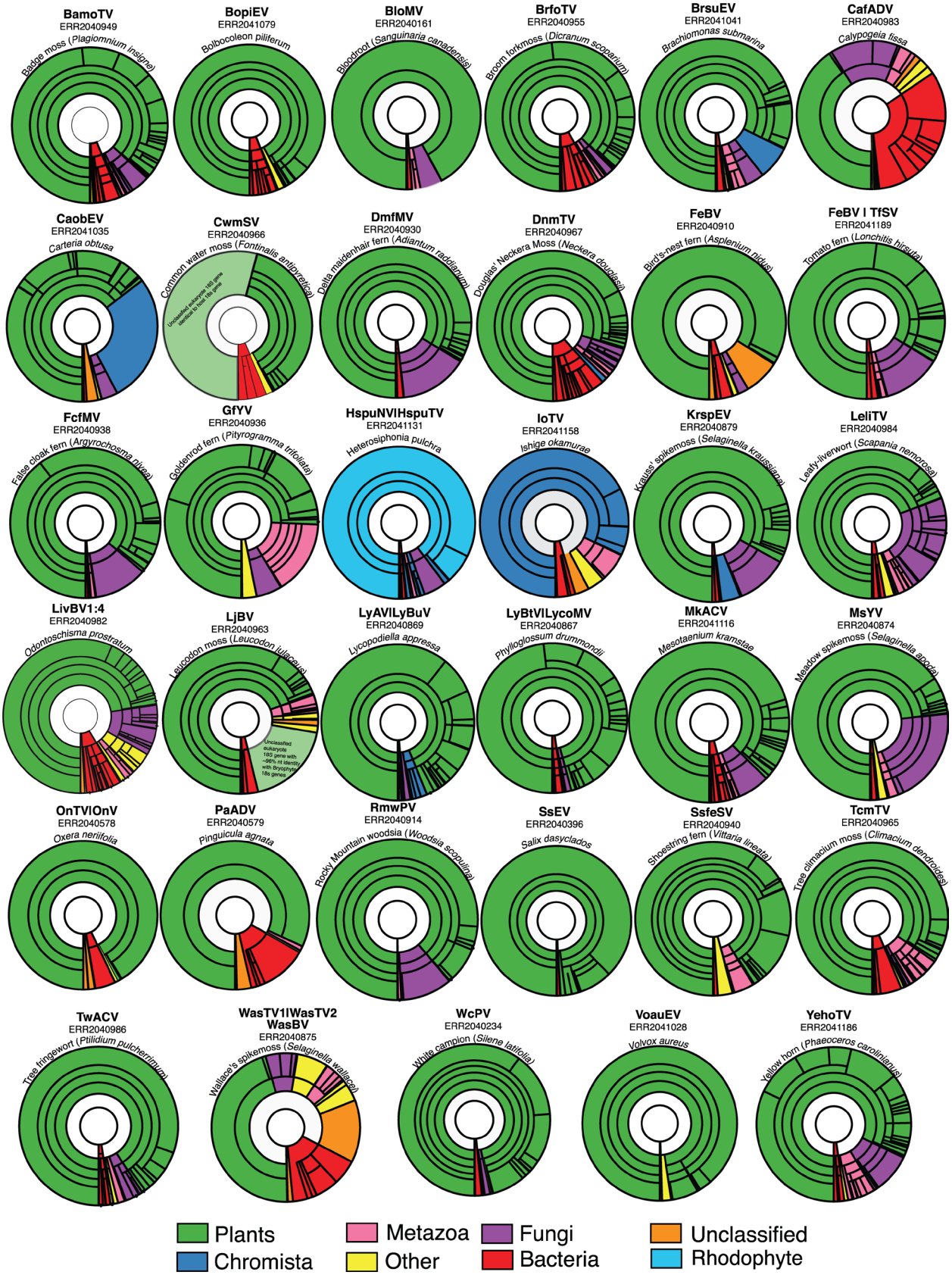

BamoTV
ERR2040949
Badge moss (Plagiomnium insigne)
BopiEV
ERR2041079
Bolbocoleon piliferum
BloMV
ERR2040161
Bloodroot (Sanguinaria canadensis)
BrfoTV
ERR2040955
Broom forkmoss (Dicranum scoparium)
BrsuEV
ERR2041041
Brachiomonas submarina
CafADV
ERR2040983
Calypogeia fissa
CaobEV
ERR2041035
Carteria obtusa
CwmSV
ERR2040966
Common water moss (Fontinalis antipyretica)
Unclassified eukaryote 18S gene identical to host 18s gene
DmfMV
ERR2040930
Delta maidenhair fern (Adiantum raddianum)
DnmTV
ERR2040967
Douglas' Neckera Moss (Neckera douglasii)
FeBV
ERR2040910
Bird's-nest fern (Asplenium nidus)
FeBV I TfSV
ERR2041189
Tomato fern (Lonchitis hirsuta)
FcfMV
ERR2040938
False cloak fern (Argyrochosma nivea)
GfYV
ERR2040936
Goldenrod fern (Pityrogramma trifoliata)
HspuNVIHspuTV
ERR2041131
Heterosiphonia pulchra
IoTV
ERR2041158
Ishige okamurae
KrspEV
ERR2040879
Krauss' spikemoss (Selaginella kraussiana)
LeliTV
ERR2040984
Leafy liverwort (Scapania nemorosa)
LivBV1:4
ERR2040982
Odontoschisma prostratum
LjBV
ERR2040963
Leucodon moss (Leucodon julaceus)
Unclassified eukaryote 18S gene with ~96% nt identity with Bryophyte 18s genes
LyAVILyBuV
ERR2040869
Lycopodiella appressa
LyBtVILycoMV
ERR2040867
Phylloglossum drummondii
MkACV
ERR2041116
Mesotaenium kramstae
MsYV
ERR2040874
Meadow spikemoss (Selaginella apoda)
OnTVIOnV
ERR2040578
Oxera neriifolia
PaADV
ERR2040579
Pinguicula agnata
RmwPV
ERR2040914
Rocky Mountain woodsia (Woodsia scopulina)
SsEV
ERR2040396
Salix dasyclados
SsfeSV
ERR2040940
Shoestring fern (Vittaria lineata)
TcmTV
ERR2040965
Tree climacium moss (Climacium dendroides)
TwACV
ERR2040986
Tree fringewort (Ptilidium pulcherrimum)
WasTV1IWasTV2
WasBV
ERR2040875
Wallace's spikemoss (Selaginella wallacei)
WcPV
ERR2040234
White campion (Silene latifolia)
VoauEV
ERR2041028
Volvox aureus
YehoTV
ERR2041186
Yellow horn (Phaeoceros carolinianus)
Plants
Chromista
Metazoa
Other
Fungi
Bacteria
Unclassified
Rhodophyte

***Figure 4.21 Taxonomic assignments of reads in select One Thousand Plant Transcriptomes Initiative (1KP) libraries.*** *Each Krona graph illustrates the relative abundance of taxa in a metatranscriptome at varying taxonomic levels. For clarity, a maximum depth of five taxonomic levels was chosen for each graph. The library Sequence Read Archive accession number, host species, and the corresponding virus of interest are annotated above each graph. Segments are highlighted based upon the species taxonomic grouping (plants = green, Chromista = blue, unclassified = orange, bacteria = red, metazoa = pink, fungi = purple, rhodophytes = light blue, other = yellow). Here "plants" encompasses the Viridiplantae. Reads without any match in the nt database are not shown.*

The detection of four vertebrate associated viruses across several libraries provided further evidence of library contamination. Sequences belonging to these viruses - Influenza A virus (16 libraries), Human mastadenovirus C (30 libraries), Human immunodeficiency virus (15 libraries) and Parainfluenza virus 5 (3 libraries) – were present at low abundance and showed little genetic variation between libraries. Notably, chordate-associated reads were only present in 66% of libraries in which these viruses were found. The failure to consistently detect potential hosts for these viruses suggests contamination during sequencing. The four vertebrate associated viruses were largely absent in libraries in which novel plant-associated viruses were discovered, except for the *Larix speciosa*, *Brachiomonas submarina*, *Climacium dendroides*, *Silene latifolia* and *Oxera neriifolia* transcriptomes.

In addition, the 1KP compared all assembled sequences to a reference set of nuclear 18S ribosomal RNA sequences from the SILVA small subunit rRNA database using BLASTn (Carpenter et al. 2019; Quast et al. 2012). Where a sample had several alignments to any other plant sequences outside of the expected source family the sample was described as having "worrisome contamination" (Carpenter et al. 2019). This applied to eleven plant libraries in which novel viruses were identified. Below, we discuss library contaminates from viewpoint of virus-host associations.

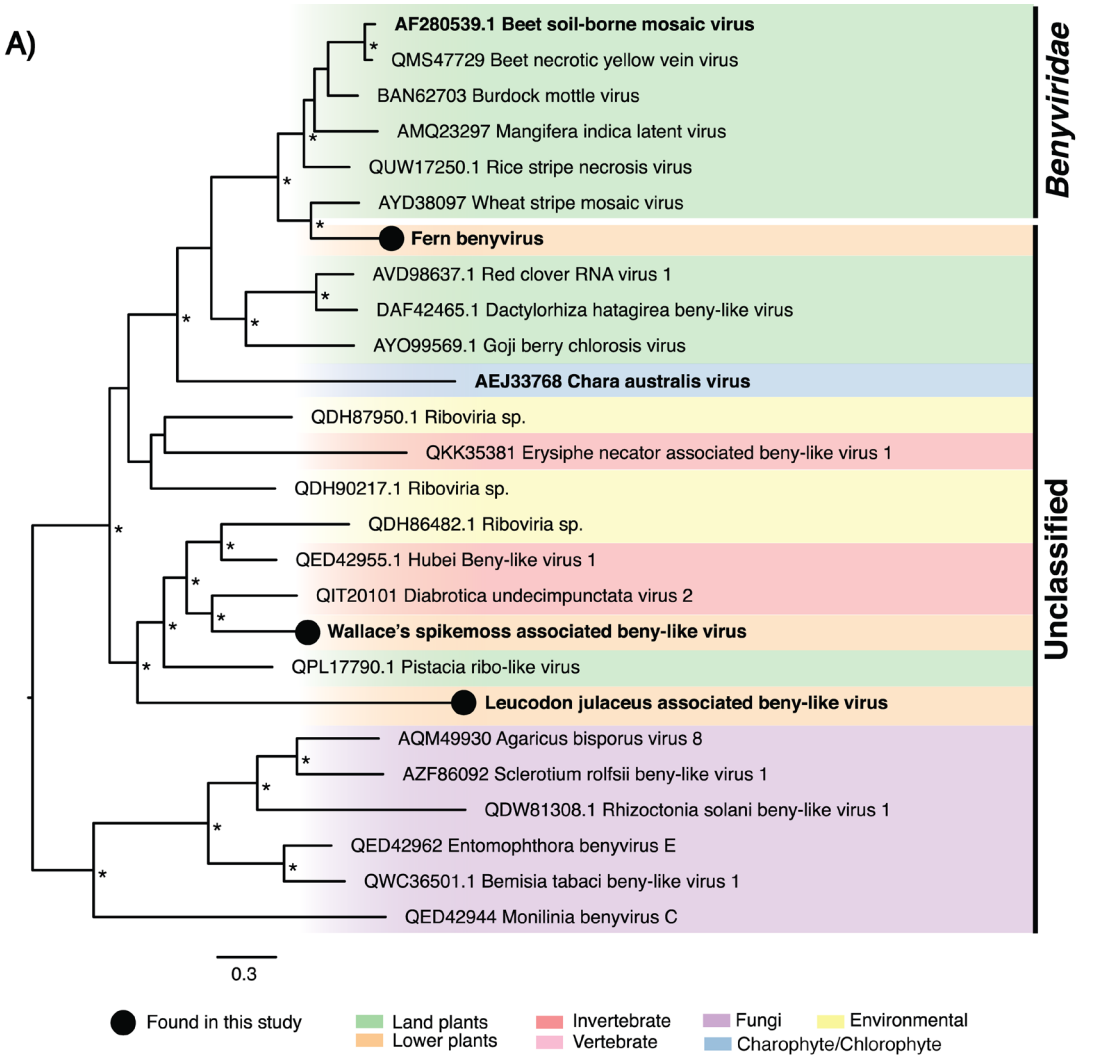
### 4.4.3 Phylogenetic analysis of identified viruses

To infer phylogenetic relationships between identified viruses, order and family-level phylogenetic trees were estimated using the highly conserved viral region that comprises the RdRp. In total, we assembled 104 RdRp contigs that likely represent novel virus species, of

which 41 were considered as unclassified or non-plant associated due to their similarities to virus groups known to infect non-plant hosts (Supplementary Table 4). Further analysis of these contigs revealed that they are likely plant-associated.

**Positive-sense single-stranded RNA [(+)ssRNA] viruses**

**(i) *Hepelivirales.*** *(a) Benyviridae.* We identified three beny-like sequences that to our knowledge represent the first benyvirid found in lower plants. The first sequence, tentatively named Fern benyvirus (FeBV), was found in both the bird's-nest fern (*Asplenium nidus*) and tomato fern (*Lonchitis hirsuta*). Together with Wheat stripe mosaic virus, FeBV represents a well-supported clade separate from the remaining plant benyviruses (Figure 4.3).


A)
AF280539.1 Beet soil-borne mosaic virus
QMS47729 Beet necrotic yellow vein virus
BAN62703 Burdock mottle virus
AMQ23297 Mangifera indica latent virus
QUW17250.1 Rice stripe necrosis virus
AYD38097 Wheat stripe mosaic virus
Fern benyvirus
AVD98637.1 Red clover RNA virus 1
DAF42465.1 Dactylorhiza hatagirea beny-like virus
AYO99569.1 Goji berry chlorosis virus
AEJ33768 Chara australis virus
QDH87950.1 Riboviria sp.
QKK35381 Erysiphe necator associated beny-like virus 1
QDH90217.1 Riboviria sp.
QDH86482.1 Riboviria sp.
QED42955.1 Hubei Beny-like virus 1
QIT20101 Diabrotica undecimpunctata virus 2
Wallace's spikemoss associated beny-like virus
QPL17790.1 Pistacia ribo-like virus
Leucodon julaceus associated beny-like virus
AQM49930 Agaricus bisporus virus 8
AZF86092 Sclerotium rolfsii beny-like virus 1
QDW81308.1 Rhizoctonia solani beny-like virus 1
QED42962 Entomophthora benyvirus E
QWC36501.1 Bemisia tabaci beny-like virus 1
QED42944 Monilinia benyvirus C
Benyviridae
Unclassified
0.3
Found in this study
Land plants
Lower plants
Invertebrate
Vertebrate
Fungi
Charophyte/Chlorophyte
Environmental

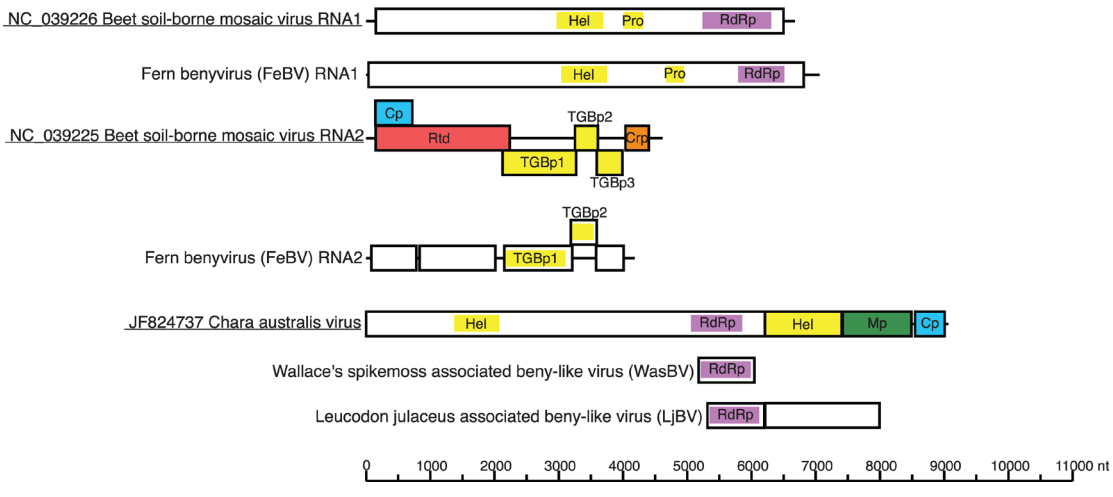

B)
NC_039226 Beet soil-borne mosaic virus RNA1
Hel
Pro
RdRp
Fern benyvirus (FeBV) RNA1
NC_039225 Beet soil-borne mosaic virus RNA2
Cp
Rtd
TGBp2
Crp
TGBp1
TGBp3
Fern benyvirus (FeBV) RNA2
JF824737 Chara australis virus
Mp
Wallace's spikemoss associated beny-like virus (WasBV)
Leucodon julaceus associated beny-like virus (LjBV)
0
1000
2000
3000
4000
5000
6000
7000
8000
9000
10000
11000 nt

***Figure 4.22 Phylogenetic relationships and genome organization of the beny-like viruses.*** *(a) ML phylogenetic tree based on the beny-like RNA-1 replicase protein shows the topological position of virus-like sequences discovered in this study (black circles) in the context of their closest relatives. Branches are highlighted to represent host clade (land plants = green, lower plants = orange, invertebrate = red, vertebrate = pink, algae = blue, fungi = purple, yellow = environmental, Chromista = light blue, rhodophytes = dark green). Here "Land plants" encompasses both angiosperms and gymnosperms while "Lower plants" includes the bryophytes, lycophytes, and ferns. 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% bootstrap support. Tip labels are bolded when the genome structure is shown on the right. (b) Genomic organization of the beny-like virus sequences identified in this study and representative species used in the phylogeny. Beet soil-borne mosaic virus RNA three and four are not pictured here. The data underlying this figure and definitions of acronyms used are presented in Supplementary Table 5.*

The triple gene block (TGB) is a hallmark gene module of the *Benyviridae* among several other virus families in the class *Alsuviricetes* (Morozov and Solovyev 2003)). In both fern libraries, proteins resembling the TGB were assembled (Figure 4.3). The TGB proteins shared ~34% amino acid identity with the TGB protein of other benyvirids. To our knowledge, this is the first TGB protein found outside of flowering plants. Phylogenetic analysis placed the TGB1 protein of FeBV basal to the *Benyviridae* (Supplementary Figure 4.1).

Two additional beny-like viruses, named here Leucodon julaceus associated beny-like virus (LjBV) and Wallace's spikemoss associated beny-like virus (WasBV) were assembled. LjBV and WasBV cluster with unclassified algae, invertebrates, fungi and soil-derived viruses forming a group basal to all plant benyvirids and potentially constitute a novel virus group (Figure 4.3). LjBV contains a second open reading frame (ORF) with no detectable homology to known sequences (Figure 4.3).

Due to the phylogenetic placement of LjBV and WasBV close to viruses infecting distant hosts (e.g., invertebrates and fungi), we investigated the potential of contamination from other eukaryotes as the source of these viruses. Of note, the Wallace's spikemoss metatranscriptome contained reads that matched various fungi orders (7% of all reads) as well

as those matching the plant-parasitic oomycete *Albugo laibachii* (7%) which makes inferring virus-host relationships challenging (Figure 4.2). Reads belonging to various fungi species accounted for 10% of the bird's-nest fern transcriptome and 12% of the tomato fern transcriptome (Figure 4.2). Despite the presence of fungi-associated reads, the phylogenetic position of FeBV suggests that FeBV is likely plant-associated (Figure 4.3). No concerning contaminants were detected in the *Leucodon julaceus* transcriptome.

**(ii) *Tymovirales.*** *(a) Betaflexiviridae.* We identified 18 virus sequences that fell in the order *Tymovirales*. Four virus transcripts were associated with the *Betaflexiviridae*. The first, Sea beet betaflexivirus (SbBV), clusters with Agapanthus virus A, an unclassified betaflexivirus (Figure 4.4). The remaining sequences denoted Iranian poppy betaflexivirus (IpBV), Linum macraei betaflexivirus (LimBV) and Lycopod associated betaflexivirus (LyBtV) resemble capilloviruses. Notably, LyBtV may extend the known host range of the *Betaflexiviridae* from angiosperms to lower plants. All sequences phylogenetically cluster with known capilloviruses and potentially represent novel virus species (Figure 4.4). The *Phylloglossum drummondii* library in which LyBtV was assembled had contamination from lycopod and dicot species (Figure 4.2). As the majority of plant-associated reads were assigned to lycophytes (50%), LyBtV has been tentatively assigned to this group.

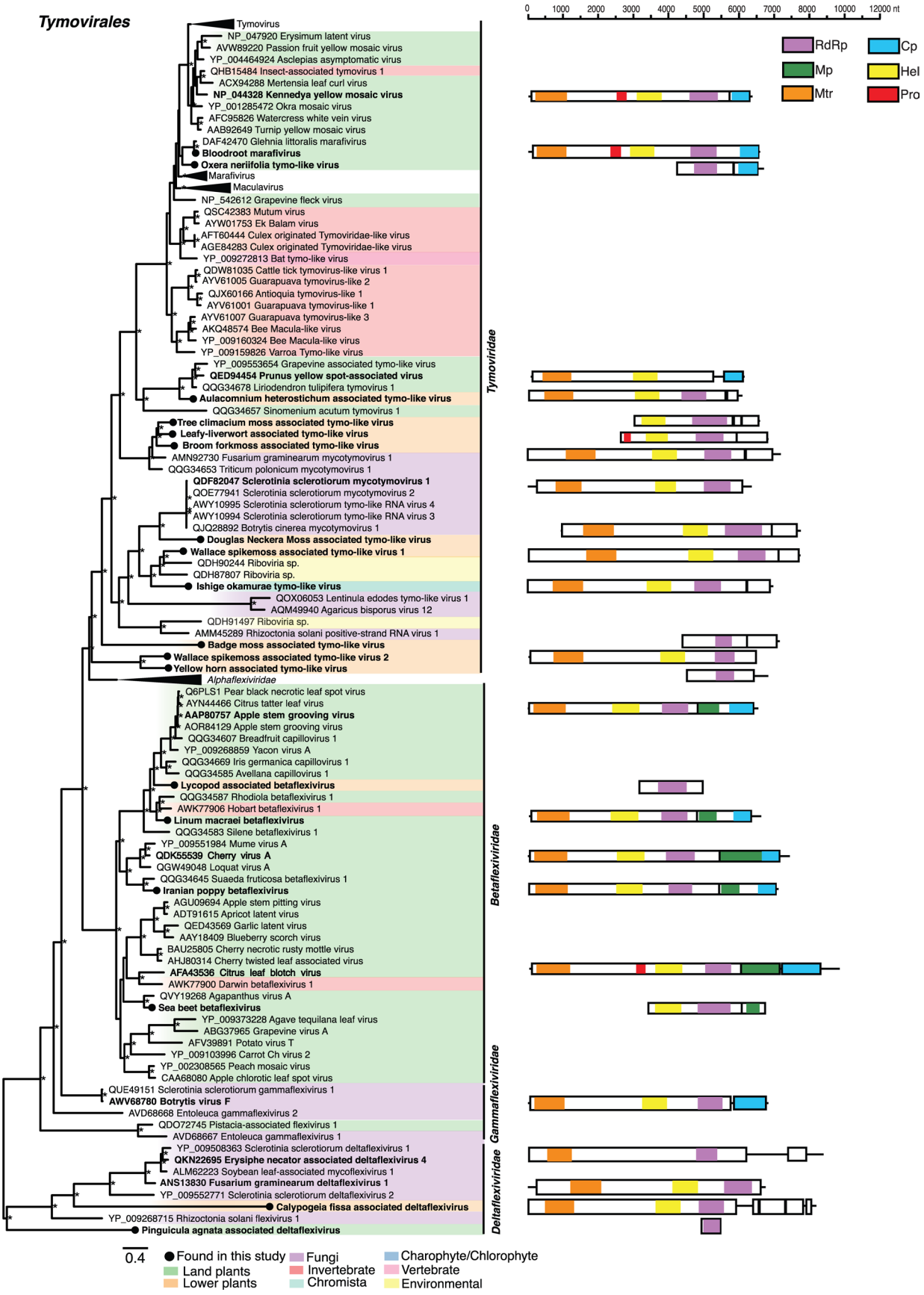

Tymovirales
Tymovirus
NP_047920 Erysimum latent virus
AVW89220 Passion fruit yellow mosaic virus
YP_004464924 Asclepias asymptomatic virus
QHB15484 Insect-associated tymovirus 1
ACX94288 Mertensia leaf curl virus
NP_044328 Kennedya yellow mosaic virus
YP_001285472 Okra mosaic virus
AFC95826 Watercress white vein virus
AAB92649 Turnip yellow mosaic virus
DAF42470 Glehnia littoralis marafivirus
Bloodroot marafivirus
Oxera neriifolia tymo-like virus
Marafivirus
Maculavirus
NP_542612 Grapevine fleck virus
QSC42383 Mutum virus
AYW01753 Ek Balam virus
AFT60444 Culex originated Tymoviridae-like virus
AGE84283 Culex originated Tymoviridae-like virus
YP_009272813 Bat tymo-like virus
QDW81035 Cattle tick tymovirus-like virus 1
AYV61005 Guarapuava tymovirus-like 2
QJX60166 Antioquia tymovirus-like 1
AYV61001 Guarapuava tymovirus-like 1
AYV61007 Guarapuava tymovirus-like 3
AKQ48574 Bee Macula-like virus
YP_009160324 Bee Macula-like virus
YP_009159826 Varroa Tymo-like virus
YP_009553654 Grapevine associated tymo-like virus
QED94454 Prunus yellow spot-associated virus
QQG34678 Liriodendron tulipifera tymovirus 1
Aulacomnium heterostichum associated tymo-like virus
QQG34657 Sinomenium acutum tymovirus 1
Tree climacium moss associated tymo-like virus
Leafy-liverwort associated tymo-like virus
Broom forkmoss associated tymo-like virus
AMN92730 Fusarium graminearum mycotymovirus 1
QQG34653 Triticum polonicum mycotymovirus 1
QDF82047 Sclerotinia sclerotiorum mycotymovirus 1
QOE77941 Sclerotinia sclerotiorum mycotymovirus 2
AWY10995 Sclerotinia sclerotiorum tymo-like RNA virus 4
AWY10994 Sclerotinia sclerotiorum tymo-like RNA virus 3
QJQ28892 Botrytis cinerea mycotymovirus 1
Douglas Neckera Moss associated tymo-like virus
Wallace spikemoss associated tymo-like virus 1
QDH90244 Riboviria sp.
QDH87807 Riboviria sp.
Ishige okamurae tymo-like virus
QOX06053 Lentinula edodes tymo-like virus 1
AQM49940 Agaricus bisporus virus 12
QDH91497 Riboviria sp.
AMM45289 Rhizoctonia solani positive-strand RNA virus 1
Badge moss associated tymo-like virus
Wallace spikemoss associated tymo-like virus 2
Yellow horn associated tymo-like virus
Alphaflexiviridae
Q6PLS1 Pear black necrotic leaf spot virus
AYN44466 Citrus tatter leaf virus
AAP80757 Apple stem grooving virus
AOR84129 Apple stem grooving virus
QQG34607 Breadfruit capillovirus 1
YP_009268859 Yacon virus A
QQG34669 Iris germanica capillovirus 1
QQG34585 Avellana capillovirus 1
Lycopod associated betaflexivirus
QQG34587 Rhodiola betaflexivirus 1
AWK77906 Hobart betaflexivirus 1
Linum macraei betaflexivirus
QQG34583 Silene betaflexivirus 1
YP_009551984 Mume virus A
QDK55539 Cherry virus A
QGW49048 Loquat virus A
QQG34645 Suaeda fruticosa betaflexivirus 1
Iranian poppy betaflexivirus
AGU09694 Apple stem pitting virus
ADT91615 Apricot latent virus
QED43569 Garlic latent virus
AAY18409 Blueberry scorch virus
BAU25805 Cherry necrotic rusty mottle virus
AHJ80314 Cherry twisted leaf associated virus
AFA43536 Citrus leaf blotch virus
AWK77900 Darwin betaflexivirus 1
QVY19268 Agapanthus virus A
Sea beet betaflexivirus
YP_009373228 Agave tequilana leaf virus
ABG37965 Grapevine virus A
AFV39891 Potato virus T
YP_009103996 Carrot Ch virus 2
YP_002308565 Peach mosaic virus
CAA68080 Apple chlorotic leaf spot virus
QUE49151 Sclerotinia sclerotiorum gammaflexivirus 1
AWV68780 Botrytis virus F
AVD68668 Entoleuca gammaflexivirus 2
QDO72745 Pistacia-associated flexivirus 1
AVD68667 Entoleuca gammaflexivirus 1
YP_009508363 Sclerotinia sclerotiorum deltaflexivirus 1
QKN22695 Erysiphe necator associated deltaflexivirus 4
ALM62223 Soybean leaf-associated mycoflexivirus 1
ANS13830 Fusarium graminearum deltaflexivirus 1
YP_009552771 Sclerotinia sclerotiorum deltaflexivirus 2
Calypogeia fissa associated deltaflexivirus
YP_009268715 Rhizoctonia solani flexivirus 1
Pinguicula agnata associated deltaflexivirus
Tymoviridae
Betaflexiviridae
Gammaflexiviridae
Deltaflexiviridae
0
1000
2000
3000
4000
5000
6000
7000
8000
10000
12000 nt
RdRp
Cp
Mp
Hel
Mtr
Pro
0.4
Found in this study
Land plants
Lower plants
Fungi
Invertebrate
Chromista
Charophyte/Chlorophyte
Vertebrate
Environmental

***Figure 4.23 Phylogenetic relationships and genome organization of the viruses within the order Tymovirales.*** *Left: ML phylogenetic tree based on the replication protein shows the topological position of virus-like sequences discovered in this study (black circles) in the context of their closest relatives. See Figure 4.3 for the colour scheme. 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% bootstrap support. Tip labels are bolded when the genome structure is shown on the right. Right: Genomic organization of the virus sequences identified in this study and representative species used in the phylogeny.*

*(b) Tymoviridae.* We identified 12 virus-like sequences that clustered within the *Tymoviridae* and related viruses. Ishige okamurae associated tymo-like virus (IoTV) was detected in the brown alga *Ishige okamurae* and likely represents the first virus in the order Tymovirales from brown algae. IoTV, along with ten sequences assembled from species in the *Anthocerotophyta* (hornworts), *Marchantiophyta* (liverworts) and other bryophytes grouped with tymo-like viruses from fungus and environmental samples (Figure 4.4). It is uncertain whether the true hosts of the novel tymo-like viruses discovered here are plants. Fungi contaminates were detected across these libraries but varied in abundance (range 1%-21%, mean = 6%). Despite their clustering with mycotymoviruses, Broom forkmoss associated tymo-like virus (BrfoTV) and Tree climacium moss associated tymo-like virus (TcmTV) were assembled from libraries with ~1% fungal reads, highlighting the inherent difficulties in host-virus assignment. Importantly, <1% of reads in Ishige okamurae transcriptome belonged to species of fungi (Figure 4.2).

We assembled two tymo-like virus sequences denoted Oxera neriifolia tymo-like virus (OnTV) and Bloodroot marafivirus (BloMV). BloMV and OnTV grouped with the unclassified Glehnia littoralis marafivirus (Figure 4.4). Marafiviruses and tymoviruses are commonly distinguished from each other based upon a highly conserved 16 nucleotide (nt) sequence known as the “tymobox” [GAGUCUGAAUUGCUUC] in tymoviruses and the “marafibox” [CA(G/A)GGUGAAUUGCUUC] in marafiviruses (Ding et al. 1990; Hammond and Ramirez 2001). While these two novel viruses cluster together, they differ in terms of genome structure and motifs. A “marafibox” like sequence is present in BloMV (CAACGCGAAUUGCUUU) (5606-5621 nt) albeit differing by several residues. This, combined with the BloMV genome likely consisting of a single large ORF, supports the assignment of BloMV as a Marafivirus. OnTV, like members of the Tymovirus genera, contains both a second ORF – likely encoding a coat protein (CP) –

and a tymobox (1493-1508 nt) (Figure 4.4). Phylogenetic analysis of the coat protein sequence places OnTV and BloMV in a clade with macula- and marafi-like viruses (Supplementary Figure 4.1).

(c) *Deltaflexiviridae.* We assembled two sequences that share similarities to members of the mycotymovirus family, *Deltaflexiviridae*. The first sequence was detected in the liverwort *Calypogeia fissa*, tentatively named Calypogeia fissa associated deltaflexivirus (CafADV) and appeared distantly related to delta- and gammaflexiviruses. A second related partial sequence, named here Pinguicula agnata virus (PaV), shared 32% amino acid identity with mycoflexivirus, Botrytis virus F. In a phylogenetic analysis with members of the Tymovirales, CafADV and PaAGV are placed with the deltaflexivirids (Figure 4.4).

It is unclear whether the source of these virus sequences is from plants or contamination from other eukaryotes. Indeed, the *C. fissa* library contained numerous contaminates including algae, fungi and bacteria representing 1%, 15% and 33% of total reads respectively (Figure 4.2). Interestingly, no fungi-associated reads were found in the *P. agnata* library suggesting a potential plant origin (Figure 4.2).

**(iii) *Picornavirales.*** *(a) Secoviridae.* We identified four sequences that shared similarities to members of the *Secoviridae* denoted Common water moss secovirus (CwmSV), Salix dasyclados secovirus (SadSV), Tomato fern secovirus (TfSV) and Shostring fern secovirus (SfSV). CwmSV, TfSV and SfSV cluster within the nepoviruses and likely represented the first seco-like virus detected in the bryophytes and ferns (Figure 4.5).

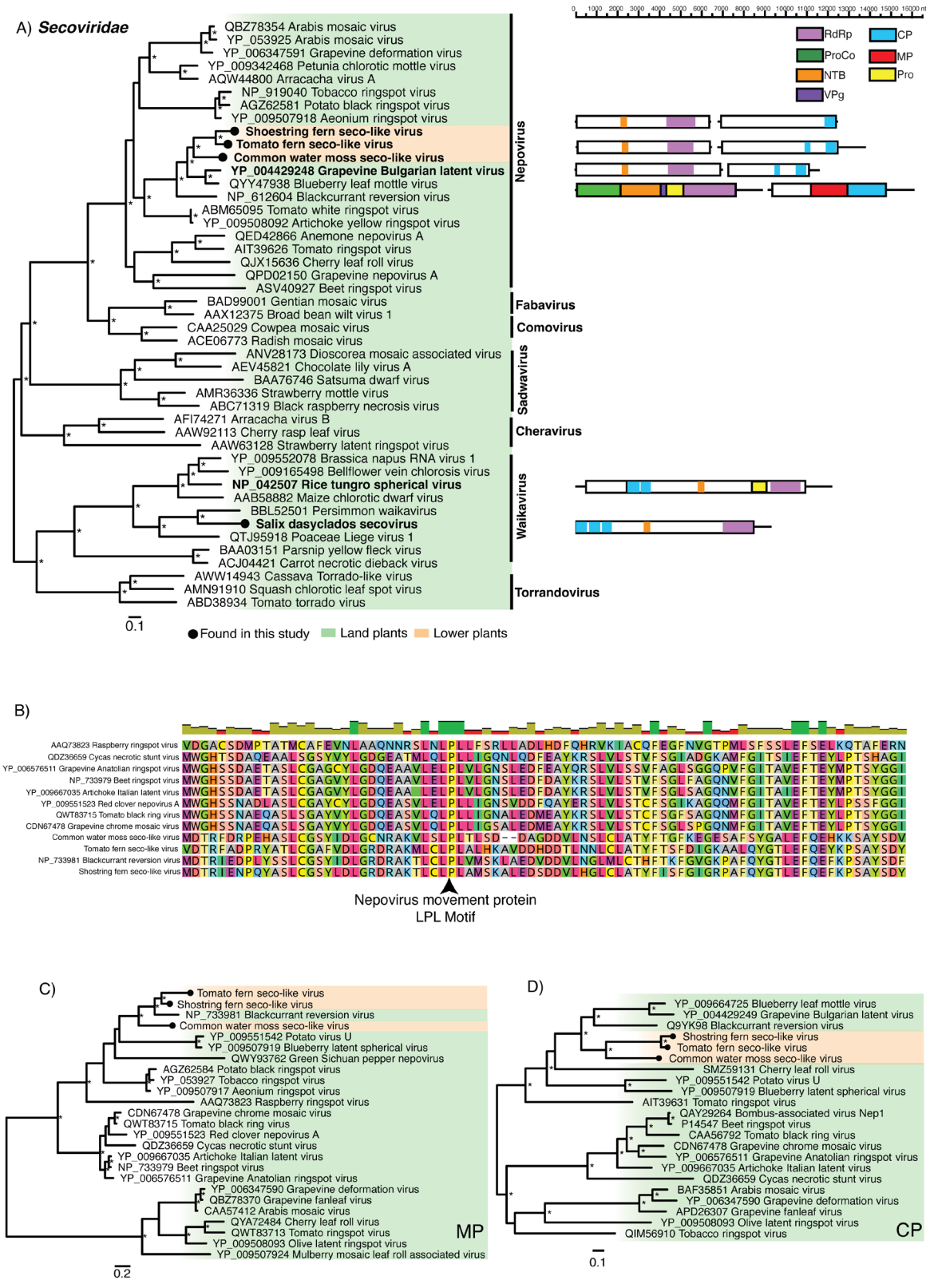



***Figure 4.24 Phylogenetic relationships, genome organization and movement protein motids of the viruses identified within the virus family Secoviridae*** *(a) Left: ML*

*phylogenetic trees based on the Pro-pol region show the topological position of virus-like sequences discovered in this study (black circles) in the context of their closest relatives. Right: Genomic organization of the seco-like sequences identified in this study and representative species used in the phylogeny. (b) Multiple amino acid sequence alignment of the 30K movement protein "LPL" motifs which are highly conserved throughout the nepoviruses. (c) Phylogenetic relationships of the Nepovirus 30K movement proteins. (d) Phylogenetic relationships of the Nepovirus coat proteins. For all trees, 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% bootstrap support. Tip labels are bolded when the genome structure is shown on the right. See Figure 4.3 for the colour scheme. Viruses discovered in this study are signified using a black circle on the tree tip.*

A putative RNA2 ORF was assembled for the three nepovirus-like sequences each containing a complete CP (Figure 4.5). The CPs fall within the nepovirus subgroup C (Figure 4.5d). While a movement protein (MP) domain was not formally detected, we predict that the region upstream of the CP contains a putative movement-like protein. For CwmSV, this region (amino acid positions 312-883) displayed sequence homology to the MP of Blackcurrant reversion virus (E-value: 5.42e-86, amino acid identity: 46%). Both TfSV and SfSV displayed similar levels of sequence identity in this region. We detected the LPL motif which is commonly found in nepovirus MPs in all three viruses (Figure 4.5b). Phylogenetic analysis of the putative MPs placed these viruses with Blackcurrant reversion virus in the genera Nepovirus (Figure 4.5c).

We found little evidence that these viruses reflect contamination by land plants or other eukaryotes. The *F. antipyretica* transcriptome was composed of reads closely related to a feather moss belonging to the order *Hypnales* to which *F. antipyretica* is also found. Furthermore, a large proportion of reads were assigned to an uncultured eukaryote 18S rRNA gene (54%) (HG421124.1) identical to the F. antipyretica 18S rRNA (AF023714.1) among other bryophyte 18S rRNA genes in a blastn search (e-value = 2e-102, nucleotide identity = 100%) (Figure 4.2). Fungi represented 12% of reads in the *Lonchitis hirsute* transcriptome. Despite this, it is unlikely that TfSV is fungi-associated as no fungal contamination was detected in the *Vittaria lineata* transcriptome in which the closely related SfSV sequence (amino acid identity: 78%) was assembled (Figure 4.2).

**(iv) *Lenarviricota.*** *(a) Mitoviridae.* We identified five virus sequences that cluster within the *Mitoviridae* – denoted Chinese swamp cypress mitovirus (CscMV), Asian bayberry mitovirus (AsbaMV), False cloak ferns mitovirus (FcfMV), Delta maidenhair fern mitovirus (DmfMV) and Lycopod associated mitovirus (LycoMV). The fern (FcfMV and DmfMV) and lycophyte (LycoMV) associated sequences cluster with the fern Azolla filiculoides mitovirus 1 and form a sister group to the plant mitoviruses and non-retroviral endogenous RNA viral elements (NERVEs) (Figure 4.6) (Nibert et al. 2018). The gymnosperm associated sequences form a sister all the plant-associated mitoviruses and NERVEs. LycoMV extends the known host range of plant mitoviruses from ferns to lycophytes. Another mito-virus sequence was detected in the chlorophyte *Bolbocoleon piliferum*, denoted Bolbocoleon piliferum mito-like virus (BopiMV). BopiMV falls basal to the mitoviruses, distinct from various unclassified mito-like viruses including the chlorophyte associated mito-like picolinusvirus (QOW97241) (Figure 4.6). All novel sequences show strong conservation of the motifs characteristic of mitovirus RdRps (Supplementary Figure 4.2) (Xie and Ghabrial 2012).

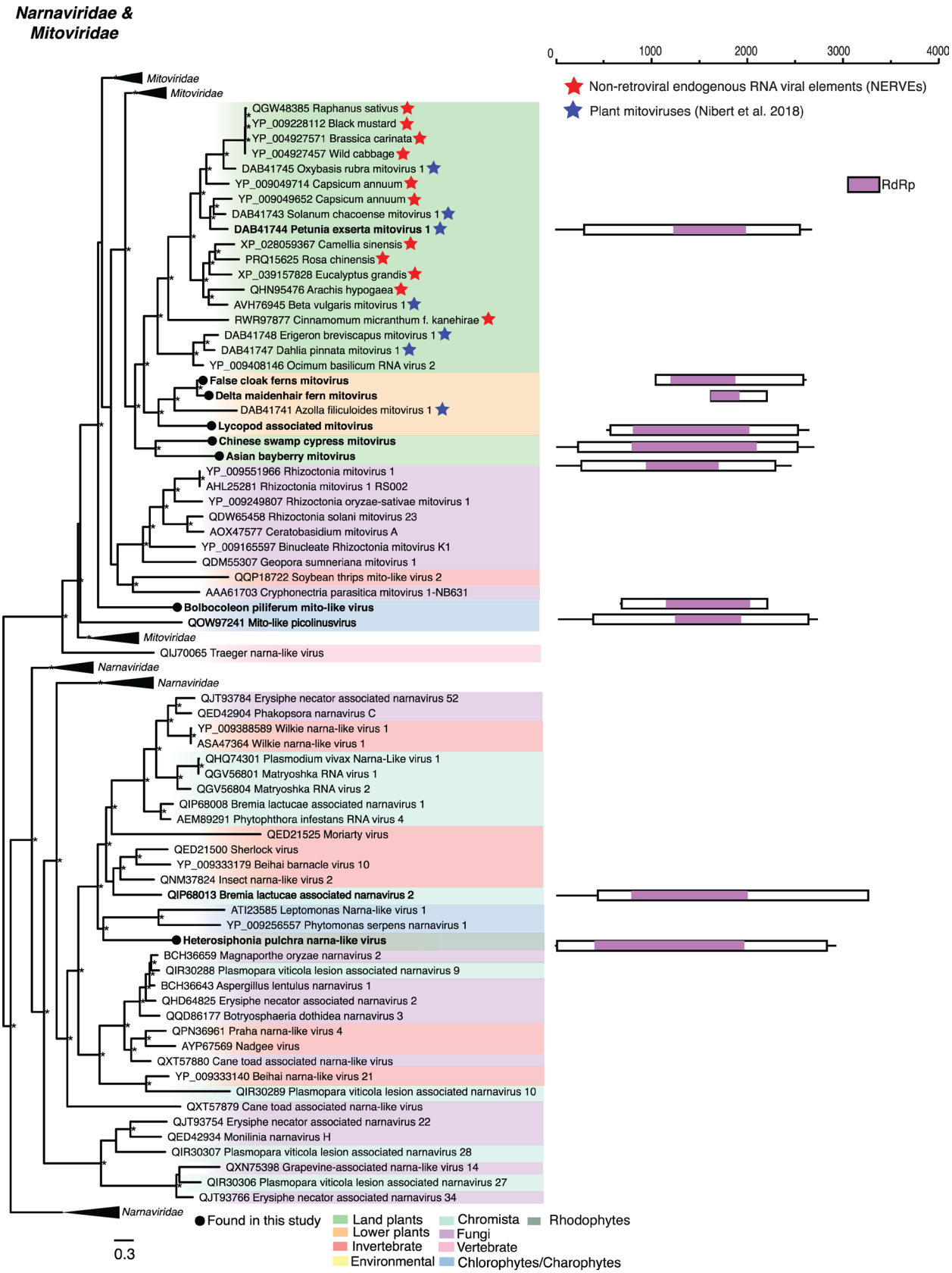

Narnaviridae &
Mitoviridae
0
1000
2000
3000
4000
Non-retroviral endogenous RNA viral elements (NERVEs)
Plant mitoviruses (Nibert et al. 2018)
RdRp
Mitoviridae
Mitoviridae
QGW48385 Raphanus sativus
YP_009228112 Black mustard
YP_004927571 Brassica carinata
YP_004927457 Wild cabbage
DAB41745 Oxybasis rubra mitovirus 1
YP_009049714 Capsicum annuum
YP_009049652 Capsicum annuum
DAB41743 Solanum chacoense mitovirus 1
DAB41744 Petunia exserta mitovirus 1
XP_028059367 Camellia sinensis
PRQ15625 Rosa chinensis
XP_039157828 Eucalyptus grandis
QHN95476 Arachis hypogaea
AVH76945 Beta vulgaris mitovirus 1
RWR97877 Cinnamomum micranthum f. kanehirae
DAB41748 Erigeron breviscapus mitovirus 1
DAB41747 Dahlia pinnata mitovirus 1
YP_009408146 Ocimum basilicum RNA virus 2
False cloak ferns mitovirus
Delta maidenhair fern mitovirus
DAB41741 Azolla filiculoides mitovirus 1
Lycopod associated mitovirus
Chinese swamp cypress mitovirus
Asian bayberry mitovirus
YP_009551966 Rhizoctonia mitovirus 1
AHL25281 Rhizoctonia mitovirus 1 RS002
YP_009249807 Rhizoctonia oryzae-sativae mitovirus 1
QDW65458 Rhizoctonia solani mitovirus 23
AOX47577 Ceratobasidium mitovirus A
YP_009165597 Binucleate Rhizoctonia mitovirus K1
QDM55307 Geopora sumneriana mitovirus 1
QQP18722 Soybean thrips mito-like virus 2
AAA61703 Cryphonectria parasitica mitovirus 1-NB631
Bolbocoleon piliferum mito-like virus
QOW97241 Mito-like picolinusvirus
Mitoviridae
QIJ70065 Traeger narna-like virus
Narnaviridae
Narnaviridae
QJT93784 Erysiphe necator associated narnavirus 52
QED42904 Phakopsora narnavirus C
YP_009388589 Wilkie narna-like virus 1
ASA47364 Wilkie narna-like virus 1
QHQ74301 Plasmodium vivax Narna-Like virus 1
QGV56801 Matryoshka RNA virus 1
QGV56804 Matryoshka RNA virus 2
QIP68008 Bremia lactucae associated narnavirus 1
AEM89291 Phytophthora infestans RNA virus 4
QED21525 Moriarty virus
QED21500 Sherlock virus
YP_009333179 Beihai barnacle virus 10
QNM37824 Insect narna-like virus 2
QIP68013 Bremia lactucae associated narnavirus 2
ATI23585 Leptomonas Narna-like virus 1
YP_009256557 Phytomonas serpens narnavirus 1
Heterosiphonia pulchra narna-like virus
BCH36659 Magnaporthe oryzae narnavirus 2
QIR30288 Plasmopara viticola lesion associated narnavirus 9
BCH36643 Aspergillus lentulus narnavirus 1
QHD64825 Erysiphe necator associated narnavirus 2
QQD86177 Botryosphaeria dothidea narnavirus 3
QPN36961 Praha narna-like virus 4
AYP67569 Nadgee virus
QXT57880 Cane toad associated narna-like virus
YP_009333140 Beihai narna-like virus 21
QIR30289 Plasmopara viticola lesion associated narnavirus 10
QXT57879 Cane toad associated narna-like virus
QJT93754 Erysiphe necator associated narnavirus 22
QED42934 Monilinia narnavirus H
QIR30307 Plasmopara viticola lesion associated narnavirus 28
QXN75398 Grapevine-associated narna-like virus 14
QIR30306 Plasmopara viticola lesion associated narnavirus 27
QJT93766 Erysiphe necator associated narnavirus 34
Narnaviridae
Found in this study
Land plants
Lower plants
Invertebrate
Environmental
Chromista
Fungi
Vertebrate
Chlorophytes/Charophytes
Rhodophytes
0.3

***Figure 4.25 Phylogenetic relationships and genome organization of the viruses within the virus families Narnaviridae and Mitoviridae*** *Left: ML phylogenetic trees based on the replication protein show the topological position of virus-like sequences discovered in this study (black circles) in the context of their closest relatives. See Figure 4.3 for the colour scheme. Blue stars signify mitovirus sequences identified in (Nibert et al. 2018). Red stars signify non-retroviral endogenous RNA viral elements (NERVEs). 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% bootstrap support. Tip labels are bolded when the genome structure is shown on the right. Right: Genomic organization of the virus sequences identified in this study and representative species used in the phylogeny.*

There is little evidence to suggest that these sequences are derived from a non-plant organism. While the FcfMV and DmfMV libraries were contaminated with fungi, (12% and 15% of reads respectively) fungi-associated reads were absent in the libraries of all other mitoviruses. As the codon UGA encodes tryptophan (Trp) in fungal mitochondria this codon assignment is also present in fungal mitoviruses (Heinze 2012; Nibert 2017; Zhang et al. 2015). In contrast, the UGA codon in plant mitochondria is a stop codon and hence absent from plant mitovirus sequences except as a stop codon (Nibert et al. 2018). The absence of internal UGA codons in these sequences is further evidence that these sequences are plant-derived (Nibert et al. 2018; Shackelton and Holmes 2008). In addition, we found no evidence through searches of the 1KP genome scaffolds and the WGS shotgun database that these sequences are mitochondrial or nuclear NERVEs. Furthermore, CscMV, AsbaMV and LycoMV contain complete RdRps and their UTRs share similarities in length and identity with plant mitoviruses.

*(b) Narnaviridae.* A partial narna-like virus sequence was identified in the rhodophyte Heterosiphonia pulchra denoted Heterosiphonia pulchra narna-like virus (HspuNV). HspuNV clusters with unclassified trypanosomatid associated viruses. While ~5% of reads in this library were associated with fungi the phylogenetic position of this virus suggests that it is not derived from fungi (Figure 4.2, Figure 4.6).

**(v) *Tolivirales.*** *(a) Tombusviridae.* An alphacarmo-like virus tentatively named Ihi tombusvirus (IhiTV) was identified in an Ihi (Portulaca molokiniensis) sample. IhiTV is phylogenetically positioned within the alphacarmoviruses (Supplementary Figure 4.3).

**(vi)** ***Patatavirales****. (a) Potyviridae.* We identified three virus-like sequences that clustered with plant viruses in the family *Potyviridae* – Traubia modesta potyvirus (TramPV), Common milkweed potyvirus (ComPV) and Salt wort potyvirus (SawPV). TramPV and ComPV shared 87% amino acid identity and may therefore represent a single virus species. The potyvirus-like sequences discovered all group with known potyviruses in a phylogenetic analysis of the Nib gene (Supplementary Figure 4.3).

**(vii)** ***Martellivirales.*** *(a) Endornaviridae.* Six alphaendorna-like virus sequences were detected in the five green algae species and one lycophyte. The green algae and lycophyte associated alphaendorna-like viruses, termed Bolbocoleon piliferum endorna-like virus (BopiEV), Volvox aureus endorna-like virus (VoauEV), Carteria obtusa associated endorna-like virus (CaobEV), Brachiomonas submarina associated endorna-like virus (BrsuEV), Staurastrum sebaldi endornavirus (SsEV) and Krauss' spikemoss associated endorna-like virus (KrspEV), fall across the alphaendornavirus phylogeny and predominately cluster with algae and fungi-associated viruses (Figure 4.7). There was little evidence of algae (non-host) or fungi contamination in the *S. sebaldi, B. piliferum* and *V. aureus* transcriptomes with <1% of all reads associated with these groups (Figure 4.2). Non-green algae contaminants were present in the *C. obtus* (28%), *B. submarina* (7%) and *S. kraussiana* (4%) transcriptomes where fungi also appeared as a notable contaminate representing 11% of all reads (Figure 4.2). To our knowledge, these sequences represent the first endornavirus associated with charophytes, chlorophytes and lycophytes although further work is needed to confirm the virus-host associations.

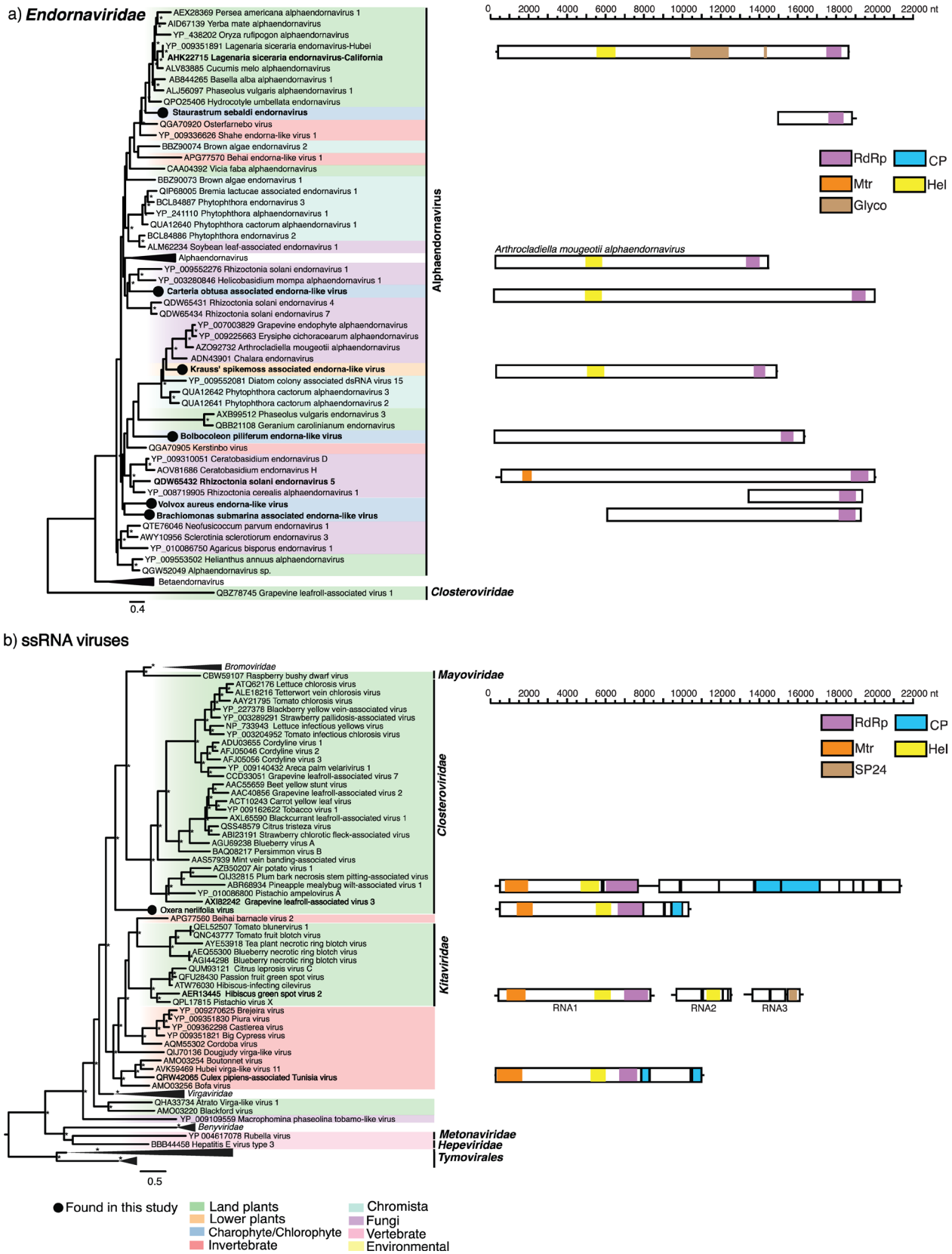

a) Endornaviridae
AEX28369 Persea americana alphaendornavirus 1
AID67139 Yerba mate alphaendornavirus
YP_438202 Oryza rufipogon alphaendornavirus
YP_009351891 Lagenaria siceraria endornavirus-Hubei
AHK22715 Lagenaria siceraria endornavirus-California
ALV83885 Cucumis melo alphaendornavirus
AB844265 Basella alba alphaendornavirus 1
ALJ56097 Phaseolus vulgaris alphaendornavirus 1
QPO25406 Hydrocotyle umbellata endornavirus
Staurastrum sebaldi endornavirus
QGA70920 Osterfarnebo virus
YP_009336626 Shahe endorna-like virus 1
BBZ90074 Brown algae endornavirus 2
APG77570 Behai endorna-like virus 1
CAA04392 Vicia faba alphaendornavirus
BBZ90073 Brown algae endornavirus 1
QIP68005 Bremia lactucae associated endornavirus 1
BCL84887 Phytophthora endornavirus 3
YP_241110 Phytophthora alphaendornavirus 1
QUA12640 Phytophthora cactorum alphaendornavirus 1
BCL84886 Phytophthora endornavirus 2
ALM62234 Soybean leaf-associated endornavirus 1
Alphaendornavirus
YP_009552276 Rhizoctonia solani endornavirus 1
YP_003280846 Helicobasidium mompa alphaendornavirus 1
Carteria obtusa associated endorna-like virus
QDW65431 Rhizoctonia solani endornavirus 4
QDW65434 Rhizoctonia solani endornavirus 7
YP_007003829 Grapevine endophyte alphaendornavirus
YP_009225663 Erysiphe cichoracearum alphaendornavirus
AZO92732 Arthrocladiella mougeotii alphaendornavirus
ADN43901 Chalara endornavirus
Krauss' spikemoss associated endorna-like virus
YP_009552081 Diatom colony associated dsRNA virus 15
QUA12642 Phytophthora cactorum alphaendornavirus 3
QUA12641 Phytophthora cactorum alphaendornavirus 2
AXB99512 Phaseolus vulgaris endornavirus 3
QBB21108 Geranium carolinianum endornavirus
Bolbocoleon piliferum endorna-like virus
QGA70905 Kerstinbo virus
YP_009310051 Ceratobasidium endornavirus D
AOV81686 Ceratobasidium endornavirus H
QDW65432 Rhizoctonia solani endornavirus 5
YP_008719905 Rhizoctonia cerealis alphaendornavirus 1
Volvox aureus endorna-like virus
Brachiomonas submarina associated endorna-like virus
QTE76046 Neofusicoccum parvum endornavirus 1
AWY10956 Sclerotinia sclerotiorum endornavirus 3
YP_010086750 Agaricus bisporus endornavirus 1
YP_009553502 Helianthus annuus alphaendornavirus
QGW52049 Alphaendornavirus sp.
Betaendornavirus
QBZ78745 Grapevine leafroll-associated virus 1
Alphaendornavirus
Closteroviridae
0.4
0 2000 4000 6000 8000 10000 12000 14000 16000 18000 20000 22000 nt
RdRp
CP
Mtr
Hel
Glyco
Arthrocladiella mougeotii alphaendornavirus
b) ssRNA viruses
Bromoviridae
CBW59107 Raspberry bushy dwarf virus
Mayoviridae
ATQ62176 Lettuce chlorosis virus
ALE18216 Tetterwort vein chlorosis virus
AAY21795 Tomato chlorosis virus
YP_227378 Blackberry yellow vein-associated virus
YP_003289291 Strawberry pallidosis-associated virus
NP_733943 Lettuce infectious yellows virus
YP_003204952 Tomato infectious chlorosis virus
ADU03655 Cordyline virus 1
AFJ05046 Cordyline virus 2
AFJ05056 Cordyline virus 3
YP_009140432 Areca palm velarivirus 1
CCD33051 Grapevine leafroll-associated virus 7
AAC55659 Beet yellow stunt virus
AAC40856 Grapevine leafroll-associated virus 2
ACT10243 Carrot yellow leaf virus
YP 009162622 Tobacco virus 1
AXL65590 Blackcurrant leafroll-associated virus 1
QSS48579 Citrus tristeza virus
ABI23191 Strawberry chlorotic fleck-associated virus
AGU69238 Blueberry virus A
BAQ08217 Persimmon virus B
AAS57939 Mint vein banding-associated virus
AZB50207 Air potato virus 1
QIJ32815 Plum bark necrosis stem pitting-associated virus
ABR68934 Pineapple mealybug wilt-associated virus 1
YP_010086800 Pistachio ampelovirus A
AXI82242 Grapevine leafroll-associated virus 3
Closteroviridae
Oxera neriifolia virus
APG77560 Beihai barnacle virus 2
QEL52507 Tomato blunervirus 1
QNC43777 Tomato fruit blotch virus
AYE53918 Tea plant necrotic ring blotch virus
AEQ55300 Blueberry necrotic ring blotch virus
AGI44298 Blueberry necrotic ring blotch virus
QUM93121 Citrus leprosis virus C
QFU28430 Passion fruit green spot virus
ATW76030 Hibiscus-infecting cilevirus
AER13445 Hibiscus green spot virus 2
QPL17815 Pistachio virus X
Kitaviridae
YP_009270625 Brejeira virus
YP_009351830 Piura virus
YP_009362298 Castlerea virus
YP 009351821 Big Cypress virus
AQM55302 Cordoba virus
QUJ70136 Dougudy virga-like virus
AMO03254 Boutonnet virus
AVK59469 Hubei virga-like virus 11
QRW42065 Culex pipiens-associated Tunisia virus
AMO03256 Bofa virus
Virgaviridae
QHA33734 Atrato Virga-like virus 1
AMO03220 Blackford virus
YP_009109559 Macrophomina phaseolina tobamo-like virus
Benyviridae
YP 004617078 Rubella virus
Metonaviridae
BBB44458 Hepatitis E virus type 3
Hepeviridae
Tymovirales
0.5
0 2000 4000 6000 8000 10000 12000 14000 16000 18000 20000 22000 nt
RdRp
CP
Mtr
Hel
SP24
RNA1
RNA2
RNA3
Found in this study
Land plants
Lower plants
Charophyte/Chlorophyte
Invertebrate
Chromista
Fungi
Vertebrate
Environmental

***Figure 4.26 Phylogenetic relationships and genome organization of the endorna-like and unclassified (+)ssRNA virus.*** *Left: ML phylogenetic trees based on the replication protein show the topological position of (a) endorna-like and (b) unclassified ssRNA virus sequences discovered in this study in the context of those obtained previously. Right: Genomic organization of the (a) endorna-like and (b) unclassified ssRNA virus sequence identified in this study and representative species used in the phylogeny. For all trees, 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% bootstrap support. Tip labels are bolded when the genome structure is shown on the right. See Figure 4.3 for the colour scheme. Viruses discovered in this study are signified using a black circle on the tree tip.*

**Unclassified.** We identified a virus-like sequence in an Oxera neriifolia library, termed Oxera neriifolia associated virus. The sequence, 10,214 nt in length, contained four ORFs. The first ORF (7,536 nt) comprised of a viral methyltransferase, helicase, and RNA polymerase while the third ORF (513 nt) most closely resembled a CP. ORF one and ORF three shared the greatest sequence similarity with Culex pipiens associated Tunisia virus (32% amino acid identity). The second and fourth ORF had no homology to known viruses. The genome organization of OnV is distinct from the other related plant virus families (Figure 4.7). OnV forms a distinct and well-supported outgroup to the Closteroviridae, Bromoviridae and Mayoviridae families. As such, OnV may potentially constitute a new virus family (Figure 4.7). We found little evidence that OnV was detected due to contamination by other eukaryotes (Figure 4.2).

**Negative-sense single-stranded RNA [(-)ssRNA] viruses**

**(i) *Bunyavirales*.** *(a) Phenuiviridae.* A phenui-like virus sequence termed *Brown algae phenui-like virus* (BralPV) was recovered from a *Sargassum thunbergii* transcriptome. The partial L segment clusters with the unclassified plant and fungi viruses (Figure 4.8A). No additional phenui-like virus segments were recovered. No concerning contaminants were detected in the *S. thunbergii* transcriptome (Figure 4.2).

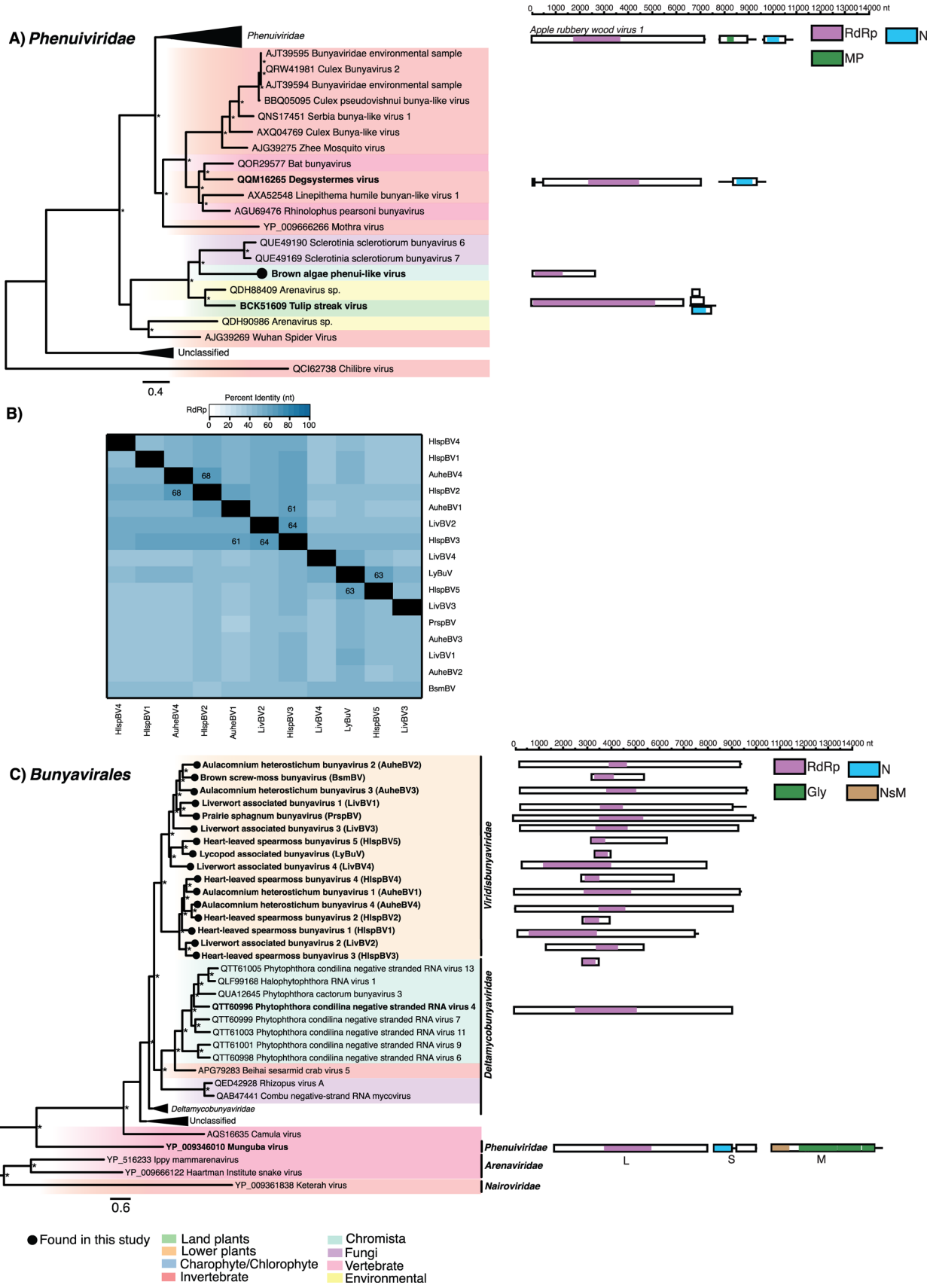

A) Phenuiviridae
Phenuiviridae
AJT39595 Bunyaviridae environmental sample
QRW41981 Culex Bunyavirus 2
AJT39594 Bunyaviridae environmental sample
BBQ05095 Culex pseudovishnui bunya-like virus
QNS17451 Serbia bunya-like virus 1
AXQ04769 Culex Bunya-like virus
AJG39275 Zhee Mosquito virus
QOR29577 Bat bunyavirus
QQM16265 Degsystermes virus
AXA52548 Linepithema humile bunyan-like virus 1
AGU69476 Rhinolophus pearsoni bunyavirus
YP_009666266 Mothra virus
QUE49190 Sclerotinia sclerotiorum bunyavirus 6
QUE49169 Sclerotinia sclerotiorum bunyavirus 7
Brown algae phenui-like virus
QDH88409 Arenavirus sp.
BCK51609 Tulip streak virus
QDH90986 Arenavirus sp.
AJG39269 Wuhan Spider Virus
Unclassified
QCI62738 Chilibre virus
0.4
0 1000 2000 3000 4000 5000 6000 7000 8000 9000 10000 11000 12000 13000 14000 nt
Apple rubbery wood virus 1
RdRp
N
MP
B)
Percent Identity (nt)
RdRp
0 20 40 60 80 100
HlspBV4
HlspBV1
AuheBV4
HlspBV2
AuheBV1
LivBV2
HlspBV3
LivBV4
LyBuV
HlspBV5
LivBV3
PrspBV
AuheBV3
LivBV1
AuheBV2
BsmBV
68
68
61
64
61
64
63
63
C) Bunyavirales
Aulacomnium heterostichum bunyavirus 2 (AuheBV2)
Brown screw-moss bunyavirus (BsmBV)
Aulacomnium heterostichum bunyavirus 3 (AuheBV3)
Liverwort associated bunyavirus 1 (LivBV1)
Prairie sphagnum bunyavirus (PrspBV)
Liverwort associated bunyavirus 3 (LivBV3)
Heart-leaved spearmoss bunyavirus 5 (HlspBV5)
Lycopod associated bunyavirus (LyBuV)
Liverwort associated bunyavirus 4 (LivBV4)
Heart-leaved spearmoss bunyavirus 4 (HlspBV4)
Aulacomnium heterostichum bunyavirus 1 (AuheBV1)
Aulacomnium heterostichum bunyavirus 4 (AuheBV4)
Heart-leaved spearmoss bunyavirus 2 (HlspBV2)
Heart-leaved spearmoss bunyavirus 1 (HlspBV1)
Liverwort associated bunyavirus 2 (LivBV2)
Heart-leaved spearmoss bunyavirus 3 (HlspBV3)
Viridisbunyaviridae
QTT61005 Phytophthora condilina negative stranded RNA virus 13
QLF99168 Halophytophthora RNA virus 1
QUA12645 Phytophthora cactorum bunyavirus 3
QTT60996 Phytophthora condilina negative stranded RNA virus 4
QTT60999 Phytophthora condilina negative stranded RNA virus 7
QTT61003 Phytophthora condilina negative stranded RNA virus 11
QTT61001 Phytophthora condilina negative stranded RNA virus 9
QTT60998 Phytophthora condilina negative stranded RNA virus 6
Deltamycobunyaviridae
APG79283 Beihai sesarmid crab virus 5
QED42928 Rhizopus virus A
QAB47441 Combu negative-strand RNA mycovirus
Deltamycobunyaviridae
Unclassified
AQS16635 Camula virus
YP_009346010 Munguba virus
YP_516233 Ippy mammarenavirus
YP_009666122 Haartman Institute snake virus
YP_009361838 Keterah virus
Phenuiviridae
Arenaviridae
Nairoviridae
0.6
0 1000 2000 3000 4000 5000 6000 7000 8000 9000 10000 11000 12000 13000 14000 nt
RdRp
N
Gly
NsM
L
S
M
Found in this study
Land plants
Lower plants
Charophyte/Chlorophyte
Invertebrate
Chromista
Fungi
Vertebrate
Environmental

***Figure 4.27 Phylogenetic relationships and genome organization of the phenui-like and bunya-like viruses*** *(a) Left: A phylogeny depicting the phenui-like virus identified in this study in the context of the Phenuiviridae. Right: Genomic organization of the virus sequences identified in this study and representative species used in the phylogeny. (b) Percent identity matrix of the novel bunya-like viruses. Identity scores are calculated from an alignment of the RdRp protein coding sequence. For clarity, the 100% identity along the diagonal has been removed. Where sequence identity is >= 60% the value is shown. (c) Left: A phylogeny depicting a novel clade of viruses related to the Deltamycobunyaviridae in the context of the Bunyavirales. Right: Genomic organization of the virus sequences identified in this study and representative species used in the phylogeny. For all trees, 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% bootstrap support. Tip labels are bolded when the genome structure is shown on the right. See Figure 4.3 for the colour scheme. Viruses discovered in this study are signified using a black circle on the tree tip.*

*(b) Viridisbunyaviridae.* We identified 16 bunya-like virus sequences from eight bryophyte libraries. Three libraries contained multiple distinct putative complete and partial viruses. The overall pairwise nucleotide identity was <70% between each sequence (Figure 4.8b). As such we consider each a different bunya-like viruses. These sequences group together to form a novel clade of unclassified bunya-like viruses distantly related to oomycete, fungi, and invertebrate viruses (Figure 4.8c). Bunyaviruses typically comprise three segments (L, M, and S), although only the L segment was recovered for these sequences. These sequences represent the first plant-associated viruses that cluster near the unofficially named *Deltamycobunyaviridae* (Nerva et al. 2019) (Figure 4.8c). As the complete coding sequences of the viruses discovered share <30% amino acid identity to the nearest relatives in the *Deltamycobunyaviridae,* they may constitute a new virus family. We tentatively name this virus family the *Viridisbunyaviridae*, *(Viridis* meaning green, while bunya is derived from the virus *order Bunyavirales* in which this clade falls within*).* There was no evidence that these sequences originated from non-plant contaminants. Host assignment was unclear for *Lycopod associated bunyavirus* and *Liverwort associated bunyavirus 1:4* as reads belonging to several lycophyte and liverwort species, respectively, were found in the source transcriptomes (Figure 4.2).

**(ii) *Mononegavirales.*** *(a) Rhabdoviridae.* We identified seven sequences that clustered with plant viruses in the family *Rhabdoviridae* denoted *Canadian violet rhabdovirus 1* (CvRV1), Canadian violet rhabdovirus 2 (CvRV2), *Common ivy rhabdovirus* (CoiRV) and *Indian pipe rhabdovirus* (InpRV), *Tree fern varicosa-like virus* (TfVV), *Monoclea gottschei varicosa-like virus* (MgVV) and *Bug moss associated rhabdo-like virus* (BmRV). Notably, TfVV and MgVV expand the host range of the rhabdoviruses from angiosperms and gymnosperm to ferns and liverworts. RNA2 segments were recovered for both viruses, TfVV RNA2 contained five genes while MgVV contained four (Figure 4.9a). Two partial segments sharing similarities to the nucleocapsid (N) of *Black grass varicosavirus-like virus* (YP_009130620.1) were found in the Indian pipe library and share 50% amino acid identity. All sequences likely represent novel species within known plant infecting genera (Figure 4.9a).

BmRV is a partial sequence (693 nt) most closely related to the unclassified Hubei rhabdo-like virus 2 (44% amino acid identity). Further evidence is needed to confirm BmRV as the first bryophyterhabdovirus, but the relatively low proportion of contaminates in this library (3% algae and 3% fungi) suggests that this virus is plant-associated (Figure 4.2). While 53% of reads in the MgVV library were fungi associated the phylogenetic position of MgVV suggests it is derived from plants (Figure 4.2, Figure 4.9a).

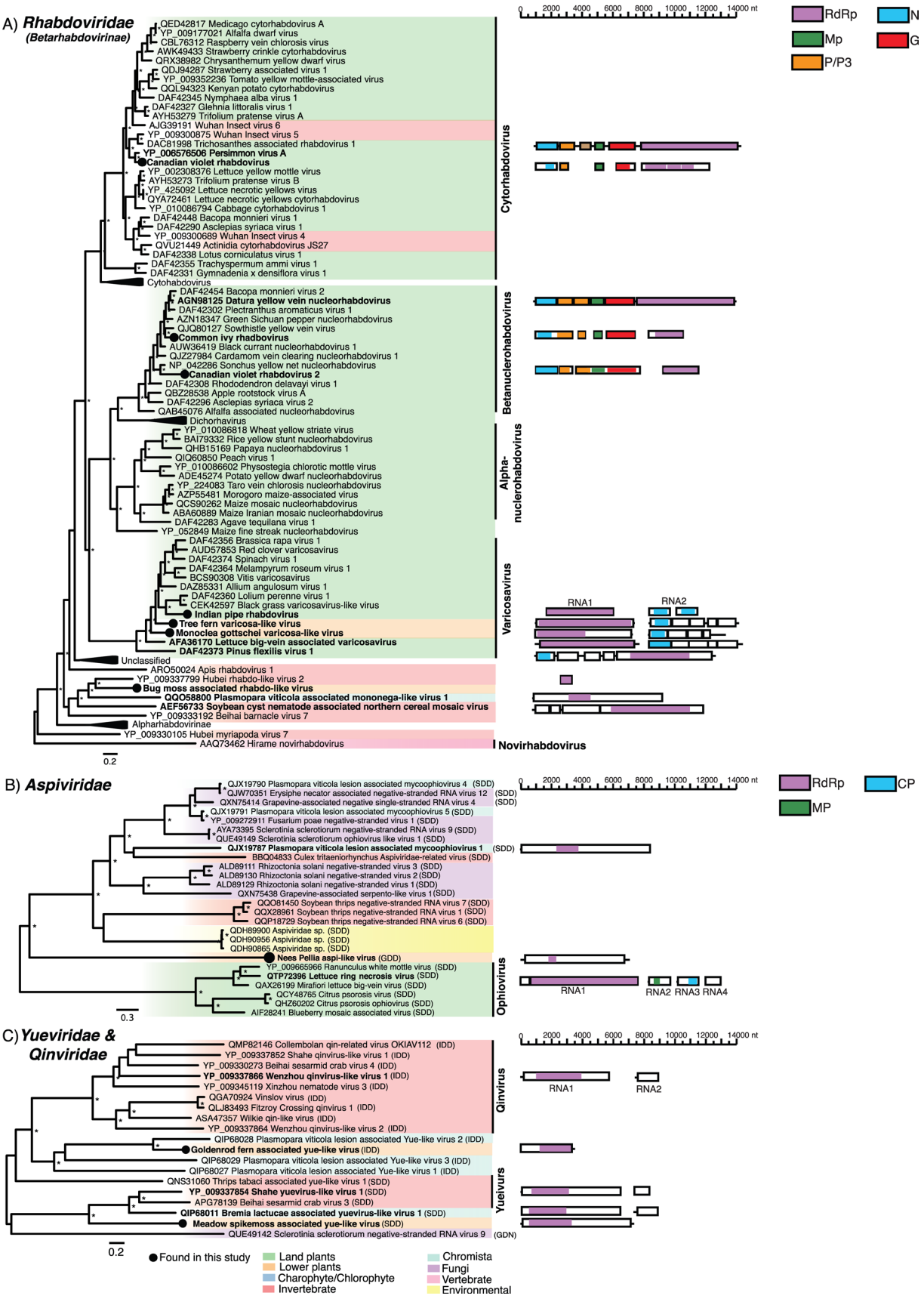

A) Rhabdoviridae
(Betarhabdovirinae)
QED42817 Medicago cytorhabdovirus A
YP_009177021 Alfalfa dwarf virus
CBL76312 Raspberry vein chlorosis virus
AWK49433 Strawberry crinkle cytorhabdovirus
QRX38982 Chrysanthemum yellow dwarf virus
QDJ94287 Strawberry associated virus 1
YP_009352236 Tomato yellow mottle-associated virus
QQL94323 Kenyan potato cytorhabdovirus
DAF42345 Nymphaea alba virus 1
DAF42327 Glehnia littoralis virus 1
AYH53279 Trifolium pratense virus A
AJG39191 Wuhan Insect virus 6
YP_009300875 Wuhan Insect virus 5
DAC81998 Trichosanthes associated rhabdovirus 1
YP_006576506 Persimmon virus A
Canadian violet rhabdovirus
YP_002308376 Lettuce yellow mottle virus
AYH53273 Trifolium pratense virus B
YP_425092 Lettuce necrotic yellows virus
QYA72461 Lettuce necrotic yellows cytorhabdovirus
YP_010086794 Cabbage cytorhabdovirus 1
DAF42448 Bacopa monnieri virus 1
DAF42290 Asclepias syriaca virus 1
YP_009300689 Wuhan Insect virus 4
QVU21449 Actinidia cytorhabdovirus JS27
DAF42338 Lotus corniculatus virus 1
DAF42355 Trachyspermum ammi virus 1
DAF42331 Gymnadenia x densiflora virus 1
Cytohabdovirus
Cytorhabdovirus
DAF42454 Bacopa monnieri virus 2
AGN98125 Datura yellow vein nucleorhabdovirus
DAF42302 Plectranthus aromaticus virus 1
AZN18347 Green Sichuan pepper nucleorhabdovirus
QJQ80127 Sowthistle yellow vein virus
Common ivy rhadbovirus
AUW36419 Black currant nucleorhabdovirus 1
QJZ27984 Cardamom vein clearing nucleorhabdovirus 1
NP_042286 Sonchus yellow net nucleorhabdovirus
Canadian violet rhabdovirus 2
DAF42308 Rhododendron delavayi virus 1
QBZ28538 Apple rootstock virus A
DAF42296 Asclepias syriaca virus 2
QAB45076 Alfalfa associated nucleorhabdovirus
Dichorhavirus
Betanucleorhabdovirus
YP_010086818 Wheat yellow striate virus
BAI79332 Rice yellow stunt nucleorhabdovirus
QHB15169 Papaya nucleorhabdovirus 1
QIQ60850 Peach virus 1
YP_010086602 Physostegia chlorotic mottle virus
ADE45274 Potato yellow dwarf nucleorhabdovirus
YP_224083 Taro vein chlorosis nucleorhabdovirus
AZP55481 Morogoro maize-associated virus
QCS90262 Maize mosaic nucleorhabdovirus
ABA60889 Maize Iranian mosaic nucleorhabdovirus
DAF42283 Agave tequilana virus 1
YP_052849 Maize fine streak nucleorhabdovirus
Alpha-nucleorhabdovirus
DAF42356 Brassica rapa virus 1
AUD57853 Red clover varicosavirus
DAF42374 Spinach virus 1
DAF42364 Melampyrum roseum virus 1
BCS90308 Vitis varicosavirus
DAZ85331 Allium angulosum virus 1
DAF42360 Lolium perenne virus 1
CEK42597 Black grass varicosavirus-like virus
Indian pipe rhabdovirus
Tree fern varicosa-like virus
Monoclea gottschei varicosa-like virus
AFA36170 Lettuce big-vein associated varicosavirus
DAF42373 Pinus flexilis virus 1
Unclassified
Varicosavirus
RNA1
RNA2
ARO50024 Apis rhabdovirus 1
YP_009337799 Hubei rhabdo-like virus 2
Bug moss associated rhabdo-like virus
QQO58800 Plasmopara viticola associated mononega-like virus 1
AEF56733 Soybean cyst nematode associated northern cereal mosaic virus
YP_009333192 Beihai barnacle virus 7
Alpharhabdovirinae
YP_009330105 Hubei myriapoda virus 7
AAQ73462 Hirame novirhabdovirus
Novirhabdovirus
0.2
0 2000 4000 6000 8000 10000 12000 14000 nt
RdRp
N
Mp
G
P/P3
B) Aspiviridae
QJX19790 Plasmopara viticola lesion associated mycoophiovirus 4 (SDD)
QJW70351 Erysiphe necator associated negative-stranded RNA virus 12 (SDD)
QXN75414 Grapevine-associated negative single-stranded RNA virus 4 (SDD)
QJX19791 Plasmopara viticola lesion associated mycoophiovirus 5 (SDD)
YP_009272911 Fusarium poae negative-stranded virus 1 (SDD)
AYA73395 Sclerotinia sclerotiorum negative-stranded RNA virus 9 (SDD)
QUE49149 Sclerotinia sclerotiorum ophiovirus like virus 1 (SDD)
QJX19787 Plasmopara viticola lesion associated mycoophiovirus 1 (SDD)
BBQ04833 Culex tritaeniorhynchus Aspiviridae-related virus (SDD)
ALD89111 Rhizoctonia solani negative-stranded virus 3 (SDD)
ALD89130 Rhizoctonia solani negative-stranded virus 2 (SDD)
ALD89129 Rhizoctonia solani negative-stranded virus 1 (SDD)
QXN75438 Grapevine-associated serpento-like virus 1 (SDD)
QQO81450 Soybean thrips negative-stranded RNA virus 7 (SDD)
QQX28961 Soybean thrips negative-stranded RNA virus 1 (SDD)
QQP18729 Soybean thrips negative-stranded RNA virus 6 (SDD)
QDH89900 Aspiviridae sp. (SDD)
QDH90956 Aspiviridae sp. (SDD)
QDH90865 Aspiviridae sp. (SDD)
Nees Pellia aspi-like virus (GDD)
YP_009665966 Ranunculus white mottle virus (SDD)
QTP72396 Lettuce ring necrosis virus (SDD)
QAX26199 Mirafiori lettuce big-vein virus (SDD)
QCY48765 Citrus psorosis virus (SDD)
QHZ60202 Citrus psorosis ophiovirus (SDD)
AIF28241 Blueberry mosaic associated virus (SDD)
Ophiovirus
RNA1
RNA2
RNA3
RNA4
0.3
0 2000 4000 6000 8000 10000 12000 14000 nt
RdRp
CP
MP
C) Yueviridae & Qinviridae
QMP82146 Collembolan qin-related virus OKIAV112 (IDD)
YP_009337852 Shahe qinvirus-like virus 1 (IDD)
YP_009330273 Beihai sesarmid crab virus 4 (IDD)
YP_009337866 Wenzhou qinvirus-like virus 1 (IDD)
YP_009345119 Xinzhou nematode virus 3 (IDD)
QGA70924 Vinslov virus (IDD)
QLJ83493 Fitzroy Crossing qinvirus 1 (IDD)
ASA47357 Wilkie qin-like virus (IDD)
YP_009337864 Wenzhou qinvirus-like virus 2 (IDD)
Qinvirus
QIP68028 Plasmopara viticola lesion associated Yue-like virus 2 (IDD)
Goldenrod fern associated yue-like virus (IDD)
QIP68029 Plasmopara viticola lesion associated Yue-like virus 3 (IDD)
QIP68027 Plasmopara viticola lesion associated Yue-like virus 1 (IDD)
QNS31060 Thrips tabaci associated yue-like virus 1 (SDD)
YP_009337854 Shahe yuevirus-like virus 1 (SDD)
APG78139 Beihai sesarmid crab virus 3 (SDD)
QIP68011 Bremia lactucae associated yuevirus-like virus 1 (SDD)
Meadow spikemoss associated yue-like virus (SDD)
QUE49142 Sclerotinia sclerotiorum negative-stranded RNA virus 9 (GDN)
Yueivurs
RNA1
RNA2
0.2
0 2000 4000 6000 8000 10000 12000 14000 nt
Found in this study
Land plants
Lower plants
Charophyte/Chlorophyte
Invertebrate
Chromista
Fungi
Vertebrate
Environmental

***Figure 4.28 Phylogenetic relationships and genome organization of select negative-strand RNA virus families.*** *Left: ML phylogenetic trees based on the RdRp region of the (a) Aspiviridae, (b) Yue- and Qinviridae and (c) Rhabdoviridae. Right: Genomic organization of the virus sequences identified in this study and representative species used in the phylogeny. For all trees, 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% bootstrap support. Tip labels are bolded when the genome structure is shown on the right. See Figure 4.3 for the colour scheme. Viruses discovered in this study are signified using a black circle on the tree tip. For trees (a) and (b), the RdRp motif C trimer of each sequence is shown in brackets at the end of the tip label.*

**(iii)** ***Serpentovirales****. (a) Aspiviridae.* We identified an aspi-like sequence termed Nees' Pellia aspi-like virus (NpAV). A complete RNA1 segment (6,989 nt) was assembled, although no other segments were recovered (Figure 4.9b). NpAV most closely resembles Rhizoctonia solani negative-stranded virus 3 (amino acid identity: 22%) and falls basal to all the unclassified aspi-like viruses including those found in fungi, invertebrates, and oomycetes. NpAV is the first aspi-like virus identified in plants outside of the angiosperms and may constitute a novel virus group (Figure 4.9b). Notably, unlike the other aspiviruses that possess a SDD sequence in motif C of the RdRp – a known signature for segmented negative-stranded RNA viruses – NpAV has a GDD sequence (Figure 4.9b).

**(iv)** ***Goujianvirales.*** *(a) Yueviridae.* A yue-like virus sequence termed Meadow spikemoss associated yue-like virus (MsYV) was found in the lycophyte *Selaginella apoda* and most closely resembles algae associated Bremia lactucae associated yuevirus-like virus 1 (amino acid identity: 26%). Phylogenetic analysis supports the assignment of MsYV as the first plant yuevirus (Figure 4.9c).

A second partial yue-like virus sequence was detected in a Pityrogramma trifoliata library and termed Goldenrod fern associated yue-like virus (GfYV). GfYV falls with a group of oomycete associated viruses. Consistent with the qin-like viruses, GfYV has an IDD (Ile-Asp-Asp) sequence motif instead of the common GDD (Gly-Asp-Asp) in the catalytic core of its RdRp, while MsYV contains SDD (Ser-Asp-Asp) in the same manner as many yue-like viruses (Figure 4.9c). The libraries from which GfYV and MsYV were assembled are contaminated with fungal

reads (5% and 21%, respectively) as such host assignment is made with caution (Figure 4.2). Reads belonging to oomycetes were not found in either library.

**Double-stranded RNA (dsRNA) viruses**

**(i) *Durnavirales.*** *(a) Amalgaviridae.* We detected five sequences that cluster with amalga-like viruses. Lycopod associated amalgavirus (LycoAV) is a partial RdRp containing a sequence that falls basal to the *Amalgaviridae* and represents the first amalga-like virus in the lycophytes (Figure 4.10). Three amalga-like sequences were discovered in green algae and rhodophyte transcriptomes and cluster with Diatom colony associated dsRNA virus 2 (Figure 4.10). As noted in the case of Bryopsis mitochondria-associated dsRNA virus and several green algae associated viruses (Charon et al. 2020) when translated into amino acids using the protozoan mitochondrial code, two overlapping ORFs are present: the first, encoding a hypothetical protein, while the second, a replicase through a –1 ribosomal frameshift (Koga et al. 2003). For two of the amalga-like sequences identified in this study – Nucleotaenium eifelense virus (NueiV) and Rhodella violacea virus (RhviV) – a similar structure was observed but we were unable to identify any ribosomal frameshift motifs in either sequence (Figure 4.10). Further work is needed to confirm if these sequences should be translated through the mitochondrial genetic code.

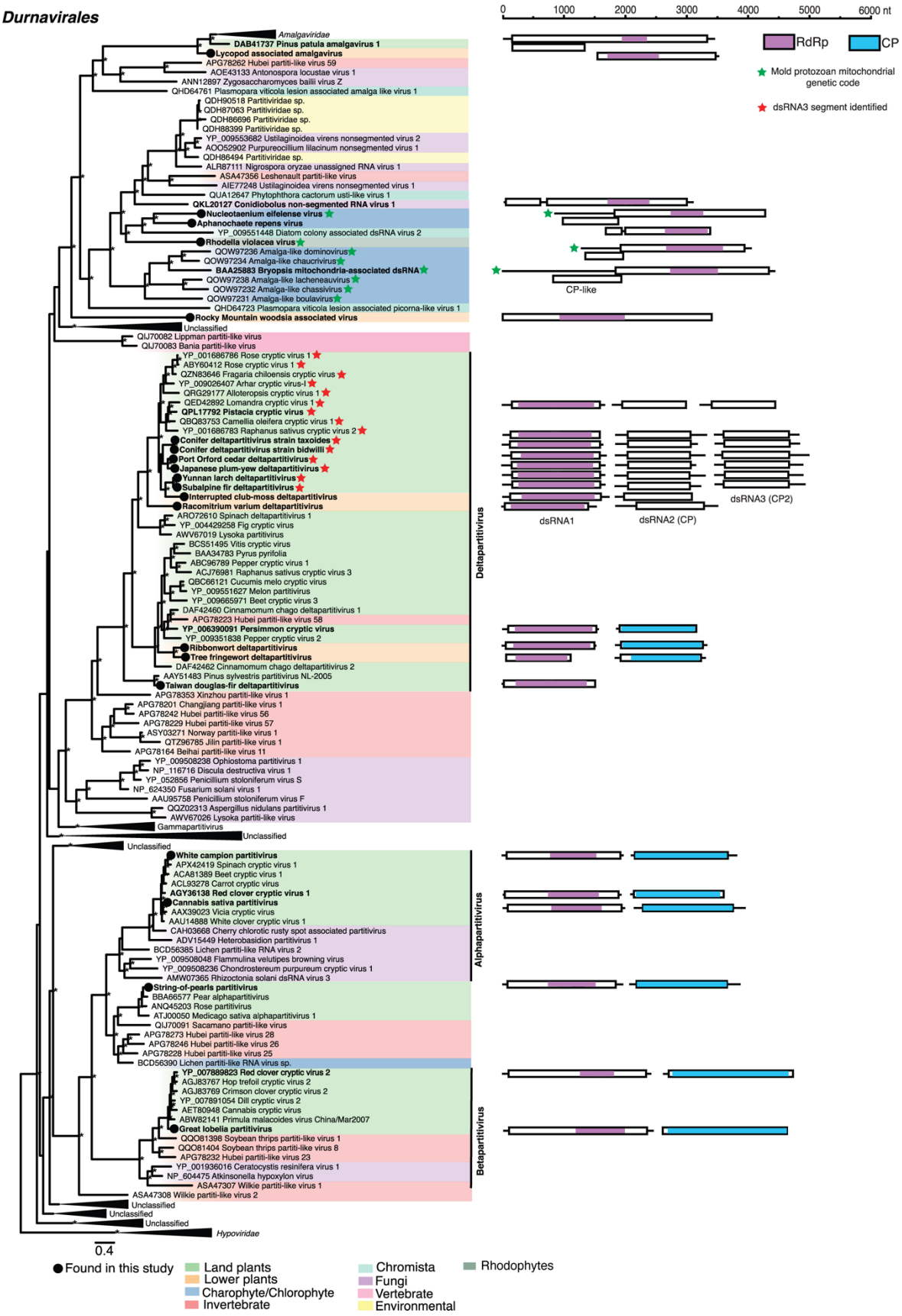

Durnavirales
Amalgaviridae
DAB41737 Pinus patula amalgavirus 1
Lycopod associated amalgavirus
APG78262 Hubei partiti-like virus 59
AOE43133 Antonospora locustae virus 1
ANN12897 Zygosaccharomyces bailii virus Z
QHD64761 Plasmopara viticola lesion associated amalga like virus 1
QDH90518 Partitiviridae sp.
QDH87063 Partitiviridae sp.
QDH86696 Partitiviridae sp.
QDH88399 Partitiviridae sp.
YP_009553682 Ustilaginoidea virens nonsegmented virus 2
AOO52902 Purpureocillium lilacinum nonsegmented virus 1
QDH86494 Partitiviridae sp.
ALR87111 Nigrospora oryzae unassigned RNA virus 1
ASA47356 Leshenault partiti-like virus
AIE77248 Ustilaginoidea virens nonsegmented virus 1
QUA12647 Phytophthora cactorum usti-like virus 1
QKL20127 Conidiobolus non-segmented RNA virus 1
Nucleotaenium eifelense virus
Aphanochaete repens virus
YP_009551448 Diatom colony associated dsRNA virus 2
Rhodella violacea virus
QOW97236 Amalga-like dominovirus
QOW97234 Amalga-like chaucrivirus
BAA25883 Bryopsis mitochondria-associated dsRNA
QOW97238 Amalga-like lacheneauvirus
QOW97232 Amalga-like chassivirus
QOW97231 Amalga-like boulavirus
QHD64723 Plasmopara viticola lesion associated picorna-like virus 1
Rocky Mountain woodsia associated virus
Unclassified
QIJ70082 Lippman partiti-like virus
QIJ70083 Bania partiti-like virus
YP_001686786 Rose cryptic virus 1
ABY60412 Rose cryptic virus 1
QZN83646 Fragaria chiloensis cryptic virus
YP_009026407 Arhar cryptic virus-I
QRG29177 Alloteropsis cryptic virus 1
QED42892 Lomandra cryptic virus 1
QPL17792 Pistacia cryptic virus
QBQ83753 Camellia oleifera cryptic virus 1
YP_001686783 Raphanus sativus cryptic virus 2
Conifer deltapartitivirus strain taxoides
Conifer deltapartitivirus strain bidwillii
Port Orford cedar deltapartitivirus
Japanese plum-yew deltapartitivirus
Yunnan larch deltapartitivirus
Subalpine fir deltapartitivirus
Interrupted club-moss deltapartitivirus
Racomitrium varium deltapartitivirus
ARO72610 Spinach deltapartitivirus 1
YP_004429258 Fig cryptic virus
AWV67019 Lysoka partitivirus
BCS51495 Vitis cryptic virus
BAA34783 Pyrus pyrifolia
ABC96789 Pepper cryptic virus 1
ACJ76981 Raphanus sativus cryptic virus 3
QBC66121 Cucumis melo cryptic virus
YP_009551627 Melon partitivirus
YP_009665971 Beet cryptic virus 3
DAF42460 Cinnamomum chago deltapartitivirus 1
APG78223 Hubei partiti-like virus 58
YP_006390091 Persimmon cryptic virus
YP_009351838 Pepper cryptic virus 2
Ribbonwort deltapartitivirus
Tree fringewort deltapartitivirus
DAF42462 Cinnamomum chago deltapartitivirus 2
AAY51483 Pinus sylvestris partitivirus NL-2005
Taiwan douglas-fir deltapartitivirus
APG78353 Xinzhou partiti-like virus 1
APG78201 Changjiang partiti-like virus 1
APG78242 Hubei partiti-like virus 56
APG78229 Hubei partiti-like virus 57
ASY03271 Norway partiti-like virus 1
QTZ96785 Jilin partiti-like virus 1
APG78164 Beihai partiti-like virus 11
YP_009508238 Ophiostoma partitivirus 1
NP_116716 Discula destructiva virus 1
YP_052856 Penicillium stoloniferum virus S
NP_624350 Fusarium solani virus 1
AAU95758 Penicillium stoloniferum virus F
QQZ02313 Aspergillus nidulans partitivirus 1
AWV67026 Lysoka partiti-like virus
Gammapartitivirus
Unclassified
Unclassified
White campion partitivirus
APX42419 Spinach cryptic virus 1
ACA81389 Beet cryptic virus 1
ACL93278 Carrot cryptic virus
AGY36138 Red clover cryptic virus 1
Cannabis sativa partitivirus
AAX39023 Vicia cryptic virus
AAU14888 White clover cryptic virus 1
CAH03668 Cherry chlorotic rusty spot associated partitivirus
ADV15449 Heterobasidion partitivirus 1
BCD56385 Lichen partiti-like RNA virus 2
YP_009508048 Flammulina velutipes browning virus
YP_009508236 Chondrostereum purpureum cryptic virus 1
AMW07365 Rhizoctonia solani dsRNA virus 3
String-of-pearls partitivirus
BBA66577 Pear alphapartitivirus
ANQ45203 Rose partitivirus
ATJ00050 Medicago sativa alphapartitivirus 1
QLJ70091 Sacamano partiti-like virus
APG78273 Hubei partiti-like virus 28
APG78246 Hubei partiti-like virus 26
APG78228 Hubei partiti-like virus 25
BCD56390 Lichen partiti-like RNA virus sp.
YP_007889823 Red clover cryptic virus 2
AGJ83767 Hop trefoil cryptic virus 2
AGJ83769 Crimson clover cryptic virus 2
YP_007891054 Dill cryptic virus 2
AET80948 Cannabis cryptic virus
ASW82141 Primula malacoides virus China/Mar2007
Great lobelia partitivirus
QQO81398 Soybean thrips partiti-like virus 1
QQO81404 Soybean thrips partiti-like virus 8
APG78232 Hubei partiti-like virus 23
YP_001936016 Ceratocystis resinifera virus 1
NP_604475 Atkinsonella hypoxylon virus
ASA47307 Wilkie partiti-like virus 1
ASA47308 Wilkie partiti-like virus 2
Unclassified
Unclassified
Unclassified
Hypoviridae
0.4
Deltapartitivirus
Alphapartitivirus
Betapartitivirus
0
1000
2000
3000
4000
5000
6000 nt
RdRp
CP
Mold protozoan mitochondrial genetic code
dsRNA3 segment identified
CP-like
dsRNA1
dsRNA2 (CP)
dsRNA3 (CP2)
Found in this study
Land plants
Lower plants
Charophyte/Chlorophyte
Invertebrate
Chromista
Fungi
Vertebrate
Environmental
Rhodophytes

***Figure 4.29 Phylogenetic relationships of the viruses within the order Durnavirales.*** *Left: ML phylogenetic trees based on the replication protein show the topological position of the Durnavirales-like sequences discovered in this study (black circles) in the context of their closest relatives. See Figure 4.3 for the colour scheme. Green stars are used to signify sequences that have been translated using the protozoan mitochondrial genetic code. Red stars are used to signify sequences for which a dsRNA3 coat protein-like segment has been described. 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% bootstrap support. Tip labels are bolded when the genome structure is shown on the right. Right: Genomic organization of the virus sequences identified in this study and representative species used in the phylogeny.*

A contig containing what appears to be a complete coding sequence (3259 nt) and RdRp motifs was assembled in the *Woodsia scopulina* transcriptome and tentatively named Rocky Mountain woodsia associated virus (RmwPV). The predicted RdRp region (918-1921 nt) of RmvPV shares similarity to both partiti-like viruses (e.g., Ustilaginoidea virens nonsegmented virus 2, 26% aa identity) and the unclassified Phytophthora infestans RNA virus 1 (42% aa identity) which likely constitutes a novel virus family (Cai et al. 2009). The resemblance RmwPV shares with two seemingly distantly related virus groups suggest its position within the *Durnavirales* should be treated with caution (Figure 4.10).

The transcriptome in which RmwPV was discovered is contaminated with fungal reads (10% (Figure 4.2). If RmwPV was derived from fungal contaminates this could potentially explain the phylogenetic placement of RmwPV (Figure 4.10). The *Lycopodiella appressa* transcriptome in which LycoAV was discovered is contaminated by reads belonging to species across various land plant groups. Reads belonging to land plants comprised 35% of plant-associated reads while lycopod associated reads comprised 65% (Figure 4.2).

(b) *Partitiviridae*. We detected 14 sequences that share a resemblance with members of the *Partitiviridae*. For each of these sequences, complete dsRNA1 and dsRNA2 segments were recovered. Ten sequences were found in non-flowering plants that cluster within the deltapartitiviruses. A clade within the deltpartitiviruses is known to encode a third segment comprising a divergent dsRNA2 full-length capsid protein with unknown function (Figure 4.10).

We identified dsRNA3 segments in related conifer associated sequences but not in those found in bryophyte and lycophyte libraries (Figure 4.10). Phylogenies estimated on the coding sequences of dsRNA2 and dsRNA3 reveal essentially the same grouping which is largely consistent with the host phylogeny (Supplementary Figure 4.4). We extend the known host range of the deltapartitiviruses to include bryophytes and lycophytes. The remaining sequences were found in eudicots and cluster with known plant partitiviruses (Figure 4.10). The white campion was judged to have contamination from ginseng and chickweed (Carpenter et al. 2019). However, the relatively low proportion of the library these contaminates compose (<1%) suggests that it is unlikely these species are the host of White campion partitivirus (WcPV) (Figure 4.2). There is no evidence that the other partiti-like sequences discovered are derived from contaminates.

**(ii) *Ghabrivirales.*** *(a) Chrysoviridae.* We identified two partial sequences that share resemblance with members of the alphachrysoviruses denoted Mesotaenium kramstae alphachyrso-like virus (MkACV) and Tree fringewort alphachyrso-like virus (TwACV) (Figure 4.11). A complete RNA2 segment was recovered for MkACV which shared similarity with the p98 of various chrysoviruses (Figure 4.11). The MkACV RNA2 segment did not contain the "PGDGXCXXHX" motif commonly found in this protein (Covelli et al. 2004). To our knowledge, these sequences represent the first chyrsoviruses in the *Marchantiophyta* and *Chlorophyta*. While reads belonging to fungi were found in the libraries MkACV and TwACV were assembled from, the phylogenetic positioning of the viruses suggest that they are plant-derived (Figure 4.2, Figure 4.11).

A) Chrysoviridae
APR73428 Isaria javanica chrysovirus 1
AZT88571 Beauveria bassiana chrysovirus 1
QKK35389 Erysiphe necator associated chrysovirus 3
ADO60926 Grapevine associated chrysovirus-1
AZT88567 Penicillium raistrickii chrysovirus 1
BCH36618 Aspergillus fumigatus chrysovirus
QGR26538 Chrysothrix chrysovirus 1
ALD89090 Macrophomina phaseolina chrysovirus 1
APG76049 Wuhan insect virus 30
ADG21213 Verticillium dahliae chrysovirus 1
YP_009667012 Colletotrichum gloeosporioides chrysovirus 1
QCY49458 Pestalotiopsis theae chrysovirus 1
QGR26536 Lepraria chrysovirus 1
AF297176 Helminthosporium victoriae 145S virus
CAH03664 Amasya cherry disease associated chrysovirus
AFE83590 Raphanus sativas chrysovirus 1
AKU48197 Brassica campestris chrysovirus 1
Tree fringewort alphachyrso-like virus
AJA37498 Persea americana chrysovirus
ACU11563 Anthurium mosaic-associated virus
AYD75725 Zea mays chrysovirus 1
Mesotaenium kramstae alphachyrso-like virus
QGA70949 Eskilstorp virus
QRD99893 Xanthi chryso-like virus
QLJ83488 Broome chryso-like virus 1
ASA47445 Shuangao chryso-like virus 1
QRD99909 Evros chryso-like virus
BBV14774 Barley aphid RNA virus 8
Betachrysovirus
0.2
Alphachrysovirus
RdRp
Cp
Protease
0 1000 2000 3000 4000 5000 6000 7000 8000 9000 10000 11000 12000 13000 nt
B) Ghabrivirales
Totiviridae
QGY72626 Plasmopara viticola lesion associated toti 1
QGY72628 Plasmopara viticola lesion associated toti 2
BAT62484 Red clover powdery mildew-associated totivirus 4
QJW70333 Erysiphe necator associated totivirus 5
QPH36700 Rhodosporidiobolus odoratus RNA virus 1
YP_009507835 Xanthophyllomyces dendrorhous virus L1B
AFH09416 Xanthophyllomyces dendrorhous virus L2
YP_007697651 Xanthophyllomyces dendrorhous virus L1A
QGY72630 Plasmopara viticola lesion associated toti 3
AZT88647 Hortaea werneckii totivirus 1
QJW70331 Erysiphe necator associated totivirus 4
QIP68084 Erysiphales associated totivirus 22
QIP68061 Erysiphales associated totivirus 10
BAT62482 Red clover powdery mildew-associated totivirus 3
QIP68080 Erysiphales associated totivirus 20
VUD77425 Umbelopsis ramanniana virus 4
QGA70771 Trichoderma koningiopsis totivirus 1
YP_009342428 Wuhan insect virus 26
QUE40494 Conidiobolus heterosporus totivirus 1
YP_009342434 Wuhan insect virus 27
QED43019 Uromyces totivirus D
AMB17466 Delisea pulchra totivirus IndA
Rhodochaete parvula toti-like virus 3
Totiviridae
AAA50321 Saccharomyces cerevisiae virus L-A
ATL63196 Saccharomyces kudriavzevii virus L-A-1082
AHF27234 Saccharomyces cerevisiae virus L-A-2
QRN76658 Saccharomyces cerevisiae virus L-A
ATL63186 Saccharomyces paradoxus virus L-A-4650
ATL63194 Saccharomyces uvarum virus L-A-10560
AYN80723 Saccharomyces paradoxus virus L-A-66
QRN76668 Torulaspora delbrueckii virus LA
QED43028 Ambrosiozyma totivirus A
KAG2734163 Scheffersomyces stipitis
YP_009507833 Tuber aestivum virus 1
Porphyridium purpureum toti-like virus 2
QNJ34609 Malassezia sympodialis mycovirus
QJA42331 Malassezia restricta virus MrV40L
QJW70339 Erysiphe necator associated totivirus 8
YP_009507831 Scheffersomyces segobiensis virus L
QJW70335 Erysiphe necator associated totivirus 6
QJW70337 Erysiphe necator associated totivirus 7
QIP68067 Erysiphales associated totivirus 13
AAB02146 Saccharomyces cerevisiae virus L-BC (La)
ANJ60743 Saccharomyces cerevisiae virus L-BC-lus
QED43040 Podosphaera totivirus A
Rhodochaete parvula toti-like virus 1
YP_009182176 Red clover powdery mildew-associated totivirus 2
Red algae totivirus 2
Rhodochaete parvula toti-like virus 2
Red algae totivirus 3
AMB17469 Delisea pulchra totivirus IndA
Porphyridium purpureum toti-like virus 1
BAM93353 Rosellinia necatrix quadrivirus 1
Victori-like viruses
ASU09856 Leishmania RNA virus 1
NP_619653 Leishmania RNA virus 1 - 4
AHJ90430 Leishmania RNA virus 1
AYD61662 Blechomonas maslovi Leishmaniavirus-like RNA virus 1
BBZ90082 Red algae totivirus 1
Gracilaria vermiculophylla toti-like virus
AOM63180 Eimeria stiedai RNA virus 1
QIJ70040 Eimeria stiedai RNA virus 1
QXV86548 Totiviridae sp.
AIW58883 Eimeria tenella RNA virus 1
QXV86392 Totiviridae sp.
AF356189 Eimeria brunetti RNA virus 1
Trichomonasvirus
BAU79488 Diatom colony associated dsRNA virus 4 genome type A
BAU79494 Diatom colony associated dsRNA virus 6
BAU79502 Diatom colony associated dsRNA virus 9 genome type B
YP_009552789 Diatom colony associated dsRNA virus 5
YP_009553336 Diatom colony associated dsRNA virus 3
YP_009551559 Diatom colony associated dsRNA virus 8
YP_009551561 Diatom colony associated dsRNA virus 13
QGY72624 Plasmopara viticola lesion associated totivirus-like 5
QJW70329 Erysiphe necator associated totivirus 2
QIP68023 Bremia lactucae associated toti-like virus 2
YP_009333465 Beihai sesarmid crab virus 7
YP_009336942 Hubei toti-like virus 5
YP_009333409 Beihai razor shell virus 4
Prasiola crispa toti-like virus
KAF4649284 Perkinsus chesapeaki
Elkhorn sea moss toti-like virus
QGA70930 Ahus virus
QLJ83490 Parry's Creek toti-like virus 1
YP_009333150 Beihai barnacle virus 15
Bolbocoleon piliferum toti-like virus
Planotaenium ohtanii toti-like virus
YP_009551502 Diatom colony associated dsRNA virus 17 genome type B
QTT60991 Phytophthora condilina RNA virus 2
Totiviridae
QDH87752 Totiviridae sp.
QDH88238 Totiviridae sp.
QDH87440 Totiviridae sp.
QDH90454 Totiviridae sp.
QDH91324 Totiviridae sp.
QDH87365 Totiviridae sp.
QIP68042 Erysiphales associated totivirus 24
Botybirnaviridae
QIP68058 Erysiphales associated toti-like virus 1
QIP68069 Erysiphales associated toti-like virus 2
YP_009182198 Red clover powdery mildew-associated totivirus 9
Heterosiphonia pulchra toti-like virus
QXL43060 Riboviria sp.
Chrysoviridae
Unclassified
0.3
Found in this study
Land plants
Lower plants
Charophyte/Chlorophyte
Invertebrate
Euglenozoa
Chromista
Fungi
Vertebrate
Environmental
Rhodophytes

***Figure 4.30 Phylogenetic relationships of the viruses within the order Ghabrivirales.*** *Left: (a) A phylogeny of the Chrysoviridae, (b) an order level phylogeny. ML phylogenetic trees based on the replication protein show the topological position of the virus-like sequences discovered in this study (black circles) in the context of their closest relatives. See Figure 4.3 for the colour scheme. 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% bootstrap support. Virus taxonomic names are labelled to the right. Right: Genomic organization of the virus sequences identified in this study and representative species used in the phylogeny.*

*(b) Totiviridae.* Thirteen sequences sharing similarities to toti-like viruses were discovered in eight red and green algae transcriptomes. All sequences share less than 50% amino acid identity across their coding sequence, as such we consider each a putative toti-like virus. Among these sequences four cluster with Delisea pulchra totivirus IndA (AMB17469.1) to form a rhodophyte associated clade basal to the totiviruses (Figure 4.11). Gracilaria vermiculophylla toti-like virus (GrveTV) along with Red algae totivirus 1 (BBZ90082) form a sister group to the protozoan infecting leishmaniaviruses (Figure 4.11). The remaining sequences are phylogenetically positioned across the tree of toti-like viruses, commonly occupying basal positions (Figure 4.11).

*Prasiola crispa* is contaminated by reads from the fungi, *Candida albicans*. *Prasiola crispa toti-like virus* (PrcrTV), clusters with unclassified protist, fungi, invertebrate and algae viruses including *Elkhorn sea moss toti-like virus* (EsmTV) (Figure 4.2, Figure 4.11). The Kappaphycus alvarezii transcriptome in which EsmTV was found showed no evidence of contamination suggesting that PrcrTV may also be derived from algae (Figure 4.2). The *Mazzaella japonica* transcriptome in which *Red algae toti-like virus 2*:*3* (RedTV2/3) were discovered was predominantly composed of reads associated with the rhodophyte genera *Chondrus*. As >99% of reads in this library belong to rhodophyte species, RedTV2 and RedTV3 have been assigned to this group. The *Porphyridium purpureum* transcriptome is highly contaminated by reads belonging to flowering plants and an unidentified cloning vector (M10197.1) (Figure 4.2). The phylogenetic positioning of the viruses discovered from this transcriptome (*Porphyridium purpureum toti-like virus* 1 & 2) point towards it being derived from rhodophytes rather than flowering plants (Figure 4.11).

### 4.4.4 Long-term virus-host evolutionary relationships

To examine the frequency of cross-species transmission and co-divergence among plant viruses, we estimated tanglegrams that depict pairs of rooted phylogenetic trees displaying the evolutionary relationship between a virus family and their hosts. This revealed cross-species transmission as the predominate macroevolutionary process predicted among all the RNA virus groups analysed (median 65%, range 46%-79%) (Figure 4.12). Cross-species transmission was most frequent in the *Betaflexiviridae* (79%) and the subfamily *Betarhabdovirinae* (79%). However, virus-host co-divergence (median 23%, range 14%-29%) and to a lesser extent duplication (i.e., speciation) (median 4.6%, range 1.4%-24%) and extinction events (median 2.9%, range 0%-11%) were detected across plant virus families (Figure 4.12). Co-divergence was most frequently predicted in the *Benyviridae* and *Tymoviridae*, representing 29% and 26% of events, respectively. Importantly, however, the results of our co-phylogenetic analysis are undoubtedly influenced by the sample of plant viruses and will likely change as the number of plant viruses identified increases.

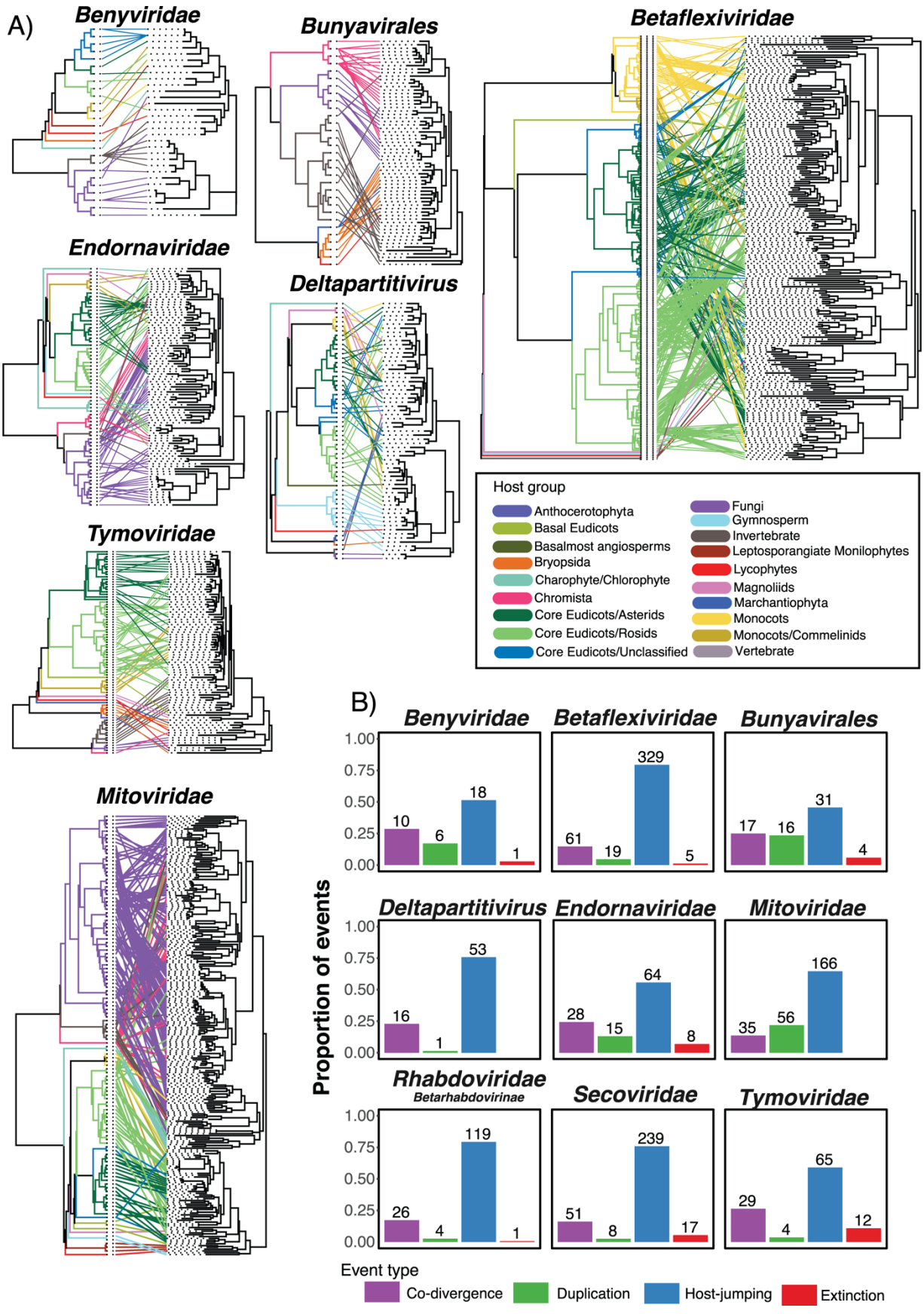

A)
Benyviridae
Bunyavirales
Betaflexiviridae
Endornaviridae
Deltapartitivirus
Tymoviridae
Mitoviridae
Host group
Anthocerotophyta
Basal Eudicots
Basalmost angiosperms
Bryopsida
Charophyte/Chlorophyte
Chromista
Core Eudicots/Asterids
Core Eudicots/Rosids
Core Eudicots/Unclassified
Fungi
Gymnosperm
Invertebrate
Leptosporangiate Monilophytes
Lycophytes
Magnoliids
Marchantiophyta
Monocots
Monocots/Commelinids
Vertebrate
B)
Benyviridae
Betaflexiviridae
Bunyavirales
Deltapartitivirus
Endornaviridae
Mitoviridae
Rhabdoviridae
Betarhabdovirinae
Secoviridae
Tymoviridae
Proportion of events
1.00
0.75
0.50
0.25
0.00
10
6
18
1
61
19
329
5
17
16
31
4
16
1
53
28
15
64
8
35
56
166
26
4
119
1
51
8
239
17
29
4
65
12
Event type
Co-divergence
Duplication
Host-jumping
Extinction

***Figure 4.31 Tanglegram and reconciliation analysis of select plant virus.*** *(a) Tanglegram of rooted phylogenetic trees for select virus groups and their hosts. Lines and branches are coloured to represent host clade. The cophylo function implemented in phytools (v0.7-80) was used to maximise the congruence between the host (left) and virus (right) phylogenies. Supplementary Figure 4.5 provides the names of the hosts and viruses along with additional tanglegrams for the Secoviridae and Rhabdoviridae. (b). Reconciliation analysis of select virus groups. Barplots illustrate the range of the proportion of possible events and are coloured by event type.*

## 4.5 Discussion

Our understanding of the interactions between plants and their viruses has been constrained by inadequate sampling across the enormous, extant diversity of plant species. Here, we provide a large-scale virus discovery project based on mining transcriptomes from across the entire breadth of the plant kingdom. In doing so we have identified 104 potentially novel virus species, considerably expanding the known host range of 13 virus families to include lower plants. In addition, we expand the host range of four other virus families to include algae (including two viruses in brown algae that fall outside the Archaeplastida lineage). We also present the first evidence of a movement protein with a predicted molecular weight of ~30 kDa (herein referred to as a “30K MP”) in a virus of non-vascular plants. Collectively, these results advance our understanding of RNA virus diversity across the Archaeplastida. It is important to note that the transcriptomes analysed were prepared with enrichment of polyA+ RNA limiting the detection of non-polyadenylated viruses. Such biases do not detract from the findings of this study but do highlight the need for further virus discovery efforts in characterising the phytovirosphere.

To date, viral surveys in basal plant lineages (namely ferns, bryophytes and algae) have revealed only the minimal occurrence of (+)RNA viruses (Charon et al. 2020; Charon et al. 2021; Dolja et al. 2020; Mushegian et al. 2016; Rousvoal et al. 2016; Vlok et al. 2019), supporting the idea that RNA viromes in angiosperms evolved as they diversified during the Cretaceous (Friis et al. 2010; Kenrick and Crane 1997). However, our results potentially challenge this paradigm as we detected the first evidence of (+)ssRNA viruses in lower plants and algae, implying that these groups are associated with older lineages of plants. Several of these virus lineages are

deep branching and sit basal to angiosperm infecting viruses (e.g., LycoAV and Lycopod associated bunyavirus) in phylogenetic trees. Other viruses discovered here occupy ambiguous positions between established plant virus families (e.g., OnV) or cluster in large numbers to form novel plant-associated clades (e.g., the *Viridisbunyaviridae* in the *Bunyavirales*). Benyviruses are typically transmitted by the root-infecting plasmodiophorids *Polymyxa betae* and *Polymyxa graminis* (Tamada et al. 1996; Valente et al. 2019). The Phytomyxids (plasmodiophorids and phagomyxids) are parasites of plants, diatoms, oomycetes and brown algae and have been shown to exhibit cross-kingdom host shifts (e.g., between angiosperms and oomycetes) (Neuhauser et al. 2014). As such, the plasmodiophorids may be a vehicle for cross-species transmission between aquatic protists and land plants (Dolja et al. 2020). FeBV, a beny-like virus identified in this study, formed a clade along with *Wheat stripe mosaic virus* distinct from members of the genus *Benyvirus*. Deciphering the evolutionary history and mode of transmission for the lower plant beny-like viruses will require further studies with particular emphasis on these taxa. Interestingly, no plasmodiophorid-associated reads were detected in any of the libraries from which we assembled a beny-like virus. LjBV and WasBV appear distantly related to the benyviruses. These viruses group with a suite of unclassified viruses assembled from a soil metatranscriptome study suggesting that, like the benyviruses, this larger group of unclassified viruses may involve soil-borne parasites like the plasmodiophorids (Starr et al. 2019). Our detection of tymovirid-like sequences in the lycophytes, bryophytes and brown algae dramatically expands the known host range of the *Tymovirales*. Several of these viruses were similar to unclassified Riboviria species assembled from a recent survey of common wild oat soil rhizosphere and detritosphere (Starr et al. 2019) (Figure 4.4). The metatranscriptome of the sequenced soil samples from the common wild oat study was largely composed of Viridiplantae, fungi, Amoebozoa, protists, nematodes, and other eukaryotes. Phylogenetic clusters to infer host associations of our viruses therefore remains challenging. Indeed, these viruses may result from contamination from other eukaryotes (e.g., fungi or invertebrates) although we found no consistent evidence among these viruses (Figure 4.2). Assuming these viruses are plant-associated, their phylogenetic pattern suggests that they may have resulted from cross-kingdom transmission events that frequent the Alsuviricetes.

The partial deltaflexi-like virus we detected in *P. agnata* (PaADV) is particularly noteworthy. The deltaflexiviruses are only known to infect fungi, although no fungi associated reads were found in the *P. agnata* metatranscriptome (Figure 4.2). The mycovirus families *Delta*- and

*Gammaflexiviridae* are thought to have been derived from the plant alpha- and betaflexivirids through cross-species transmission (Dolja et al. 2020; Ghabrial et al. 2015)). As such PaAGV could potentially represent an intermediate between the plant and fungi flexiviruses or perhaps a more recent fungus to plant transmission. As only a fragment of the polymerase gene was assembled for this virus future work should confirm the presence of PaAGV and its phylogenetic position.

Through the analysis of mitoviruses-like, non-retroviral endogenous RNA viral elements (NERVEs), it was argued that the origin of plant mitovirus NERVEs was a single horizontal transfer from a fungal mitovirus before the origin of vascular plants in the early Silurian, ~400 MYA (Bruenn et al. 2015). Evidence of contemporary mitoviruses in flowering plants and a fern have challenged this view, suggesting that a lineage of plant rather than fungal mitoviruses are the immediate ancestors of plant mitovirus NERVEs (Nibert et al. 2018). Indeed, plant-to-fungus transmission would eliminate code conflicts between fungi and plant mitochondrial genetic codes (Shackelton and Holmes 2008). Herein, we demonstrate the existence of a lower plant-associated sister clade to the angiosperm mitoviruses and NERVEs. This clade includes a clubmoss associated mitovirus, the most basal plant mitovirus sequence to date. This finding aligns with the estimation of the origin of plant mitovirus NERVEs occurring as early as the evolution of the clubmoss (90). The recent finding of mitoviruses in green algae – including BopiMV in this study – highlight the broad host range of mitoviruses (Charon et al. 2020; Charon et al. 2021). The phylogenetic position of these viruses and the absence of NERVEs from these groups suggest that they are not the ancestors of land plant mitoviruses and NERVEs.

We identified 16 bunya-like viruses assembled from six non-vascular plant libraries including bryophyte, and lycophyte species. These viruses form a novel clade within the *Bunyavirales* and represent the first viruses in this order to be associated with lower plants. This clade likely represents a novel virus family which we have tentatively named the *Viridisbunyaviridae.* Several libraries contained up to five distinct viruses (each sharing <70% nucleotide identity). Virus co-infections are frequently observed in plants and have been reported in the closest relatives of these viruses, the *Deltamycobunyaviridae* (Botella and Jung 2021; Moreno Goncalves and Lopez-Moya 2019). As with previous studies we were only able to recover the bunyavirus L segment (Botella et al. 2020; Botella and Jung 2021). Further studies are needed

to recover the missing small and medium-sized segments and to confirm the presence of mixed infections in plants.

Through the discovery of lower plant-associated viruses, we have gained insights into how the genome structure and composition of contemporary flowering plants viruses have evolved. The detection of secovirid-like sequences in bryophytes and ferns represents the first occurrence of plant secoviruses outside of angiosperms and the first evidence of a 30K MP homolog in non-vascular plants. These proteins aid the cell-to-cell movement of viruses in plants. For example, the MP of *Cucumber mosaic virus* increases the size exclusion limit of plasmodesmata allowing virus particles to pass through cell walls (Su et al. 2010). To date, homologs of 30K MP have only been detected in plant viruses infecting angiosperms to the lycophytes (Mushegian and Elena 2015; Mushegian et al. 2016). Further work is needed to confirm the presence and function of 30K MPs in viruses infecting the bryophytes and other lower plants.

Our discovery of six tri-segmented deltapartitivirus species provides insights into the evolution of the deltapartitivirus dsRNA3 segment. dsRNA3 segments have been found in several alpha- and deltapartitiviruses infecting flowering plants (Chen et al. 2006; Kumar et al. 2017; Sabanadzovic and Ghanem-Sabanadzovic 2008; Wu et al. 2020). These segments typically encode seemingly full-length capsid protein or in the case of alphapartitivirus *Rosellinia necatrix partitivirus 2*, a truncated version of the RdRp which may serve as an interfering RNA (Chiba et al. 2013). There is some debate as to the source of dsRNA3 segments, particularly whether they are satellite viruses that co-opt the RdRp of the co-infecting helper viruses or that the additional segment is a result of coinfection of two different plant partitiviruses and the second RdRp-encoding segment is lost after the initial infection (Nibert et al. 2014). For the first time, we find dsRNA3 segments in conifer associated viruses but not in those found in lower plants including bryophytes and lycophytes. The absence of dsRNA3 in non-vascular plants means that it is possible that this segment evolved after the divergence of vascular and non-vascular plants in the Silurian period (Harrison and Morris 2018). It is possible that dsRNA3 segments exist for the non-vascular plant infecting deltapartitiviruses but was not detected due to the large degree of divergence between this segment and reference sequences (including those found in this study). However, the dsRNA1 and dsRNA2 segments of the putative lower plant deltapartitiviruses shared >50% aa identity with the tri-segmented deltapartitiviruses -

well above the detection limit for tools such as Diamond BLASTx (Buchfink et al. 2015). dsRNA3 segments typically appear no more divergent than dsRNA2 segments so it is unlikely that we would detect both the dsRNA1 and dsRNA2 segments without detecting dsRNA3. Further work is needed to confirm the presence of deltapartitivirus dsRNA3 segments.

Finally, the recently discovered gymnosperm varicose-like *Pinus flexilis virus 1* in the family *Rhabdoviridae* contains an unsegmented genome organisation that differs from the typical bi-segmented structure of the varicosaviruses (Bejerman et al. 2021; Walker et al. 2018). We find the bi-segmented structure in varicosavirus-like viruses for the first time in ferns and species in the Marchantiophyta (TfVV and MgVV) which predate the gymnosperms.

The data generated under the 1KP were not explicitly created for virus discovery, such that there are important caveats associated with the methods and metatranscriptomic data mined for virus contigs. For instance, as axenic cultures are not a viable option in most instances, the 1KP samples are commonly contaminated by nucleic acids belonging to bacterial, fungal, and insect species. We addressed this by using a combination of host/virus abundance measurements and phylogenetic analyses to improve the accuracy of virus-host assignments. For most of the viruses described, phylogenetic placement within plant infecting virus families strongly supports their association with plants. However, several of the viruses found in algae and lower plants were associated with lineages known to infect invertebrates and fungi or unclassified viruses recovered from environmental samples. The association between the viruses of lower plants and algae with that of fungi and invertebrate viruses may reflect the absence of algal and lower plant viruses in reference sequence databases. Experimental confirmation is needed to formally assign the viruses discovered in this study to their hosts.

The average sequencing depth of the 1KP libraries was 1.99 Gb of sequence per sample (range 1.3-3.0), lower than many other virus discovery studies (Hao et al. 2018; Shates et al. 2019; Shi et al. 2016). Sequencing depth has been shown to correlate with the ability to detect viruses present at low abundance (Maclot et al. 2020; Visser et al. 2016). Further, a large proportion of the virus transcripts detected were from viruses whose full-length genomic or subgenomic mRNAs were polyadenylated at the 3′ end (Supplementary Table 4, Figure 4.1). Although this was anticipated (i.e., the libraries generated by the 1KP initiative were prepared

from polyA+ RNA), it limited the detection of non-polyadenylated viruses and may have contributed to the lack of phycodnavirus sequences detected in algae (Visser et al. 2016). To reduce the computational burden of assembly, we attempted to remove host-associated reads before contig assembly by mapping them to the host scaffolds provided by the 1KP initiative. While this step reduces the occurrence of false-positive virus detection it also risks removing virus reads, particularly reverse-transcribing plant viruses (Llorens et al. 2009). While we frequently detected transcripts associated with the reverse-transcribing family *Caulimoviridae,* no members of the *Metaviridae* or *Pseudoviridae* were detected.

### 4.6 Data availability

The raw One Thousand Plant Transcriptomes Initiative sequence reads are available at BioProject PRJEB21674. All viral genomes and corresponding sequences assembled in this study have been deposited in the European Nucleotide Archive at EMBL-EBI under accession number PRJEB52092 (https://www.ebi.ac.uk/ena/browser/view/PRJEB52092) (Supplementary Table 4).

### 4.7 Acknowledgements

We thank Richard Miller for computational support and Justine Charon for advice on alignments. This work would not have been possible without the data generously provided by the One Thousand Plant Transcriptomes Initiative. J.C.O.M. was supported by the William Macleay Microbiological Research Fund from the Linnean Society of NSW. R.V.G. was funded by a Discovery Early Career Researcher Award (DECRA) (DE170100208) and E.C.H. was funded by an ARC Australian Laureate Fellowship (FL170100022) from the Australian Research Council. J.L.G. was funded by a New Zealand Royal Society Rutherford Discovery Fellowship (RDF-20-UOO-007).

### 4.8 References


Altschul, S. F., et al. (1990), 'Basic local alignment search tool', *Journal of Molecular Biology,* 215 (3), 403-10.

Anderson, J. T. (2016), 'Plant fitness in a rapidly changing world', *New Phytologist,* 210 (1), 81-87.

Bejerman, N., Dietzgen, R. G., and Debat, H. (2021), 'Illuminating the Plant Rhabdovirus Landscape through Metatranscriptomics Data', *Viruses,* 13 (7), 1304.

Botella, L. and Jung, T. (2021), 'Multiple Viral Infections Detected in Phytophthora condilina by Total and Small RNA Sequencing', *Viruses,* 13 (4), 620.

Botella, L., et al. (2020), 'Marine Oomycetes of the Genus Halophytophthora Harbor Viruses Related to Bunyaviruses', *Frontiers in Microbiology,* 11 (1467).

Brister, J. R., et al. (2015), 'NCBI viral genomes resource', *Nucleic Acids Research,* 43, 571-7.

Bruenn, J. A., Warner, B. E., and Yerramsetty, P. (2015), 'Widespread mitovirus sequences in plant genomes', *PeerJ,* 3.

Brunkard, J. O. and Zambryski, P. C. (2017), 'Plasmodesmata enable multicellularity: new insights into their evolution, biogenesis, and functions in development and immunity', *Current Opinion in Plant Biology,* 35, 76-83.

Buchfink, B., Xie, C., and Huson, D. H. (2015), 'Fast and sensitive protein alignment using DIAMOND', *Nature Methods,* 12 (1), 59-60.

Bushnell, B. (2014), 'BBMap: a fast, accurate, splice-aware aligner', (Lawrence Berkeley National Lab (LBNL), Berkeley, CA.).

Cai, G., et al. (2009), 'A novel virus of the late blight pathogen, Phytophthora infestans, with two RNA segments and a supergroup 1 RNA-dependent RNA polymerase', *Virology,* 392 (1), 52-61.

Capella-Gutiérrez, S., Silla-Martínez, J. M., and Gabaldón, T. (2009), 'trimAl: a tool for automated alignment trimming in large-scale phylogenetic analyses', *Bioinformatics,* 25 (15), 1972-73.

Carpenter, E. J., et al. (2019), 'Access to RNA-sequencing data from 1,173 plant species: The 1000 Plant transcriptomes initiative (1KP)', *GigaScience,* 8 (10).

Charon, J., Murray, S., and Holmes, E. C. (2021), 'Revealing RNA virus diversity and evolution in unicellular algae transcriptomes', *Virus Evolution,* 7 (2).

Charon, J., et al. (2020), 'Metatranscriptomic identification of diverse and divergent RNA viruses in green and Chlorarachniophyte algae cultures', *Viruses,* 12 (10).

Chen, L., et al. (2006), 'Complete nucleotide sequences and genome characterization of double-stranded RNA 1 and RNA 2 in the Raphanus sativus-root cv. Yipinghong', *Archives of Virology,* 151 (5), 849-59.

Chiba, S., et al. (2013), 'Effects of Defective Interfering RNA on Symptom Induction by, and Replication of, a Novel Partitivirus from a Phytopathogenic Fungus, Rosellinia necatrix', *Journal of Virology,* 87 (4), 2330-41.

Christenhusz, M. J. and Byng, J. W. (2016), 'The number of known plants species in the world and its annual increase', *Phytotaxa,* 261 (3), 201-17.

Clausen, P. T., Aarestrup, F. M., and Lund, O. (2018), 'Rapid and precise alignment of raw reads against redundant databases with KMA', *BMC Bioinformatics,* 19 (1), 1-8.

Conow, C., et al. (2010), 'Jane: a new tool for the cophylogeny reconstruction problem', *Algorithms for Molecular Biology,* 5 (1), 1-10.

Covelli, L., et al. (2004), 'Cherry chlorotic rusty spot and Amasya cherry diseases are associated with a complex pattern of mycoviral-like double-stranded RNAs. I. Characterization of a new species in the genus Chrysovirus', *Journal of General Virology,* 85 (11), 3389-97.

Dawes, C. (2016), 'Chapter 4 - Macroalgae Systematics', in Joël Fleurence and Ira Levine (eds.), *Seaweed in Health and Disease Prevention* (San Diego: Academic Press), 107-48.

Ding, S., et al. (1990), 'The tymobox, a sequence shared by most tymoviruses: its use in molecular studies of tymoviruses', *Nucleic Acids Research,* 18 (5), 1181-87.

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

Dolja, V. V., Krupovic, M., and Koonin, E. V. (2020), 'Deep roots and splendid boughs of the global plant virome', *Annual Review of Phytopathology,* 58, 23-53.

Friis, E. M., Pedersen, K. R., and Crane, P. R. (2010), 'Diversity in obscurity: fossil flowers and the early history of angiosperms', *Philosophical Transactions of the Royal Society B: Biological Sciences,* 365 (1539), 369-82.

Gertz, E. M., et al. (2006), 'Composition-based statistics and translated nucleotide searches: improving the TBLASTN module of BLAST', *BMC Biology,* 4 (1), 1-14.

Ghabrial, S. A., et al. (2015), '50-plus years of fungal viruses', *Virology,* 479-480, 356-68.

Gibbs, A. J., et al. (2011), 'The enigmatic genome of Chara australis virus', *Journal of General Virology,* 92 (11), 2679-90.

Gilbert, K. B., et al. (2019), 'Hiding in plain sight: New virus genomes discovered via a systematic analysis of fungal public transcriptomes', *PLOS One,* 14 (7), e0219207.

Greninger, A. L. (2018), 'A decade of RNA virus metagenomics is (not) enough', *Virus Research,* 244, 218-29.

Grimwood, R. M., Holmes, E. C., and Geoghegan, J. L. (2021), 'A Novel Rubi-Like Virus in the Pacific Electric Ray (Tetronarce californica) Reveals the Complex Evolutionary History of the Matonaviridae', *Viruses,* 13 (4).

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 (8), 1494-512.

Hammond, R. and Ramirez, P. (2001), 'Molecular characterization of the genome of Maize rayado fino virus, the type member of the genus Marafivirus', *Virology,* 282 (2), 338-47.

Han, G.-Z. (2019), 'Origin and evolution of the plant immune system', *New Phytologist,* 222 (1), 70-83.

Han, S., et al. (2002), 'Nucleotide sequence and taxonomy of Cycas necrotic stunt virus', *Archives of Virology,* 147 (11), 2207-14.

Hao, X., et al. (2018), 'Discovery of plant viruses from tea plant (Camellia sinensis (L.) O. Kuntze) by metagenomic sequencing', *Frontiers in Microbiology,* 9, 2175.

Harrison, C. J. and Morris, J. L. (2018), 'The origin and early evolution of vascular plant shoots and leaves', *Philosophical Transactions of the Royal Society B: Biological Sciences,* 373 (1739), 20160496.

Hatcher, E. L., et al. (2017), 'Virus Variation Resource–improved response to emergent viral outbreaks', *Nucleic Acids Research,* 45 (D1), D482-D90.

Heinze, C. (2012), 'A novel mycovirus from Clitocybe odora', *Archives of Virology,* 157 (9), 1831-34.

Jin, Y. and Qian, H. (2019), 'V.PhyloMaker: an R package that can generate very large phylogenies for vascular plants', *Ecography,* 42 (8), 1353-59.

Johnson, M. T., et al. (2012), 'Evaluating methods for isolating total RNA and predicting the success of sequencing phylogenetically diverse plant transcriptomes', *PLOS One,* 7 (11), e50226.

Kalyaanamoorthy, S., et al. (2017), 'ModelFinder: fast model selection for accurate phylogenetic estimates', *Nature Methods,* 14 (6), 587-89.

Kearse, M., et al. (2012), 'Geneious Basic: an integrated and extendable desktop software platform for the organization and analysis of sequence data', *Bioinformatics,* 28 (12), 1647-49.

Kenrick, P. and Crane, P. R. (1997), 'The origin and early evolution of plants on land', *Nature,* 389 (6646), 33-39.

Koga, R., Horiuchi, H., and Fukuhara, T. (2003), 'Double-stranded RNA replicons associated with chloroplasts of a green alga, Bryopsis cinicola', *Plant Molecular Biology,* 51 (6), 991-99.

Kumar, S., et al. (2017), 'Molecular characterization of a novel cryptic virus infecting pigeonpea plants', *PLOS One,* 12 (8), e0181829.

Langmead, B. and Salzberg, S. L. (2012), 'Fast gapped-read alignment with Bowtie 2', *Nature Methods,* 9 (4), 357-59.

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 & Microbe,* 22 (3), 387-99.e6.

Lay, C. L. (2021), 'biolumber/littlegenomes: First release'.

Leebens-Mack, J. H., et al. (2019), 'One thousand plant transcriptomes and the phylogenomics of green plants', *Nature,* 574 (7780), 679-85.

Leinonen, R., et al. (2010), 'The sequence read archive', *Nucleic Acids Research,* 39, 19-21.

Letunic, I. and Bork, P. (2019), 'Interactive Tree Of Life (iTOL) v4: recent updates and new developments', *Nucleic Acids Research,* 47 (W1), W256-W59.

Li, B. and Dewey, C. N. (2011), 'RSEM: accurate transcript quantification from RNA-Seq data with or without a reference genome', *BMC Bioinformatics,* 12 (1), 323.

Llorens, C., et al. (2009), 'Network dynamics of eukaryotic LTR retroelements beyond phylogenetic trees', *Biology Direct,* 4 (1), 41.

Maclot, F., et al. (2020), 'Illuminating an ecological blackbox: using high throughput Sequencing to characterize the plant virome across scales', *Frontiers in Microbiology,* 11, 2575.

Marcelino, V. R., et al. (2020), 'CCMetagen: comprehensive and accurate identification of eukaryotes and prokaryotes in metagenomic data', *Genome Biology,* 21 (1), 103.

Mifsud, J. C. O. (2020), 'Explorations of the plant virosphere', Masters dissertation.

Mihara, T., et al. (2016), 'Linking Virus Genomes with Host Taxonomy', *Viruses,* 8 (3), 66.

Miller, A. K., et al. (2021), 'Slippery when wet: cross-species transmission of divergent coronaviruses in bony and jawless fish and the evolutionary history of the Coronaviridae', *Virus Evolution,* 7 (2).

Mirdita, M., Steinegger, M., and Söding, J. (2019), 'MMseqs2 desktop and local web server app for fast, interactive sequence searches', *Bioinformatics,* 35 (16), 2856-58.

Moreno Goncalves, A. B. and Lopez-Moya, J. J. (2019), 'When viruses play team sports: mixed infections in plants', *Phytopathology,* 110 (1), 29-48.

Morozov, S. Y. and Solovyev, A. G. (2003), 'Triple gene block: modular design of a multifunctional machine for plant virus movement', *Journal of General Virology,* 84 (6), 1351-66.

Mushegian, A., Shipunov, A., and Elena, S. F. (2016), 'Changes in the composition of the RNA virome mark evolutionary transitions in green plants', *BMC Biology,* 14 (1), 68.

Mushegian, A. R. and Elena, S. F. (2015), 'Evolution of plant virus movement proteins from the 30K superfamily and of their homologs integrated in plant genomes', *Virology,* 476, 304-15.

Nayfach, S., et al. (2021), 'CheckV assesses the quality and completeness of metagenome-assembled viral genomes', *Nature Biotechnology,* 39 (5), 578-85.

Nerva, L., et al. (2019), 'Isolation, molecular characterization and virome analysis of culturable wood fungal endophytes in esca symptomatic and asymptomatic grapevine plants', *Environmental Microbiology,* 21 (8), 2886-904.

Neuhauser, S., et al. (2014), 'Cross-kingdom host shifts of phytomyxid parasites', *BMC Evolutionary Biology,* 14 (1), 33.

Nguyen, L.-T., et al. (2015), 'IQ-TREE: a fast and effective stochastic algorithm for estimating maximum-likelihood phylogenies', *Molecular Biology and Evolution,* 32 (1), 268-74.

Nibert, M. L. (2017), 'Mitovirus UGA (Trp) codon usage parallels that of host mitochondria', *Virology,* 507, 96-100.

Nibert, M. L., Pyle, J. D., and Firth, A. E. (2016), 'A +1 ribosomal frameshifting motif prevalent among plant amalgaviruses', *Virology,* 498, 201-08.

Nibert, M. L., et al. (2018), 'Evidence for contemporary plant mitoviruses', *Virology,* 518, 14-24.

Nibert, M. L., et al. (2014), 'Taxonomic reorganization of family Partitiviridae and other recent progress in partitivirus research', *Virus Research,* 188, 128-41.

Ondov, B. D., Bergman, N. H., and Phillippy, A. M. (2011), 'Interactive metagenomic visualization in a Web browser', *BMC Bioinformatics,* 12 (1), 385.

Paradis, E. and Schliep, K. (2019), 'ape 5.0: an environment for modern phylogenetics and evolutionary analyses in R', *Bioinformatics,* 35 (3), 526-28.

Parry, R., et al. (2020), 'Divergent Influenza-Like Viruses of Amphibians and Fish Support an Ancient Evolutionary Association', *Viruses,* 12 (9), 1042.

Quast, C., et al. (2012), 'The SILVA ribosomal RNA gene database project: improved data processing and web-based tools', *Nucleic Acids Research,* 41 (D1), D590-D96.

Rambaut, A. and Drummond, A. (2012), 'FigTree: Tree figure drawing tool, version 1.4.0', (Institute of Evolutionary Biology, University of Edinburgh).

Revell, L. J. (2012), 'phytools: an R package for phylogenetic comparative biology (and other things)', *Methods in Ecology and Evolution,* 3 (2), 217-23.

Roossinck, M. J., Martin, D. P., and Roumagnac, P. (2015), 'Plant virus metagenomics: advances in virus discovery', *Phytopathology,* 105 (6), 716-27.

Rousvoal, S., et al. (2016), 'Mutant swarms of a totivirus-like entities are present in the red macroalga Chondrus crispus and have been partially transferred to the nuclear genome', *Journal of Phycology,* 52 (4), 493-504.

Sabanadzovic, S. and Ghanem-Sabanadzovic, N. A. (2008), 'Molecular characterization and detection of a tripartite cryptic virus from rose', *Journal of Plant Pathology*, 287-93.

Santichaivekin, S., et al. (2020), 'eMPRess: a systematic cophylogeny reconciliation tool', *Bioinformatics,* 16 (37), 2481–82.

Sayers, E. W., et al. (2021), 'GenBank', *Nucleic Acids Research,* 49 (D1), D92-D96.

Shackelton, L. A. and Holmes, E. C. (2008), 'The role of alternative genetic codes in viral evolution and emergence', *Journal of Theoretical Biology,* 254 (1), 128-34.

Shates, T. M., et al. (2019), 'Addressing Research Needs in the Field of Plant Virus Ecology by Defining Knowledge Gaps and Developing Wild Dicot Study Systems', *Frontiers in Microbiology,* 9.

Shi, M., et al. (2016), 'Redefining the invertebrate RNA virosphere', *Nature,* 540 (7634), 539.

Short, S. M., et al. (2020), 'Diversity of Viruses Infecting Eukaryotic Algae', *Current Issues in Molecular Biology,* 39, 29-61.

Sidharthan, V. K., Kalaivanan, N. S., and Baranwal, V. K. (2021), 'Discovery of putative novel viruses in the transcriptomes of endangered plant species native to India and China', *Gene,* 786, 145626.

Sievers, F., et al. (2011), 'Fast, scalable generation of high-quality protein multiple sequence alignments using Clustal Omega', *Molecular Systems Biology,* 7 (1), 539.

Smith, S. A. and Brown, J. W. (2018), 'Constructing a broadly inclusive seed plant phylogeny', *American Journal of Botany,* 105 (3), 302-14.

Starr, E. P., et al. (2019), 'Metatranscriptomic reconstruction reveals RNA viruses with the potential to shape carbon cycling in soil', *Proceedings of the National Academy of Sciences,* 116 (51), 25900.

Su, S., et al. (2010), 'Cucumber Mosaic Virus Movement Protein Severs Actin Filaments to Increase the Plasmodesmal Size Exclusion Limit in Tobacco ', *The Plant Cell,* 22 (4), 1373-87.

Tamada, T., et al. (1996), 'High resolution analysis of the readthrough domain of beet necrotic yellow vein virus readthrough protein: a KTER motif is important for efficient transmission of the virus by Polymyxa betae', *Journal of General Virology,* 77 (7), 1359-67.

Valente, J. B., et al. (2019), 'A novel putative member of the family Benyviridae is associated with soilborne wheat mosaic disease in Brazil', *Plant Pathology,* 68 (3), 588-600.

Valverde, R. A. and Sabanadzovic, S. (2009), 'A novel plant virus with unique properties infecting Japanese holly fern', *Journal of General Virology,* 90 (10), 2542-49.

Veliceasa, D., et al. (2006), 'Searching for a new putative cryptic virus in Pinus sylvestris L', *Virus Genes,* 32 (2), 177-86.

Visser, M., et al. (2016), 'Next-generation sequencing for virus detection: covering all the bases', *Virology Journal,* 13 (1), 85.

Vlok, M., Gibbs, A. J., and Suttle, C. A. (2019), 'Metagenomes of a freshwater charavirus from British Columbia provide a window into ancient lineages of viruses', *Viruses,* 11 (3).

Walker, P. J., et al. (2018), 'ICTV virus taxonomy profile: Rhabdoviridae', *Journal of General Virology,* 99 (4), 447-48.

Wickham, H., et al. (2019), 'Welcome to the Tidyverse', *Journal of Open Source Software,* 4 (43), 1686.

Wren, J. D., et al. (2006), 'Plant virus biodiversity and ecology', *PLoS Biology,* 4 (3), 80.

Wu, L. P., et al. (2020), 'Complete genomic sequence of tea-oil camellia deltapartitivirus 1, a novel virus from Camellia oleifera', *Archives of Virology,* 165 (1), 227-31.

Xie, J. and Ghabrial, S. A. (2012), 'Molecular characterizations of two mitoviruses co-infecting a hyovirulent isolate of the plant pathogenic fungus Sclerotinia sclerotiorum', *Virology,* 428 (2), 77-85.

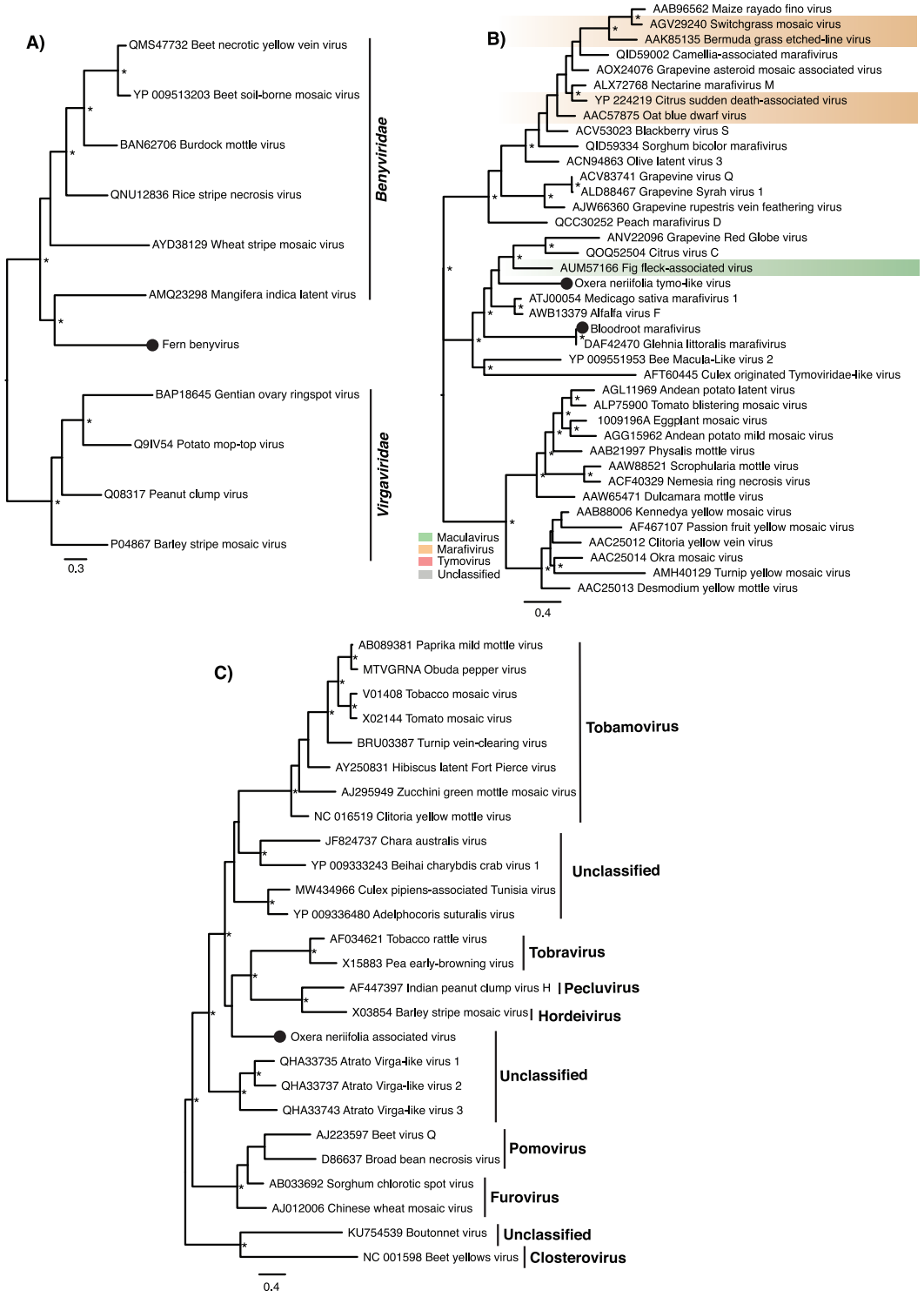
Yang, S., et al. (2020), 'Virome of riverside phytocommunity ecosystem of an ancient canal'.

Zanne, A. E., et al. (2014), 'Three keys to the radiation of angiosperms into freezing environments', *Nature,* 506 (7486), 89-92.

Zhang, T., et al. (2015), 'Full genome sequence of a putative novel mitovirus isolated from Rhizoctonia cerealis', *Archives of Virology,* 160 (7), 1815-18.

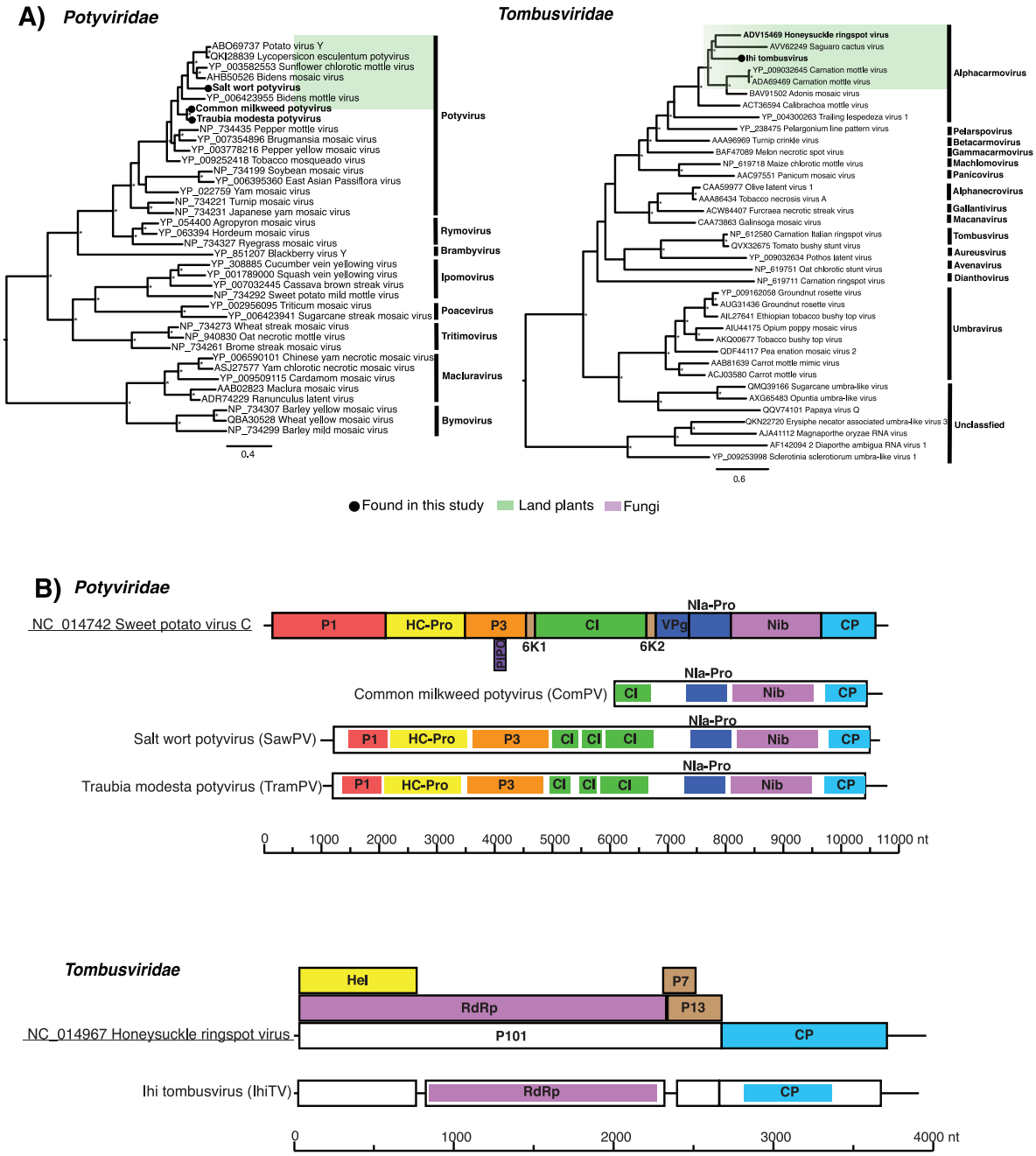
Zimmermann, L., et al. (2018), 'A Completely Reimplemented MPI Bioinformatics Toolkit with a New HHpred Server at its Core', *Journal of Molecular Biology,* 430 (15), 2237-43.

## 4.9 Supplementary Information



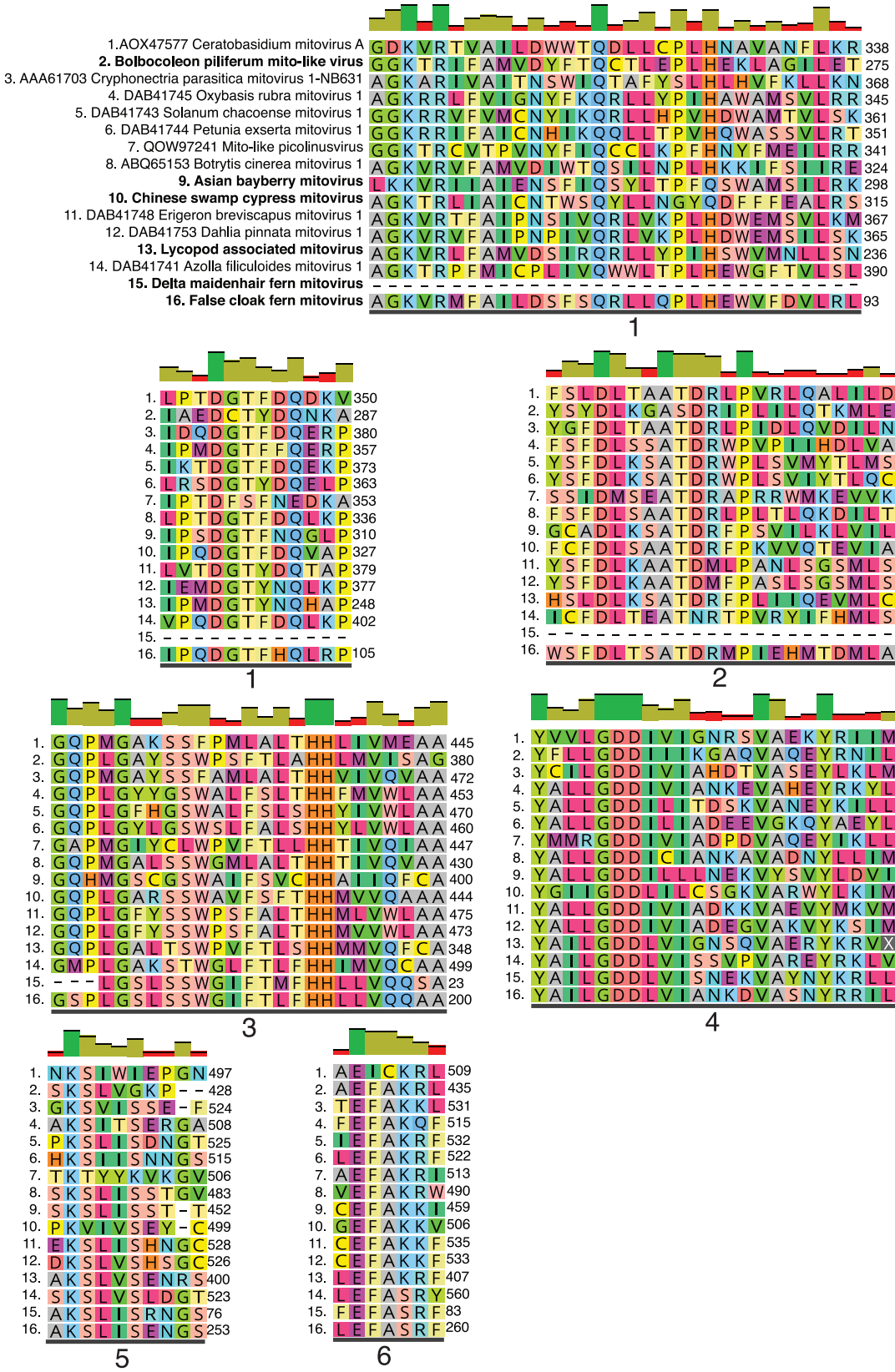

***Supplementary Figure 4.1.*** *(a) Phylogram of the triple gene block (TGB) protein 1. ML phylogenetic trees show the topological position of the newly discovered TGB sequence in the*

*tomato fern (black circle) in the context of the closest relatives. (b) Phylogram of the Tymoviridae virus coat proteins (CP). ML phylogenetic trees show the topological position of the newly discovered CP sequences in (black circle) in the context of the closest relatives. Branches are highlighted to represent virus taxonomy (Maculavirus = green, Marafivirus = orange, Tymovirus = red and unclassified = grey). For each colour, a lighter shade signifies that this virus is related to but has not formally been assigned to this genus. (c) Phylogram of the Oxera neriifolia associated virus coat protein (CP). ML phylogenetic trees show the topological position of the newly discovered CP sequence (black circle) in the context of the closest relatives. For all trees, branches are scaled to the number of amino acid substitutions per site and trees were mid-point rooted for clarity only. Numbers at the nodes indicate bootstrap support over 70% (1000 replicates).*



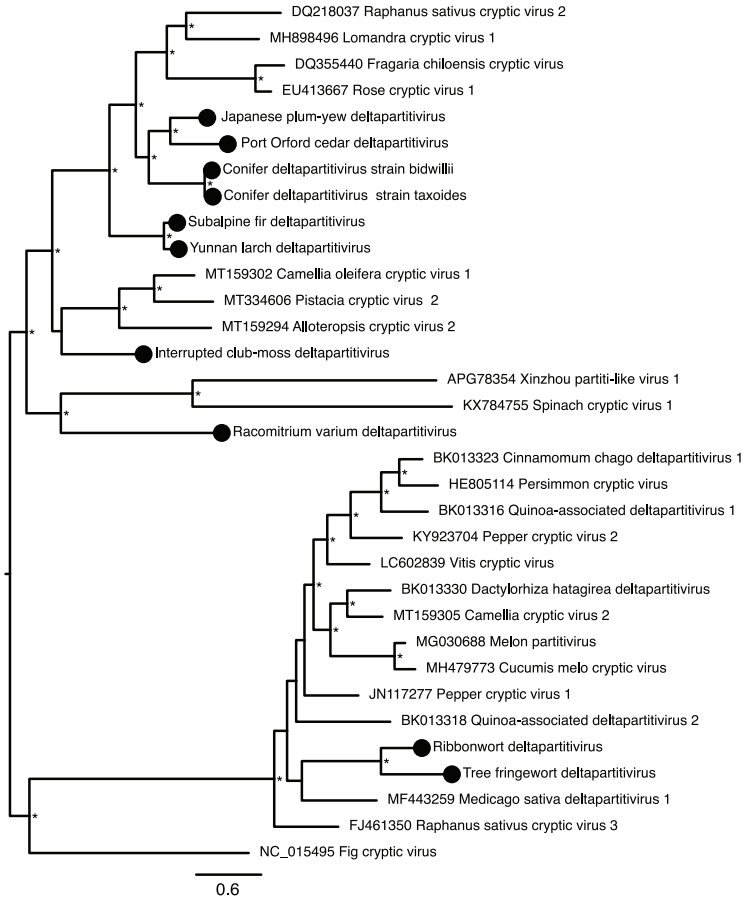

***Supplementary Figure 4.2.*** *(a) Phylogenetic relationships of the viruses identified within the virus families Potyviridae and Tombusviridae. ML phylogenetic trees based upon alignments of the amino acid sequences of the RdRp protein show the topological position of discovered virus-like sequences (black circles) from this study in the context of their closest relatives. See Figure 3 for the colour scheme. 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% bootstrap support. (b) Genomic organization of the Potyviridae and*

*Tombusviridae sequences identified in this study and representative species used in the phylogeny.*

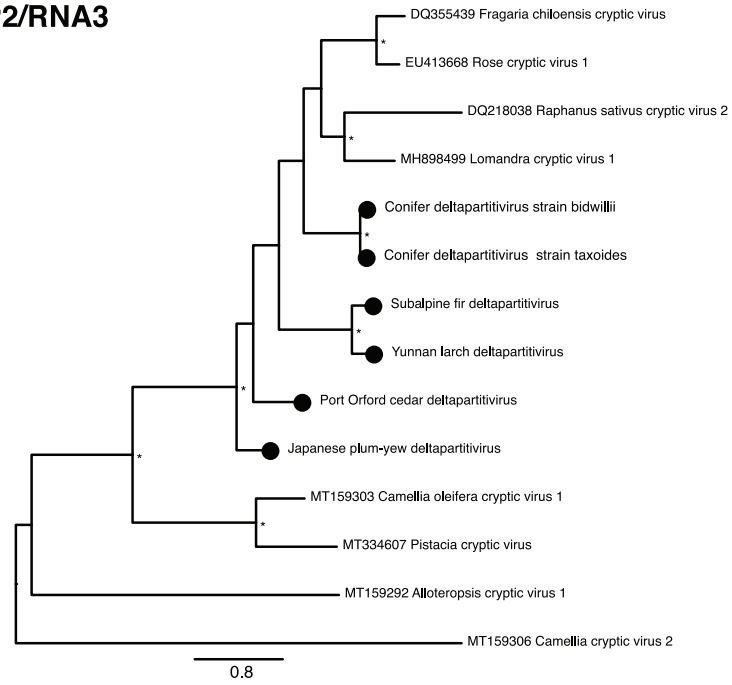

1.AOX47577 Ceratobasidium mitovirus A
2. Bolbocoleon piliferum mito-like virus
3. AAA61703 Cryphonectria parasitica mitovirus 1-NB631
4. DAB41745 Oxybasis rubra mitovirus 1
5. DAB41743 Solanum chacoense mitovirus 1
6. DAB41744 Petunia exserta mitovirus 1
7. QOW97241 Mito-like picolinusvirus
8. ABQ65153 Botrytis cinerea mitovirus 1
9. Asian bayberry mitovirus
10. Chinese swamp cypress mitovirus
11. DAB41748 Erigeron breviscapus mitovirus 1
12. DAB41753 Dahlia pinnata mitovirus 1
13. Lycopod associated mitovirus
14. DAB41741 Azolla filiculoides mitovirus 1
15. Delta maidenhair fern mitovirus
16. False cloak fern mitovirus
1
2
3
4
5
6

***Supplementary Figure 4.3.*** *Multiple sequence alignment conserved amino acid motifs in RNA-dependent RNA polymerase (RdRp) regions of the mitoviruses discovered in this study along with reference mitoviruses. The bar above each residue is green if 100% of residues in that column are identical, green-brown if they are 30%-99%, and red if under 30%. The numbers under each section correspond to regions containing motifs identified in (72).*

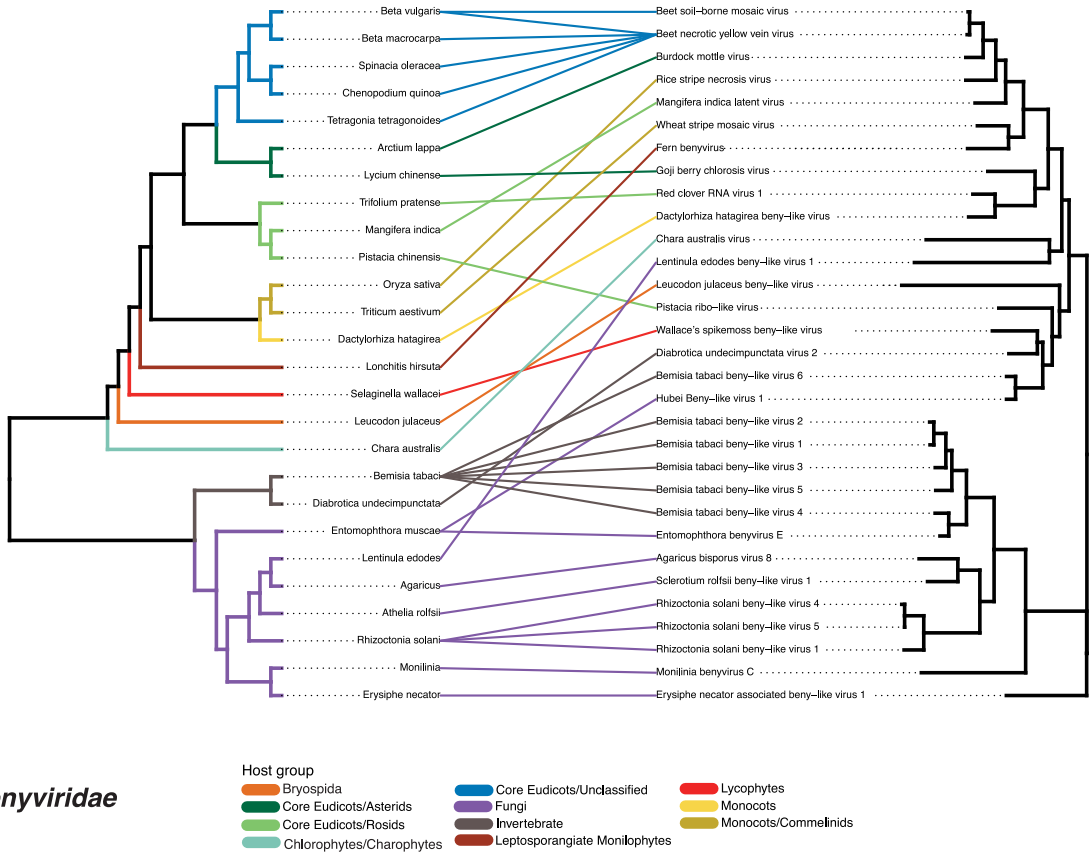
A) **CP1/RNA2**




B) **CP2/RNA3**



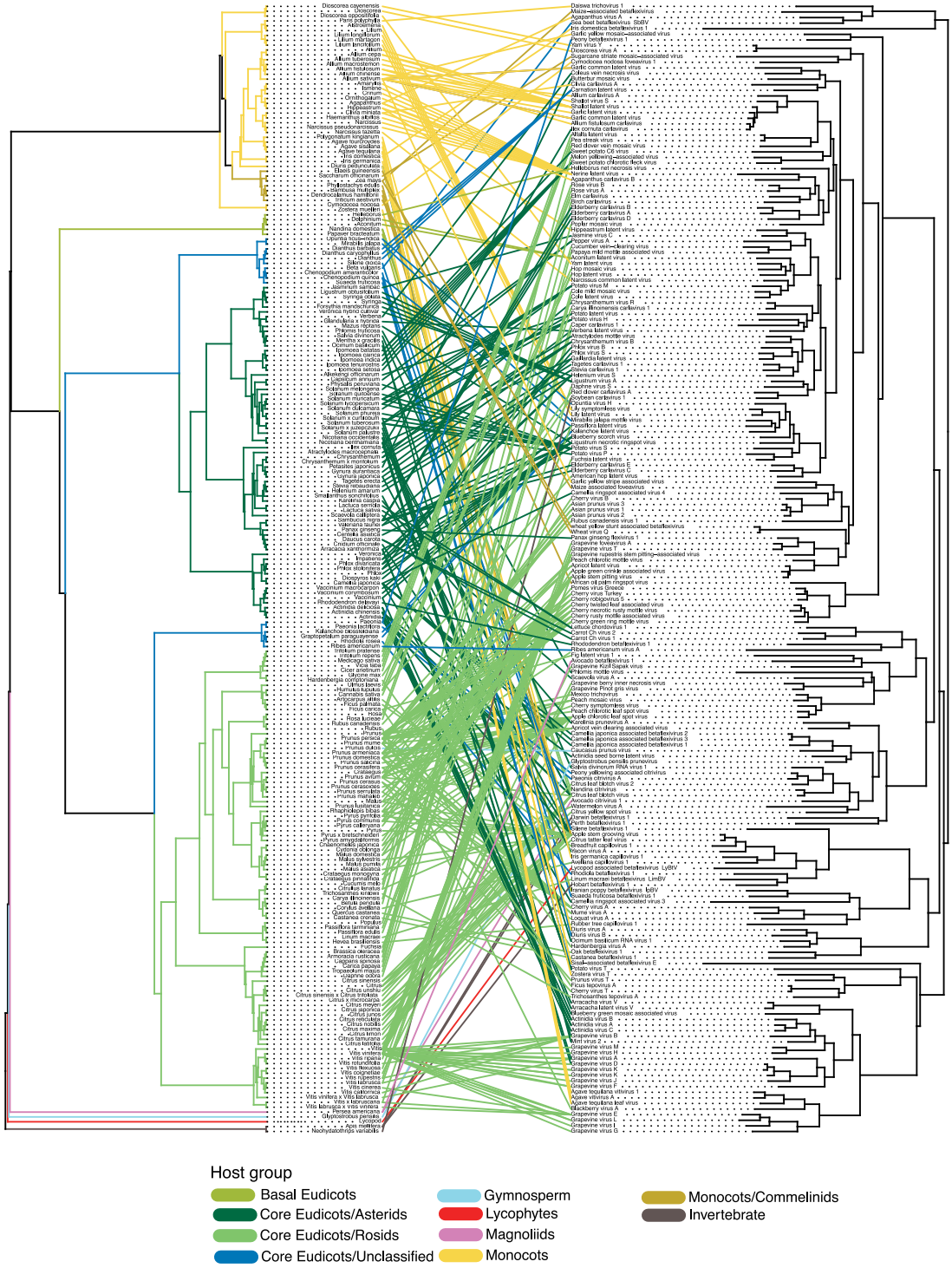
***Supplementary Figure 4.4.*** *Phylogram of the deltapartitii-like virus (a) coat protein/RNA2 (CP) and (b) RNA3/coat protein 2. ML phylogenetic trees show the topological position of the newly discovered CP sequences in (black circle) in the context of the closest relatives. All branches are scaled to the number of amino acid substitutions per site and trees were mid-point rooted for clarity only. Numbers at the nodes indicate bootstrap support over 70% (1000 replicates).*




***Benyviridae***

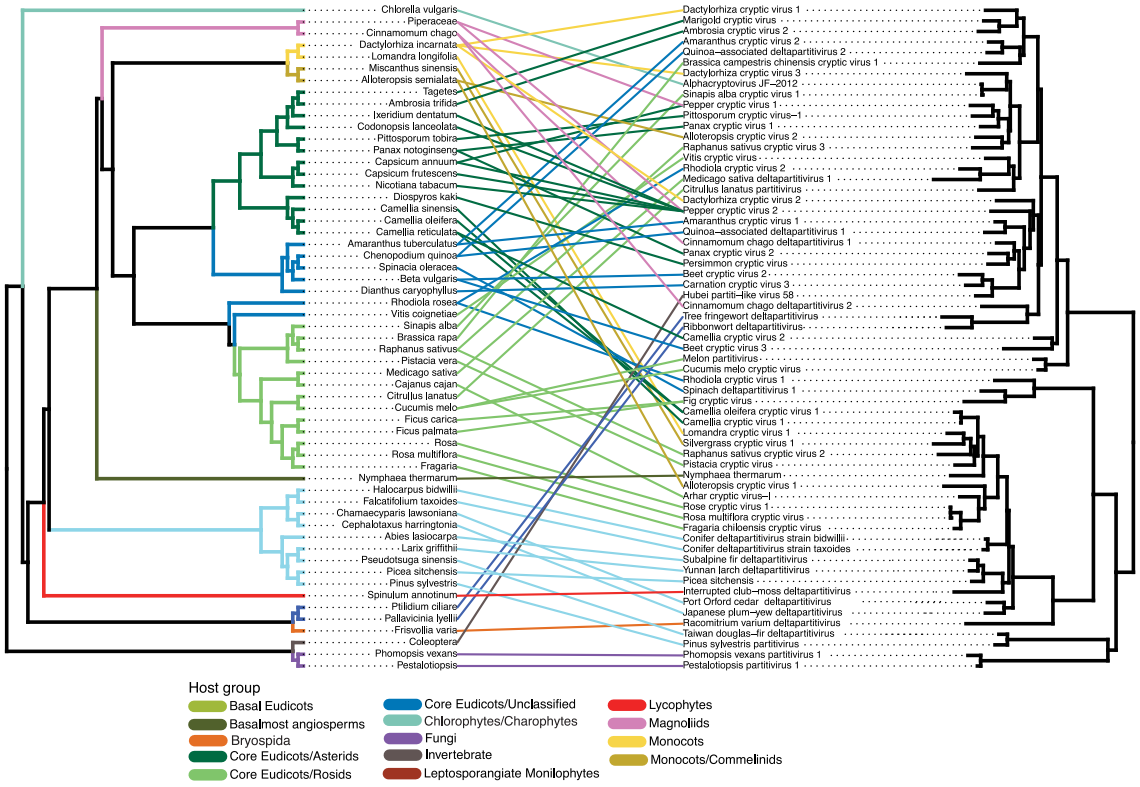
Host group

- Bryospida
- Core Eudicots/Asterids
- Core Eudicots/Rosids
- Chlorophytes/Charophytes
- Core Eudicots/Unclassified
- Fungi
- Invertebrate
- Leptosporangiate Monilophytes
- Lycophytes
- Monocots
- Monocots/Commelinids

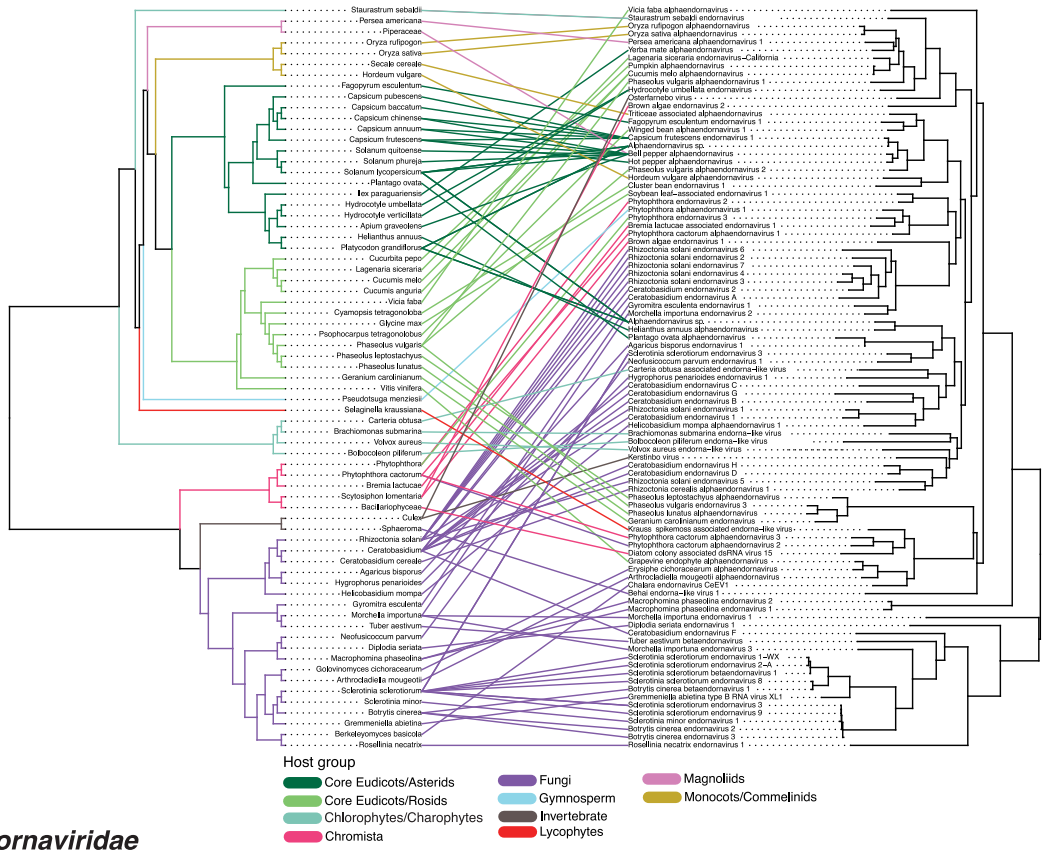



## *Betaflexiviridae*

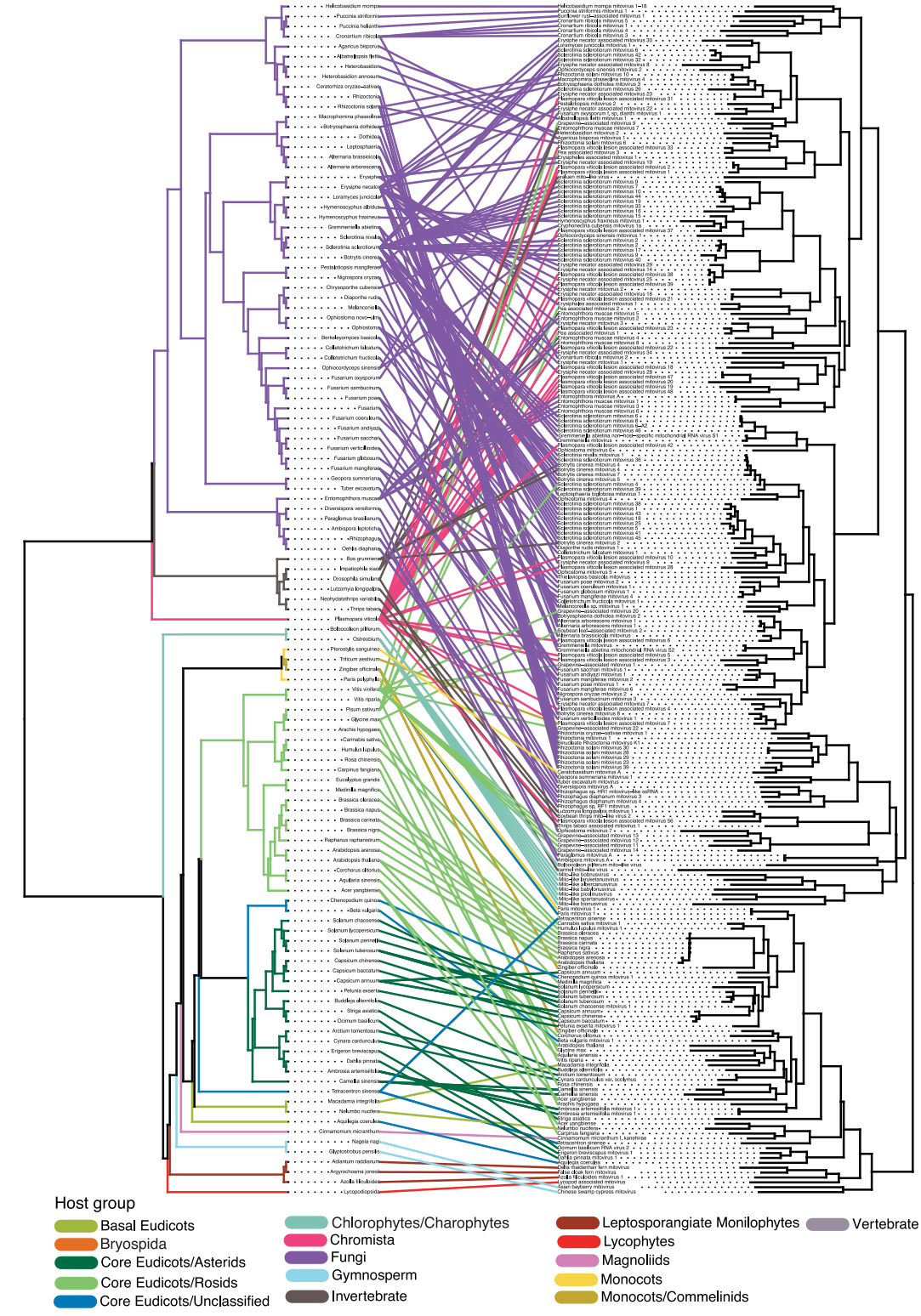
## *Bunyavirales/Viridisbunyaviridae*



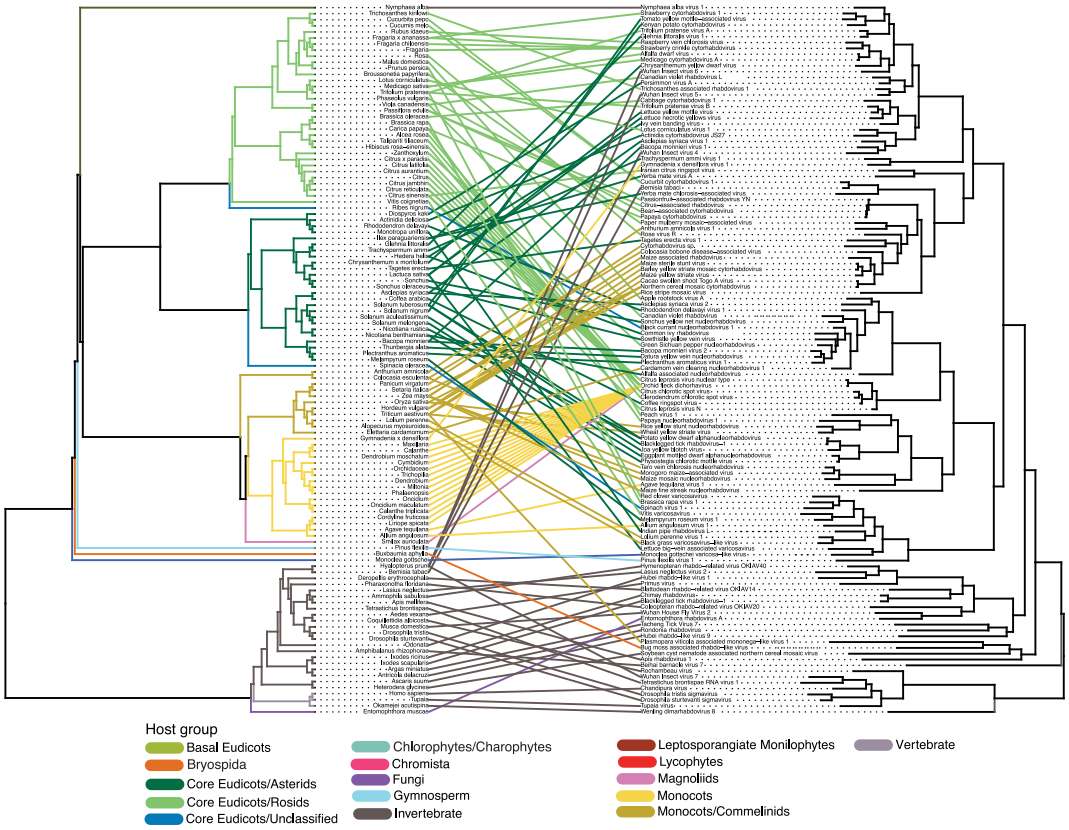

## *Deltapartitivirus*



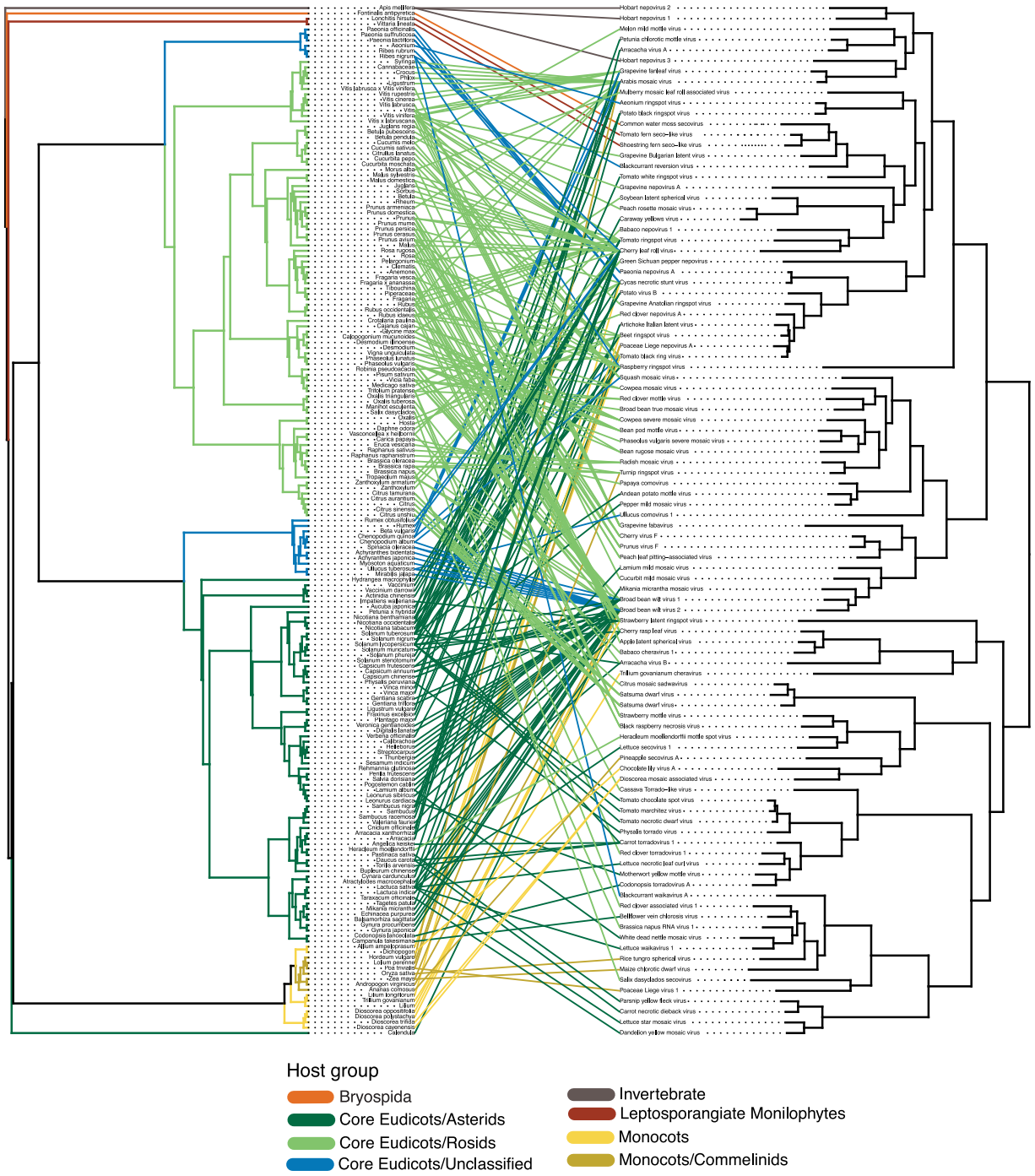

***Endornaviridae***



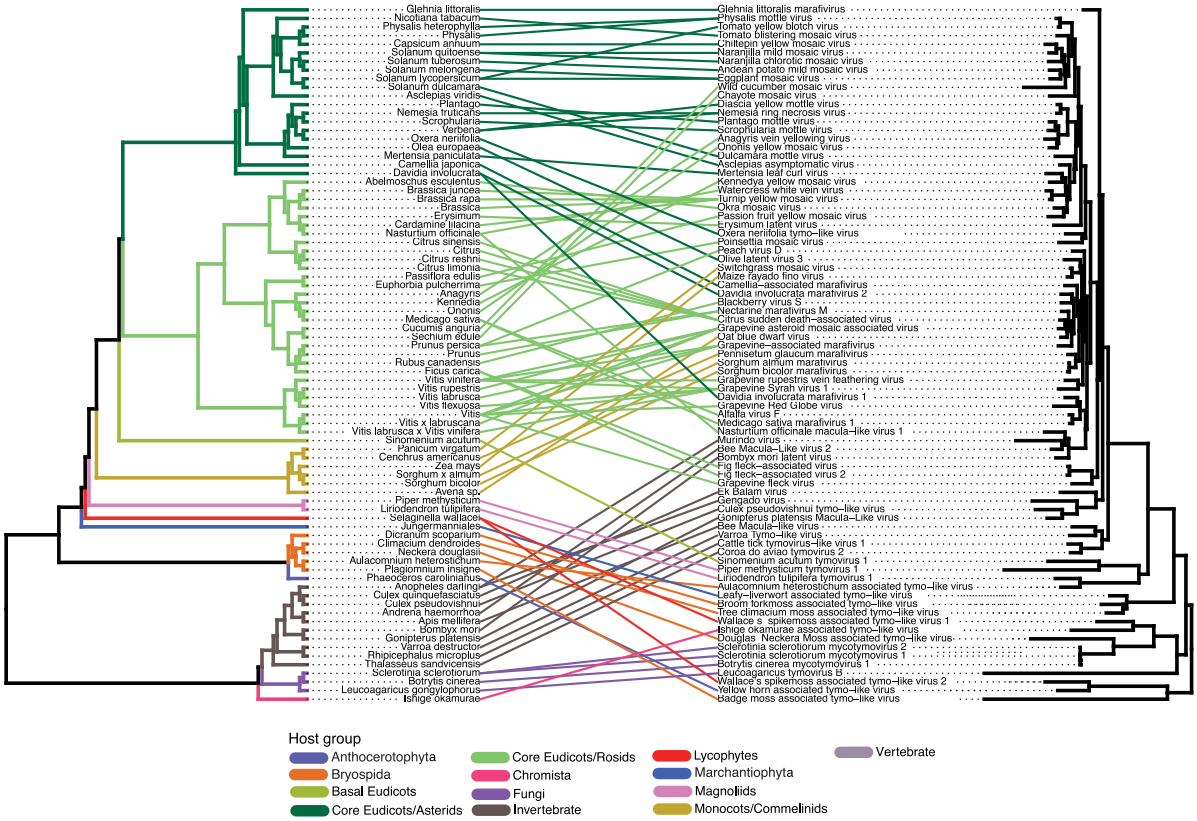

***Mitoviridae***




***Rhabdoviridae/Betarhabdovirinae***




## *Secoviridae*




***Tymoviridae***

***Supplementary Figure 4.5****. Tanglegram of rooted phylogenetic trees for select virus families and their hosts. Lines and branches are coloured to represent host clade. The cophylo function implemented in phytools (v0.7-80) was used to maximise the congruence between the host (left) and virus (right) phylogenies.*

*The following Supplementary tables are available online:*

https://journals.asm.org/doi/suppl/10.1128/jvi.00260-22/suppl_file/jvi.00260-22-s0001.pdf

***Supplementary Table 1.*** *Clade assignment for all One Thousand Plant Transcriptomes Initiative (1KP) species for which a virus was detected.*

***Supplementary Table 2.*** *Summary information for each One Thousand Plant Transcriptomes Initiative (1KP) libraries analysed.*

***Supplementary Table 3.*** *Proportion of transcripts and abundance assigned to each plant virus family.*

***Supplementary Table 4.*** *Summary table of the viruses discovered in this study*

***Supplementary Table 5.*** *Genome annotation information underlying the annotation graphs*

# Chapter 5. Transcriptome mining extends the host range of the *Flaviviridae* to non-bilaterians

Jonathon C.O. Mifsud[1], Vincenzo A. Costa[1], Mary E. Petrone[1], Ezequiel M. Marzinelli[2,3,4], Edward C. Holmes[1]*, Erin Harvey[1]

7. Sydney Institute for Infectious Diseases, School of Medical Sciences, The University of Sydney, Sydney, NSW 2006, Australia.
8. School of Life and Environmental Sciences, The University of Sydney, Sydney, NSW 2006, Australia.
9. Sydney Institute of Marine Science, 19 Chowder Bay Rd, Mosman, NSW, 2088 Australia.
10. Singapore Centre for Environmental Life Sciences Engineering, Nanyang Technological University, Singapore, 637551 Singapore.

### 5.1 Abstract

The flavivirids (family *Flaviviridae*) are a group of positive-sense RNA viruses that include well-documented agents of human disease. Despite their importance and ubiquity, the time-scale of flavivirid evolution is uncertain. An ancient origin, spanning millions of years, is supported by their presence in both vertebrates and invertebrates and by the identification of a flavivirus-derived endogenous viral element in the peach blossom jellyfish genome (*Craspedacusta Sowerby,* phylum Cnidaria), implying that the flaviviruses arose early in the evolution of the Metazoa. To date, however, no exogenous flavivirid sequences have been identified in these hosts. To help resolve the antiquity of the *Flaviviridae,* we mined publicly available transcriptome data across the Metazoa. From this, we expanded the diversity within the family through the identification of 32 novel viral sequences and extended the host range of the pestiviruses to include amphibians, reptiles, and ray-finned fish. Through cophylogenetic analysis we found cross-species transmission to be the predominate macroevolutionary event across the non-vectored flavivirid genera (median, 68%), including a cross-species transmission event between bats and rodents, although long-term virus-host co-divergence was still a regular occurrence (median, 23%). Notably, we discovered flavivirus-like sequences in basal metazoan species, including the first associated with Cnidaria. This sequence formed a basal lineage to the genus *Flavivirus* and was closer to arthropod and crustacean flaviviruses than those in the tamanavirus group that include a variety of invertebrate and vertebrate viruses. Combined, these data attest an ancient origin of the flaviviruses, likely close to the emergence of the metazoans 750–800 million years ago.

## 5.2 Introduction

The flavivirids (family *Flaviviridae)* are a group of positive-sense single-stranded RNA viruses comprising the genera *Flavivirus*, *Pestivirus*, *Pegivirus*, and *Hepacivirus.* These viruses include well-documented agents of human and livestock disease, including dengue virus, hepatitis C virus, yellow fever virus, Zika virus, and Bovine viral diarrhea virus 1. Reflecting their regular occurrence as pathogens, our understanding of flavivirid biology is necessarily skewed towards a subset of metazoan hosts, particularly those known to experience overt disease or act as reservoirs for these viruses, impeding our ability to understand the evolutionary history of this family. Currently available data suggests that all established genera, with the exception of the genus *Flavivirus*, are vertebrate-infecting viruses and do not require an arthropod vector for transmission (Simmonds et al. 2017).

The genus *Flavivirus*, can itself be divided into four groups defined by phylogenetic position and host range; the (i) mosquito-borne flaviviruses (MBFV), (ii) tick-borne flaviviruses (TBFV), (iii) insect-specific flaviviruses (ISF) and (iv) vertebrate-specific flaviviruses, also known as the “no known vector” (NKV) flaviviruses (Blitvich and Firth 2017; Simmonds et al. 2017). A wide diversity of more divergent “flavi-like” viruses have also been identified, including a group associated with crustaceans and decapods as well as the tamanaviruses (after Tamana bat virus (TABV)), which contains viruses from a broad range of vertebrate and invertebrate species (Costa et al. 2021; Geoghegan et al. 2018; Le Lay et al. 2020; Parry et al. 2019; Price 1978; Shi et al. 2018; Skoge et al. 2018; Soto et al. 2020). Another clade of related flavi-like viruses was recently identified in free-living parasitic flatworms (order Tricladida) (Dheilly et al. 2022).

Metagenomic surveys have identified flavivirid sequences with diverse genome structures, straying from the single 9–13 kb polyprotein that previously appeared to be canonical for the family. This expanded diversity includes a group of novel predominantly arthropod associated viruses – the Jingmenviruses – that are both segmented and perhaps multicomponent (Ladner et al. 2016; Qin et al. 2014; Shi et al. 2016; Simmonds et al. 2017). Metagenomic studies have also expanded the host range of hepaci-, pesti- and pegiviruses in non-mammalian hosts, including the discovery of hepaci-and pegiviruses in birds (Chang et al. 2021; Goldberg et al. 2019; Porter et al. 2020; Zhang et al. 2022), hepaci- and pesti-like viruses in cartilaginous fish

(Chondrichthyes) (Shi et al. 2018), and hepaciviruses in reptiles and bony fish (Osteichthyes) (Costa et al. 2022; Harding et al. 2022; Porter et al. 2020; Shi et al. 2018).

The identification of flavivirid sequences in marine invertebrate and basal vertebrate lineages has led to suggestions that the evolution of the *Flaviviridae* may follow that of the metazoans through virus-host co-divergence over time-scales of hundreds of millions of years (Bamford et al. 2022; Lensink et al. 2022; Shi et al. 2018). This, in turn, has stimulated questions regarding their host range and mode of transmission, while the complex evolutionary history of the flaviviruses and related sequences has been highlighted by their broad host range and sequence diversity. For example, the large phylogenetic gap between the cartilaginous fish and mammalian pestiviruses suggests that related viruses in bony fish, amphibians, reptiles, and birds exist but have yet to be sampled. The identification of flaviviruses in freshwater and marine crustaceans and a flavivirus-derived endogenous viral element (EVE) in the peach blossom jellyfish genome (*Craspedacusta Sowerby*, phylum Cnidaria) (Bamford et al. 2022) points towards an aquatic origin for the flaviviruses and highlights their long evolutionary association with the Metazoa. In particular, the cnidarian EVE suggests the existence of exogenous cnidarian flaviviruses. These are of importance for understanding the evolution of the *Flaviviridae*, as cnidarians, which include jellyfish, sea anemones, and corals, are an early branching lineage of the metazoans thought to have originated 700 million years ago (Erwin 2015). The phylogeny of the Metazoa can itself be divided into two major groups: those with bilateral body symmetry, the bilaterians, that comprise 99% of all animal species and, basal to them, the non-bilaterians, that include all the early diverging metazoan lineages – the Cnidaria, Placozoa, Porifera, and Ctenophora. Because non-bilaterians lack the body plan and circulatory system of vertebrates, it is possible that viruses of these hosts use an alternate mode of cell-to-cell transmission. To date, however, no flavivirids have been identified in these early diverging metazoan phyla.

Transcriptome mining is a proven method of virus discovery that leverages previous investment in metagenomics (Dheilly et al. 2022; Edgar et al. 2022; Greninger 2018; Grimwood et al. 2021; Iwamoto et al. 2021; Mifsud et al. 2022; Miller et al. 2021; Olendraite et al. 2022; Paraskevopoulou et al. 2021; Parry et al. 2019). To understand the host range of flavivirid sequences throughout the Metazoa and hence more accurately determine the age of the *Flaviviridae*, we used the Serratus RNA-dependent RNA polymerase (RdRp) search

(https://www.serratus.io/explorer/rdrp) to mine the Sequence Read Archive (SRA) database for novel flavivirid sequences. To supplement this analysis, total RNA-sequencing data of the tunicate *Botrylloides leachii* was generated and screened to identify additional flavivirid sequences.

## 5.3 Methods

### 5.3.1 Screening of SRAs for flavivirid-like sequences

The Serratus RdRp search and palmID analysis suite (Babaian and Edgar 2022; Edgar et al. 2022) were used to identify data sets within the SRA (as of May 2022) that contain signatures of novel flavivirid-like sequences. This search was limited to the family *Flaviviridae* with a threshold score of >= 50 (for an explanation of the Serratus classifier score see https://github.com/ababaian/serratus/wiki/.summary-Reports). The *de novo* transcriptome assemblies available at the National Center for Biotechnology Information (NCBI) Transcriptome Shotgun Assembly (TSA) Database (https://www.ncbi.nlm.nih.gov/genbank/tsa/) (as of June, 2021) were also screened using the translated Basic Local Alignment Search Tool (tBLASTn) algorithm under default scoring parameters and the BLOSUM45 matrix. Amino acid sequences from representatives of the four *Flaviviridae* genera along with the related Jingmenviruses were used as queries for the palmID and TSA database searches (Supplementary Table 1a). All novel virus sequences discovered were then used as queries in further SRA and TSA searches. The SRA and TSA search range was limited to Eukaryotes (NCBI taxonomic identifier (taxid 2759) excluding the Viridiplantae (taxid 33090). Invertebrate data sets were limited to aquatic species as terrestrial invertebrate SRAs has been previously examined (Dheilly et al. 2022).

### 5.3.2 Tunicate collection, RNA extraction, and metagenomic next-generation sequencing

The tunicate *Botrylloides leachii* was collected by divers wearing surgical gloves at 0.5-3m depth at the pier-pilings in Chowder Bay, Sydney, Australia (site description in Marzinelli (2012)), on 24 November, 2021. Sections of colonies were detached from the substratum using sterile tweezers, which were rinsed in 80% ethanol between samples, and brought to the surface, where they were placed in sterile cryogenic tubes. Samples were stored in liquid nitrogen on site and then transferred to a -80°C freezer until extraction. Total RNA was extracted using the RNeasy Plus Mini Kit (Qiagen, Hilden, Germany) as previously described in

(Geoghegan et al. 2021). These libraries were constructed using the Truseq Total RNA Library Preparation Protocol (Illumina). Host ribosomal RNA (rRNA) was depleted with the Ribo-Zero Plus Kit (Illumina) and paired-end sequencing (150 bp) was performed on the NovaSeq 6000 platform (Illumina). Library construction and metatranscriptomic sequencing were performed by the Australian Genome Research Facility (AGRF).

### 5.3.3 Identification of novel flavivirid genomes

Raw FASTQ files for all libraries that contained flavivirid-like sequences were obtained through the European Nucleotide Archive (https://www.ebi.ac.uk/ena/browser/home). Adapter removal and quality trimming were conducted using Trimmomatic (v0.38) with parameters SLIDINGWINDOW:4:5, LEADING:5, TRAILING:5 and MINLEN:25 (Bolger et al. 2014). To recover full-length virus sequences, raw reads were assembled *de novo* into contigs using MEGAHIT (v1.2.9) (Li et al. 2015). The assembled contigs were then compared to the NCBI non-redundant protein database (as of August, 2021) and a custom *Flaviviridae* protein database using DIAMOND BLASTx (v2.0.9) with an e-value threshold of $1\times10^{-5}$ (Buchfink et al. 2015). To identify highly divergent sequences a custom *Flaviviridae* protein database was regularly updated with the novel viruses identified.

### 5.3.4 Genome extension and annotation

Sequence reads were mapped onto virus-like contigs using Bbmap (v37.98) and areas of heterogeneous coverage were manually checked using Geneious (v11.0.9) (Bushnell 2014; Kearse et al. 2012). Where possible, the extremities of contigs were manually extended and re-submitted to read mapping until the contig appeared complete or no overhanging extremities were observed. Sequences of vector origin were detected using VecScreen (https://www.ncbi.nlm.nih.gov/tools/vecscreen/) and removed. Contig abundances were calculated using the RNA-Seq by Expectation Maximization (RSEM) software (v1.3.0) (Li and Dewey 2011). GetORF from EMBOSS (v6.6.0) was used to predict open reading frames (ORFs) (Rice et al. 2000). To annotate protein functional domains, the InterProScan software package (v5.56) was used with the TIGRFAMs (v15.0), SFLD (v4.0), PANTHER (v15.0), SuperFamily (v1.75), PROSITE (v2022_01), CDD (v3.18), Pfam (v34.0), SMART (v7.1), PRINTS (v42.0), and CATH-Gene3D databases (v4.3.0) (Jones et al. 2014). Genome diagrams were constructed using a manually curated selection of predicted functional domains and visualised using gggenomes (Hackl et al. 2024).

### 5.3.5 Detection of endogenous virus elements

To screen for endogenous viral elements (EVEs) within the viral-like contigs, the putative virus-like nucleotide sequence was compared to the corresponding host genome (where available) and a subset of the whole-genome shotgun (WGS) contig database (as of October 2022) using the TBLASTN algorithm with an E value cutoff of $1 \times 10^{-20}$. In addition, the virus-like sequences were checked for host gene contamination using the contamination function implemented in CheckV (v0.8.1) (Nayfach et al. 2021). All EVEs were removed from subsequent analyses.

### 5.3.6 Assessment of library composition

Taxonomic identification for the contigs assembled for each library was obtained by aligning them to the custom NCBI nt database using the KMA aligner and the CCMetagen program (Clausen et al. 2018; Marcelino et al. 2020). In the case of the cigar comb jelly flavivirus, where raw reads are not publicly available, contigs from the corresponding transcriptome shotgun assembly (TSA: GHXY01000001:GHXY01366104) were used as input. Virus abundance was calculated by counting the number of nucleotides matching the reference sequence with an additional correction for template length (the default parameter in KMA). Krona graphs were created using the KMA and CCMetagen methods and further edited in Adobe Illustrator (https://www.adobe.com) (Clausen et al. 2018; Marcelino et al. 2020).

Virus sequences identified in this study were named using a combination of the host common name—if known—and the appropriate *Flaviviridae* genera (e.g., harrimaniidae flavivirus). Virus-host assignments were made using a combination of host/virus abundance measurements and phylogenetic analyses. Where <80% of host abundance was associated with the target species of the library, the possibility of alternative hosts was considered. In this case, the other organisms comprising this library were examined to determine if they might represent the source of the virus sequence. For instance, given the known host range of the flavivirids it is more likely that these sequences are derived from metazoan species than from bacteria, fungi, or archaea. As such, metazoan species were given greater weighting when assigning putative virus-host assignments. Where host assignment proved difficult to assign with accuracy, the suffix “associated” was added to the host name to signify this (e.g., digyalum oweni associated virus). Where the taxonomic position of a virus was ambiguous, the suffix “-like” was used (e.g., African cichlid flavi-like virus).

### 5.3.7 Phylogenetic analysis

Phylogenetic trees of the putative flavivirid sequence identified here were inferred using a maximum likelihood approach. Translated virus contigs were aligned with known flavivirid protein sequences from NCBI/GenBank using MAFFT (v7.402) employing the generalised affine gap cost algorithm (Katoh and Standley 2013; Sayers et al. 2021). Poorly aligned regions were removed using trimAl (v1.2) with a gap threshold ranging from 0.7 to 0.9 and a variable conserve value (Capella-Gutiérrez et al. 2009). All phylogenetic trees were estimated using IQ-TREE2. Branch support was calculated using 1,000 bootstrap replicates with the UFBoot2 algorithm and an implementation of the SH-like approximate likelihood ratio test within IQ-TREE2 (Guindon et al. 2010; Hoang et al. 2017). Selection of the best-fit model of amino acid substitution was determined using the Akaike information criterion (AIC), the corrected AIC, and the Bayesian information criterion with the ModelFinder function in IQ-TREE 2 (Kalyaanamoorthy et al. 2017; Minh et al. 2020). The trimming methods, alignment lengths and phylogenetic models chosen in this analysis are outlined in Supplementary Table 5.1b. Phylogenetic trees were annotated using the R packages phytools (v1.0-3), and ggtree (v3.3.0.9) and further edited in Adobe Illustrator (https://www.adobe.com) (Revell 2012; Yu et al. 2017).

### 5.3.8 Assessment of cross-species virus transmission

To visualise the relative occurrence of cross-species transmission and virus-host codivergence across the *Flaviviridae* we analysed the cophylogenetic relationship between viruses and their hosts. Host cladograms were created using the phyloT software, a phylogenetic tree generator based on NCBI taxonomy (http://phylot.biobyte.de/). Virus-host associations were obtained from the NCBI virus (Brister et al. 2015; Hatcher et al. 2017) and Virus-Host database (release 213) (Mihara et al. 2016) (accessed 14 September 2022). Tanglegrams that graphically represent the correspondence between host and virus trees were created using the R packages phytools (1.0-3) (Revell 2012) and ape (v5.6-2) (Paradis and Schliep 2019). The virus phylogenies used in the cophylogenies were constructed as described above. The relative frequencies of cross-species transmission versus virus-host codivergence were quantified using the Jane package that employs a maximum parsimony approach to establish the best “map” of the virus phylogeny onto the host phylogeny (Conow et al. 2010). The cost of duplication, host-jumping, and extinction event types were set to 1.0, while virus-host

codivergence was set to zero as it was considered the null event. The number of generations and the population size were set to 100. Jane was chosen over its successor eMPRess (Santichaivekin et al. 2020), as it allows a virus to be associated with multiple host species and handles polytomies (Santichaivekin et al. 2020). For a multi-host virus, each association was represented as a polytomy in the virus phylogeny. A co-phylogenetic analysis of the genus *Flavivirus*, was not conducted as vector-borne viruses, that have both invertebrate and invertebrate hosts, are problematic to incorporate into analyses of this kind.

### 5.4 Results

Screening of transcriptomes revealed the presence of flavivirid-like sequences in 154 sequencing libraries within the SRA and TSA databases as well as one newly generated sequencing library from tunicates. The assembly and mining of these sequencing libraries identified 32 novel virus-like sequences, which were subsequently assigned as hepaci-like (20), flavi-like (7), pesti-like (4), and as unclassified flavivirial-like sequences (1) (Table 5.1, Figure 5.1). These virus-like sequences were predominately found in metazoan transcriptomes belonging to aquatic species (amphibians, bony fish, cnidarians, comb jellies, crustaceans, and hemichordates), although some were also found in land-dwelling vertebrates (birds, primates, and rodents). One virus-like sequence was assembled from a non-metazoan, alveolate library. No pegi-like virus sequences were found. We now examine each genus in turn.

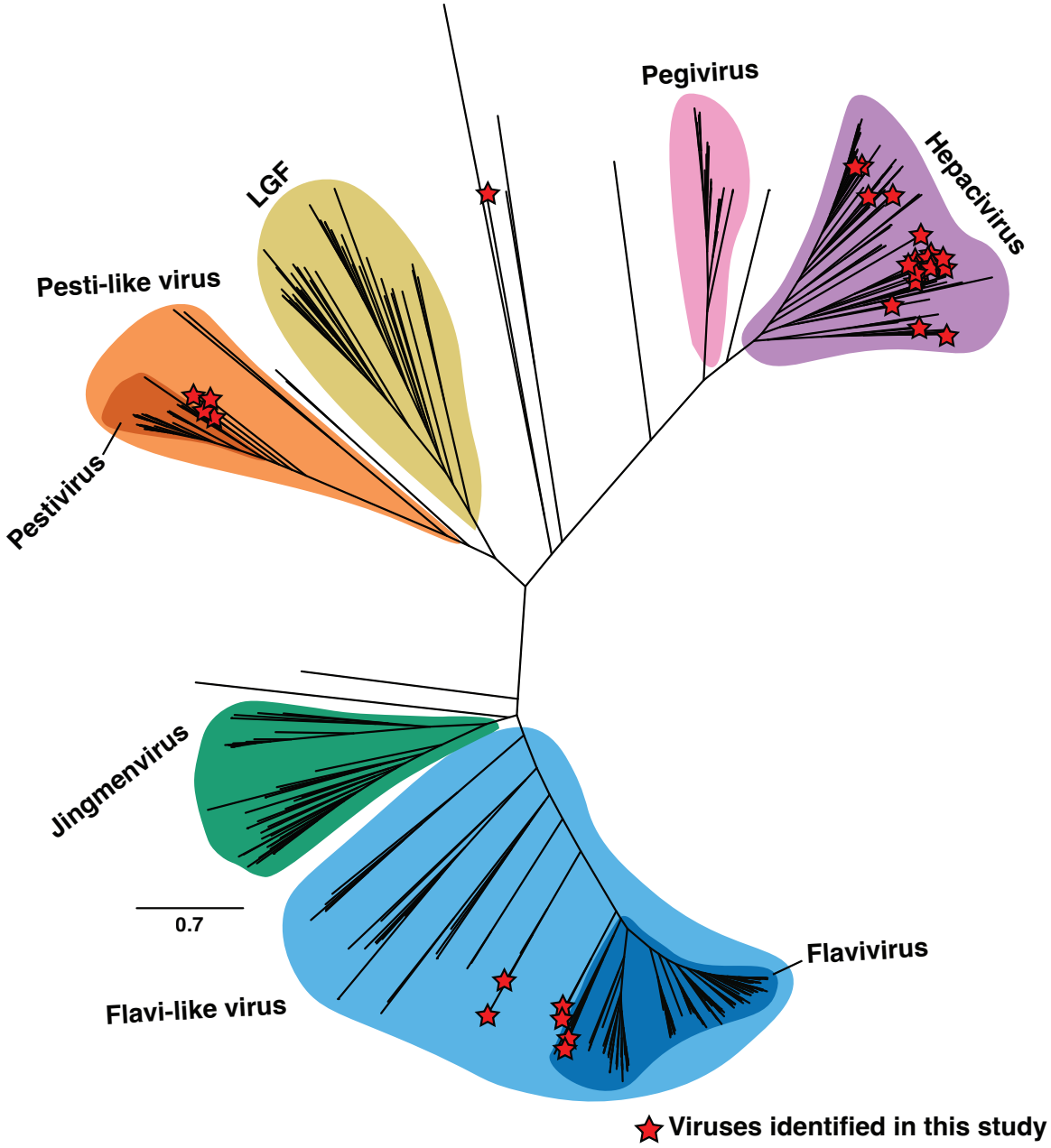



***Figure 5.32 Phylogeny of the Flaviviridae.*** *Unrooted maximum likelihood phylogenetic tree of the flavivirid sequences based on the conserved amino acid in the RdRp (NS5). All branches are scaled according to the number of amino acid substitutions per site. Established genera and notable clades that are yet to be ratified by ICTV are highlighted. Novel virus sequences identified in this study are displayed with a red star. LGF refers to the "large genome flaviviruses".*

***Table 5.1. Summary table of the viruses discovered in this study***

| Virus name | Abbreviation | Virus classification | Sample organism | SRA/TSA | Sequence length (bp) | % Std abundance | Closest Blast Hit (BLASTx) | % Ident (BLASTx) | *E*-value (BLASTx) | Tissue | Reference |
|---|---|---|---|---|---|---|---|---|---|---|---|
| African cichlid flavi-like virus | AfciFV | Flavi-like | Cyprichromis microlepidotus | SRR9688165 | 10,364 | 0.01 | Salmon flavivirus (QJU12405.1) | 38 | 0 | Lower pharyngeal jaw | El Taher et al. (2021) |
| Chowder bay tunicate–associated flavi-like virus | CbtuFV | Flavi-like | *B. leachii* | PRJEB57836 | 10,455 | 0.001 | Flavivirus sp.(QOQ37358.1) | 29 | 1.00E−80 | Whole body | NA |
| Cigar comb jelly flavi-like virus | CcjeFV | Flavi-like | *Beroe forskalii* | TSA:GHXY01073473.1 | 346 | NA | Salmon flavivirus (QJU12405.1) | 36 | 2.77E−07 | Whole body | Schultz et al. (2020) |
| Cigar comb jelly flavi-like virus | CcjeFV | Flavi-like | *B. forskalii* | TSA:GHXY01284671.1 | 226 | NA | Salmon flavivirus (QJU12405.1) | 51 | 2.66E−29 | Whole body | Schultz et al. (2020) |
| Cnidaria flavivirus | CnidFV | Flavivirus | *Eunicea flexuosa* | SRR12876665 | 10,729 | 0.004 | Kadam virus (YP 009345035.1) | 28 | 2.47E−312 | Mixed | Prada and Hellberg (2021) |
| Harrimaniidae flavivirus | HarFV | Flavivirus | Harrimaniidae sp | SRR1695462 | 10,951 | 0.003 | Rio Bravo virus (NP 620,044.1) | 28 | 2.89E−278 | Mixed | Cannon et al. (2014) |
| Photeros flavivirus | PhoFV | Flavivirus | *Photeros* sp | SRR10811635 | 11,290 | 0.003 | Crangon crangon flavivirus (QCH00713.1) | 38 | 0 | Whole body | Hensley et al. (2019) |
| Sea-firefly flavivirus | SefiFV | Flavivirus | *Vargula hilgendorfii* | DRR171178 | 10,930 | 0.016 | Crangon crangon flavivirus (QCH00713.1) | 37 | 0 | Whole body | Bessho-Uehara et al. (2020) |

| Virus name | Abbreviation | Virus classification | Sample organism | SRA/TSA | Sequence length (bp) | % Std abundance | Closest Blast Hit (BLASTx) | % Ident (BLASTx) | *E*-value (BLASTx) | Tissue | Reference |
|---|---|---|---|---|---|---|---|---|---|---|---|
| Brown Julie hepacivirus | BrjuHV | Hepacivirus | *Julidochromis dickfeldi* | SRR9688628 | 2160 | 0.0007 | Guangxi houndshark hepacivirus (AVM87259.1) | 41 | 1.08E−140 | Liver | El Taher et al. (2021) |
| Brown Julie hepacivirus | BrjuHV | Hepacivirus | *J. dickfeldi* | SRR9688628 | 2990 | 0.001 | Rodent hepacivirus (AGC52826.1) | 43 | 1.08E−111 | Liver | El Taher et al. (2021)) |
| Cape dune mole-rat hepacivirus | CdmrHV | Hepacivirus | *Bathyergus suillus* | SRR2141210 | 6224 | 0.002 | Bat hepacivirus (QQM18105.1) | 42 | 0 | Mixed | Davies et al. (2015) |
| Catfish hepacivirus | CatfHV | Hepacivirus | *G. macromaculatus* | SRR6426015 | 9865 | 0.005 | Guangxi houndshark hepacivirus (AVM87259.1) | 32 | 0 | Liver | Jiang et al. (2019) |
| Caudopunctatus cichlid hepacivirus | CaciHV | Hepacivirus | *Neolamprologus toae* | SRR9680779 | 7165 | 0.009 | Guangxi houndshark hepacivirus (AVM87256.1) | 34 | 4.49E−284 | Liver | El Taher et al. (2021)) |
| Caudopunctatus cichlid hepacivirus | CaciHV | Hepacivirus | *N. toae* | SRR9680779 | 2385 | 0.003 | Rodent hepacivirus (AGC52826.1) | 43 | 2.96E−116 | Liver | El Taher et al. (2021) |
| Crab-eating macaque hepacivirus | CremHV | Hepacivirus | *Macaca fascicularis* | SRR6032465 | 9855 | 0.031 | Guereza hepacivirus (YP 009325369.1) | 57 | 0 | Whole blood | Speranza et al. (2018) |
| Cyprichromis hepacivirus | CyprHV | Hepacivirus | *Cyprichromis* sp. 'dwarf jumbo' AB-2019 | SRR9688149 | 8678 | 0.012 | Guangxi houndshark hepacivirus (AVM87256.1) | 33 | 0 | Liver | El Taher et al. (2021) |
| Dewindti cichlid hepacivirus | DeciHV | Hepacivirus | *Aulonocranus dewindti* | SRR9689044 | 1583 | 0.0009 | Rodent hepacvirus (ATP66829.1) | 50 | 5.08E−127 | Liver | El Taher et al. (2021) |

| Virus name | Abbreviation | Virus classification | Sample organism | SRA/TSA | Sequence length (bp) | % Std abundance | Closest Blast Hit (BLASTx) | % Ident (BLASTx) | *E*-value (BLASTx) | Tissue | Reference |
|---|---|---|---|---|---|---|---|---|---|---|---|
| Dewindti cichlid hepacivirus | DeciHV | Hepacivirus | *A. dewindti* | SRR9689044 | 1345 | 0.0005 | Hepacivirus L (YP 009679023.1) | 44 | 4.95E−99 | Liver | El Taher et al. (2021) |
| Featherfin cichlid hepacivirus | FeciHV | Hepacivirus | *Cyathopharynx furcifer* | SRR9680031 | 1070 | 0.0004 | Nanhai ghost shark hepacivirus 1 (AVM87556.1) | 44 | 3.73E−71 | Liver | El Taher et al. (2021) |
| Freshwater butterflyfish hepacivirus | FrbuHV | Hepacivirus | *Pantodon buchholzi* | SRR5997808 | 8422 | 0.004 | Nanhai ghost shark hepacivirus 1 (AVM87556.1) | 29 | 4.40E−275 | Viscera mixture | Sun et al. (2016) |
| Gold head compressiceps hepacivirus | GhcoHV | Hepacivirus | *Altolamprologus compressiceps* | SRR9688991 | 2553 | 0.002 | Guangxi houndshark hepacivirus (AVM87259.1) | 43 | 1.00E−176 | Liver | El Taher et al. (2021) |
| Gold head compressiceps hepacivirus | GhcoHV | Hepacivirus | *A. compressiceps* | SRR9688991 | 1603 | 0.002 | Hepacivirus L (YP 009679023.1) | 43 | 3.59E−97 | Liver | El Taher et al. (2021) |
| Greater mouse-eared bat hepacivirus | GmebHV | Hepacivirus | *Myotis myotis* | SRR2063918 | 3300 | 0.046 | Rodent hepacivirus (QLM02863.1) | 71 | 0 | Pharyngeal and anal swab | Wu et al. (2012) |
| Greater mouse-eared bat hepacivirus | GmebHV | Hepacivirus | *M. myotis* | SRR2063918 | 2981 | 0.046 | Rodent hepacivirus (QLM02863.1) | 63 | 0 | Pharyngeal and anal swab | Wu et al. (2012) |
| Neolamprologus buescheri hepacivirus | NebuHV | Hepacivirus | *N. buescheri* | SRR9681059 | 2105 | 0.002 | Wenling shark virus (YP 009179227.1) | 45 | 4.57E−74 | Liver | El Taher et al. (2021) |

| Virus name | Abbreviation | Virus classification | Sample organism | SRA/TSA | Sequence length (bp) | % Std abundance | Closest Blast Hit (BLASTx) | % Ident (BLASTx) | *E*-value (BLASTx) | Tissue | Reference |
|---|---|---|---|---|---|---|---|---|---|---|---|
| Neolamprologus buescheri hepacivirus | NebuHV | Hepacivirus | *N. buescheri* | SRR9681059 | 2294 | 0.002 | Rodent hepacvirus (ATP66829.1) | 40 | 1.32E−144 | Liver | El Taher et al. (2021) |
| Neolamprologus savoryi hepacivirus | NesaHV | Hepacivirus | *N. savoryi* | SRR9680861 | 5451 | 0.007 | Wenling shark virus (YP 009179227.1) | 45 | 2.95E−93 | Liver | El Taher et al. (2021) |
| Neolamprologus savoryi hepacivirus | NesaHV | Hepacivirus | *N. savoryi* | SRR9680861 | 1223 | 0.001 | Wenling shark virus (YP 009179227.1) | 34 | 9.32E−231 | Liver | El Taher et al. (2021) |
| Northern treeshrew hepacivirus | NotrHV | Hepacivirus | *Tupaia belangeri* | DRR155087 | 9208 | 0.00003 | Hepacivirus P (YP 009553586.1) | 56 | 0 | Liver | Sanada et al. (2019) |
| Variabilichromis moorii hepacivirus | VamoHV | Hepacivirus | *V. moorii* | SRR9689131 | 4358 | 0.006 | Guangxi houndshark hepacivirus (AVM87259.1) | 36 | 1.82E−231 | Liver | El Taher et al. (2021) |
| *Neolamprologus savoryi hepacivirus* | VamoHV | Hepacivirus | *V. moorii* | SRR9689131 | 2857 | 0.004 | Rodent hepacivirus (AGC52826.1) | 44 | 4.71E−113 | Liver | El Taher et al. (2021) |
| Viviparous lizard hepacivirus | ViliHV | Hepacivirus | *Zootoca vivipara* | SRR11262332 | 11,862 | 0.001 | Duck hepacivirus (QKT21547.1) | 27 | 9.26E−263 | Whole body | Yurchenko et al. (2020) |
| White sucker hepacivirus | WhsuHV | Hepacivirus | *Catostomus commersonii* | SRR2912516 | 10,120 | 0.001 | Guangxi houndshark hepacivirus (AVM87256.1) | 31 | 0 | Liver | Hahn et al. (2016) |
| Xenotilapia hepacivirus | XenoHV | Hepacivirus | *Xenotilapia* sp. | SRR9670143 | 2596 | 0.003 | Rodent hepacivirus (QLM02868.1) | 40 | 3.32E−118 | Liver | El Taher et al. (2021) |

| Virus name | Abbreviation | Virus classification | Sample organism | SRA/TSA | Sequence length (bp) | % Std abundance | Closest Blast Hit (BLASTx) | % Ident (BLASTx) | *E*-value (BLASTx) | Tissue | Reference |
|---|---|---|---|---|---|---|---|---|---|---|---|
| | | | 'spilopterus north' AB-2019 | | | | | | | | |
| Xenotilapia hepacivirus | XenoHV | Hepacivirus | *Xenotilapia* sp. 'spilopterus north' AB-2019 | SRR9670143 | 4117 | 0.002 | Guangxi houndshark hepacivirus (AVM87256.1) | 35 | 6.73E−211 | Liver | El Taher et al. (2021) |
| Zebra finch hepacivirus 1 | ZefiHV1 | Hepacivirus | *Taeniopygia guttata* | SRR7031415 | 2262 | 0.0001 | Duck hepacivirus (QDF44087.1) | 52 | 2.74E−180 | Liver | NA |
| Zebra finch hepacivirus 1 | ZefiHV1 | Hepacivirus | *T. guttata* | SRR7031415 | 3396 | 0.0001 | Jogalong virus (QHD25538.1) | 49 | 0 | Ovary | NA |
| Zebra finch hepacivirus 2 | ZefiHV2 | Hepacivirus | *T. guttata* | SRR6896649 | 1245 | 0.0002 | Jogalong virus (QHD25538.1) | 51 | 3.85E−67 | Liver | Biederman et al. (2018) |
| Zebra finch hepacivirus 2 | ZefiHV2 | Hepacivirus | *T. guttata* | SRR6896649 | 1842 | 0.0001 | Jogalong virus (QHD25538.1) | 53 | 1.64E−209 | Ovary | Biederman et al. (2018) |
| Cayenne caecilian pestivirus | CacaPV | Pestivirus | *Typhlonectes compressicauda* | SRR5591415, SRR5591441 | 1391 | 0.0000 | Bovine viral diarrhoea virus 2 (AAC64056.1) | 58 | 1.59E−128 | Kidney | Torres-Sánchez et al. (2019) |
| Frog pestivirus | FrogPV | Pestivirus | *L. catesbeianus* | SRR5810423 | 15,334 | 0.012 | Classical swine fever virus (NP 777,506) | 28 | 1.35E−256 | Ventral skin | Eskew et al. (2018) |

| Virus name | Abbreviation | Virus classification | Sample organism | SRA/TSA | Sequence length (bp) | % Std abundance | Closest Blast Hit (BLASTx) | % Ident (BLASTx) | *E*-value (BLASTx) | Tissue | Reference |
|---|---|---|---|---|---|---|---|---|---|---|---|
| Glass knifefish pestivirus | GlknPV | Pestivirus | *Eigenmannia virescens* | SRR6675258 | 14,199 | 0.013 | Pronghorn antelope pestivirus (YP 009026415.1) | 29 | 4.28E−232 | Muscle | Thompson et al. (2018) |
| Transcaucasian sand viper pestivirus | TrsaPV | Pestivirus | *Vipera transcaucasiana* | SRR12802473 | 3054 | 0.0001 | Bamboo rat pestivirus (UMO75502.1) | 25 | 6.51E−50 | Venom gland | Xie et al. (2022) |
| Transcaucasian sand viper pestivirus | TrsaPV | Pestivirus | *V. transcaucasiana* | SRR12802473 | 2120 | 0.0001 | Atypical porcine pestivirus (QBJ01589.1) | 53 | 1.65E−188 | Venom gland | Xie et al. (2022) |
| Digyalum oweni-associated virus | DiowV | Unclassifed pesti-like virus | *Digyalum oweni* | SRR9888042 | 3715 | 0.0003 | Xinzhou spider virus 3 (YP 009254746.1) | 24 | 4.09E−13 | Individual gut cell | Janouškovec et al. (2019) |
| Digyalum oweni-associated virus | DiowV | Unclassifed pesti-like virus | *Digyalum oweni* | SRR9888042 | 4594 | 0.0002 | Soybean thrips virus 4 (QPZ88419.1) | 25 | 8.94E−15 | Individual gut cell | Janouškovec et al. (2019) |

### 5.4.1 Genus Flavivirus

We identified seven putative flavi-like virus sequences, including cnidaria flavivirus (CnidFV) and cigar comb jelly flavi-like virus (CcjeFV) in libraries of the early diverging metazoan phyla Cnidaria and Ctenophora, harrimaniidae flavivirus (HarFV) in an acorn worm (Enteropneusta), photeros flavivirus (PhoFV) and sea-firefly flavivirus (SefiFV) in marine ostracods, Chowder Bay tunicate-associated flavivirus in tunicates (CbtuFV), and African cichlid flavivirus (AfciFV) in a cichlid fish (Figure 5.2). For all but one of these sequences (CcjeFV), complete genome sequences ranging in length from 10,364 to 11,290 nucleotides were assembled. CcjeFV consists of two partial RdRp fragments, 346 and 226 bp in length.

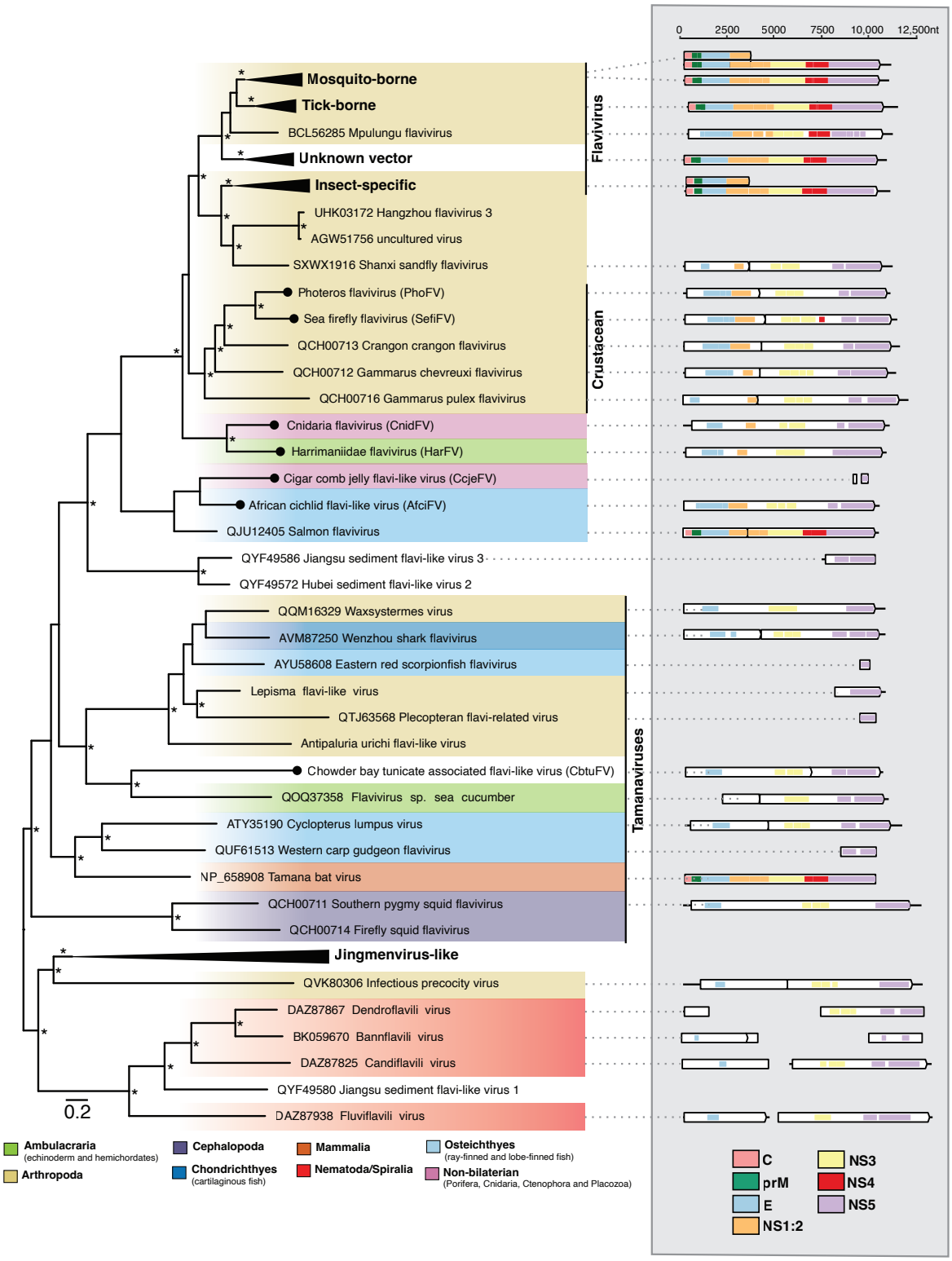



***Figure 5.33 Phylogenetic relationships of the flavi-like viruses identified in this study.*** *(Left) Phylogenetic relationships of the flavi- and jingmenviruses. ML phylogenetic trees based on the conserved amino acid in the RdRp (NS5) show the topological position of virus-like sequences discovered in this study (black circles) in the context of their closest relatives. Branches are highlighted to represent host clade (Ambulacraria = green,*

*Arthropoda = khaki, Cephalopoda = purple, Chondrichthyes = light blue, Mammalia = orange, Nematoda/Spiralia = red, Osteichthyes = dark blue, non-bilaterian = light purple). All branches are scaled to the number of amino acid substitutions per site, and trees were midpoint rooted for clarity only. An asterisk indicates node support where SH-aLRT >= 80% and UFboot >= 95%. (Right) Genomic organization of the virus sequences identified in this study and representative species used in the phylogeny. The data underlying this figure and definitions of acronyms used are presented in Supplementary Table 2.*

A range of genome structures were observed and found to be largely consistent with those found in this genus. For example, PhoFV and SefiFV, like the other viruses identified in marine crustaceans, are predicted to contain a programmed −1 ribosomal frameshift on a "slippery" heptanucleotide sequence downstream of the NS1 region (Parry et al. 2019) (Figure 2, Supplementary Figure 5.1). However, CbtuFV was predicted to contain two ORFs, with the NS4/5 region encoded on the second open reading frame, although no "slippery" heptanucleotide motifs could be detected (Figure 5.2). The remaining full-length sequences were predicted to contain a single ORF. Virus domains consistent with this genus were detected across all sequences (Figure 5.2).

Phylogenetic analyses of the conserved NS5 region place the ostracod sequences (PhoFV, SefiFV) within a larger diversity of marine crustacean flaviviruses. Two sequences, CnidFV and HarFV, fell basal to all classified members of the genus *Flavivirus* along with the crustacean flaviviruses (Figure 5.2). Notably, these sequences appear closer in phylogenetic position and amino acid identity to tick, insect-specific and crustacean flaviviruses than those viruses in the more divergent tamanavirus clade. The flavivirus-derived EVEs identified in the Cnidaria fell in approximately the same phylogenetic location as CnidFV and SefiFV (Supplementary Figure 5.2). CcjeFV and AfciFV were placed phylogenetically with salmon flavivirus (QJU12405.1), although unlike salmon flavivirus AfciFV consists of a single ORF.

### 5.4.2 Genus *Pestivirus*

We identified four pesti-like virus sequences in amphibians, reptiles, and bony fish (Table 5.1). Two full genomes – glass knifefish pestivirus (GlknPV) and frog pestivirus (FrogPV) – were recovered, ranging from 14,199 to 15,334 bp in length, in addition to two partial genomes, Transcaucasian sand viper pestivirus (FrogPV) and Cayenne caecilian pestivirus (CacaPV) (Figure 5.3). These sequences exhibit more sequence similarity with mammalian

pestiviruses than those associated with cartilaginous fish, with an average of 28% versus 24% amino acid identity across the complete polyprotein. This is reflected in the phylogenetic positioning of the novel pesti-like viruses based on the conserved NS5 region (Figure 5.3). The newly identified reptile and amphibian pesti-like virus sequences, FrogPV and CacaPV, form a sister group to those found in rodents, bats, and pigs, while the sequence discovered in fish, GlknPV, fell basal to this group but remained as a sister group to those viruses from cartilaginous fish (Shi et al. 2018). The topology of the pestivirus phylogeny varied depending on whether the NS3 or NS5 domains were used in the analysis. In particular, FrogPV formed a sister lineage to the known pestiviruses in a phylogeny based upon the NS3 region (Figure 5.3, Supplementary Figure 5.3).

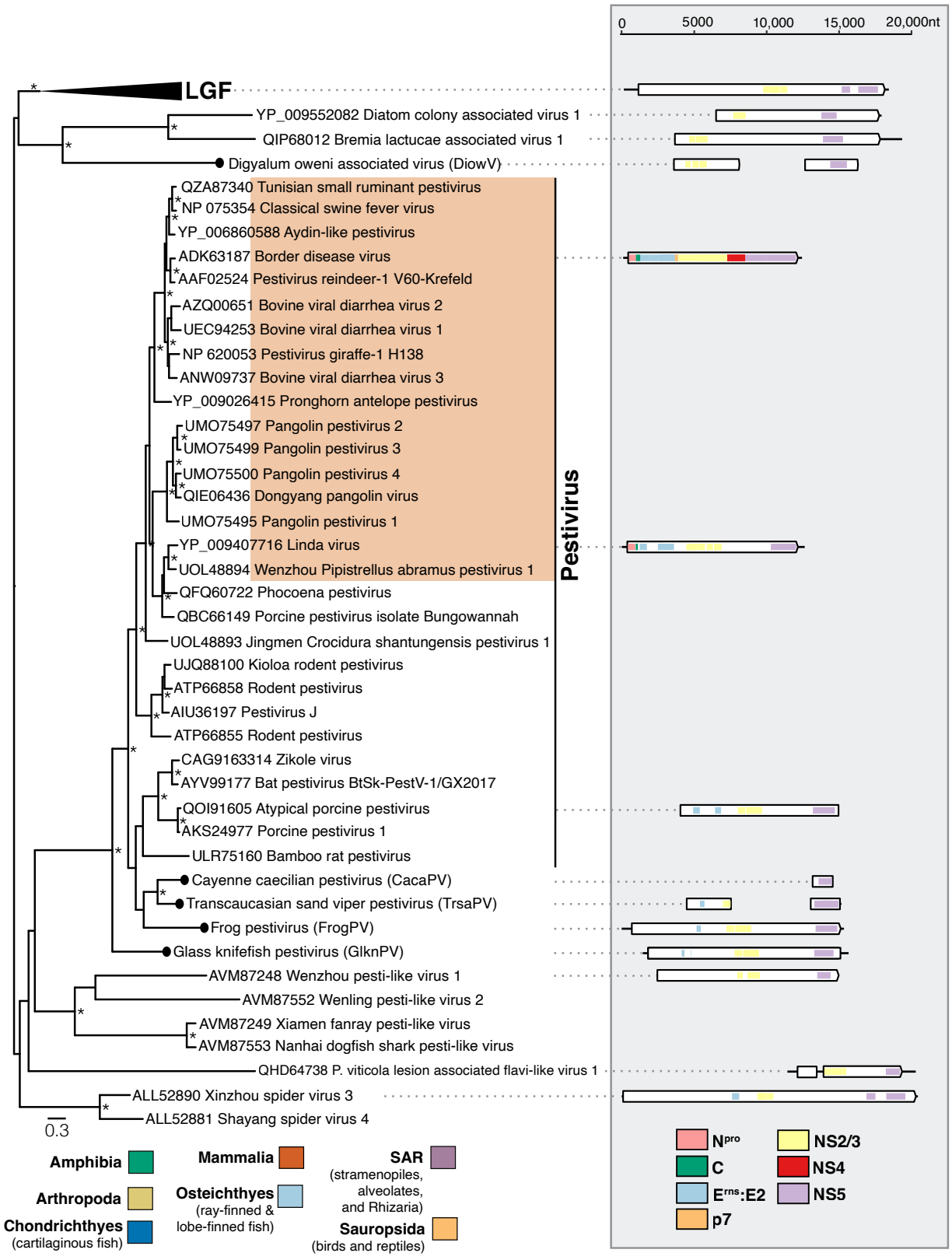



***Figure 5.34 Phylogenetic relationships of the pesti-like viruses identified in this study.*** *(Left) Phylogenetic relationships of the pestiviruses and unclassified relatives. ML phylogenetic trees based on the conserved amino acid in the RdRp (NS5) show the topological position of virus-like sequences discovered in this study (black circles) in the*

*context of their closest relatives. The colour scheme is as found in Figure 5.2, with the following exceptions, Amphibia = green, Sauropsida = light orange, SAR = light purple. All branches are scaled to the number of amino acid substitutions per site, and trees were midpoint rooted for clarity only. An asterisk indicates node support where SH-aLRT >= 80% and UFboot >= 95%. LGF refers to the "large genome flaviviruses". Non-novel sequences without NCBI accession were obtained from Wu et al. (2020) (Right) Genomic organization of the virus sequences identified in this study and representative species used in the phylogeny. The data underlying this figure and definitions of acronyms used are presented in Supplementary Table 2.*

### 5.4.3 Genus *Hepacivirus*

We identified 20 novel hepacivirus sequences, of which 14 were found in ray-finned fish (Actinopterygii), expanding on the two hepaciviruses previously identified in this group (Figure 4). The remaining sequences (n = 6) add to the known diversity of bat, avian, primate, rodent, and treeshrew hepaciviruses (Figure 5.4). Of the novel hepaciviruses, five complete genomes were assembled, ranging from 9,208 to 11,862 bp in length (Figure 5.4). Partial genome sequences containing at least the NS3 and NS5 domains were assembled for the remaining sequences, with the exception of featherfin cichlid hepacivirus for which only the NS5 region could be assembled (Figure 4). Of note, greater mouse-eared bat hepacivirus (GmebHV) was assembled from a library generated for the analysis of bat viromes (Wu et al. 2012) and shares 70% amino acid identity with rodent hepacivirus (QLM02863.1).

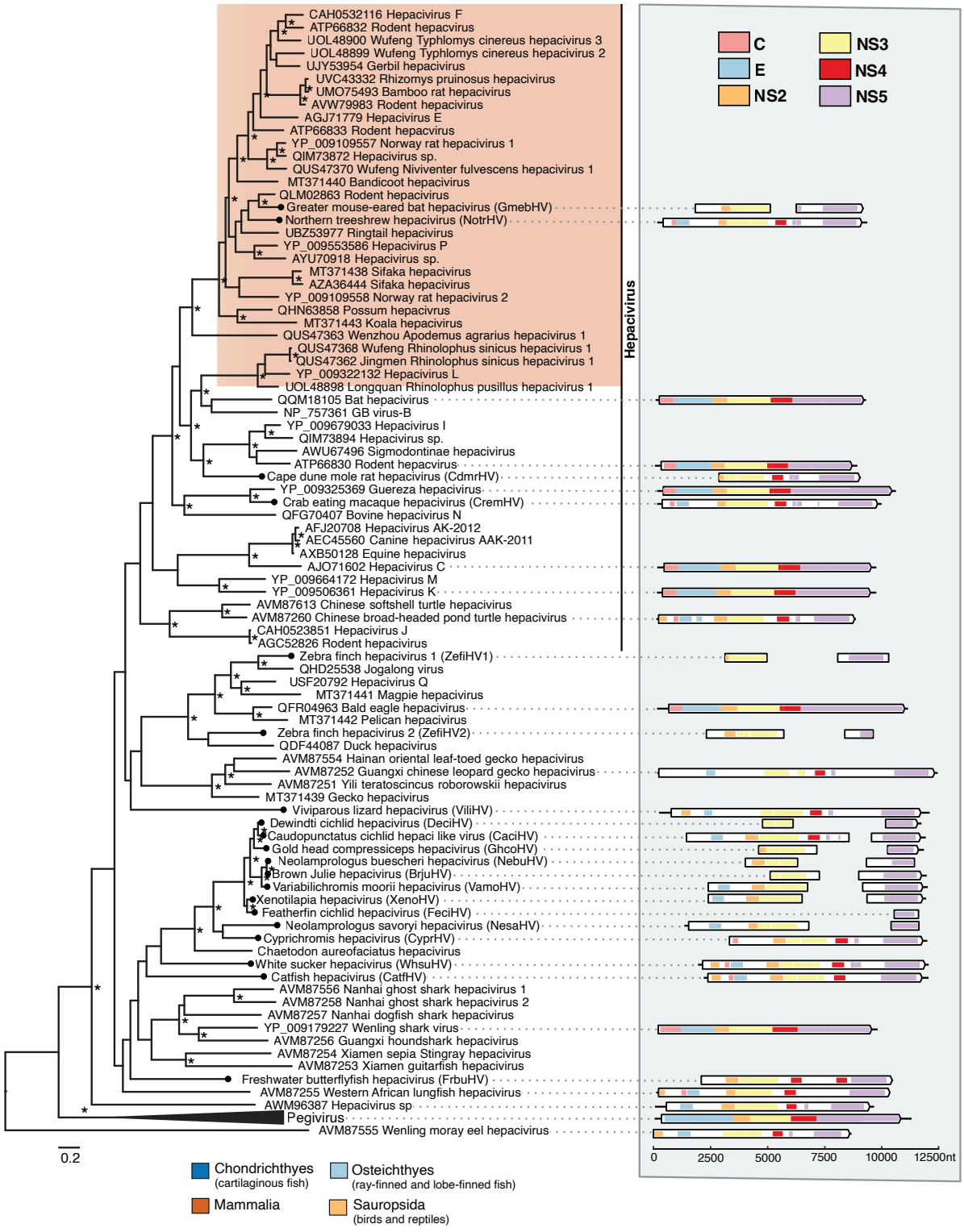



***Figure 5.35 Phylogenetic relationships of the hepaciviruses viruses identified in this study.*** *(Left) Phylogenetic relationships of the 'pegi-hepaci' clade. ML phylogenetic trees based on the conserved amino acid in the RdRp (NS5) show the topological position of virus-like sequences discovered in this study (black circles) in the context of their closest relatives. The colour scheme is as found in Figure 5.2, with the following exception,*

*Sauropsida = light orange. All branches are scaled to the number of amino acid substitutions per site, and trees were midpoint rooted for clarity only. An asterisk indicates node support where SH-aLRT >= 80% and UFboot >= 95%. (Right) Genomic organization of the virus sequences identified in this study and representative species used in the phylogeny. The data underlying this figure and definitions of acronyms used are presented in Supplementary Table 2.*

### 5.4.4 An unclassified flavivirid-like virus

In addition to the viruses that fell within established genera, we identified a partial flavi-like virus sequence termed digyalum oweni associated virus (DiowV) in *Digyalum oweni*, a species of parasitic protist belonging to the phylum Apicomplexa. Two contigs were assembled from this library, 3,689 and 4,577 bp in length and predicted to contain the NS3 and NS5 domains, respectively (Figure 5.3). DiowV shares the greatest amino acid identity (24%) with the Xinzhou spider virus 3 (YP_009254746) among other large genome flaviviruses (LGF). When included in the 'pesti-LGF' tree, DiowV, along with diatom colony associated virus 1 (YP_009552082) and bremia lactucae associated virus 1 (QIP68012), forms a sister group to the LGF. However, in the family-wide tree, these sequences, along with Snake River alfalfa virus (ON669064), fall outside of the 'pesti-LGF' lineage and basal to the 'pegi-hepaci' group, although these branches receive poor bootstrap support (Figure 5.1).

### 5.4.5 Genetic composition of sequencing libraries

Metagenomic sequencing libraries are often comprised of organisms in addition to the target host, which can complicate virus-host assignment. To quantify the composition of these libraries and improve virus host assignments we utilised the KMA and CCMetagen tools (Figure 5.5). For 20 of the libraries, over 80% of eukaryotic contigs were assigned to the expected target host of the sequencing library (median, 90%; range, 0% to 98%). In the case of the *E. flexuosa* (family *Plexauridae*) library in which CnidFV was assembled, a genus of unicellular microalgae, *Symbiodinium* (phylum Dinoflagellata), represented 64% of all contigs (Figure 5.5). In this library, soft corals (order Alcyonacea, phylum Cnidaria), which include *E. flexuosa*, represented 63% of metazoan abundance, while tunicates and bony fish represented 13% and 10% of abundance, respectively. Despite *Plexauridae* comprising 60% of cnidarian abundance, other soft coral families were also detected, including the *Ellisellidae*, *Nephtheidae*, *Acanthogorgiidae*, and *Nidaliidae*, each representing ~10% of cnidarian abundance. Likewise, the tunicate library from which CbtuFV was assembled

comprised reads belonging to various marine organisms, including Bryozoa, Cnidaria, and crustaceans, representing an average of 8% abundance each.

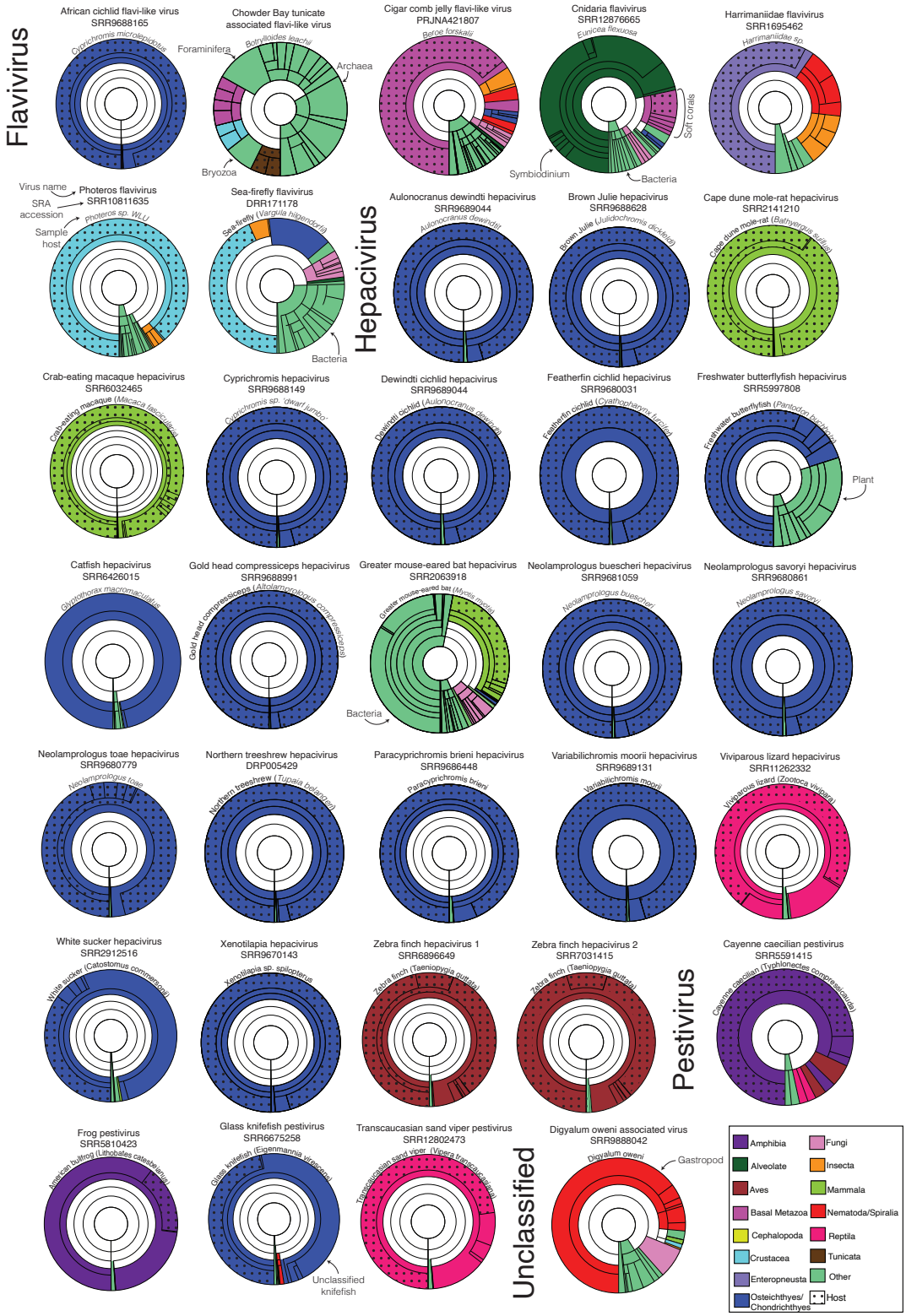



***Figure 36.5 Taxonomic assignments of contigs in sequencing libraries.*** *Each Krona graph illustrates the relative abundance of taxa in a metatranscriptome at varying taxonomic levels. For clarity, a maximum depth of five taxonomic levels was chosen for each graph.*

*The library Sequence Read Archive accession number, host species, and the corresponding virus of interest are annotated above each graph. Segments are highlighted based on the species' taxonomic grouping. Dots have been used to signify where contigs have been taxonomically assigned within the same family as the host species. Contigs without any matches in the database are not shown.*

Contigs belonging to catfish (order Siluriformes) comprised 95% of the *Glyptothorax macromaculatus* library from which catfish hepacivirus (CatfHV) was assembled, although it uncertain to which family of catfish this sample belonged. Likewise, the American bullfrog (*Lithobates catesbeianus)* transcriptome comprised 60% contigs associated with fork-tongued frogs (*Dicroglossidae*) and 17% associated with true frogs (*Ranidae*), including *L. catesbeianus*. No host-associated contigs were detected in the *Digyalum oweni* library in which DiowV was assembled. Instead, 64% of the library is composed of contigs associated with marine gastropod molluscs.

**5.4.6 Long-term virus-host evolutionary relationships**

To examine the frequency of four macroevolutionary events (i.e., co-divergence, duplication, host-switching, and extinction) among the *Flaviviridae*, we estimated cophylogenies to quantify the evolutionary relationship between the 'pegi-hepaci' and pestivirus clades and their hosts (Figure 5.6; members of the genus *Flavivirus,* were excluded because of the high frequency of vector-borne viruses). In accordance with earlier studies (Geoghegan et al. 2017), this analysis revealed that cross-species transmission was the most common evolutionary event across the 'pegi-hepaci' and pestivirus clades, representing 65% and 71% of events, respectively (Supplementary Figure 5.4). Two viruses, GmebHV and freshwater butterflyfish hepacivirus (FrbuHV) identified in this study, present notable exceptions (Figure 5.6). GmebHV is distinct from known bat hepaciviruses (Hepacivirus K, Hepacivirus L and Hepacivirus M), and instead groups with those viruses found in rodents, shrews, sloths, and raccoons (Figure 5.4). FrbuHV, along with Western African lungfish hepacivirus and Wenling moray eel hepacivirus, fell basal to those viruses identified in cartilaginous fish.

Importantly, despite the widespread occurrence of cross-species transmission, virus-host co-divergence was also predicted to have occurred relatively frequently across the 'pegi-hepaci' and *Pestivirus* clades, representing 22% and 23% of all events, respectively. For these groups, duplication events were more uncommon, representing 10% and 6% of total

events (Supplementary Figure 5.4). Extinction events were rarely predicted, representing 4% of events in the 'pegi-hepaci' clade, while no extinction events were detected in the Pestivirus cophylogeny.

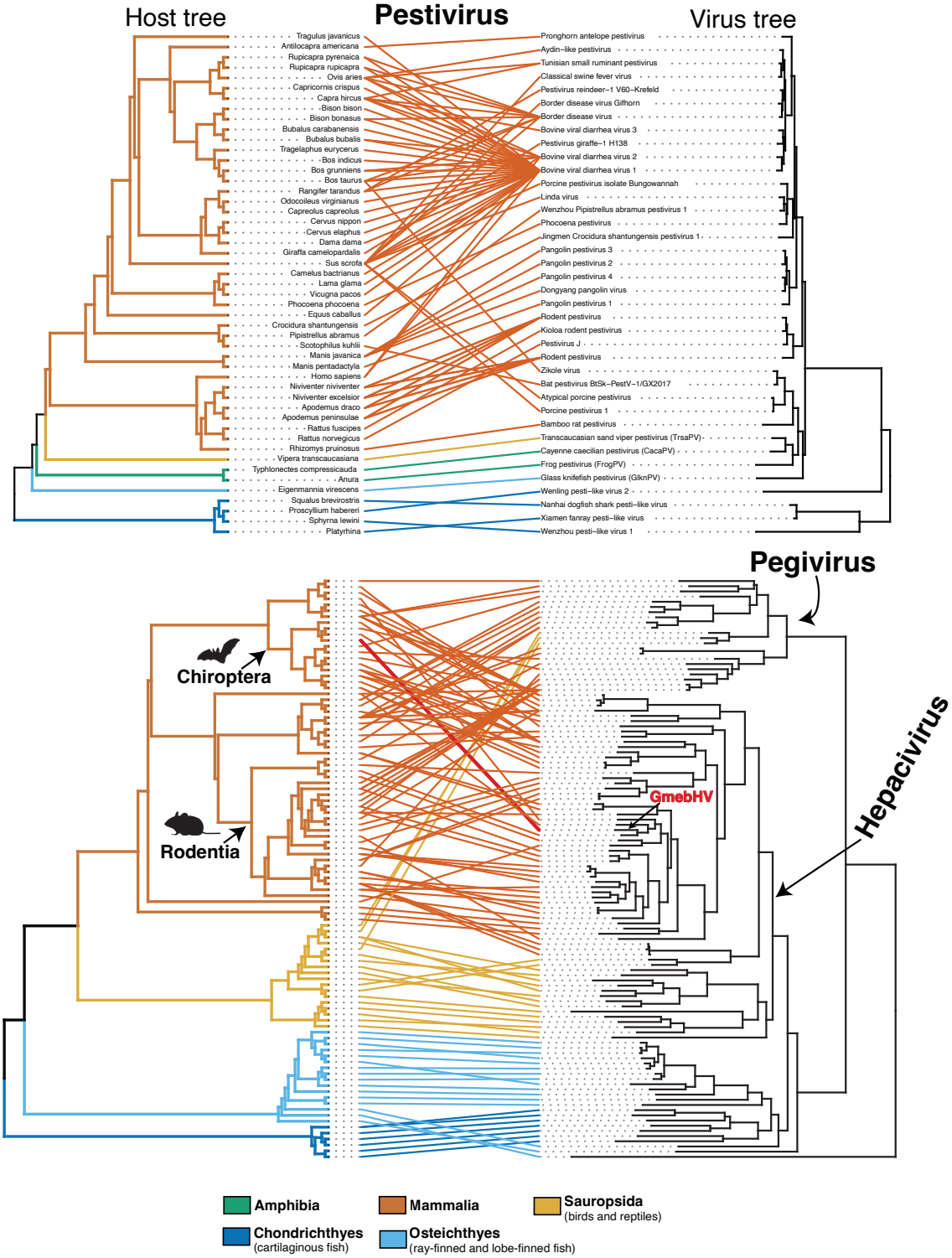



***Figure 5.37 Tanglegram of rooted phylogenetic trees for representative virus groups and their hosts***. *Branches of the host tree (left) and lines are coloured to represent the host*

*clade. The colour scheme is as found in Figure 5.2, with the following exceptions, Amphibia = green, Sauropsida = light orange. All branches on the virus tree are scaled to the number of amino acid substitutions per site, and both trees were midpoint rooted for clarity only. Greater mouse-eared bat hepacivirus (GmebHV) is highlighted in red. Images were obtained from* http://phylopic.org *under Public Domain Dedication. Supplementary Figure 5.5 provides the names of the hosts and viruses for the 'pegi-hepaci' cophylogeny.*

### 5.5 Discussion

Through transcriptome mining we identified 32 novel flavivirid sequences across the Metazoa, including the first flavivirus-like sequences in non-bilaterians, pestivirus-like sequences in amphibians, reptiles, and bony fish, as well as a range of vertebrate hepaciviruses. Hence, this work provides further evidence of the long-term associations between the *Flaviviridae* and Metazoa and highlights the vast number of viruses that remain undiscovered.

The Cnidaria are a primitive and basal phylum of Metazoa. Based on the identification of a flavivirus-like sequence in a cnidarian sample (CnidFV) we suggest that the origins of this group of viruses likely extends much further back in time than previous estimates and closer to the emergence of the metazoans 750–800 million years ago (Erwin 2015). This conclusion is supported by the earlier finding of a flavivirus-derived EVE in the Cnidaria (Bamford et al. 2022). Notably, CnidFV and the cnidrian EVE are more closely related to members of the genus *Flavivirus* than are the tamana/flavi-like viruses, suggesting that these groups, including the jingmenviruses, are evolutionarily distinct (Bamford et al. 2022). As such, we suggest that the tamana/flavi-like viruses should be given a distinct taxonomic classification within the *Flaviviridae*. However, it is clear difficult to fully resolve the evolutionary history of the flavivirids with our current understanding of their diversity, although it appears that the origins of this group lie in aquatic environments (Lensink et al. 2022).

It is important to note that host assignment of the non-bilaterian flaviviruses is tentative as these sequences are extremely divergent and they have only rarely been sampled. Due to the detection of several cnidarian species in addition to the target species, the octocoral *E. flexuosa* in library SRR12876665, we have assigned the resulting virus sequence as cnidaria flavivirus (CnidFV). The high abundance of Symbiodinium in this library is unsurprising given that octocoral-Symbiodinium mutualism is well known (van de Water et al. 2018). However,

the phylogenetic placement of this virus with those found in a marine acorn worm suggests it is more likely associated with *E. flexuosa* than Symbiodinium. While CnidFV and the EVE are relatively closely related to each other, there is substantial genetic divergence between these sequences. This has been previously observed with crocidura pestivirus and a Crocidura EVE and may reflect divergent evolution since the historic endogenization event (Li et al. 2022).

The discovery of Wenzhou pesti-like virus 1, Wenling pesti-like virus 2, Xiamen fanray pesti-like virus and Nanhai dogfish shark pesti-like virus in cartilaginous fish marked the expansion of the pestiviruses from warm-blooded mammals to basal vertebrate species, suggesting that these viruses infect a range of vertebrate lineages (Shi et al. 2018). For the first time, we identified pesti-like viruses in reptiles, amphibians, and bony fish, extending the host range of these viruses to encompass all vertebrate classes with the exception of Aves. The deep evolutionary association between pestiviruses and vertebrates is further reflected in the clear pattern of pestivirus-host codivergence among the viruses identified in this study. As a result, we anticipate that novel pestiviruses will be found infecting a wider diversity of vertebrates and that their known host range largely reflects where sampling efforts have been directed to date. Additionally, frog pestivirus was identified in the ventral skin of the American bullfrog, although other species of frog were detected in this library. Within the study in which this library was generated, the American bullfrog appeared resistant to the fungal pathogen *Batrachochytrium dendrobatidis* (Bd) (Eskew et al. 2018). Co-infection with Bd and ranaviruses is frequently observed in frogs, but whether the interactions between these pathogens are antagonistic or facilitative is currently unclear (Bosch et al. 2020). If Bd and pestiviruses are found to commonly co-infect frogs, future efforts should be directed towards studying their interactions.

We identified 20 novel hepacivirus sequences, among which a clade of cichlid-associated hepacivirus sequences is notable. This clade was derived from a study of Lake Tanganyika, a freshwater lake shared by Tanzania, the Democratic Republic of the Congo, Burundi, and Zambia that is known for its high diversity of endemic cichlid species (El Taher et al. 2021; Koblmüller et al. 2008). Importantly, the fish and reptile hepaciviruses identified in this study were predominately associated with samples of liver tissue, suggesting hepatotropism is likely a universal feature of these viruses across vertebrates (Smith et al. 2016).

Bats and rodents harbour a large diversity of hepaciviruses and are thought to have played an important role in their global spread and broader evolutionary history (Bletsa et al. 2021; de Souza et al. 2019; Drexler et al. 2013; Epstein et al. 2010; Kapoor et al. 2013; Quan et al. 2013). We identified GmebHV, which falls within a clade of rodent, sloth, and raccoon hepaciviruses. The clear relatedness between GmebHV and rodent hepacivirus (QLM02863), combined with evidence from our co-evolutionary analyses, suggests that this sequence might represent a cross-species transmission event between bats and rodents. Similarly, ancestral state reconstructions have previously shown that cross-species transmission from rodents is likely the source of the sloth and ringtail hepaciviruses (Jo et al. 2022; Moreira-Soto et al. 2020). In this case, we cannot resolve the direction of virus transmission with any certainty or whether other species are involved.

In very broad terms, we find that the hepaci-, pesti-, and pegiviruses cluster with the phylogeny of their hosts, with the relevant frequent cross-species virus transmission events only occurring within host classes (i.e., Mammalia, Sauropsida and Chondrichthyes). The exceptions were FrbuHV, Western African lungfish hepacivirus, and Wenling moray eel hepacivirus that fell basal to those viruses identified in cartilaginous fish (although the phylogenetic position of these viruses should be treated with caution as the relevant nodes have weak bootstrap support; Figure 4, Figure 6). The clear phylogenetically defined barriers between host classes may reflect differences in receptor binding and cell entry mechanisms among distantly related hosts (Parrish et al. 2008). Host ecology also likely contributes to these barriers, particularly as physical separation means that fewer cross-species virus transmission events are expected to occur between marine and land vertebrates than among land vertebrates (Luis et al. 2015). Together, this suggests that more cross-species transmission occurs among closely related hosts, which may have also resulted in the apparent loss of the signal of co-divergence signal within relatively well-sampled taxonomic groups (e.g., mammals). At deeper taxonomic levels, we found clear evidence for virus-host co-divergence, particularly in lower vertebrates, which is consistent with previous findings (Geoghegan et al. 2017; Hartlage et al. 2016; Porter et al. 2020; Shi et al. 2018). However, it is also apparent that the results of our cophylogenetic analysis are influenced by the sample of virus diversity used and will likely change as more viruses are identified. In addition, virus phylogenies were estimated using the RdRp (NS5) alone. It is possible that differences in the phylogenies of the entire polyprotein or NS3 region would produce different estimates of the frequencies of co-divergence and host jumping.

Wenling moray eel hepacivirus (AVM87555) forms a sister group to the 'pegi-hepaci' lineage, although this may be artifactual due to recombination or extreme rate variation (Porter et al. 2020). If the position of Wenling moray eel hepacivirus is correct, this suggests that a common ancestor of the 'pegi-hepaci' lineage may have existed in an aquatic environment. This notion is supported by the recent finding of 'pegi-hepaci' derived EVE in a marine mollusc (Bamford et al. 2022). The apparent lack of pegiviruses in aquatic vertebrate and invertebrate species in this study does not equate to their absence in these organisms due to the current depth of SRA libraries available.

Another notable observation from this study was the identification of a flavivirus in non-bilaterians, which raises additional questions on the ancestral mode of flavivirus transmission. Non-bilaterians lack the circulatory system of vertebrates, suggesting that an alternative mode of cell-to-cell virus transmission may exist in these animals (Bamford et al. 2022).

In sum, through a broad-scale survey of publicly available and newly generated transcriptome data, we revealed a wide diversity of flavivirid sequences in undersampled metazoan species. In doing so, we provide additional information for an ancient origin of the flaviviruses, likely closer to the emergence of the metazoans some 750–800 million years ago, and hence for the long-term association between the *Flaviviridae* and the Metazoa as a whole.

### 5.6 Data availability

All tunicate sequence reads are available on the NCBI Sequence Read Archive (SRA) under BioProject PRJEB57836. All viral genomes and corresponding sequences assembled in this study have been deposited in the European Nucleotide Archive at EMBL-EBI under BioProject PRJEB57836. The sequences, alignments, phylogenetic trees and the custom *Flaviviridae* database generated in this study is available at https://github.com/JonathonMifsud/Transcriptome_mining_extends_the_host_range_of_the_Flaviviridae_to_non-bilaterians

### 5.7 Acknowledgements

We thank A.H. McGrath for assistance in field sampling. We acknowledge the University of Sydney's high-performance computing cluster Artemis for providing the computing

resources used for this study. This work would not be possible without the sequencing data that has been generously shared and we are grateful to the NCBI SRA team for their management of this important resource. E.C.H. is supported by an NHMRC Australian Laureate Fellowship (FL170100022). J.C.O.M. is supported by a Research Training Program (RTP) Scholarship. E.M.M. received funding from the Australian Research Council (DP180104041).

### 5.8 References


Babaian, A. and Edgar, R. (2022), 'Ribovirus classification by a polymerase barcode sequence', *PeerJ,* 10, e14055.

Bamford, C. G., et al. (2022), 'Comparative analysis of genome-encoded viral sequences reveals the evolutionary history of flavivirids (family Flaviviridae)', *Virus Evolution,* 8 (2).

Bessho-Uehara, M., et al. (2020), 'Kleptoprotein bioluminescence: Parapriacanthus fish obtain luciferase from ostracod prey', *Sci Adv,* 6 (2), eaax4942.

Biederman, M. K., et al. (2018), 'Discovery of the First Germline-Restricted Gene by Subtractive Transcriptomic Analysis in the Zebra Finch, Taeniopygia guttata', *Current Biology,* 28 (10), 1620-27.e5.

Bletsa, M., et al. (2021), 'Molecular detection and genomic characterization of diverse hepaciviruses in African rodents', *Virus Evolution,* 7 (1).

Blitvich, B. J. and Firth, A. E. (2017), 'A review of flaviviruses that have no known arthropod vector', *Viruses,* 9 (6), 154.

Bolger, A. M., Lohse, M., and Usadel, B. (2014), 'Trimmomatic: a flexible trimmer for Illumina sequence data', *Bioinformatics,* 30 (15), 2114-20.

Bosch, J., et al. (2020), 'Single infection with Batrachochytrium dendrobatidis or Ranavirus does not increase probability of co-infection in a montane community of amphibians', *Scientific Reports,* 10 (1), 21115.

Brister, J. R., et al. (2015), 'NCBI viral genomes resource', *Nucleic Acids Research,* 43, 571-7.

Buchfink, B., Xie, C., and Huson, D. H. (2015), 'Fast and sensitive protein alignment using DIAMOND', *Nature Methods,* 12 (1), 59-60.

Bushnell, B. (2014), 'BBMap: a fast, accurate, splice-aware aligner', (Lawrence Berkeley National Lab (LBNL), Berkeley, CA.).

Cannon, Johanna T., et al. (2014), 'Phylogenomic Resolution of the Hemichordate and Echinoderm Clade', *Current Biology,* 24 (23), 2827-32.

Capella-Gutiérrez, S., Silla-Martínez, J. M., and Gabaldón, T. (2009), 'trimAl: a tool for automated alignment trimming in large-scale phylogenetic analyses', *Bioinformatics,* 25 (15), 1972-73.

Chang, W.-S., Rose, K., and 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.

Clausen, P. T., Aarestrup, F. M., and Lund, O. (2018), 'Rapid and precise alignment of raw reads against redundant databases with KMA', *BMC Bioinformatics,* 19 (1), 1-8.

Conow, C., et al. (2010), 'Jane: a new tool for the cophylogeny reconstruction problem', *Algorithms for Molecular Biology,* 5 (1), 1-10.

Costa, V. A., et al. (2021), 'Metagenomic sequencing reveals a lack of virus exchange between native and invasive freshwater fish across the Murray–Darling Basin, Australia', *Virus Evolution,* 7 (1).

Costa, V. A., et al. (2022), 'Limited Cross-Species Virus Transmission in a Spatially Restricted Coral Reef Fish Community', *bioRxiv*, 2022.05.17.492384.

Davies, K. T., et al. (2015), 'Family Wide Molecular Adaptations to Underground Life in African Mole-Rats Revealed by Phylogenomic Analysis', *Molecular Biology and Evolution,* 32 (12), 3089-107.

de Souza, W. M., et al. (2019), 'A Novel Hepacivirus in Wild Rodents from South America', *Viruses,* 11 (3), 297.

Dheilly, N. M., et al. (2022), 'A World of Viruses Nested within Parasites: Unraveling Viral Diversity within Parasitic Flatworms (Platyhelminthes)', *Microbiology Spectrum,* 10 (3), e00138-22.

Drexler, J. F., et al. (2013), 'Evidence for novel hepaciviruses in rodents', *PLoS pathogens,* 9 (6), e1003438.

Edgar, R. C., et al. (2022), 'Petabase-scale sequence alignment catalyses viral discovery', *Nature,* 602 (7895), 142-47.

El Taher, A., et al. (2021), 'Gene expression dynamics during rapid organismal diversification in African cichlid fishes', *Nature Ecology Evolution,* 5 (2), 243-50.

Epstein, J. H., et al. (2010), 'Identification of GBV-D, a novel GB-like flavivirus from old world frugivorous bats (Pteropus giganteus) in Bangladesh', *PLoS pathogens,* 6 (7), e1000972.

Erwin, D. H. (2015), 'Early metazoan life: divergence, environment and ecology', *Philosophical Transactions of the Royal Society B: Biological Sciences,* 370 (1684), 20150036.

Eskew, E. A., et al. (2018), 'Gene expression differs in susceptible and resistant amphibians exposed to Batrachochytrium dendrobatidis', *Royal Society Open Science,* 5 (2), 170910.

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 (2), e1006215.

Geoghegan, J. L., et al. (2018), 'Hidden diversity and evolution of viruses in market fish', *Virus Evolution,* 4 (2).

Geoghegan, J. L., et al. (2021), 'Virome composition in marine fish revealed by meta-transcriptomics', *Virus Evolution,* 7 (1).

Goldberg, T. L., et al. (2019), 'Multidecade mortality and a homolog of hepatitis C virus in bald eagles (Haliaeetus leucocephalus), the National Bird of the USA', *Scientific Reports,* 9 (1), 1-12.

Greninger, A. L. (2018), 'A decade of RNA virus metagenomics is (not) enough', *Virus Research,* 244, 218-29.

Grimwood, R. M., Holmes, E. C., and Geoghegan, J. L. (2021), 'A Novel Rubi-Like Virus in the Pacific Electric Ray (Tetronarce californica) Reveals the Complex Evolutionary History of the Matonaviridae', *Viruses,* 13 (4).

Guindon, S., et al. (2010), 'New Algorithms and Methods to Estimate Maximum-Likelihood Phylogenies: Assessing the Performance of PhyML 3.0', *Systematic Biology,* 59 (3), 307-21.

Hackl, T., et al. (2024), 'gggenomes: effective and versatile visualizations for comparative genomics', *arXiv preprint arXiv:2411.13556*.

Hahn, C. M., et al. (2016), 'Transcriptome discovery in non-model wild fish species for the development of quantitative transcript abundance assays', *Comparative Biochemistry and Physiology Part D: Genomics and Proteomics,* 20, 27-40.

Harding, E. F., et al. (2022), 'Revealing the uncharacterised diversity of amphibian and reptile viruses', *ISME Communications,* 2 (1), 95.

Hartlage, A. S., Cullen, J. M., and Kapoor, A. (2016), 'The Strange, Expanding World of Animal Hepaciviruses', *Annual Review of Virology,* 3 (1), 53-75.

Hatcher, E. L., et al. (2017), 'Virus Variation Resource–improved response to emergent viral outbreaks', *Nucleic Acids Research,* 45 (D1), D482-D90.

Hensley, N. M., et al. (2019), 'Phenotypic evolution shaped by current enzyme function in the bioluminescent courtship signals of sea fireflies', *Proceedings of the Royal Society B,* 286 (1894), 20182621.

Hoang, D. T., et al. (2017), 'UFBoot2: Improving the Ultrafast Bootstrap Approximation', *Molecular Biology and Evolution,* 35 (2), 518-22.

Iwamoto, M., et al. (2021), 'Identification of novel avian and mammalian deltaviruses provides new insights into deltavirus evolution', *Virus Evolution,* 7 (1), veab003.

Janouškovec, J., et al. (2019), 'Apicomplexan-like parasites are polyphyletic and widely but selectively dependent on cryptic plastid organelles', *Elife,* 8.

Jiang, W., et al. (2019), 'Insights into Body Size Evolution: A Comparative Transcriptome Study on Three Species of Asian Sisoridae Catfish', *International Journal of Molecular Sciences,* 20 (4).

Jo, W. K., et al. (2022), 'Natural co-infection of divergent hepatitis B and C virus homologues in carnivores', *Transboundary and Emerging Diseases,* 69 (2), 195-203.

Jones, P., et al. (2014), 'InterProScan 5: genome-scale protein function classification', *Bioinformatics,* 30 (9), 1236-40.

Kalyaanamoorthy, S., et al. (2017), 'ModelFinder: fast model selection for accurate phylogenetic estimates', *Nature Methods,* 14 (6), 587-89.

Kapoor, A., et al. (2013), 'Identification of Rodent Homologs of Hepatitis C Virus and Pegiviruses', *MBio,* 4 (2), e00216-13.

Katoh, K. and Standley, D. M. (2013), 'MAFFT multiple sequence alignment software version 7: improvements in performance and usability', *Molecular Biology and Evolution,* 30 (4), 772-80.

Kearse, M., et al. (2012), 'Geneious Basic: an integrated and extendable desktop software platform for the organization and analysis of sequence data', *Bioinformatics,* 28 (12), 1647-49.

Ladner, J. T., et al. (2016), 'A Multicomponent Animal Virus Isolated from Mosquitoes', *Cell Host & Microbe,* 20 (3), 357-67.

Le Lay, C., et al. (2020), 'Unmapped RNA virus diversity in termites and their symbionts', *Viruses,* 12 (10), 1145.

Lensink, M. J., Li, Y., and Lequime, S. (2022), 'Aquatic Flaviviruses', *Journal of Virology,* 96 (17), e00439-22.

Li, B. and Dewey, C. N. (2011), 'RSEM: accurate transcript quantification from RNA-Seq data with or without a reference genome', *BMC Bioinformatics,* 12 (1), 323.

Li, D., et al. (2015), 'MEGAHIT: an ultra-fast single-node solution for large and complex metagenomics assembly via succinct de Bruijn graph', *Bioinformatics,* 31 (10), 1674-76.

Li, Y., et al. (2022), 'Endogenous Viral Elements in Shrew Genomes Provide Insights into Pestivirus Ancient History', *Molecular Biology and Evolution,* 39 (10), msac190.

Luis, A. D., et al. (2015), 'Network analysis of host–virus communities in bats and rodents reveals determinants of cross-species transmission', *Ecology Letters,* 18 (11), 1153-62.

Marcelino, V. R., et al. (2020), 'CCMetagen: comprehensive and accurate identification of eukaryotes and prokaryotes in metagenomic data', *Genome Biology,* 21 (1), 103.

Marzinelli, E. M. (2012), 'Artificial structures influence fouling on habitat-forming kelps', *Biofouling,* 28 (3), 339-49.

Mifsud, J. C. O., et al. (2022), 'Transcriptome Mining Expands Knowledge of RNA Viruses across the Plant Kingdom', *Journal of Virology,* 0 (0), e00260-22.

Miller, A. K., et al. (2021), 'Slippery when wet: cross-species transmission of divergent coronaviruses in bony and jawless fish and the evolutionary history of the Coronaviridae', *Virus Evolution,* 7 (2).

Minh, B. Q., et al. (2020), 'IQ-TREE 2: new models and efficient methods for phylogenetic inference in the genomic era', *Molecular Biology and Evolution,* 37 (5), 1530-34.

Moreira-Soto, A., et al. (2020), 'Cross-order host switches of hepatitis C-related viruses illustrated by a novel hepacivirus from sloths', *Virus Evolution,* 6 (2), veaa033.

Nayfach, S., et al. (2021), 'CheckV assesses the quality and completeness of metagenome-assembled viral genomes', *Nature Biotechnology,* 39 (5), 578-85.

Olendraite, I., Brown, K., and Firth, A. E. (2022), 'Identification of RNA virus-derived RdRp sequences in publicly available transcriptomic datasets', *bioRxiv*, 2022.10.18.512700.

Paradis, E. and Schliep, K. (2019), 'ape 5.0: an environment for modern phylogenetics and evolutionary analyses in R', *Bioinformatics,* 35 (3), 526-28.

Paraskevopoulou, S., et al. (2021), 'Viromics of extant insect orders unveil the evolution of the flavi-like superfamily', *Virus Evolution,* 7 (1).

Parrish, C. R., et al. (2008), 'Cross-species virus transmission and the emergence of new epidemic diseases', *Microbiology and Molecular Biology Reviews,* 72 (3), 457-70.

Parry, R., Asgari, S., and Williams, B. R. G. (2019), 'Discovery of Novel Crustacean and Cephalopod Flaviviruses: Insights into the Evolution and Circulation of Flaviviruses between Marine Invertebrate and Vertebrate Hosts', *Journal of Virology,* 93 (14), e00432-19.

Porter, A. F., et al. (2020), 'Novel hepaci- and pegi-like viruses in native Australian wildlife and non-human primates', *Virus Evolution,* 6 (2).

Prada, C. and Hellberg, M. E. (2021), 'Speciation-by-depth on coral reefs: Sympatric divergence with gene flow or cryptic transient isolation?', *Journal of Evolutionary Biology,* 34 (1), 128-37.

Price, J. L. (1978), 'Isolation of Rio Bravo and a hitherto undescribed agent, Tamana bat virus, from insectivorous bats in Trinidad, with serological evidence of infection in bats and man', *American Journal of Tropical Medicine and Hygiene,* 27 (1), 153-61.

Qin, X.-C., et al. (2014), 'A tick-borne segmented RNA virus contains genome segments derived from unsegmented viral ancestors', *Proceedings of the National Academy of Sciences,* 111 (18), 6744-49.

Quan, P.-L., et al. (2013), 'Bats are a major natural reservoir for hepaciviruses and pegiviruses', *Proceedings of the National Academy of Sciences,* 110 (20), 8194-99.

Revell, L. J. (2012), 'phytools: an R package for phylogenetic comparative biology (and other things)', *Methods in Ecology and Evolution,* 3 (2), 217-23.

Sanada, T., et al. (2019), 'Construction of complete Tupaia belangeri transcriptome database by whole-genome and comprehensive RNA sequencing', *Scientific Reports,* 9 (1), 12372.

Santichaivekin, S., et al. (2020), 'eMPRess: a systematic cophylogeny reconciliation tool', *Bioinformatics,* 16 (37), 2481–82.

Sayers, E. W., et al. (2021), 'GenBank', *Nucleic Acids Research,* 49 (D1), D92-D96.

Schultz, D. T., et al. (2020), 'Conserved novel ORFs in the mitochondrial genome of the ctenophore Beroe forskalii', *PeerJ,* 8, e8356.

Shi, M., et al. (2016), 'Divergent viruses discovered in arthropods and vertebrates revise the evolutionary history of the Flaviviridae and related viruses', *Journal of Virology,* 90 (2), 659-69.

Shi, M., et al. (2018), 'The evolutionary history of vertebrate RNA viruses', *Nature,* 556 (7700), 197.

Simmonds, P., et al. (2017), 'ICTV virus taxonomy profile: Flaviviridae', *The Journal of General Virology,* 98 (1), 2.

Skoge, R. H., et al. (2018), 'New virus of the family Flaviviridae detected in lumpfish (Cyclopterus lumpus)', *Archives of Virology,* 163 (3), 679-85.

Smith, D. B., et al. (2016), 'Proposed update to the taxonomy of the genera Hepacivirus and Pegivirus within the Flaviviridae family', *The Journal of General Virology,* 97 (11), 2894.

Soto, E., et al. (2020), 'First Isolation of a Novel Aquatic Flavivirus from Chinook Salmon (Oncorhynchus tshawytscha) and Its In Vivo Replication in a Piscine Animal Model', *Journal of Virology,* 94 (15).

Speranza, E., et al. (2018), 'A conserved transcriptional response to intranasal Ebola virus exposure in nonhuman primates prior to onset of fever', *Science Translational Medicine,* 10 (434), eaaq1016.

Sun, Y., et al. (2016), 'Fish-T1K (Transcriptomes of 1,000 Fishes) Project: large-scale transcriptome data for fish evolution studies', *GigaScience,* 5 (1).

Thompson, A., et al. (2018), 'Rapid evolution of a voltage-gated sodium channel gene in a lineage of electric fish leads to a persistent sodium current', *PLoS Biology,* 16 (3), e2004892.

Torres-Sánchez, M., et al. (2019), 'Multi-tissue transcriptomes of caecilian amphibians highlight incomplete knowledge of vertebrate gene families', *DNA Research,* 26 (1), 13-20.

Wu, Z., et al. (2012), 'Virome Analysis for Identification of Novel Mammalian Viruses in Bat Species from Chinese Provinces', *Journal of Virology,* 86 (20), 10999-1012.

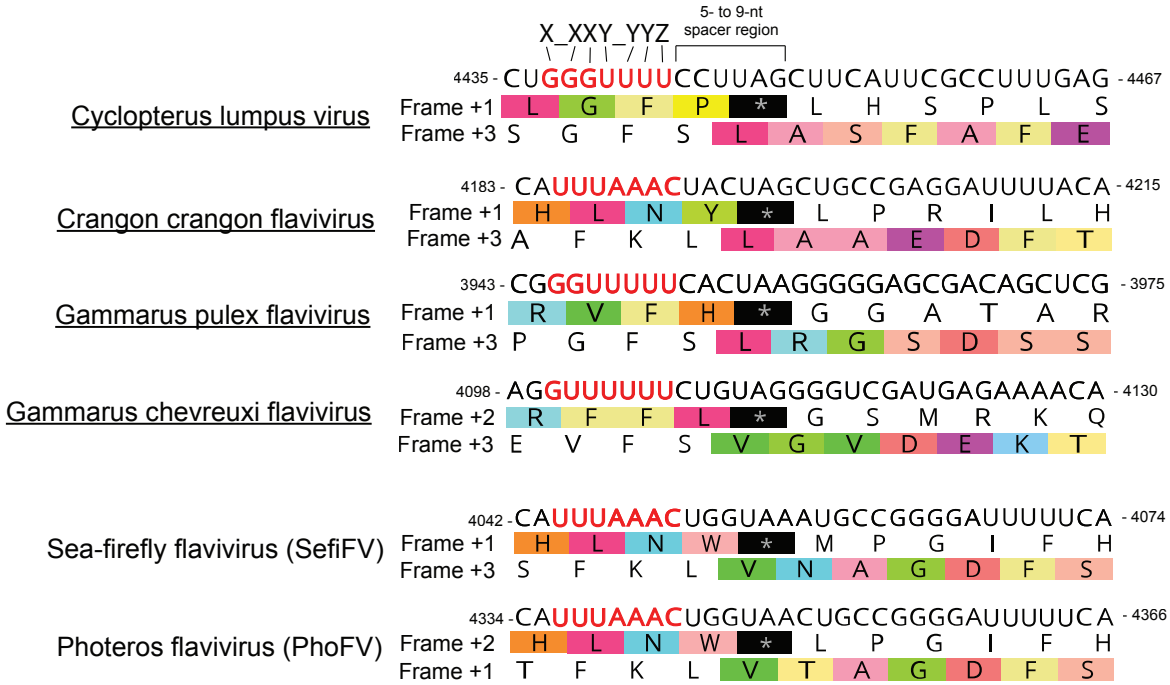
Xie, B., et al. (2022), 'Dynamic genetic differentiation drives the widespread structural and functional convergent evolution of snake venom proteinaceous toxins', *BMC Biology,* 20 (1), 4.

Yu, G., et al. (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.

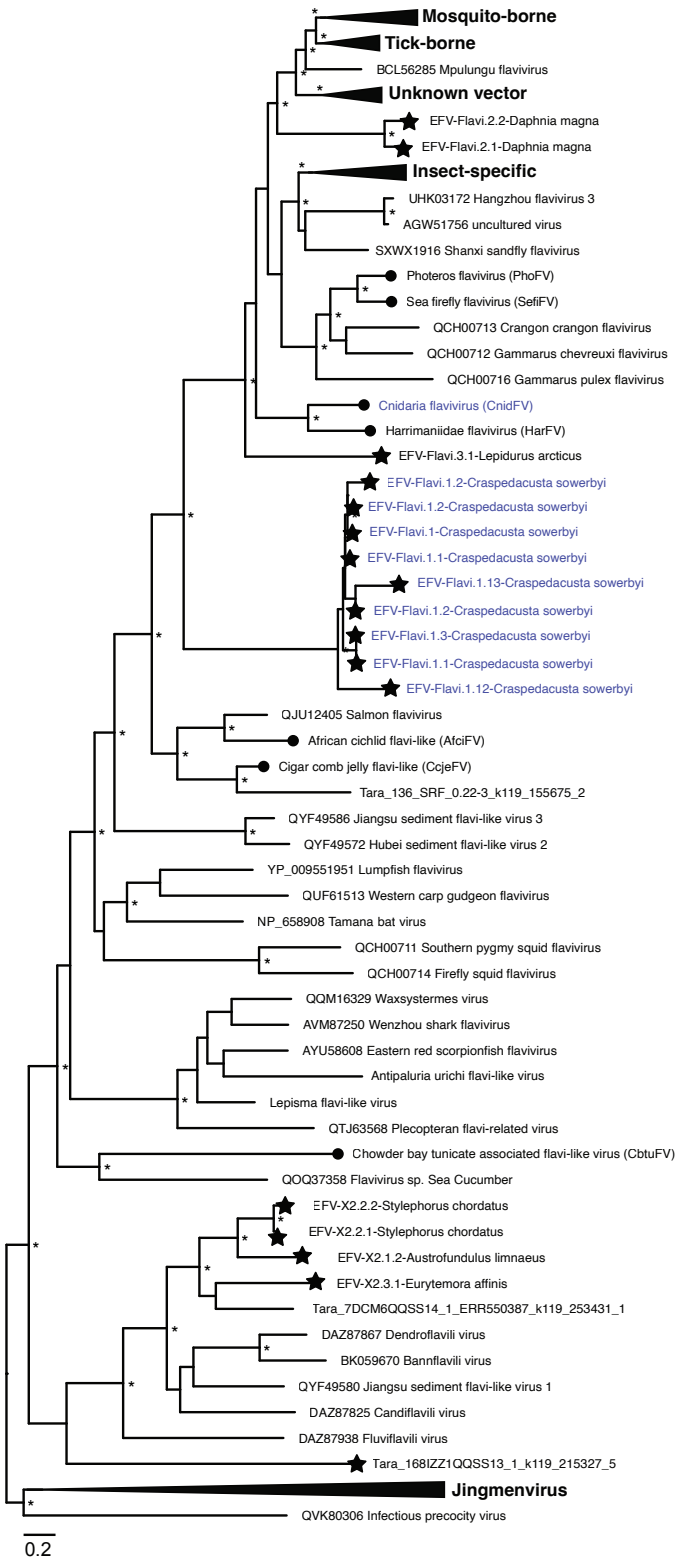
Yurchenko, A. A., Recknagel, H., and Elmer, K. R. (2020), 'Chromosome-Level Assembly of the Common Lizard (Zootoca vivipara) Genome', *Genome Biology and Evolution,* 12 (11), 1953-60.

Zhang, X.-L., et al. (2022), 'A Highly Divergent Hepacivirus Identified in Domestic Ducks Further Reveals the Genetic Diversity of Hepaciviruses', *Viruses,* 14 (2), 371.

## 5.9 Supplementary Figures

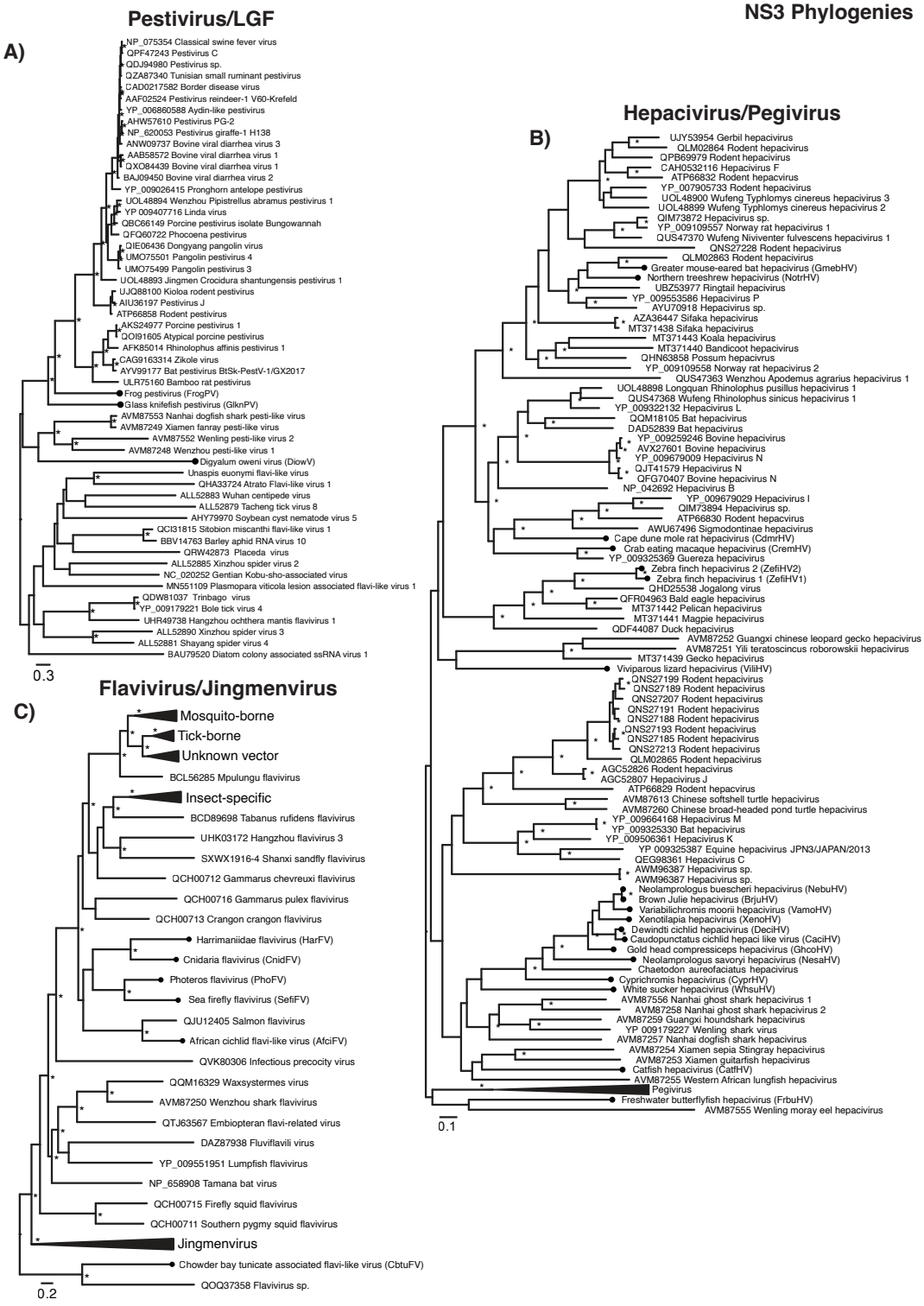



***Supplementary Figure 5.1. Alignment of the predicted structured programmed ribosomal frameshift (–1 PRF) regions identified with reference sequences.*** *Slippery heptanucleotide regions are highlighted in red. Predicted translated protein regions are indicated in colour.*



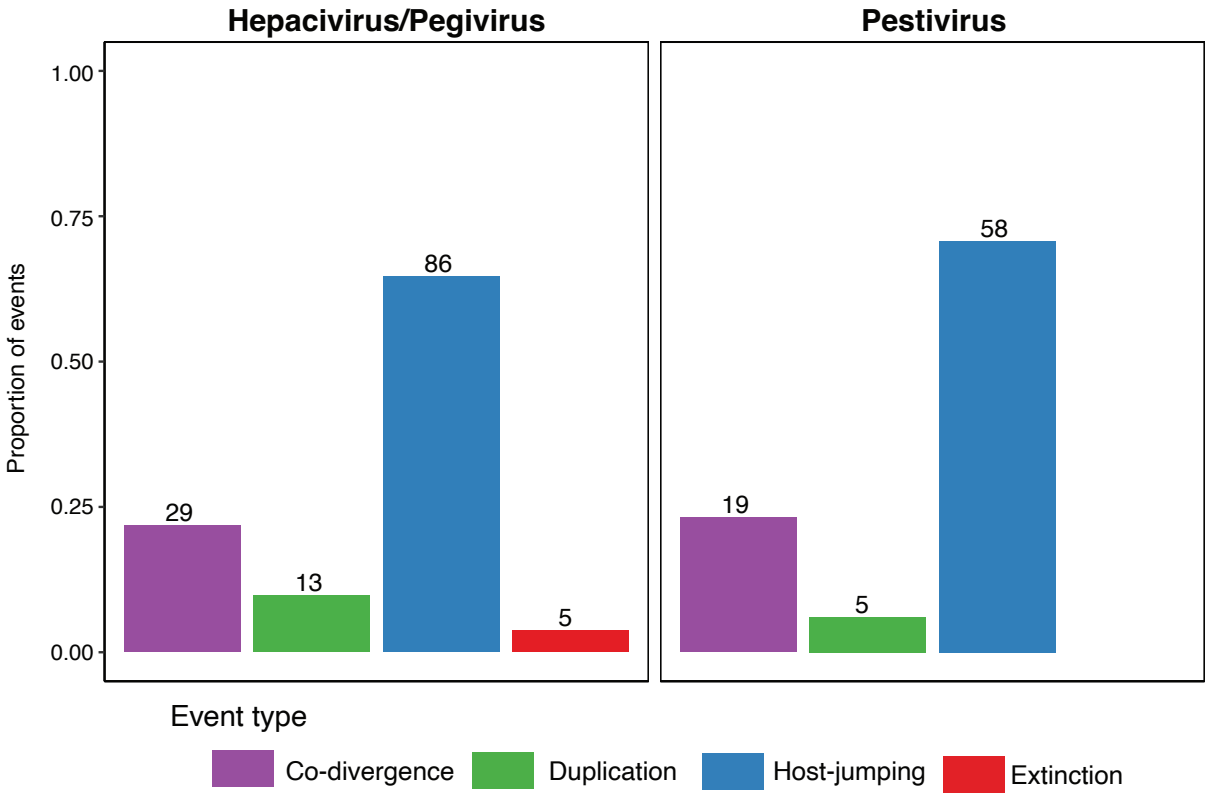

***Supplementary Figure 5.2. Phylogenetic relationships of the flavi-like viruses and flavivirus-derived EVEs.*** *Phylogenetic relationships of the 'flavi-jingmen' clade including the EVE identified in Bamford et al. (2022) (black stars). ML phylogenetic trees based on the conserved amino acid in the NS5 region show the topological position of virus-like sequences discovered in this study (black circles) in the context of their closest relatives. All*

*branches are scaled to the number of amino acid substitutions per site, and trees were midpoint rooted for clarity only. An asterisk indicates node support where SH-aLRT >= 80% and UFboot >= 95%.*

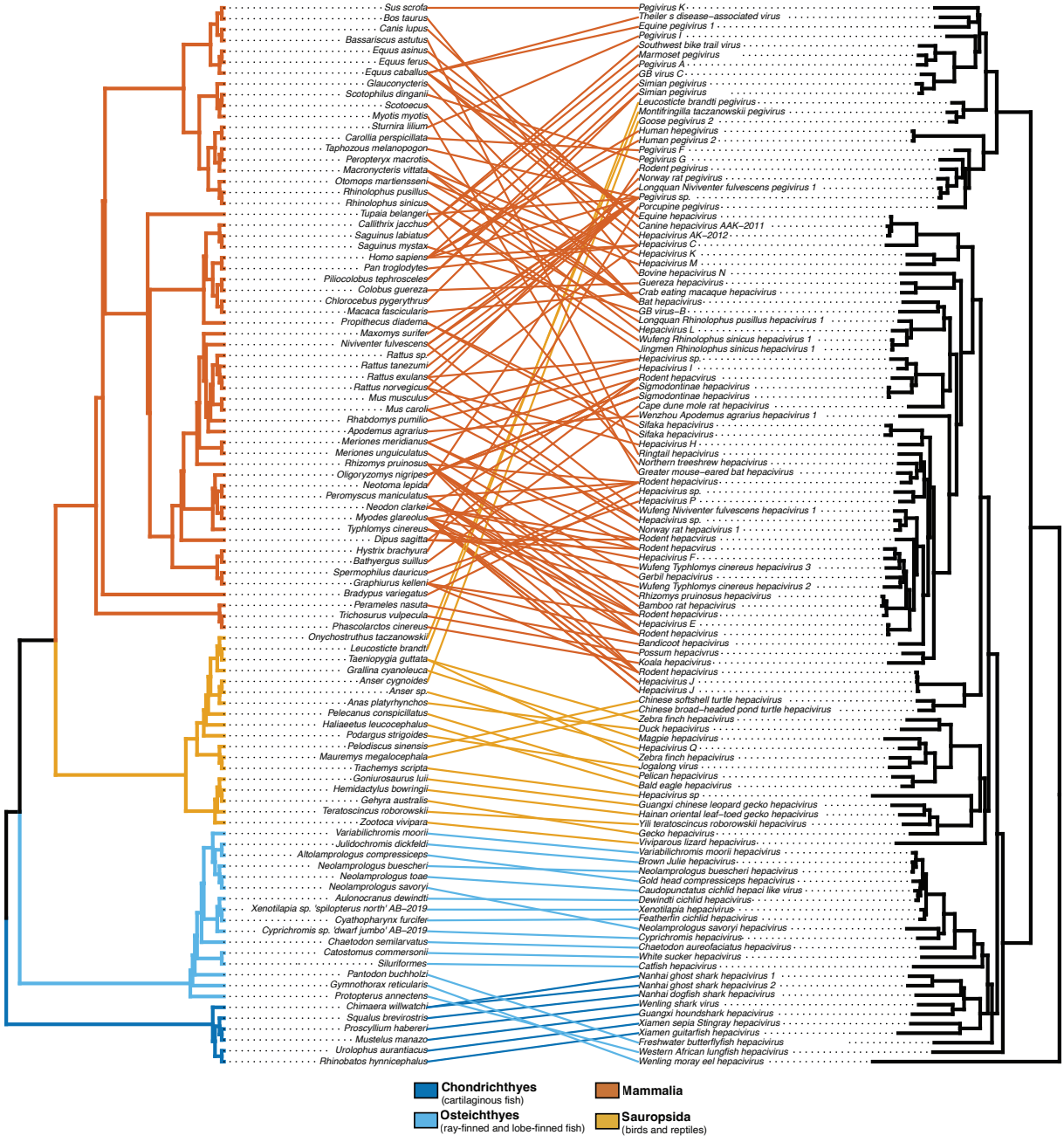



***Supplementary Figure 5.3. Alternative phylograms based on the conserved NS3 region.***

*Phylogenetic relationships of the (a) 'pesti-LGF' clades, (b) 'pegi-hepaci', and (c) 'flavi-jingmen'. ML phylogenetic trees based on the conserved amino acid in the NS3 region show the topological position of virus-like sequences discovered in this study (black circles) in the context of their closest relatives. All branches are scaled to the number of amino acid substitutions per site, and trees were midpoint rooted for clarity only. An asterisk indicates node support where SH-aLRT >= 80% and UFboot >= 95%.*




***Supplementary Figure 5.4. Reconciliation analysis.*** *Reconciliation analysis of select virus groups. Bar plots illustrate the range of the proportion of possible events and are coloured by event type.*




***Supplementary Figure 5.5. Tanglegram of the 'pegi-hepaci' clade and their hosts with species labels present.*** *Branches of the host tree (left) and lines are coloured to represent the host clade. All branches on the virus tree are scaled to the number of amino acid substitutions per site, and both trees were midpoint rooted for clarity only.*

The following Supplementary tables are available as .xlsx files:
https://oup.silverchair-cdn.com/oup/backfile/Content_public/Journal/ve/9/1/10.1093_ve_veac124/1/veac124_sup

p.zip?Expires=1735725726&Signature=TrwQsLj~pnII56AqO-nno7liQ5C5rP7nC4SmxjmxGIxjhITyoSK9NPjDACtwhKjfBTNFW95ZJB4VVO5x6uecqeQL4MMZm-DoW03FR-MOEWz1P5uzHksV90H2wUGPoWbk4uaPIgWcd7h8nlOSMjHG5rzT9ieGpEZ6o6P3qpD5RHwikxgW730E8tzgAjKjzrtyPcoONzV~Gyu~ywzliyS8dOmQRFsMnkVTA0uYYkUyyqm6zCQDtKJ82beHG~edu6gPbtcqOIrKnGNG~ow8nYvGjQKuQWYejfRd3p1A4dO6IMW0DVM1jqx6gCxrticvpSE3a6jU0b-vGTJog4r52x43XA__&Key-Pair-Id=APKAIE5G5CRDK6RD3PGA

**Supplementary Table 1.** (A) GenBank accessions of the sequences used as queries for the TSA and Serratus palmID searches. (B) Trim and alignments parameters used to construct phylogenetic trees.

**Supplementary Table 2.** Genome annotation data and acronyms underlying the annotation graphs.

# Chapter 6. Mapping glycoprotein structure reveals *Flaviviridae* evolutionary history

Jonathon C.O. Mifsud[1], Spyros Lytras[2,3], Michael R. Oliver[2], Kamilla Toon[2], Vincenzo A. Costa[1], Edward C. Holmes[1,4], Joe Grove[2]

1. Sydney Institute for Infectious Diseases, School of Medical Sciences, The University of Sydney, Sydney, NSW 2006, Australia.
2. MRC-University of Glasgow Centre for Virus Research, Glasgow, G61 1QH, UK.
3. Division of Systems Virology, Department of Microbiology and Immunology, The Institute of Medical Science, The University of Tokyo, Tokyo, Japan
4. Laboratory of Data Discovery for Health Limited, Hong Kong SAR, China.

## 6.1 Abstract

Viral glycoproteins drive membrane fusion in enveloped viruses and determine host range, tissue tropism and pathogenesis (Grove and Marsh 2011). Despite their importance, there is a fragmentary understanding of glycoproteins within the *Flaviviridae* (Simmonds et al. 2017), a large virus family that include pathogens such as hepatitis C, dengue and Zika viruses, and numerous other human, animal and emergent viruses. For many flaviviruses the glycoproteins have not yet been identified, for others, such as the hepaciviruses, the molecular mechanisms of membrane fusion remain uncharacterised (Rey and Lok 2018). Here we combine phylogenetic analyses with protein structure prediction to survey glycoproteins across the entire *Flaviviridae*. We find class II fusion systems, homologous to the Orthoflavivirus E glycoprotein in most species, including highly divergent jingmenviruses and large genome flaviviruses. However, the E1E2 glycoproteins of the hepaciviruses, pegiviruses and pestiviruses are structurally distinct, may represent a novel class of fusion mechanism, and are strictly associated with infection of vertebrate hosts. By mapping glycoprotein distribution onto the underlying phylogeny, we reveal a complex evolutionary history marked by the capture of bacterial genes and potentially inter-genus recombination. These insights, made possible through protein structure prediction, refine our understanding of viral fusion mechanisms and reveal the events that have shaped the diverse virology and ecology of the *Flaviviridae*.

## 6.2 Introduction

The *Flaviviridae* (Simmonds et al. 2017) is a highly diverse family of enveloped positive-sense RNA viruses that includes important pathogens of humans (for example, dengue virus, Zika virus and hepatitis C virus) and other animals (for example, classical swine fever virus and bovine viral diarrhoea virus), as well as many viruses that pose emerging threats to human health (for example, West Nile virus, Alongshan virus and Haseki tick virus (Hubálek and Halouzka 1999; Kartashov et al. 2023; Wang et al. 2019). The *Flaviviridae* is currently classified into four genera: Orthoflavivirus, Pestivirus, Pegivirus and Hepacivirus (Postler et al. 2023). In recent years a remarkable diversity of novel flaviviruses with varied genome structures have been discovered, including the jingmenvirus group, which are unique in being both segmented and potentially multicomponent (Ladner et al. 2016; Qin et al. 2014). Another group, tentatively known as the large genome flaviviruses (LGF) are primarily associated with invertebrates (Paraskevopoulou et al. 2021), but have also been linked to plants (Kobayashi et al. 2013) and vertebrates (Kartashov et al. 2023). LGFs have genomes up to 39.8kb in length, challenging previous assumptions about the maximum

genome size achievable by RNA viruses that lack proofreading mechanisms (Ferron et al. 2021; Petrone et al. 2024). The jingmenviruses and LGFs have yet to receive taxonomic ratification, and consensus is lacking regarding their placement within the flavivirus phylogeny.

Previous efforts to reconstruct the evolutionary history of large and diverse families of RNA viruses such as the *Flaviviridae* have relied largely on the phylogenetic analysis of highly conserved viral proteins, most notably the RNA-dependent RNA polymerase (Paraskevopoulou et al. 2021; Shi et al. 2016a) (RdRp). Although of considerable utility, the functions and features that define virus biology and pathogenesis are typically encoded by highly divergent sequences outside of the conserved replication machinery. In these regions, it is difficult to detect deep sequence homology and hence perform reliable multiple sequence alignment (MSA) or phylogenetic analysis. As a consequence, our understanding of long-term virus evolution is generally based on the analysis of a single protein (the RdRp), such that we lack an understanding of genome-wide relationships and hence of the genesis and evolution of viral genera and species.

Glycoproteins are likely to be important determinants of phenotypic characteristics across the *Flaviviridae*. They are essential for virus entry, influence host range and spillover potential, and are primary targets for host immune responses. However, glycoproteins have yet to be identified and/or classified for many species in the *Flaviviridae* (Rey and Lok 2018). Owing to high levels of sequence divergence, this cannot be resolved by even the most sensitive of sequence-based approaches (Garry and Garry 2020; Petrone et al. 2024), and classical structural biology lacks the speed and scalability to sample enough species. This knowledge gap limits the investigation of molecular mechanisms, which in turn hinders the development of interventions such as vaccines.

Here we have augmented conventional phylogenetics with machine learning-enabled protein structure prediction to comprehensively map glycoprotein structures across the *Flaviviridae*. This provides an evolutionary and genomic-scale perspective of the entire family, revealing molecular signatures that define the diverse virology and ecology found within the *Flaviviridae*.

## 6.3 Methods

### 6.3.1 Compilation of *Flaviviridae* sequence set

#### (i) Retrieval of flavivirus genomes

Flavivirus sequences were collected using the search phrase "*Flaviviridae* taxid 11050 and Unclassified *Flaviviridae* taxid 38144" in the NCBI Virus Database on 15 December 2022. The search was complemented by referencing sequences from Mifsud et al. (2022) and supplemented with sequences from the NCBI nucleotide database using the search phrase "flavi[All Fields] OR pesti[All Fields] OR hepaci[All Fields] OR pegi[All Fields] AND viruses[filter]" on the same date. Additional sequences were later retrieved from publications that had sequences not available in GenBank at the time (Costa et al. 2022; Debat and Bejerman 2023; Dunay et al. 2023; Elbadry et al. 2023; Guo et al. 2023; Hou et al. 2024; Kong et al. 2022; Perveen et al. 2023).

#### (ii) Sequence set curation

Sequences were clustered to a 95% nucleotide identity threshold to approximate a species-level distinction, excluding the LGF tick-associated clade. Clustering was performed using CD-HIT (v4.6.1) (Fu et al. 2012) with non-default parameters "cd-hit-est -c 0.95 -n 9". Subsequently, the clustered sequence set was manually curated by removing incomplete coding regions. Sequences shorter than 2,000 nucleotides in length were removed, with the exception of the jingmenviruses where segments are known to be <2000 nucleotides in length. These nucleotide sequences were translated using the Geneious Prime Find ORFs tool (v2022.0) (https://www.geneious.com/) (Kearse et al. 2012) and along with protein sequences aligned to annotated reference sequences (where available) using MAFFT FFT-NS-I X2 (v7.402) to assess genome completeness (Katoh and Standley 2013). This was complemented by predicting conserved domains using the InterProScan software package (v5.56-89.0) with the SFLD (v4.0), PANTHER (v17.0), SuperFamily (v1.75), PROSITE (v2022_01), CDD (v3.18), Pfam (v34.0), SMART (v7.1), PRINTS (v42.0), and CATH-Gene3D databases (v4.3.0) (Jones et al. 2014). Sequences determined to contain partial coding sequences were removed from the subsequent analyses.

#### (iii) Discovery of novel LGF sequences

Tick-associated LGFs are of particular interest due to the recently reported association between Haseki tick virus and tick-borne infectious disease in humans (Kartashov et al. 2023). To identify related viruses, we screened the Sequence Read Archive (SRA) RdRp microassemblies generated by Serratus (Edgar et al. 2022) using DIAMOND BLASTx

(v2.0.9) (Buchfink et al. 2021) (*e*-value threshold of $10^{-5}$ and the "--ultra-sensitive" flag) with Haseki tick virus (UTQ11742) as the query. An *e*-value threshold of $1.6\times10^{-15}$ was established to restrict the number of libraries for reassembly to a manageable quantity. This threshold was determined based on the organism associated with the SRA library and the percent identity values. The 319 SRA libraries that meet this threshold were processed following the BatchArtemisSRAMiner pipeline (Mifsud 2023). In brief, raw FASTQ files were retrieved using Kingfisher (v0.3.0) (https://github.com/wwood/kingfisher-download), quality trimming and adapter removal using Trimmomatic (v0.38) (Bolger et al. 2014) with parameters SLIDINGWINDOW:4:5, LEADING:5, TRAILING:5, and MINLEN:25 and de novo assembly using MEGAHIT (v1.2.9) (Li et al. 2015) with default parameters. The assembled contigs were compared to the NCBI non-redundant protein database (as of March 2023) and a custom *Flaviviridae* protein database using DIAMOND BLASTx as described above. All novel flaviviruses predicted to contain complete coding sequences identified by this method (including those outside of the LGF group) were included in phylogenetic analyses.

### 6.3.2 Structure prediction and homology search

#### (i) Systematic protein structure prediction

We adopted a strategy to overcome incomplete and ambiguous genome annotations, and generate sequence lengths that are amenable to rapid inference of structure. *Flaviviridae* polyprotein amino acid sequences were broken into sequential 300-residue blocks with a 100-residue overlap. However, most polyproteins are not equally divisible by 300, therefore, we set the final sequence block to cover the final 300 residues of the polyprotein, irrespective of overlap with the penultimate block. This resulted in 16,463 sequence blocks from 561 species (558 from the *Flaviviridae* and 3 from the Tombusvirus outgroup). Structures were predicted for each sequence using the ColabFold (v1.5.1) implementation of AlphaFold2 (v2.3) (Mirdita et al. 2022), with default settings but only generating a single model per target, performed using Google Colab cloud computing. Structural inference was also performed with ESMFold (v1) (Lin et al. 2023) (using the 3 billion parameter ESM-2 model), on local compute (Nvidia V100 GPU+32GB vRAM). This resulted in a total of 32,926 structural models. Custom Python scripts were used to break up sequences for folding and extract metrics from outputs (that is, pLDDT confidence and MSA depth). For inference of putative mature protein sequences the SignalP server (v6.0) was used to predict the junctions between viral proteins (Teufel et al. 2022). For custom ColabFold inference, whole polyprotein sequences of the Bole tick virus group were aligned using MAFFT, MUSCLE (v5.1) (Edgar 2022), and subalignments covering only the putative protein sequences were

converted to the.a3m format and used as input for ColabFold structure prediction (Mirdita et al. 2022). All predicted structures and summary statistics are included in the associated Zenodo repository (https://doi.org/10.5281/zenodo.11092288) (Mifsud et al. 2024). Representative structural superpositions were performed using FATCAT (v2.0) (Li et al. 2020). All structural visualizations were prepared for publication using UCSF ChimeraX (Meng et al. 2023).

**(ii) Structural homology searches**

We used Foldseek in exhaustive search mode to cross compare the *Flaviviridae* protein foldome with a library of reference structures drawn from the protein database and ColabFold models of particular targets (see below). Foldseek was set to output *e* values, structurally aligned amino acid sequences, % identity of aligned residues, bit score and lddt structural similarity, with an *e*-value cut-off of 0.1 to eliminate low probability hits and reduce the size of the output datafile. To interrogate the output data, the lowest *e*-value scores for any given species against any given reference structure were extracted using a custom python script. Where multiple references were used for a single protein the lowest *e* value against any given species was chosen. This data was plotted against sequence-based phylogenies using the Interactive Tree Of Life (Letunic and Bork 2021). Representative hits (Supplementary Figure 6.1) were selected manually to reflect the levels of similarity and divergence in structure and sequence. All reference structures are included in the underlying data, the following experimental structures were used from the PDB: 6ZQI (Spondweni virus E and prM), 1L9K (DENV-2 MTase), 5F3Z (DENV-3 RdRp), 7QRF (TBEV E and prM), 7V1E (Omsk haemorrhagic fever virus MTase), 7T6X (HCV E1 and E2), 6VYB (SARS-CoV-2 spike, negative control), 2YQ2 (BVDV E2) and 4DVK(BVDV $E^{rns}$) (de la Peña et al. 2022; Egloff et al. 2002; El Omari et al. 2013; Jia et al. 2022; Krey et al. 2012; Noble et al. 2016; Renner et al. 2021; Vaney et al. 2022; Walls et al. 2020). To increase reference coverage, additional ColabFold models of DENV-1 prM E and diverse Hepacivirus, Pegivirus and Pestivirus E1 and E2 (from Oliver et al. (2023)) were also used. ColabFold or ESMFold structures of BTV4 proteins were used in downstream Foldseek analysis and as references in the assembly of continuous glycoprotein structures for the structural phylogeny work.

**(iii) Sequence homology search benchmarking**

To demonstrate the increased sensitivity achieved through structure prediction approaches, we conducted two benchmarking analyses. In the first analysis we recapitulated the

Foldseek search by querying the 300-residue blocks against the cognate protein sequences underlying the reference structure database using DIAMOND BLASTp (*e*-value threshold of 0.1 and the "--ultra-sensitive" flag) (Buchfink et al. 2021). We then filtered the results to select the block with the lowest *e* value for each flavivirus and reference sequence pair. The second analysis involved annotating the complete *Flaviviridae* polyprotein sequences using the InterProScan software package (v5.65-97.0) with the AntiFam (v7.0), FunFAM (4.3.0), MobiDBLite (v2.0), NCBIfam (v13.0), SFLD (v4.0), PANTHER (v18.0), SuperFamily (v1.75), CDD (v3.20), Pfam (v36.0), SMART (v9.0), PRINTS (v42.0) and PIRSF (v2023_05) databases (Jones et al. 2014). As *e* values are specific to each InterPro database and each utilises their own *e*-value post-processing, direct comparisons are not feasible. Consequently, as advised, all matches were considered tentative hits (https://interproscan-docs.readthedocs.io/en/latest/FAQ.html).

#### 6.3.3 Phylogenetic analysis

**(i) NS5b phylogeny**

The evolutionary relationships among the *Flaviviridae* were inferred using maximum likelihood phylogenies derived from MSAs of the highly conserved NS5b region (which encodes the RdRp). This region was extracted from each sequence by aligning polyprotein sequence subsets according to their taxonomy and using both pre-existing and newly generated NS5b annotations from InterProScan as a guide. As alignment and trimming parameters have been shown to influence the topology of the *Flaviviridae* (Dong et al. 2021) we compared several methods resulting in 225 phylogenies. In brief, flavivirus sequences were aligned using MAFFT, MUSCLE (v5.1) (Edgar 2022) and Clustal Omega (v1.2.4) (Sievers et al. 2011) with default parameters. Ambiguously aligned regions were removed using trimAl (v1.4.1) (Capella-Gutiérrez et al. 2009) with 8 conservation thresholds (that is, minimum percentage of alignment columns to retain): 5, 7.5, 10, 12.5, 15, 17.5, 20 and 25; and 3 gap thresholds (that is, the minimum fraction of sequences without a gap needed to keep a column): 0.7, 0.8, and 0.9—as well as the automated parameter selection method gappyout.

All maximum likelihood phylogenetic trees were estimated using IQ-TREE 2 (v2.1.0) (Minh et al. 2020). Selection of the best-fit model of amino acid substitution was inferred for a subset of phylogenies using the ModelFinder function in IQ-TREE 2 (Kalyaanamoorthy et al. 2017). In addition to the model chosen by ModelFinder (LG+F+R10) two additional models, the Le-Gascual model (LG) and FLAVI (Le and Vinh 2020) were compared. Branch support was

calculated using 1,000 bootstrap replicates with the UFBoot2 algorithm and an implementation of the SH-like approximate likelihood ratio test within IQ-TREE 2 (Guindon et al. 2010; Hoang et al. 2017). To root the phylogeny, three members of the *Tombusviridae* family were chosen given their remote sequence similarity to the NS5 region of the *Flaviviridae* (Gilmer et al. 2017; Paraskevopoulou et al. 2021). Phylogenetic trees were annotated using the R packages ape (v5.6.2) (Paradis and Schliep 2019), phytools (v1.5-1) (Revell 2024), and ggtree (v3.3.0.9) (Yu et al. 2017) and further edited in Adobe Illustrator. Genome diagrams were constructed using a manually curated selection of predicted functional domains and visualised using gggenomes (v0.9.8.9) (Thomas Hackl 2022).

For each virus sequence, host information was pulled from the corresponding GenBank 'host' field using Rentrez (v1.2.3) (Winter 2017) and standardised using Taxize (v0.9.1) (Chamberlain and Szöcs 2013). Vector status, defined as 'yes', 'no' or 'potentially', was assigned by first querying the Arbovirus Catalog (https://wwwn.cdc.gov/arbocat/). Where a taxon was identified as an 'Arbovirus' by the Arbovirus Catalog it was assigned 'yes', otherwise for those listed as 'potential arboviruses', 'probable arboviruses', or those not present in the catalogue, literature on this taxa was reviewed for evidence of vector association. Three main criteria were considered: (1) whether the virus replicated in both invertebrate and vertebrate cells; (2) the phylogenetic position of the virus—that is, is the virus in the middle of an insect-specific clade; and (3) consensus among the literature on the possibility of the virus being vectored. The assigned vector status for each taxon and the underlying evidence for this is provided in Supplementary Table 1.

(ii) **Evaluation of topological concordance**

To determine the most robust NS5b phylogeny, alignments (pre- and post- trimming) (n = 225) were examined for the presence of canonical RdRp motifs, misalignments, and overall pairwise identity and length. The resultant tree topology and branch support were examined in FigTree (v1.4.4) (Rambaut and Drummond 2012). This analysis was combined with comparisons of genome composition and to previous *Flaviviridae* phylogenies (Bamford et al. 2022; Dong et al. 2021; Paraskevopoulou et al. 2021) to identify the most concordant topology across the multiple parameters tested. To supplement this, the R package treespace (v1.1.4.2) (Jombart et al. 2017) was used to conduct a principal component analysis (PCA) with the goal of identifying clusters of similar trees and assessing whether the selected topology is consistent with the median topology. Accordingly, Kendall–Colijn

distance was calculated for each tree and used for the PCA, with two principal components retained (Kendall and Colijn 2016). To identify discrete clusters of related trees, pairwise distances were mapped into four clusters using hierarchical clustering (Ward's method) (Legendre and Legendre 2012). Manual and distance-based inspection revealed that the alignment method drove variation in tree topology and branch lengths between phylogenies. Specifically, tree topologies and their corresponding phylogenetic distances derived from Clustal Omega were frequently topologically discordant compared to those generated by MAFFT and MUSCLE; as such, these phylogenies were excluded and the PCA was recalculated. Geometric median trees were generated from each cluster and alignment method and used to inform the selection of the final phylogeny. This phylogeny was aligned using MUSCLE with a trimAl consensus and gap threshold value of 5 and 0.9, respectively, and based on the LG+F+R10 amino acid substitution model. We further conducted an extensive stratified MUSCLE alignment analysis to validate the robustness of our NS5b alignment and resulting phylogenies which considered variations in hidden Markov model (HMM) parameters and guide tree merge orders (Supplementary Note 6.1 and Supplementary Figure 6.2-7).

(iii) **RNase T2 phylogeny**

To infer the evolutionary history of the RNase T2 protein, sequences were obtained from the GenBank protein database for conserved domains using the queries "taxid 238513" and "taxid 238220", literature searches (Rolland et al. 2020; Saberi et al. 2018) , these were supplemented with structurally homologous protein clusters identified using the AlphaFold database Foldseek clusters server (Barrio-Hernandez et al. 2023).

To identify unannotated RNase T2-like sequences in virus genomes, a NCBI web protein BLAST (https://blast.ncbi.nlm.nih.gov/Blast.cgi) was used with RNase T2 sequences used as a query against the clustered non-redundant protein database (as of June 2023) (Steinegger and Söding 2017), using the BLOSUM45 matrix and with taxonomy limited to the group 'Viruses' (taxid:10239). The HMM search web server (v2.41.2) (Potter et al. 2018) was used to identify additional viral T2 RNase-like sequences. An alignment of RNase T2 sequences was used as a query against the Reference Proteomes, UniProtKB, SwissProt and PDB databases (as of June 2023), with results again limited to 'Viruses' (taxid:10239). This was further repeated for the PDB, SCOPe, SMART, Pfam, UniProt-SwissProt-Viral, PHROG, COG and KOG databases using the HHpred web server (Gabler et al. 2020; Zimmermann et al. 2018) as of April 2024 and the Uniclust30 using HHsearch (v3.3.0) (Mirdita et al. 2017; Steinegger et al. 2019). For all methods, if new virus sequences

were detected, they were manually inspected for the presence of RNase T2 motifs and, in turn, used as queries. To estimate the RNase T2 phylogeny, non-viral sequences were clustered at 80% amino acid identity using DIAMOND cluster (v.2.0.9) (Buchfink et al. 2023) with default parameters and aligned with the viral sequences using MAFFT and a maximum likelihood phylogeny as described above.

**(iv) Structure-guided glycoprotein phylogenies**

We implemented the approach described previously (Puente-Lelievre et al. 2024) to infer glycoprotein phylogenies based on both structural and sequence homology. Owing to the arbitrary fragmentation of protein sequences into the 300-residue blocks, our predicted structures represent overlapping truncated segments of the true glycoproteins. To generate full glycoprotein structures we filtered our Foldseek results by an *e*-value cut-off of 0.001 and selected the E, E1 or E2 reference structure that had the highest bit score value for any protein block of each query virus. This reference was then used for determining the putative coordinates of each glycoprotein in the viruses' polyproteins, defined as the start position of the earliest block's Foldseek hit to the reference and the end position of the latest block's hit. This process yielded 247 E, 190 E1 and 189 E2 protein sequences, the majority of which appeared to be full length glycoproteins, but with a minority of truncated sequences likely due to low-accuracy structure prediction in the protein foldome. The structure of the full glycoprotein sequences was predicted using ColabFold and ESMFold as described above, but with five ColabFold models produced for each target. The most confident prediction based on their average pLDDT values was chosen for downstream analysis.

We modified the FAMSA alignment program (Deorowicz et al. 2016) to use the Foldseek 3Di character substitution matrix as described previously (Puente-Lelievre et al. 2024) (https://github.com/nmatzke/3diphy). We then converted our predicted full glycoprotein structures to 3Di sequences with the Foldseek 'structureto3didescriptor' option and used the modified FAMSA aligner to infer structural, 3Di sequence alignments of the E, E1 and E2 protein sets. These MSAs are based on the homology between the 3Di characters corresponding to each protein residue and should represent the overall structural homology between these proteins. Consistent with the methodology of Puente-Lelievre et al. we used trimAl (Capella-Gutiérrez et al. 2009) with a gap threshold of 35% to create trimmed versions of the 3Di MSAs. In addition to the 3Di character alignments, we replaced the 3Di characters with the protein amino acid residues in both the complete and trimmed versions

of the MSAs. This resulted in a total of four MSAs (3Di, trimmed 3Di, amino acid, trimmed amino acid) for each E, E1 and E2. Modelfinder (Kalyaanamoorthy et al. 2017) implemented in IQ-TREE 2 (Minh et al. 2020) was used to determine the best substitution model for each alignment. All possible models were tested, including the custom 3Di substitution model (-mset Blosum62,Dayhoff,DCMut,JTT,JTTDCMut,LG,Poisson,Poisson+FQ,Poisson,PMB,WAG,EX2,EX3,EHO,EX_EHO,3DI -mfreq FU,F -mrate E,G,R). Selected substitution models for all alignments are included in the associated Zenodo repository (https://doi.org/10.5281/zenodo.11092288). Phylogenetic trees based on each MSA were inferred using IQ-TREE 2 (v2.2.2.6) (Minh et al. 2020) under each corresponding best-fit substitution model, with node support assessed using 1000 ultrafast bootstrap replicates (Hoang et al. 2017). Finally, we performed phylogenetic inferences based on both 3Di character and amino acid homology by combining the corresponding pairs of 3Di and amino acid MSAs and performing a partition model IQ-TREE 2 phylogenetic inference (Chernomor et al. 2016), in which the two partitions correspond to the 3Di and the amino acid MSAs and each partition uses the best-fit substitution model of its corresponding MSA. The contribution of each partition in the combined MSA phylogenetic inferences was determined based on the partition-wise log-likelihoods, inferred with IQ-TREE's -wpl option. Manual inspection and analysis of partition contribution was used to select trees for display (all resulting phylogenetic trees are provided in Supplementary Figure 6.8–10). In brief, the 3Di partition had a consistently larger contribution to the joint phylogenetic inference compared to the amino acid partition, although using both alignments (instead of either alone) generally aids phylogenetic reconstruction (Puente-Lelievre et al. 2024). However, in the case of the E glycoprotein phylogeny, the contribution of the amino acid partition was markedly lower than that of the 3Di partition (Supplementary Table 2). Moreover, we found clear evidence of long branch attraction in the amino acid only reconstructions of the E phylogeny (Supplementary Figure 6.8), and reasoned that these artefacts may carry over into the combined 3Di–amino acid reconstruction. Therefore, the 3Di-based phylogeny was selected for the E protein whereas 3Di–amino acid trees were used for E1 and E2.

As a point of comparison, additional structural phylogenies were generated from our custom full length glycoprotein structures using FoldTree (Moi et al. 2023). For each structure set (E, E1 and E2 protein sets) phylogenies were inferred using three metrics FoldTree, LDDT and TM-score with default parameters (see Supplementary Figure 6.11). However, given the limitations associated with the use of neighbour joining methods on

structural distances (outline in Puente-Lelievre et al. (2023)) we reasoned that the 3Di-guided approach, outlined above, is likely to yield more robust results.

## 6.4 Results

### 6.4.1 *Flaviviridae* contains three major clades

Understanding the evolution of molecular features across the *Flaviviridae* requires a proper gauge of the phylogenetic and genomic diversity of this family. To achieve this, we first constructed a comprehensive data set of flavivirus sequences, which after clustering and manual curation, comprised 458 flavivirus genomes with complete coding sequences, including 11 that were novel taxa identified in this study (Supplementary Table 3). We next inferred a robust family-level phylogenetic tree for these data. Using conserved NS5 gene sequences that encode the RdRp, we applied various sequence alignment methods, quality trimming protocols and amino acid substitution models, to infer a total of 225 phylogenetic trees for this family (Supplementary Table 4). Using distance-based approaches and manual inspection of the alignments and trees, we identified a phylogeny, denoted Tree 18 (Supplementary Figure 6.12-13, Supplementary Table 4), that appeared to best represent the consensus topology of the *Flaviviridae*. Specifically, the topological placement of the major *Flaviviridae* clades in Tree 18 was consistent with 93% of phylogenies derived from MUSCLE and MAFFT MSAs, although this percentage dropped to 70% when Clustal Omega was included.

Our best-fit RdRp phylogeny supported the division of the *Flaviviridae* into three distinct clades: (1) an Orthoflavivirus/jingmenvirus group (that also contains ‘orthoflavirus-like’ viruses—for example, Cnidaria flavivirus and Tamana bat virus); (2) a clade comprising the large genome flaviviruses and members of the genus Pestivirus; and (3) a Pegivirus/Hepacivirus clade (Figure 6.1a) Regardless of whether the tree was unrooted or rooted on a *Tombusviridae* outgroup, the LGF/Pestivirus and Orthoflavivirus/jingmenvirus groups clustered together and formed a sister group to the Pegivirus/Hepacivirus clade. The Orthoflavivirus/jingmenvirus clade had the largest number of taxa (n = 182), followed by Pegivirus/Hepacivirus (*n*=157) and LGF/Pestivirus (*n*=119). All novel taxa (n = 11) identified in this study fell within the LGF/Pestivirus clade.

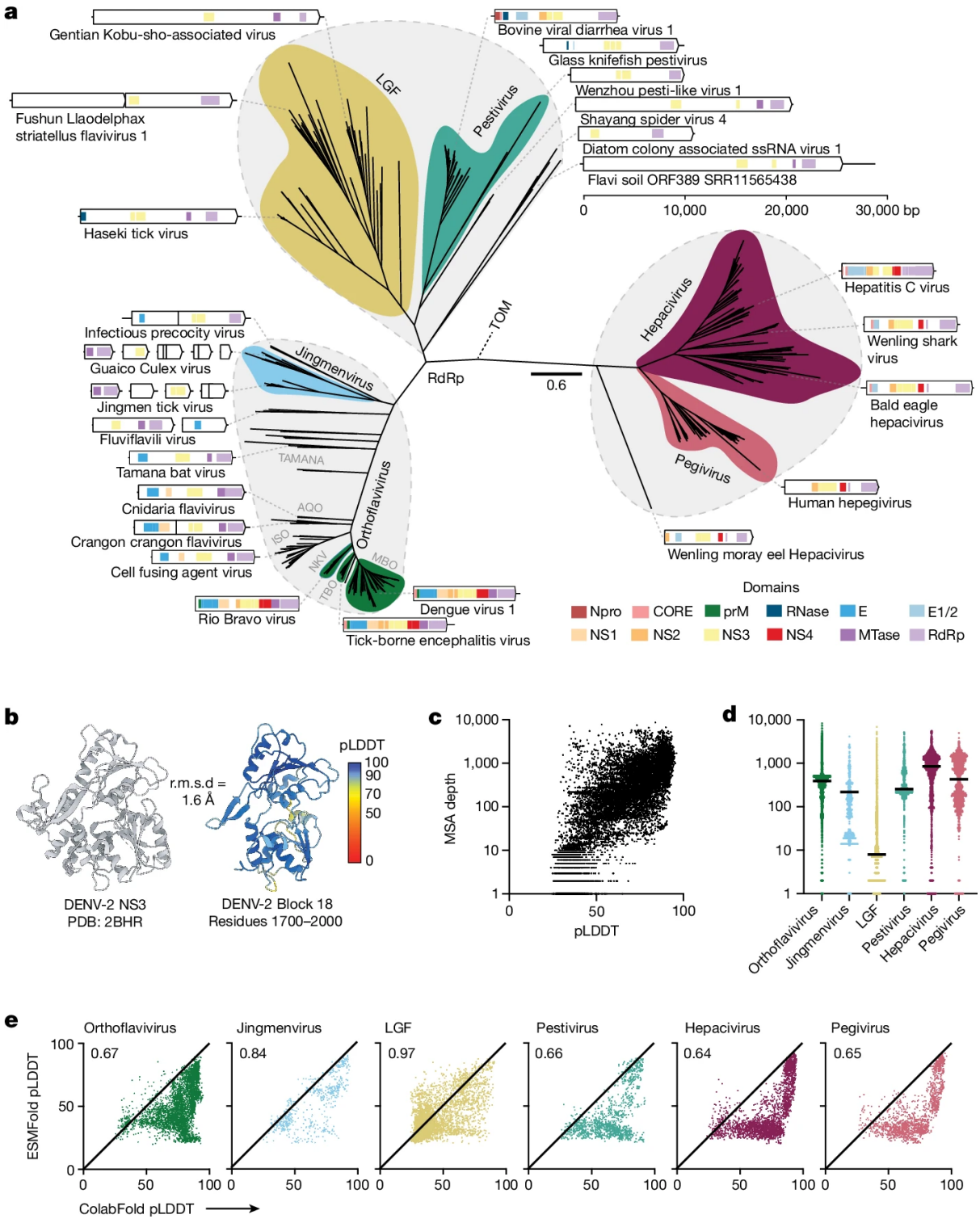



***Figure 6.38 Generation of a protein foldome for the Flaviviridae****. (a), RdRp phylogeny reveals three major lineages within the Flaviviridae: (1) Orthoflavivirus/jingmenvirus (including orthoflavivirus-like—for example, Tamana bat virus); (2) LGF/Pestivirus; and (3) Hepacivirus/Pegivirus. An unrooted tree is shown, with the tombusviruses (TOM) representing the outgroup taxa and a scale bar denoting the number of amino acid substitutions per site. Genome organization is provided for exemplar species, with*

*annotations based on InterProScan searches. (b), Crystal structure of DENV-2 NS3 (left) shown alongside a ColabFold predicted structure for the corresponding region of the polyprotein (right). These structures superpose with a root mean square deviation (r.m.s.d.) of 1.6Å. The predicted structure is colour-coded by per residue confidence scores (predicted local distance difference test (pLDDT)), as indicated in the bar. (c), Scatter plot of MSA depth and prediction confidence (pLDDT). (d), MSA depths for each sequence block in each genus or subclade, colour-coded as in a (orthoflavivirus-like viruses are included with the orthoflaviviruses). The mean is shown as a solid black line (*n *= 4,754 (Orthoflavivirus), 704 (jingmenvirus), 4,358 (LGF), 1,278 (Pestivirus), 2,904 (Hepacivirus) and 1,623 (Pegivirus) sequence blocks). e, Scatter plots representing prediction confidence (pLDDT) for ColabFold and ESMFold for each sequence block in each genus or subclade. Numerical values provide the performance ratio between the protein structure prediction methods; values below 1 indicate better performance by ColabFold.*

### 6.4.2 Building a *Flaviviridae* protein foldome

We next aimed to explore protein functionality across the *Flaviviridae* using machine learning-enabled protein structure prediction. All flaviviruses encode polyproteins that undergo proteolytic maturation to liberate the constituent viral proteins. However, incomplete and ambiguous genome annotations combined with extensive sequence divergence make it very difficult to reliably identify the regions encoding each mature protein in all species. We therefore took a genome-agnostic approach, in which polyprotein coding sequences were split into sequentially overlapping 300-residue blocks for structural inference by two leading prediction models—ColabFold–AlphaFold2 (hereafter referred to as ColabFold) and ESMFold (Jumper et al. 2021; Lin et al. 2023; Mirdita et al. 2022). This provided a comprehensive survey of protein structure across the *Flaviviridae* (458 species, more than 16,000 sequence blocks and more than 33,000 predicted structures), referred to here as the ‘protein foldome’.

As protein structural prediction has yet to be systematically applied in virology, we first evaluated folding performance. ColabFold performed extremely well for many virus species (for example, dengue virus 2; Figure 6.1b and Supplementary Figure 6.14a-c). However, its accuracy is directly proportional to the depth of the MSAs that guide structural inference (Lee et al. 2023), with shallow MSAs producing low confidence predictions (Figure 6.1c). This becomes particularly problematic for the LGF, which are poorly sampled and

consequently underrepresented in sequence databases, resulting in consistently shallow MSAs (Figure 6.1d).

Structural inference by ESMFold is driven by a protein language model and does not require MSAs, but is less accurate than ColabFold. A comparison of folding confidence demonstrated that ColabFold consistently outperformed ESMFold across the three major *Flaviviridae* clades (Figure 6.1e and Supplementary Figure 6.14d). However, for the LGF, ESMFold yields informative predictions from sequences for which ColabFold fails, and this proved important for downstream analysis.

### 6.4.3 *Flaviviridae* glycoprotein discovery

Orthoflaviviruses, including yellow fever virus (the canonical species for the group), tick-borne encephalitis virus (TBEV) and dengue virus (DENV), are predominantly vector-borne, although exceptions exist (Blitvich and Firth 2017). These viruses possess the E glycoprotein, a prototypical class II fusion protein. Structurally and functionally homologous class II fusion proteins have been identified both in viruses (for example, Gc in the bunyaviruses) and in eukaryotes (for example, HAP2 in plants, protists and invertebrates) and they are expected to share a common ancestor (Dessau and Modis 2013; Fédry et al. 2017; Guardado-Calvo and Rey 2021; Kielian and Rey 2006; Rey et al. 1995). In viruses, class II fusion proteins are accompanied by a partner glycoprotein that is responsible for regulation and/or chaperoning of the fusogenic component. In the orthoflaviviruses this function is performed by the small glycoprotein prM (Li et al. 2008).

Identifying membrane fusion mechanisms in the hepaciviruses, pegiviruses and pestiviruses has proved more challenging. These viruses possess E1 and E2 glycoproteins that work in concert to achieve pH-dependent membrane fusion. The sequences of E1E2 bear no similarity to prM/E, and experimental structures of E2 from prototypical hepaciviruses and pestiviruses reveals folds that are broadly dissimilar from one another and from the E glycoprotein in orthoflaviviruses (Aitkenhead et al. 2024; El Omari et al. 2013; Khan et al. 2014; Kong et al. 2013; Li et al. 2013). Recent cryo-electron microscopy analyses suggest that E1 adopts a unique fold, unlike that of any other known protein (de la Peña et al. 2022; Metcalf et al. 2023).

Whether E1E2 represents a novel and as yet uncharacterised fusion mechanism or a highly divergent iteration of a class II system remains to be determined. Understanding the

distribution and characteristics of glycoproteins across the *Flaviviridae* would likely provide insights on this. To achieve this we performed pairwise Foldseek (van Kempen et al. 2023) structure similarity searches against the *Flaviviridae* protein foldome using a custom reference library comprising selected experimental glycoprotein structures from the Protein Data Bank (PDB) and published ColabFold models (Oliver et al. 2023).

To benchmark our approach we performed parallel analyses using state-of-the-art sequence-based approaches (DIAMOND and InterProScan (Buchfink et al. 2021; Jones et al. 2014), which did not detect deep homology, even for highly conserved targets (Supplementary Figure 6.15). By contrast, Foldseek demonstrated strong sensitivity, successfully detecting unambiguous structural homology between Pegivirus/Hepacivirus and Pestivirus E1 even though they share only 10–15% amino acid sequence identity (Figure 6.2b,d). Relative to E1, the Pegivirus/Hepacivirus and Pestivirus E2 are structurally divergent (Oliver et al. 2023), nonetheless, Foldseek identified reciprocal structural similarity focussed on the C-terminal portion of E2 where sequence identity ranged from 8.5 to 15% (Figure 6.2b,e). The distribution of E1 and E2 were in near-perfect correlation, consistent with mechanistic interdependence, and we found no evidence for E1E2-like folds outside of Pegivirus/Hepacivirus and Pestivirus groups.

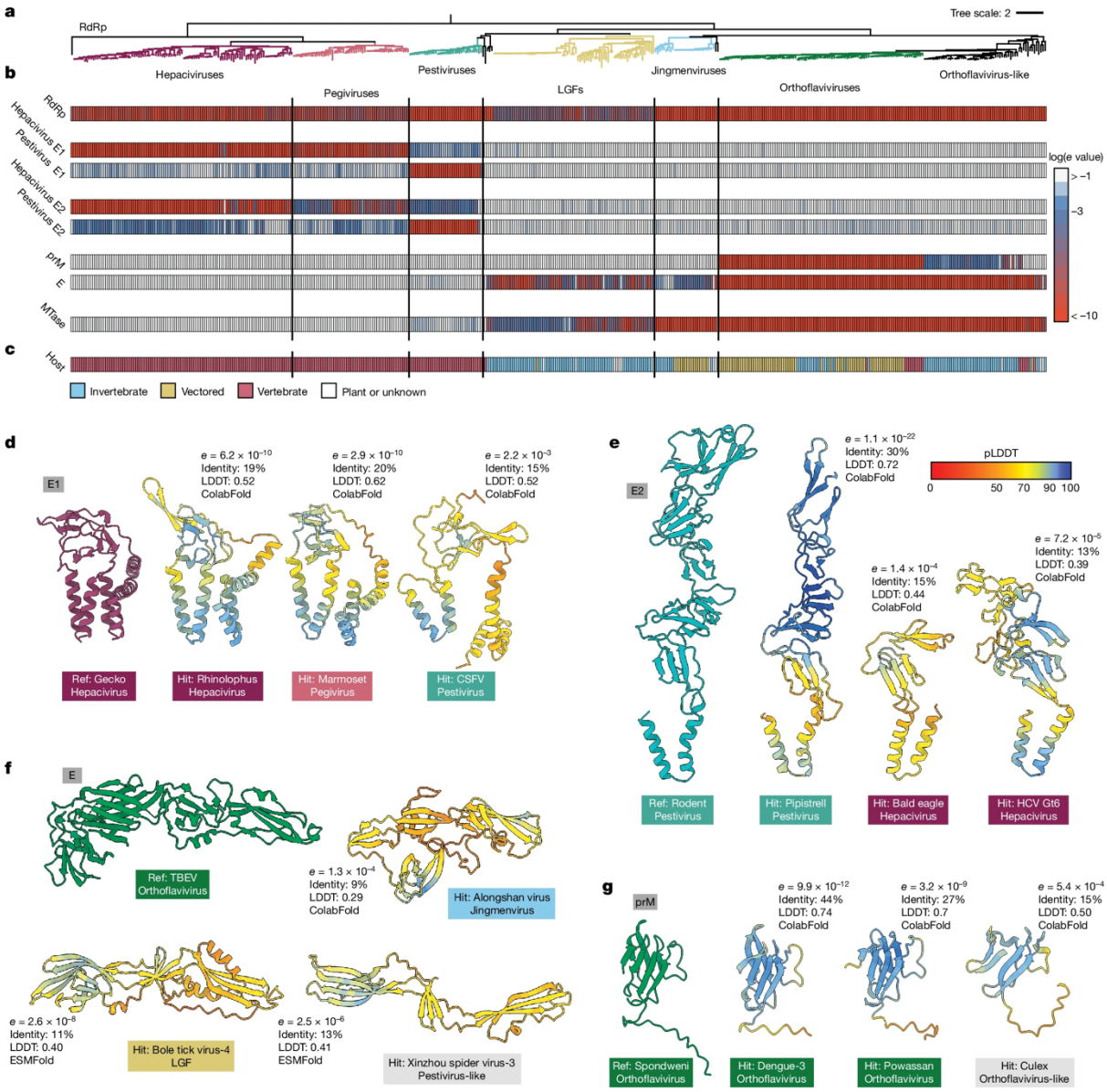



***Figure 6.39 Discovery of glycoproteins across the Flaviviridae.*** *(a), RdRp phylogeny rooted on the tombusviruses (removed for visualization), with each genus or subclade colour-coded as in Figure 6.1a. (b), Foldseek structure-based homology e-value heat maps for the stated references, colour-coded as shown in the key. In the case of E1, E2, E, prM and MTase the values represent summary e values after comparison with a range of relevant reference structures (Methods). (c), Host species tropism for each virus. 'Vectored' refers to those assigned as 'Yes' or 'Potentially' in Supplementary Table 1. Vertical lines within the heat map demark divisions between major clades. (d–g), Representative reference structures and Foldseek hits for E1 (d), E2 (e), E (f) and prM (g). For each hit only the Foldseek-aligned residues are shown for any given structure, metrics provide e value, sequence identity, structural alignment score (local distance difference test (LDDT), ranging from 0 to 1) and protein structure prediction method. Predicted structures are colour-coded*

*by pLDDT confidence scores, as shown in the key. In (d,e), the reference structures are previously published ColabFold models (Oliver, Toon et al. 2023), In (f,g), experimental structures are used (PDB: 7QRF and 6ZQI, respectively). CSFV, classical swine fever virus; HCV, hepatitis C virus*

We mapped structural homologues of E glycoprotein to the orthoflaviviruses, orthoflavivirus-like, jingmenviruses, LGF and pestivirus-like species that sit basal to the classical Pestivirus genus (for example, Xinzhou spider virus 3) (Figure 6.2b,f). For the most divergent sequences (for example, LGF and pestivirus-like viruses) detection required ESMFold structures, emphasizing the value of using complementary prediction methods (Supplementary Figure 6.16). A notable exception was a group of viruses of unknown hosts discovered in environmental samples (Chen et al. 2022; Hou et al. 2024; Urayama et al. 2016) (for example, Inner Mongolia sediment flavi-like virus 3) for which no glycoproteins were identified (Supplementary Figure 6.17). Whether these represent species without structural proteins or partial genomes remains to be determined.

For most E glycoprotein homologues, the predicted structures were sufficient to identify the hydrophobic fusion loop at the tip of domain II, which inserts into host membranes and is central to the class II fusion mechanism (Supplementary Figure 6.18). However, the fusion loop was absent from the jingmenvirus E homologues, suggesting substantial mechanistic divergence in these viruses. We were only able to detect the prM partner glycoprotein within the orthoflaviviruses and some orthoflavivirus-like viruses (Figure 6.2b,g). A critical function of prM is occluding the fusion loop of E during particle maturation and we may yet expect to find orthologous partners in other clades.

### 6.4.4 Glycoproteins follow ecological niche

We could assign either E1E2 or E glycoproteins to the vast majority of species in the *Flaviviridae* (Figure 6.2b). Their distributions divide the family broadly in two, although this division is incongruent with the RdRp phylogeny, suggesting a complex evolutionary history. In particular, the pestiviruses and LGF, which represent sister clades on the RdRp tree, possess E1E2 and E, respectively. Although mapping glycoproteins was the primary focus of our study, we also compared the foldome against other *Flaviviridae* proteins from the PDB. In doing so we observed that all species with a methyltransferase (MTase) also possess an E glycoprotein homologue, and all species with E1E2 (that is, the hepaciviruses, pegiviruses and pestiviruses) lack MTase (Figure 6.2b and Supplementary Figure 6.1).

Viruses with MTase undergo cap-dependent translation, whereas those without MTase rely on an internal ribosome entry site (Arhab et al. 2020) (IRES). We reasoned that E–MTase and E1E2–IRES may represent divergent co-adaptations to particular ecological niches, so we compared virus–host associations across the phylogeny (Figure 6.2c). Viruses with E–MTase infect a variety of hosts, including those that are transmitted between vertebrates by invertebrate vectors (for example, DENV by *Aedes* mosquitoes). By contrast, E1E2–IRES was strictly correlated with vertebrate hosts. This suggests that the gain of E1E2 and an IRES, with the concomitant loss of E and MTase, represent a molecular commitment to the vertebrate niche. Moreover, on the basis of the underlying RdRp phylogeny, this commitment to vertebrate infection is likely to have occurred twice in the *Flaviviridae*, once for Pegivirus/Hepacivirus and once for the pestiviruses.

### 6.4.5 LGFs harbour novel and acquired proteins

Our structure-guided approach can offer new insights into divergent and/or poorly characterised viruses such as those found in the LGF. Whereas the majority of LGF species are likely to infect invertebrates, there is evidence that one subclade—the bole tick virus group—are capable of tick-borne infection of mammals, including humans (Kartashov et al. 2023). This group may represent an emergent threat to public health and warrants closer scrutiny.

Focusing on Bole tick virus 4 (BTV4), we examined the N-terminal portion of the polyprotein proximal to the E glycoprotein homologue. The N-terminal structural proteins of *Flaviviridae* polyproteins are typically processed by host signal peptidases to liberate the mature proteins. We used cleavage site prediction to identify five putative protein coding sequences (labelled A–E; Figure 6.3a). Protein structures were predicted using three approaches—ESMFold, ColabFold and ColabFold with manually curated MSAs. For each sequence, we provide the highest confidence model produced by any of these methods (Figure 6.3a).

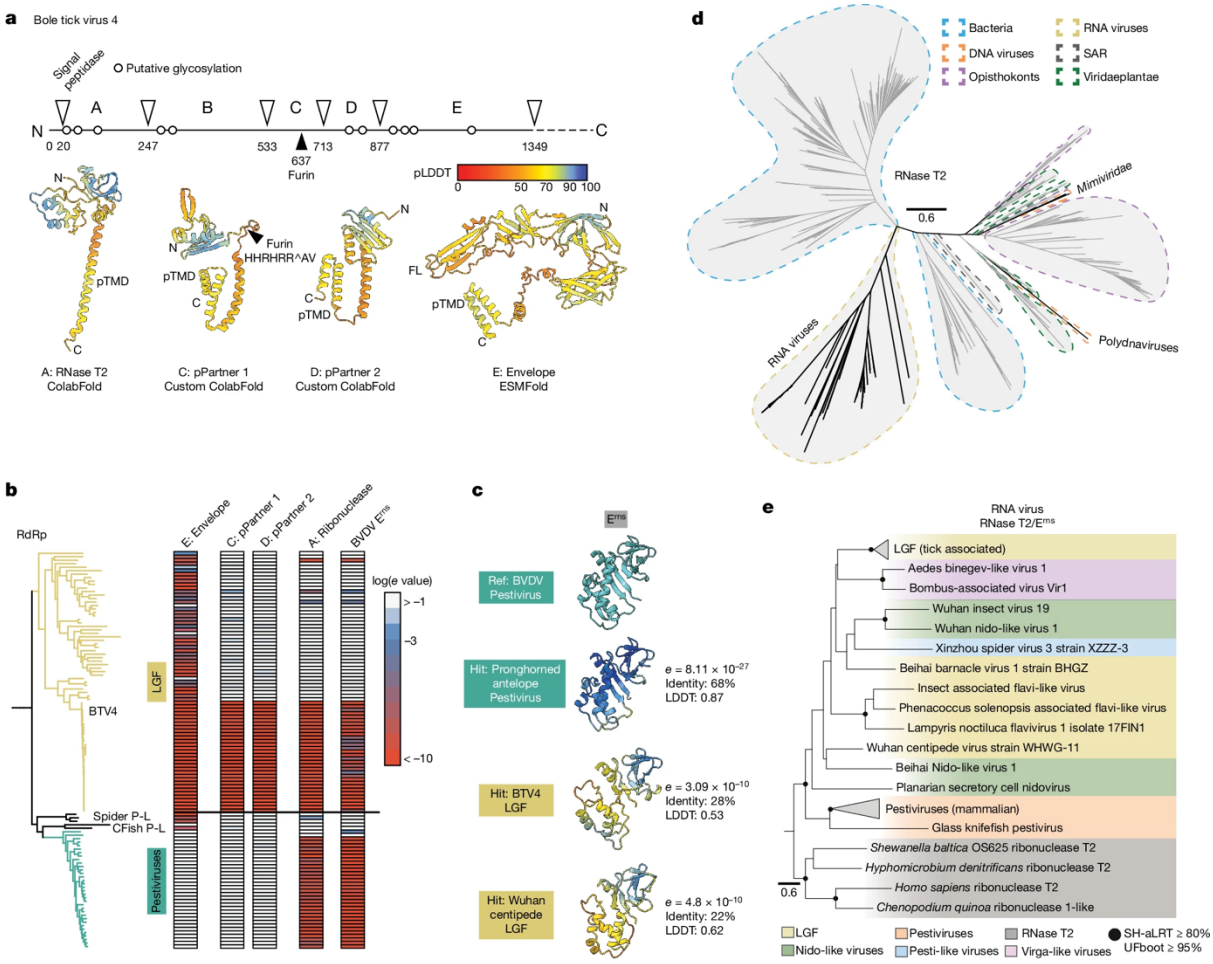



***Figure 6.40 Novel and acquired proteins in a large genome flavivirus.*** *(a), N-terminal glycoproteins from BTV4. Linear representation of the BTV4 polyprotein displays location of putative glycosylation and the predicted signal peptidase cleavage sites that delineate five mature proteins (labelled A–E). For each of these mature proteins, the highest confidence models are shown from three prediction methods (ColabFold, ColabFold with custom MSAs and ESMFold). Protein B only yielded low-confidence models and is not shown. Each protein contains a putative transmembrane domain (pTMD), protein C contains a canonical furin cleavage site, and the conserved fusion loop (FL) of E is also annotated. (b), LGF/Pestivirus lineage RdRp phylogeny and Foldseek e-value heat maps for the stated reference structures. Annotations provide the location of BTV4 (reference), the spider pestivirus-like viruses (Spider P-L), and the cartilaginous fish pestivirus-like viruses (CFish P-L). (c), Example Foldseek hits against an experimental structure of bovine viral diarrhoea virus Erns ribonuclease (PDB: 4DVK). BVDV, bovine viral diarrhoea virus. (d), Ribonuclease T2 (RNase T2) sequence phylogeny, with domains of life and viruses colour-coded as shown in the key. The scale bar indicates phylogenetic distance as number of substitutions per site. This protein has been independently acquired once by RNA viruses and twice by DNA viruses (in Mimiviridae and polydnaviruses). The RNA virus clade is nested within bacterial*

*instances of the RNase T2, suggesting a single horizontal gene transfer event. e, Phylogeny of the RNA virus RNase T2/Erns clade rooted on non-viral sequences, with viral clades colour-coded as shown in the key. An uncollapsed version containing all tip labels is provided in Supplementary Figure 6.19* (Edgar 2022)*.*

The largest and most C-terminal protein is the E glycoprotein homologue. ESMFold produces a model in which the transmembrane domain and fusion loop are juxtaposed; this is consistent with experimental structures of the post-fusion conformation of E (Kielian and Rey 2006). Directly upstream of the E homologue are two smaller proteins for which custom ColabFold yielded the optimal predictions. Although neither protein shares direct homology with prM, they both have a similar organization, consisting of a small globular domain anchored by a putative transmembrane domain. Given their direct proximity to the E homologue, we suggest that these are partner proteins that provide chaperoning to the class II fusogen. Indeed, the putative partner 1 possesses a furin cleavage site, analogous to prM, that would enable proteolysis during secretion, akin to the maturation of Orthoflavivirus particles (Li et al. 2008). Protein coding sequence B yielded low confidence predictions from each folding approach (not shown). The most N-terminal sequence was identified as a T2 family ribonuclease (RNase T2), with homologues across the tree of cellular life (MacIntosh 2011).

We used Foldseek to investigate the distribution of these protein structures across the LGF/Pestivirus clade (Figure 6.3b). Homologues of BTV4 E glycoprotein were detected throughout the LGF and in pestivirus-like viruses identified in spiders and cartilaginous fish that fall basal to members of the classical genus Pestivirus. Therefore, using a proximal reference (BTV4 E glycoprotein), we provide evidence for the loss of E and gain of E1E2 at the genesis of the pestiviruses. By contrast, the putative partner proteins were confined to the Bole tick virus subclade, and structural similarity searches against current protein databases (for example, PDB and AlphaFoldDB) revealed no homologues. Therefore, these proteins are likely adaptive features, specific to these viruses.

BTV4 RNase T2 has homologues throughout the Bole tick virus subclade and, notably, across the genus Pestivirus, where the homology maps to the $E^{rns}$ ribonuclease. Foldseek searches against pestivirus $E^{rns}$ revealed a reciprocal distribution of homology (Figure 6.3b,c). Phylogenetically, the LGF/pestivirus $E^{rns}$ form a deep branch amongst homologous RNase T2 sequences from viruses, bacteria, plants and animals (Figure 6.3d,e). Together,

this indicates that $E^{rns}$ originated in a distant ancestor of the pestiviruses and LGFs, probably from a single horizontal gene transfer of a bacterial RNase T2. Moreover, the distribution of $E^{rns}$ is broadly concordant with the RdRp phylogeny, suggesting that $E^{rns}$ has been continuously retained in certain species and lost in others (Figure 6.3e), rather than undergoing genetic exchange within the clade. Further instances of RNase T2 from nidovirus-like and virgavirus-like viruses were also nested within the $E^{rns}$ tree, indicating onward genetic transfer to other RNA viruses.

### 6.4.6 Evolutionary history of the *Flaviviridae*

Our approach enabled the discovery of glycoproteins (and other features) across the entire *Flaviviridae*. To better understand the evolutionary events that gave rise to this distribution of molecular characteristics we utilised a method (Puente-Lelievre et al. 2024) that leverages structural conservation (as represented in the Foldseek 3Di alphabet; Supplementary Figure 6.20) to guide and augment traditional amino acid-based evolutionary analyses. This revealed consensus-level glycoprotein sequence similarities, indicative of shared ancestries, that can be estimated through phylogenetic modelling (Figure 6.4a-c and Supplementary Figure 6.8–11).

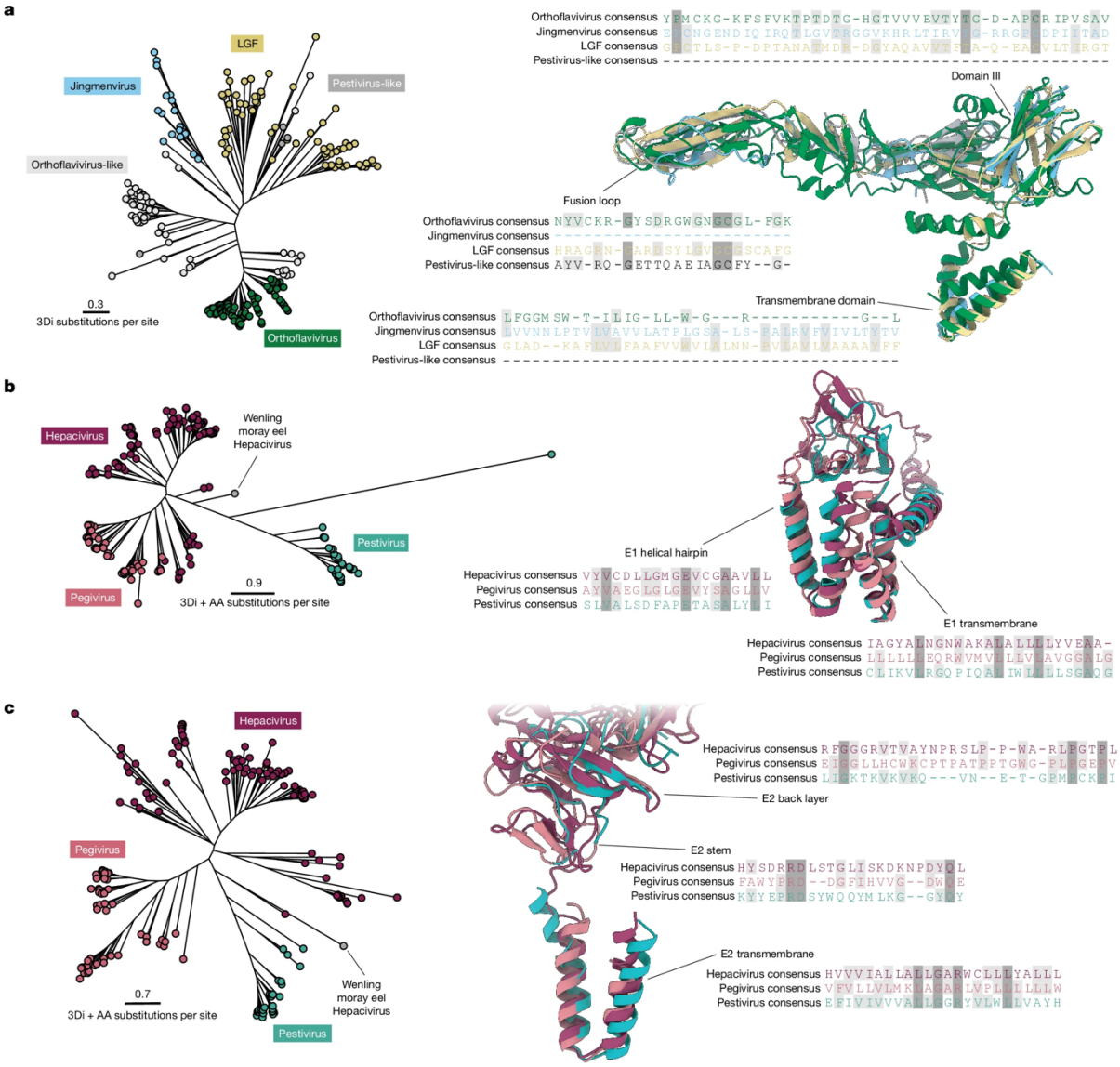



***Figure 6.41 Structurally informed phylogenetics.*** *(a), Left, 3Di-based E structural phylogeny. The scale bar indicates the number of 3Di character substitutions per site (see Methods for details of tree selection). Right, representative structures superposed using flexible FATCAT (Li, Jaroszewski et al. 2020) with a ColabFold model of West Nile virus E protein as reference (green). Structures are colour-coded as in the phylogeny. The protein alignments provide structurally aligned consensus-level amino acid sequences for the fusion loop, domain III and transmembrane domain. Conserved residues are highlighted. (b), Left, combined 3Di and amino acid-based E1 structural phylogeny. The scale bar indicates the number of 3Di and amino acid character substitutions per site. Right, representative structures are superposed with Hepacivirus F E1 protein. Alignments demonstrate consensus-level homology in the E1 helical hairpin and transmembrane domain. Structures are colour-coded as in the phylogeny. (c), Left, combined 3Di and amino acid-based structural phylogeny of E2 protein. Right, representative structures are superposed with*

*Hepacivirus F E2. Consensus-level homology in E2 back layer, stem and transmembrane domain are provided. The basal Wenling moray eel Hepacivirus is marked in both the E1 and E2 trees. AA, amino acid.*

The optimal E glycoprotein phylogeny largely reflects the RdRp tree, with E homologues from orthoflaviviruses, orthoflavirus-like viruses jingmenviruses and LGFs distributed across various subclades (Figure 6.4a). Notably, the E protein homologues in pestivirus-like viruses of spiders fell within the LGF glycoprotein clade, similar to the RdRp tree topology (Figure 6.3b). This again suggests that the acquisition of E1E2, accompanied by loss of E, was a defining event in the emergence of the pestiviruses from an LGF-like progenitor.

Both the E1 and E2 phylogenies indicate a common glycoprotein ancestry in Pegivirus/Hepacivirus and Pestivirus groups (Figure 6.4b,c), even though they are paraphyletic in the RdRp phylogeny (Figure 6.1a). Of note, the Wenling moray eel hepacivirus (which is basal to the pegivirus/hepacivirus RdRp lineage) sits at the intersection of the Pegivirus/Hepacivirus and Pestivirus E1 and E2 clades, consistent with deep ancestry. We could not detect any significant structural homology between E1E2 and E (Figure 6.2), or identify intermediate forms between these glycoprotein systems, further suggesting they are mechanistically distinct. We therefore propose that E1E2 represents a novel class of fusion protein. Moreover, the structural and sequence conservation within E1 and the basal portion of E2 (Figure 6.2 and 4b,c and Supplementary Figure 6.20) suggests a mechanistic role requiring experimental exploration.

On the basis of our combined analyses, we propose a *Flaviviridae* evolutionary history shaped by gains and losses of defining protein functions, as summarised in Figure 6.5. The most parsimonious interpretation of the data is that Orthoflavivirus/jingmenvirus and LGF/Pestivirus clades (lineage 1) arose from an ancestor that possessed the E glycoprotein and performed cap-dependent translation, necessitating MTase. By contrast, the Pegivirus/Hepacivirus clade (lineage 2) arose from an ancestor that possessed E1E2 glycoproteins and lacked MTase, implying reliance on IRES-dependent translation.

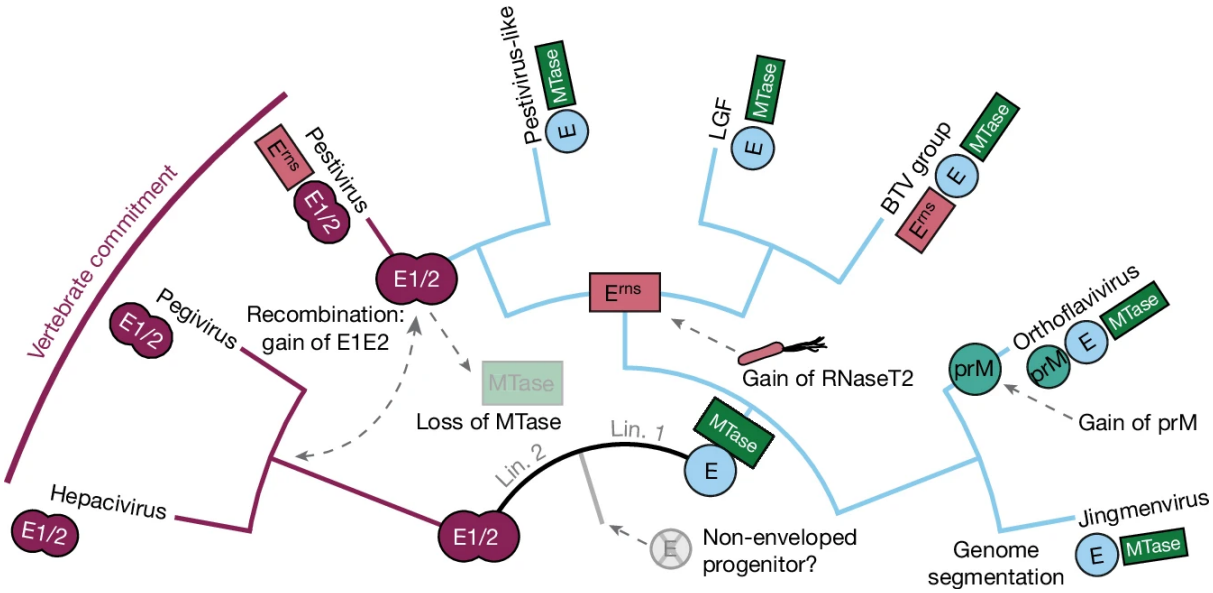



***Figure 6.42 Proposed evolutionary history of the Flaviviridae**. Illustrative cladogram showing the key protein acquisition and loss events across the major Flaviviridae clades. The two major lineages are labelled (Lin. 1 and Lin. 2), near the root. Each clade, displayed as a tip, is annotated with symbols representing the presence of key proteins. Branches are highlighted to denote the lineage-specific presence of envelope protein E (in light blue) or E1/E2 (in maroon). Major nodes are emphasised with larger symbols to infer the ancestral emergence of each protein within the Flaviviridae. Dashed lines and arrows denote the loss or gain of specific proteins, highlighting potential recombination events and gene transfers. Image of Aquifex pyrophilus by Guillaume Dera; CC BY 1.0 (https://creativecommons.org/licenses/by/1.0/).*

Compared with lineage 2 (where the hepaciviruses and pegiviruses share relatively conserved RdRp and glycoproteins), lineage 1 has undergone extensive diversification. Genome segmentation occurred in the jingmenviruses, with coincident divergence in their E glycoprotein (Garry and Garry 2020), including the apparent loss of its canonical fusion loop. The orthoflaviviruses gained prM, a partner to the E glycoprotein, probably derived from a host chaperonin (Vaney et al. 2022). A sister lineage gives rise to the LGF and Pestivirus clades, in which an ancestral species gained RNase T2 from bacteria. Whereas the LGF, including basal pesti-like viruses, possess E glycoprotein, all pestiviruses possess E1E2 glycoproteins, homologous to those found in the hepaciviruses and pegiviruses. This indicates a switch in glycoprotein systems through an inter-genus horizontal gene transfer; with concomitant loss of MTase. In lineages one and two, the presence of E1E2 (and IRES-dependent translation) is strictly associated with viruses of vertebrates, suggesting a molecular commitment to ecological niche.

The characteristics of the common ancestor of the entire *Flaviviridae* remain speculative and, on the basis of the taxonomic distribution of infected hosts and the existence of endogenous viral elements, may have originated over 900 million years ago (Bamford et al. 2022). This ancestor possibly contained core NS3 and NS5 proteins and potentially an MTase. Although the absence of RrmJ-like methyltransferases (as found in the *Flaviviridae)* in other members of the phylum Kitrinoviricota (Mushegian 2022), hints that MTase might have been acquired at the base of lineage 1. It remains unclear whether this progenitor possessed an envelope, therefore necessitating fusion glycoproteins; although it is noteworthy that, with the exception of *Togaviridae*, *Matonaviridae* and *Flaviviridae*, all Kitrinoviricota families (n = 21) are non-enveloped.

Ultimately, the origins of the E and E1E2 glycoproteins remain uncertain. We cannot exclude the possibility of hidden E glycoprotein horizontal gene transfer within the *Flaviviridae*. For instance, alternative phylogenetic models to the one in Figure 6.4a place the jingmenvirus E within the LGF clade (although this may be an artefact of long branch attraction; see Methods and Supplementary Figure 6.8). Moreover, we have observed apparent ‘genetic piracy’ by LGFs (for example, RNase T2 (Figure 6.3) and in the recent work of Petrone et al. (2024) and therefore these viruses may have acquired their glycoproteins by horizontal gene transfer from within the *Flaviviridae* or beyond. Resolving these questions will probably require the discovery of further novel species and the inclusion of diverse taxa from beyond the *Flaviviridae*, such as the *Matonaviridae*, *Togaviridae* and *Peribunyaviridae*, which also possess class II fusion proteins. However, these analyses, across a wide diversity of the RNA virosphere, are likely to challenge even the highly sensitive structure-driven approaches outlined here.

### 6.5 Discussion

The limited ability of sequence-based methods to detect deep homology has resulted in significant ambiguity regarding the distribution and classification of fusion glycoproteins across the *Flaviviridae*. Our work, using protein structure prediction, has discovered previously unknown glycoproteins in more than 100 species and reveal unambiguous structural and sequence similarity between E1E2 in the hepaciviruses, pegiviruses and pestiviruses, indicative of inter-genus genetic exchange. The absence of homology between the E glycoprotein (class II fusion mechanism) and E1E2, even in basal species, provides the strongest evidence yet of a novel fusion mechanism in the Pegivirus/Hepacivirus and Pestivirus groups. Through comparison to host tropism, we

found that E1E2 is strictly correlated with infection of vertebrates, suggesting a molecular commitment to virological niche.

Beyond biological insights, our work demonstrates that protein structure prediction and structure-guided homology searches outperform the gold standard sequence-based approaches to provide unprecedented clarity to the evolution of viruses. Whereas AlphaFold-based methods offer unparalleled accuracy (Lee et al. 2023), protein language model-based systems such as ESMFold may be more capable of exploring the 'viral dark matter' revealed by metatranscriptomics. In sum, our study offers a new state-of-the-art approach for understanding the diversity and distribution of protein functions throughout the virosphere.

### 6.6 Data availability

All underlying data, including sequences and structures, are available on *Zenodo* at https://doi.org/10.5281/zenodo.11092288. The virus sequences assembled from SRA mining in this study are available in the Third-Party Annotation Section of the DDBJ/ENA/GenBank databases under the accession numbers TPA: BK067806–BK067816. Original code and scripts related to the phylogenetic analyses are available at https://doi.org/10.5281/zenodo.11092288.

### 6.7 Acknowledgements

The authors acknowledge the University of Sydney's high-performance computing cluster, Artemis, for providing the computing resources used for this study. E.C.H. is supported by a National Health and Medical Research Council Investigator award (GNT2017197). J.C.O.M. is supported by the Australian Government's Research Training Program Scholarship. J.G., S.L. and M.R.O. are supported by the Wellcome Trust and Royal Society, through a Sir Henry Dale Fellowship (107653/Z/15/Z). K.T. was supported by a Lord Kelvin Adam Smith Fellowship from the University of Glasgow. J.G. is also supported by the MRC-University of Glasgow Centre for Virus Research core support from the Medical Research Council/UKRI (MC_UU_00034/1) and by a Medical Research Foundation Emerging Leaders Prize (MRF-ELP-VAH-23-107).

### 6.8 References


Aitkenhead, H., et al. (2024), 'Structural comparison of typical and atypical E2 pestivirus glycoproteins', *Structure,* 32 (3).

Arhab, Y., et al. (2020), 'Dissemination of Internal Ribosomal Entry Sites (IRES) Between Viruses by Horizontal Gene Transfer', *Viruses-Basel,* 12 (6).

Babaian, A. and Edgar, R. (2022), 'Ribovirus classification by a polymerase barcode sequence', *PeerJ,* 10, e14055.

Bamford, C. G., et al. (2022), 'Comparative analysis of genome-encoded viral sequences reveals the evolutionary history of flavivirids (family Flaviviridae)', *Virus Evolution,* 8 (2).

Barrio-Hernandez, I., et al. (2023), 'Clustering predicted structures at the scale of the known protein universe', *Nature,* 622 (7983), 637.

Blitvich, B. J. and Firth, A. E. (2017), 'A review of flaviviruses that have no known arthropod vector', *Viruses,* 9 (6), 154.

Bolger, A. M., Lohse, M., and Usadel, B. (2014), 'Trimmomatic: a flexible trimmer for Illumina sequence data', *Bioinformatics,* 30 (15), 2114-20.

Buchfink, B., Reuter, K., and Drost, H.-G. (2021), 'Sensitive protein alignments at tree-of-life scale using DIAMOND', *Nature Methods,* 18 (4), 366-68.

Buchfink, B., et al. (2023), 'Sensitive clustering of protein sequences at tree-of-life scale using DIAMOND DeepClust', *bioRxiv*, 2023.01.24.525373.

Capella-Gutiérrez, S., Silla-Martínez, J. M., and Gabaldón, T. (2009), 'trimAl: a tool for automated alignment trimming in large-scale phylogenetic analyses', *Bioinformatics,* 25 (15), 1972-73.

Chamberlain, S. A. and Szöcs, E. (2013), 'taxize: taxonomic search and retrieval in R', *F1000Research,* 2, 191.

Charon, J., et al. (2022), 'RdRp-scan: A bioinformatic resource to identify and annotate divergent RNA viruses in metagenomic sequence data', *Virus Evolution,* 8 (2), veac082.

Chen, Y. M., et al. (2022), 'RNA viromes from terrestrial sites across China expand environmental viral diversity', *Nature Microbiology,* 7 (8), 1312.

Chernomor, O., von Haeseler, A., and Minh, B. Q. (2016), 'Terrace Aware Data Structure for Phylogenomic Inference from Supermatrices', *Systematic Biology,* 65 (6), 997-1008.

Costa, V. A., et al. (2022), 'Limited Cross-Species Virus Transmission in a Spatially Restricted Coral Reef Fish Community', *bioRxiv*, 2022.05.17.492384.

Dahan, J., et al. (2022), 'A novel flavi-like virus in alfalfa (Medicago sativa L.) crops along the Snake River Valley', *Viruses,* 14 (6), 1320.

de la Peña, A. T., et al. (2022), 'Structure of the hepatitis C virus E1E2 glycoprotein complex', *Science,* 378 (6617), 263-69.

Debat, H. and Bejerman, N. (2023), 'Two novel flavi-like viruses shed light on the plant-infecting koshoviruses', *Archives of Virology,* 168 (7), 184.

Deorowicz, S., Debudaj-Grabysz, A., and Gudys, A. (2016), 'FAMSA: Fast and accurate multiple sequence alignment of huge protein families', *Scientific Reports,* 6.

Dessau, M. and Modis, Y. (2013), 'Crystal structure of glycoprotein C from Rift Valley fever virus', *Proceedings of the National Academy of Sciences of the United States of America,* 110 (5), 1696-701.

Dong, X., et al. (2021), 'A Novel Virus of Flaviviridae Associated with Sexual Precocity in Macrobrachium rosenbergii', *mSystems,* 6 (3), e00003-21.

Dunay, E., et al. (2023), 'Viruses in sanctuary chimpanzees across Africa', *American Journal of Primatology,* 85 (1), e23452.

Edgar, R. C. (2022), 'Muscle5: High-accuracy alignment ensembles enable unbiased assessments of sequence homology and phylogeny', *Nature Communications,* 13 (1), 6968.

Edgar, R. C., et al. (2022), 'Petabase-scale sequence alignment catalyses viral discovery', *Nature,* 602 (7895), 142-47.

Egloff, M. P., et al. (2002), 'An RNA cap (nucleoside-2′-O-)-methyltransferase in the flavivirus RNA polymerase NS5: crystal structure and functional characterization', *EMBO Journal,* 21 (11), 2757-68.

El Omari, K., et al. (2013), 'Structure of a Pestivirus Envelope Glycoprotein E2 Clarifies Its Role in Cell Entry', *Cell Reports,* 3 (1), 30-35.

Elbadry, M. A., et al. (2023), 'Diversity and Genetic Reassortment of Keystone Virus in Mosquito Populations in Florida', *Am J Trop Med Hyg,* 108 (6), 1256-63.

Fédry, J., et al. (2017), 'The Ancient Gamete Fusogen HAP2 Is a Eukaryotic Class II Fusion Protein', *Cell,* 168 (5), 904.

Ferron, F., et al. (2021), 'The enzymes for genome size increase and maintenance of large (+)RNA viruses', *Trends in Biochemical Sciences,* 46 (11), 866-77.

Fu, L., et al. (2012), 'CD-HIT: accelerated for clustering the next-generation sequencing data', *Bioinformatics,* 28 (23), 3150-52.

Gabler, F., et al. (2020), 'Protein Sequence Analysis Using the MPI Bioinformatics Toolkit', *Current Protocols in Bioinformatics,* 72 (1), e108.

Garry, C. E. and Garry, R. E. (2020), 'Proteomics Computational Analyses Suggest That the Envelope Glycoproteins of Segmented Jingmen Flavi-Like Viruses are Class II Viral Fusion Proteins ( $\beta$ -Penetrenes) with Mucin-Like Domains', *Viruses-Basel,* 12 (3).

Gilmer, D., Ratti, C., and Consortium, I. R. (2017), 'ICTV Virus taxonomy profile: Benyviridae', *The Journal of General Virology,* 98 (7), 1571.

Grove, J. and Marsh, M. (2011), 'The cell biology of receptor-mediated virus entry', *Journal of Cell Biology,* 195 (7), 1071-82.

Guardado-Calvo, P. and Rey, F. A. (2021), 'The Viral Class II Membrane Fusion Machinery: Divergent Evolution from an Ancestral Heterodimer', *Viruses-Basel,* 13 (12).

Guindon, S., et al. (2010), 'New algorithms and methods to estimate maximum-likelihood phylogenies: assessing the performance of PhyML 3.0', *Systematic Biology,* 59 (3), 307-21.

Guo, G., et al. (2023), 'Virome Analysis Provides an Insight into the Viral Community of Chinese Mitten Crab Eriocheir sinensis', *Microbiology Spectrum,* 11 (4), e0143923.

Hoang, D. T., et al. (2017), 'UFBoot2: Improving the Ultrafast Bootstrap Approximation', *Molecular Biology and Evolution,* 35 (2), 518-22.

Hou, X., et al. (2024), 'Using artificial intelligence to document the hidden RNA virosphere', *Cell,* 187 (24), 6929-42.e16.

Hubálek, Z. and Halouzka, J. (1999), 'West Nile fever -: a reemerging mosquito-borne viral disease in Europe', *Emerging Infectious Diseases,* 5 (5), 643-50.

Jia, H. X., et al. (2022), 'Crystal Structures of Flavivirus NS5 Guanylyltransferase Reveal a GMP-Arginine Adduct', *Journal of Virology,* 96 (14).

Jombart, T., et al. (2017), 'treespace: Statistical exploration of landscapes of phylogenetic trees', *Molecular Ecology Resources,* 17 (6), 1385-92.

Jones, P., et al. (2014), 'InterProScan 5: genome-scale protein function classification', *Bioinformatics,* 30 (9), 1236-40.

Jumper, J., et al. (2021), 'Highly accurate protein structure prediction with AlphaFold', *Nature,* 596 (7873), 583-89.

Kalyaanamoorthy, S., et al. (2017), 'ModelFinder: fast model selection for accurate phylogenetic estimates', *Nature Methods,* 14 (6), 587-89.

Kartashov, M. Y., et al. (2023), 'Novel Flavi-like virus in ixodid ticks and patients in Russia', *Ticks and Tick-Borne Diseases,* 14 (2), 102101.

Katoh, K. and Standley, D. M. (2013), 'MAFFT multiple sequence alignment software version 7: improvements in performance and usability', *Molecular Biology and Evolution,* 30 (4), 772-80.

Kearse, M., et al. (2012), 'Geneious Basic: an integrated and extendable desktop software platform for the organization and analysis of sequence data', *Bioinformatics,* 28 (12), 1647-49.

Kendall, M. and Colijn, C. (2016), 'Mapping phylogenetic trees to reveal distinct patterns of evolution', *Molecular Biology and Evolution,* 33 (10), 2735-43.

Khan, A. G., et al. (2014), 'Structure of the core ectodomain of the hepatitis C virus envelope glycoprotein 2', *Nature,* 509 (7500), 381.

Kielian, M. and Rey, F. A. (2006), 'Virus membrane-fusion proteins: more than one way to make a hairpin', *Nature Reviews Microbiology,* 4 (1), 67-76.

Kobayashi, K., et al. (2013), 'Gentian Kobu-sho-associated virus: a tentative, novel double-stranded RNA virus that is relevant to gentian Kobu-sho syndrome', *Journal of General Plant Pathology,* 79 (1), 56-63.

Kong, L., et al. (2013), 'Hepatitis C Virus E2 Envelope Glycoprotein Core Structure', *Science,* 342 (6162), 1090-94.

Kong, Y., et al. (2022), 'Metatranscriptomics Reveals the Diversity of the Tick Virome in Northwest China', *Microbiol Spectrum,* 10 (5), e0111522.

Krey, T., et al. (2012), 'Crystal Structure of the Pestivirus Envelope Glycoprotein Erns and Mechanistic Analysis of Its Ribonuclease Activity', *Structure,* 20 (5), 862-73.

Ladner, J. T., et al. (2016), 'A Multicomponent Animal Virus Isolated from Mosquitoes', *Cell Host & Microbe,* 20 (3), 357-67.

Le, T. K. and Vinh, L. (2020), 'FLAVI: An Amino Acid Substitution Model for Flaviviruses', *Journal of Molecular Evolution,* 88 (5), 445-52.

Lee, S. W., et al. (2023), 'Petascale Homology Search for Structure Prediction', *bioRxiv*.

Legendre, P. and Legendre, L. (2012), *Numerical ecology* (Elsevier).

Letunic, I. and Bork, P. (2021), 'Interactive Tree Of Life (iTOL) v5: an online tool for phylogenetic tree display and annotation', *Nucleic Acids Research,* 49 (W1), W293-W96.

Li, D., et al. (2015), 'MEGAHIT: an ultra-fast single-node solution for large and complex metagenomics assembly via succinct de Bruijn graph', *Bioinformatics,* 31 (10), 1674-76.

Li, L., et al. (2008), 'The flavivirus precursor membrane-envelope protein complex: Structure and maturation', *Science,* 319 (5871), 1830-34.

Li, Y., et al. (2013), 'Crystal structure of glycoprotein E2 from bovine viral diarrhea virus', *Proceedings of the National Academy of Sciences of the United States of America,* 110 (17), 6805-10.

Li, Z., et al. (2020), 'FATCAT 2.0: towards a better understanding of the structural diversity of proteins', *Nucleic Acids Research,* 48 (W1), W60-W64.

Lin, Z., et al. (2023), 'Evolutionary-scale prediction of atomic-level protein structure with a language model', *Science,* 379 (6637), 1123-30.

MacIntosh, G. C. (2011), 'RNase T2 Family: Enzymatic Properties, Functional Diversity, and Evolution of Ancient Ribonucleases', in A. W. Nicholson (ed.), *ribonucleases* (26), 89-114.

Meng, E. C., et al. (2023), 'UCSF ChimeraX: Tools for structure building and analysis', *Protein Science,* 32 (11).

Metcalf, M. C., et al. (2023), 'Structure of engineered hepatitis C virus E1E2 ectodomain in complex with neutralizing antibodies', *Nature Communications,* 14 (1).

Mifsud, J. C. O. (2023), 'BatchArtemisSRAMiner: v1.0.0', (Zenodo).

Mifsud, J. C. O., et al. (2022), 'Transcriptome mining extends the host range of the Flaviviridae to non-bilaterians', *Virus Evolution,* 9 (1).

Mifsud, J. C. O., et al. (2023), 'Transcriptome mining extends the host range of the Flaviviridae to non-bilaterians', *Virus Evolution,* 9 (1).

Mifsud, J. C. O., et al. (2024), 'Underlying data for "Mapping glycoprotein structure reveals Flaviviridae evolutionary history" (Version v2)', (Zenodo).

Minh, B. Q., et al. (2020), 'IQ-TREE 2: new models and efficient methods for phylogenetic inference in the genomic era', *Molecular Biology and Evolution,* 37 (5), 1530-34.

Mirdita, M., et al. (2017), 'Uniclust databases of clustered and deeply annotated protein sequences and alignments', *Nucleic Acids Research,* 45 (D1), D170-D76.

Mirdita, M., et al. (2022), 'ColabFold: making protein folding accessible to all', *Nature Methods,* 19 (6), 679-+.

Moi, D., et al. (2023), 'Structural phylogenetics unravels the evolutionary diversification of communication systems in gram-positive bacteria and their viruses', *bioRxiv*.

Mushegian, A. (2022), 'Methyltransferases of Riboviria', *BioMolecules,* 12 (9).

Noble, C. G., et al. (2016), 'A Conserved Pocket in the Dengue Virus Polymerase Identified through Fragment-based Screening', *Journal of Biological Chemistry,* 291 (16), 8541-48.

Oliver, M. R., et al. (2023), 'Structures of the Hepaci-, Pegi-, and Pestiviruses envelope proteins suggest a novel membrane fusion mechanism', *PLoS Biology,* 21 (7).

Paradis, E. and Schliep, K. (2019), 'ape 5.0: an environment for modern phylogenetics and evolutionary analyses in R', *Bioinformatics,* 35 (3), 526-28.

Paraskevopoulou, S., et al. (2021), 'Viromics of extant insect orders unveil the evolution of the flavi-like superfamily', *Virus Evolution,* 7 (1).

Perveen, N., et al. (2023), 'Virome diversity of Hyalomma dromedarii ticks collected from camels in the United Arab Emirates', *Veterinary World,* 16 (3), 439-48.

Petrone, M. E., et al. (2024), 'A ~40-kb flavi-like virus does not encode a known error-correcting mechanism', *Proceedings of the National Academy of Sciences,* 121 (30), e2403805121.

Postler, T. S., et al. (2023), 'Renaming of the genus Flavivirus to Orthoflavivirus and extension of binomial species names within the family Flaviviridae', *Archives of Virology,* 168 (9).

Potter, S. C., et al. (2018), 'HMMER web server: 2018 update', *Nucleic Acids Research,* 46 (W1), W200-W04.

Puente-Lelievre, C., et al. (2023), 'Tertiary-interaction characters enable fast, model-based structural phylogenetics beyond the twilight zone', *bioRxiv*.

Puente-Lelievre, C., et al. (2024), 'Tertiary-interaction characters enable fast, model-based structural phylogenetics beyond the twilight zone', *bioRxiv*, 2023.12.12.571181.

Qin, X.-C., et al. (2014), 'A tick-borne segmented RNA virus contains genome segments derived from unsegmented viral ancestors', *Proceedings of the National Academy of Sciences,* 111 (18), 6744-49.

Rambaut, A. and Drummond, A. (2012), 'FigTree: Tree figure drawing tool, version 1.4.0', (Institute of Evolutionary Biology, University of Edinburgh).

Renner, M., et al. (2021), 'Flavivirus maturation leads to the formation of an occupied lipid pocket in the surface glycoproteins', *Nature Communications,* 12 (1).

Revell, L. J. (2024), 'phytools 2.0: an updated R ecosystem for phylogenetic comparative methods (and other things)', *PeerJ,* 12.

Rey, F. A. and Lok, S. M. (2018), 'Common Features of Enveloped Viruses and Implications for Immunogen Design for Next-Generation Vaccines', *Cell,* 172 (6), 1319-34.

Rey, F. A., et al. (1995), 'The Envelope Glycoprotein from Tick-Borne Encephalitis-Virus at 2 Angstrom Resolution', *Nature,* 375 (6529), 291-98.

Rolland, C., La Scola, B., and Levasseur, A. (2020), 'How Tupanvirus Degrades the Ribosomal RNA of Its Amoebal Host? The Ribonuclease T2 Track', *Frontiers in Microbiology,* 11, 1691.

Saberi, A., et al. (2018), 'A planarian nidovirus expands the limits of RNA genome size', *PLOS Pathogens,* 14 (11), e1007314.

Schwery, O. and O'Meara, B. C. (2016), 'MonoPhy: a simple R package to find and visualize monophyly issues', *PeerJ Computer Science,* 2, e56.

Shi, M., et al. (2016a), 'Divergent viruses discovered in arthropods and vertebrates revise the evolutionary history of the Flaviviridae and related viruses', *Journal of Virology,* 90 (2), 659-69.

Shi, M., et al. (2016b), 'Redefining the invertebrate RNA virosphere', *Nature,* 540 (7634), 539.

Sievers, F., et al. (2011), 'Fast, scalable generation of high-quality protein multiple sequence alignments using Clustal Omega', *Molecular Systems Biology,* 7 (1), 539.

Simmonds, P., et al. (2017), 'ICTV virus taxonomy profile: Flaviviridae', *The Journal of General Virology,* 98 (1), 2.

Steinegger, M. and Söding, J. (2017), 'MMseqs2 enables sensitive protein sequence searching for the analysis of massive data sets', *Nature Biotechnology,* 35 (11), 1026-28.

Steinegger, M., et al. (2019), 'HH-suite3 for fast remote homology detection and deep protein annotation', *BMC Bioinformatics,* 20 (1).

Teufel, F., et al. (2022), 'SignalP 6.0 predicts all five types of signal peptides using protein language models', *Nature Biotechnology,* 40 (7), 1023-+.

Thomas Hackl, M. J. A. (2022), 'gggenomes: A Grammar of Graphics for Comparative Genomics', (R package version 0.9.5.9000).

Urayama, S. I., Takaki, Y., and Nunoura, T. (2016), 'FLDS: A Comprehensive dsRNA Sequencing Method for Intracellular RNA Virus Surveillance', *Microbes and Environments,* 31 (1), 33-40.

van Kempen, M., et al. (2023), 'Fast and accurate protein structure search with Foldseek', *Nature Biotechnology*, 1-4.

Vaney, M. C., et al. (2022), 'Evolution and activation mechanism of the flavivirus class II membrane-fusion machinery', *Nature Communications,* 13 (1).

Walls, A. C., et al. (2020), 'Structure, Function, and Antigenicity of the SARS-CoV-2 Spike Glycoprotein', *Cell,* 181 (2), 281.

Wang, Z. D., et al. (2019), 'A New Segmented Virus Associated with Human Febrile Illness in China', *New England Journal of Medicine,* 380 (22), 2116-25.

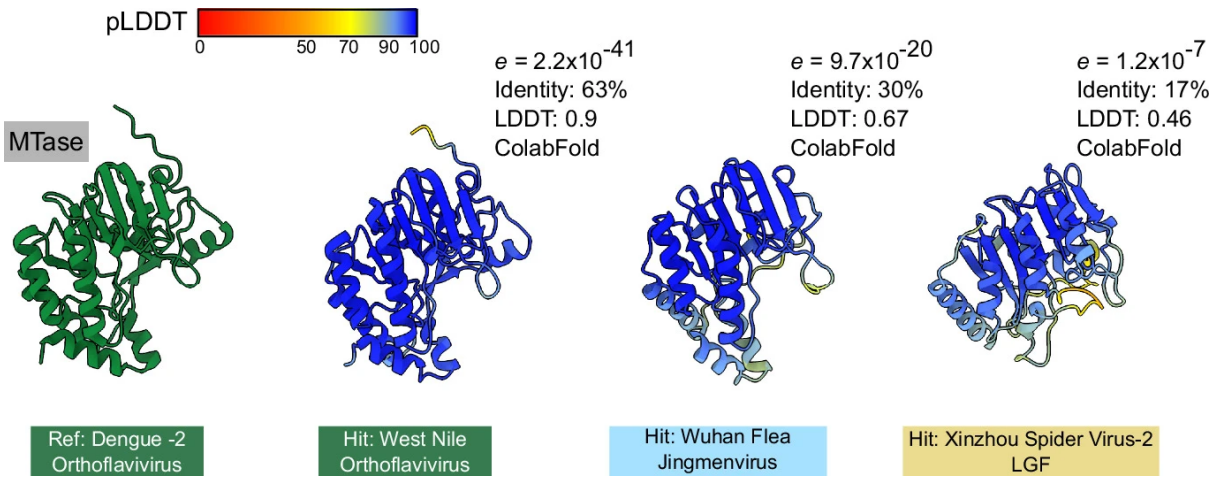
Winter, D. J. (2017), 'rentrez: An R package for the NCBI eUtils API', (PeerJ Preprints).

Young, C., Meng, S., and Moshiri, N. (2022), 'An Evaluation of Phylogenetic Workflows in Viral Molecular Epidemiology', *Viruses,* 14 (4), 774.

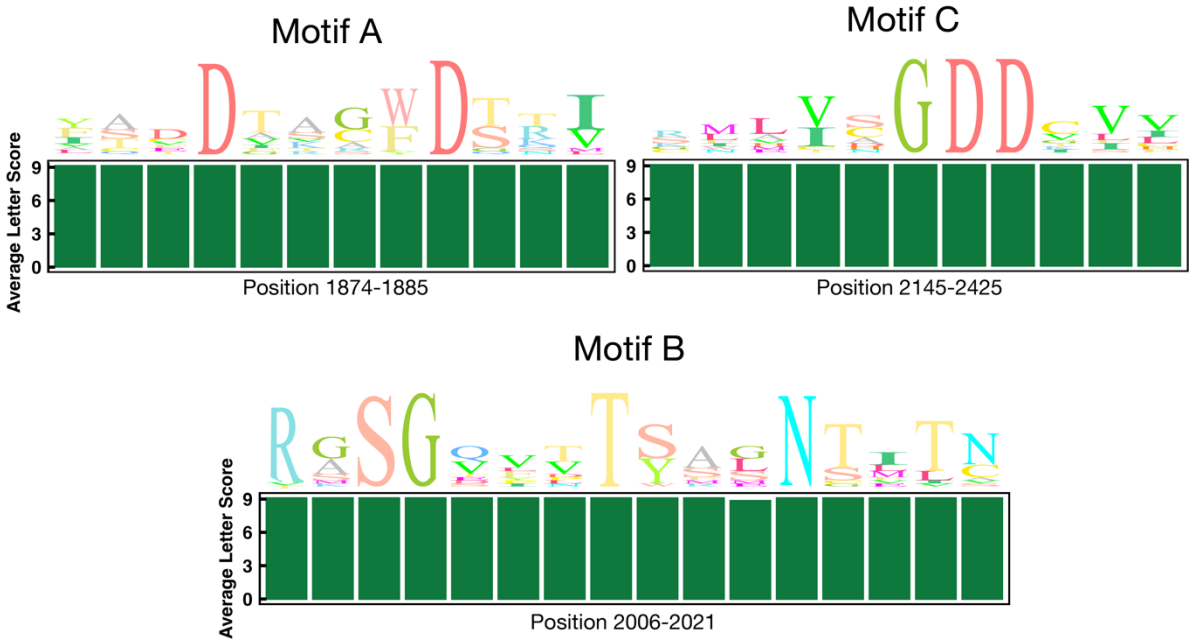
Yu, G., et al. (2017), 'ggtree: an R package for visualization and annotation of phylogenetic trees with their covariates and other associated data', *Methods in Ecology and Evolution,* 8 (1), 28-36.

Zimmermann, L., et al. (2018), 'A Completely Reimplemented MPI Bioinformatics Toolkit with a New HHpred Server at its Core', *Journal of Molecular Biology,* 430 (15), 2237-43.

## 6.9 Supplementary Information

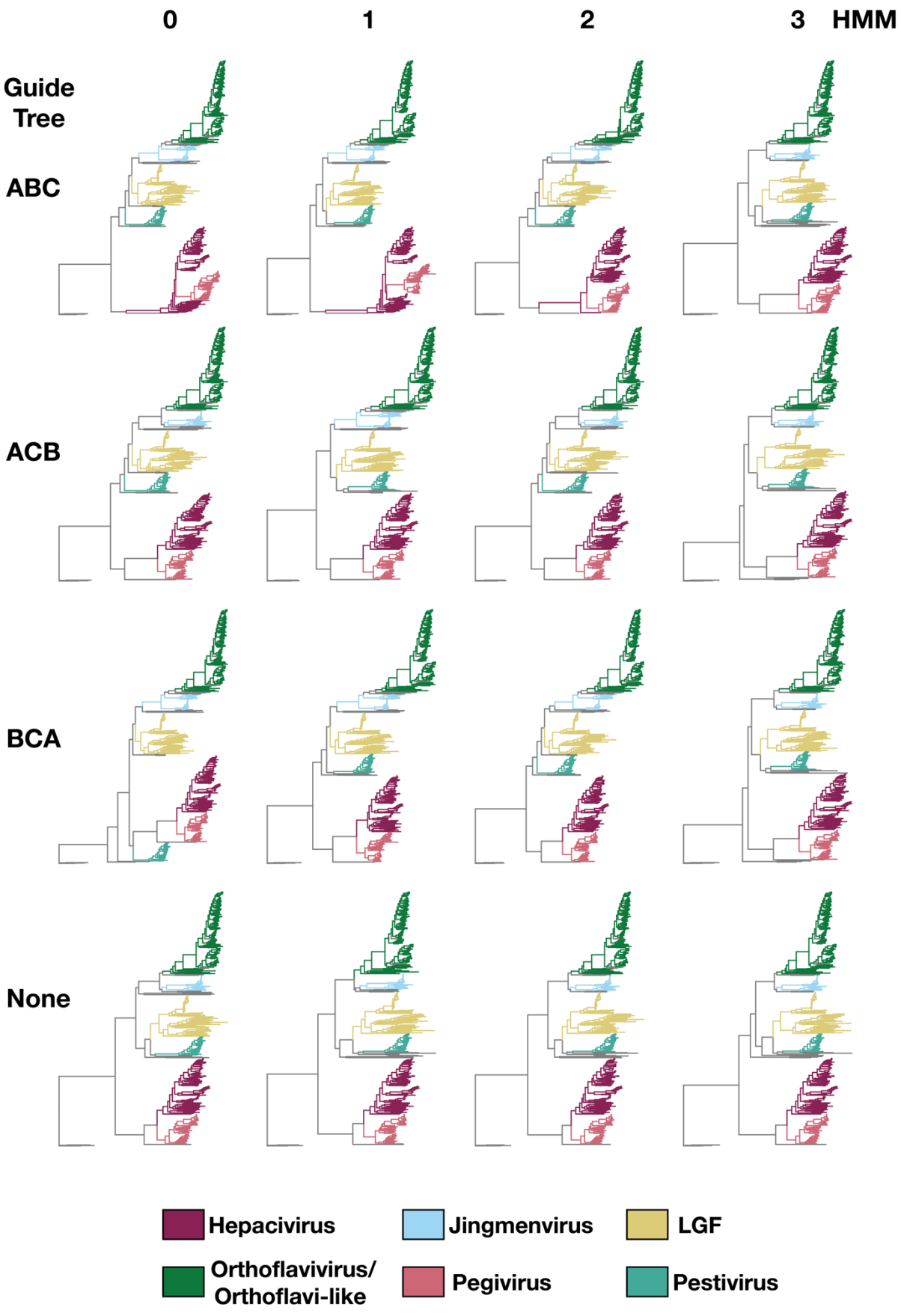



***Supplementary Figure 6.1. Foldseek detection of methyltransferase in diverse viruses.*** *DENV-2 reference structure and Foldseek hits for MTase. For each hit, only the Foldseek-aligned residues are shown, metrics provide e-value, sequence identity (%), structural alignment score (LDDT, ranging from 0 to 1), and protein structure prediction method. Predicted structures are colour-coded by pLDDT confidence scores, as shown in the key.*



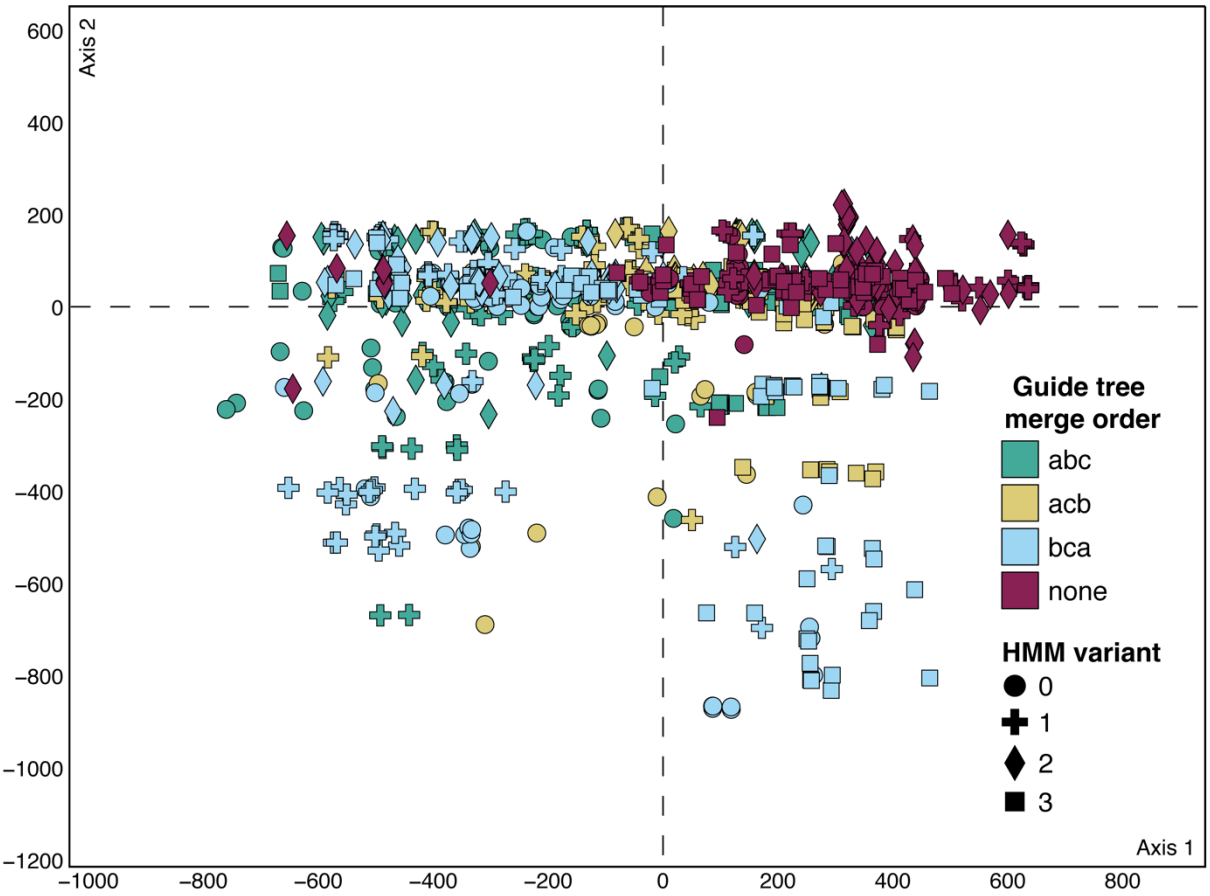

***Supplementary Figure 6.2. Letter confidence measurements for the NS5b palmprint motifs.*** *An average of letter confidence scores for the NS5b palmprint motifs (n = 448 sequences), which ranges from 0 (low confidence) to 9 (high confidence). Confidence was calculated by comparing the position of each amino acid in the MUSCLE replicates to a reference alignment, untrimmed Tree 18 (flaviviridae_ns5_untrimmed_MUSCLE_20230630).*

*The position of each motif in the untrimmed Tree 18 MUSCLE alignment is depicted under the barplot.*

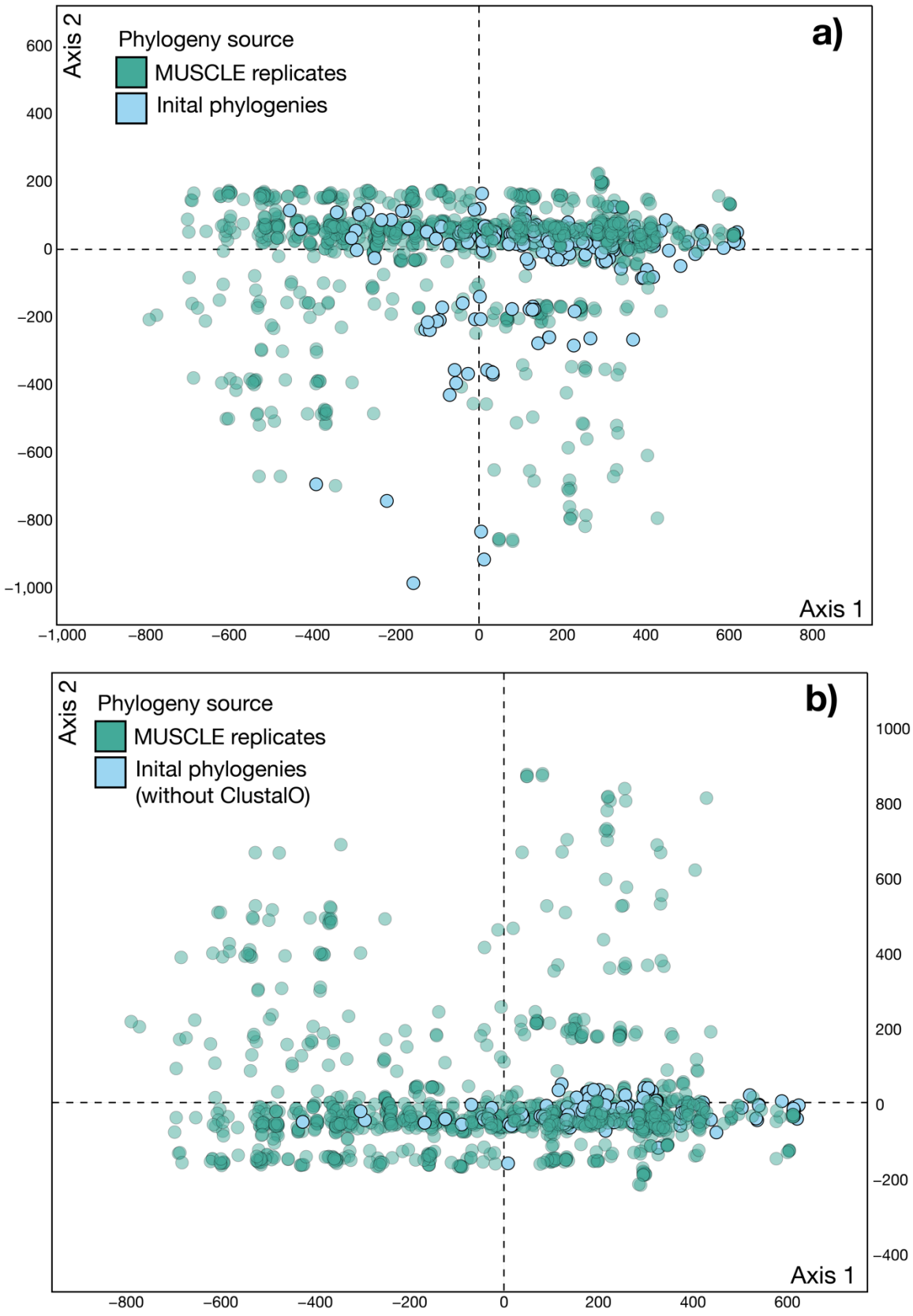



***Supplementary Figure 6.3. NS5b phylogenies estimated from the stratified MUSCLE replicates.*** *Guide tree merge order is held fixed on one axis, and HMM parameters are held*

*fixed on the other axis. Branches are coloured by Flaviviridae clade as defined in Supplementary Table 3. Tips corresponding to unclassified sequences (e.g., FJUN, PLUN and HPUN) were coloured black. Phylogenies were inferred using IQTREE 2 and the LG+F+R10 amino acid substitution model. Phylogenies are rooted on the Tombusvirus outgroup.*

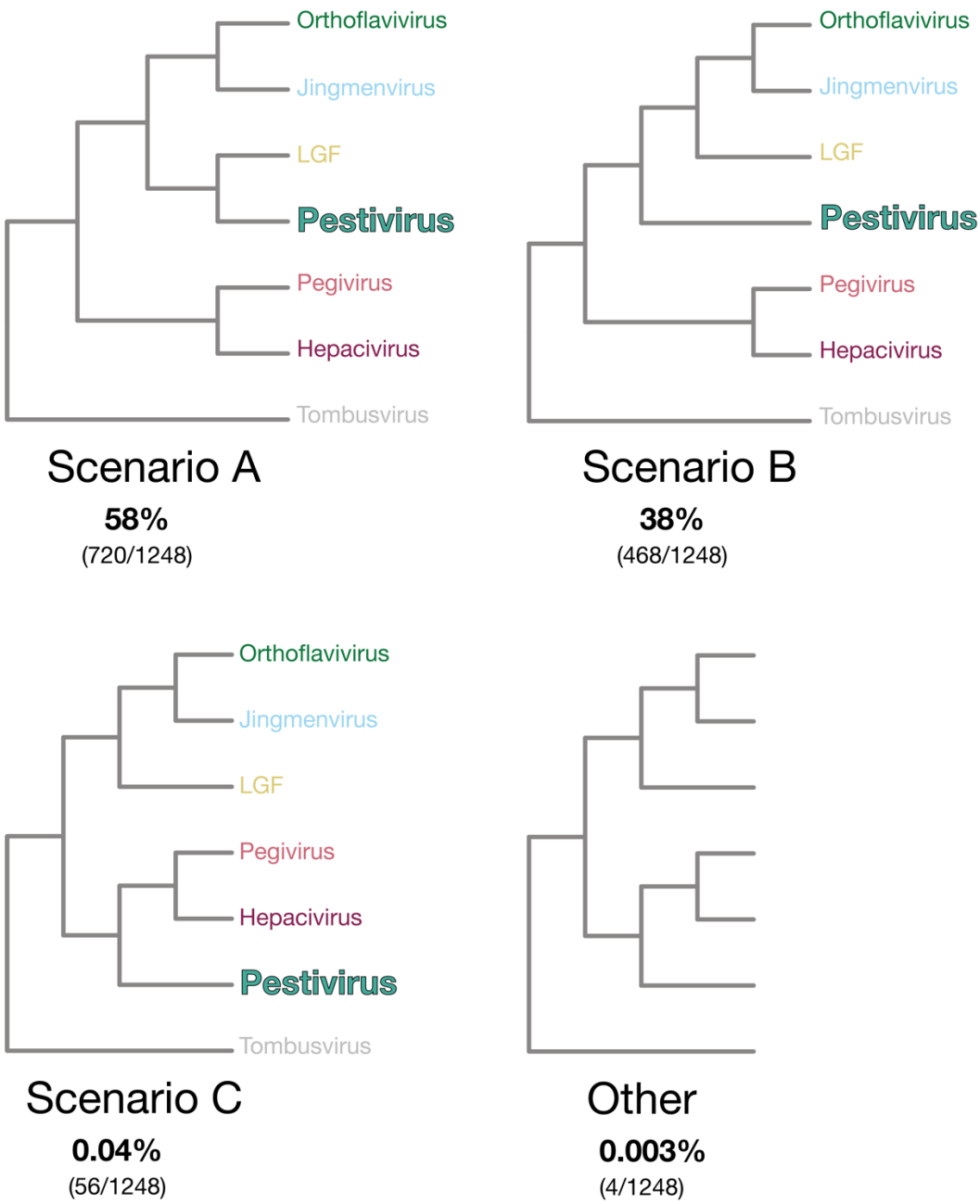



***Supplementary Figure 6.4. Two-dimensional MDS plot of the NS5b stratified MUSCLE replicate phylogenies.*** *Points are coloured by guide tree merge order and shapes indicate HMM parameters. Phylogeny none.1_cons17.5_gt0.7_20240420_LG.treefile was excluded from the scatterplot as it was a clear outlier.*

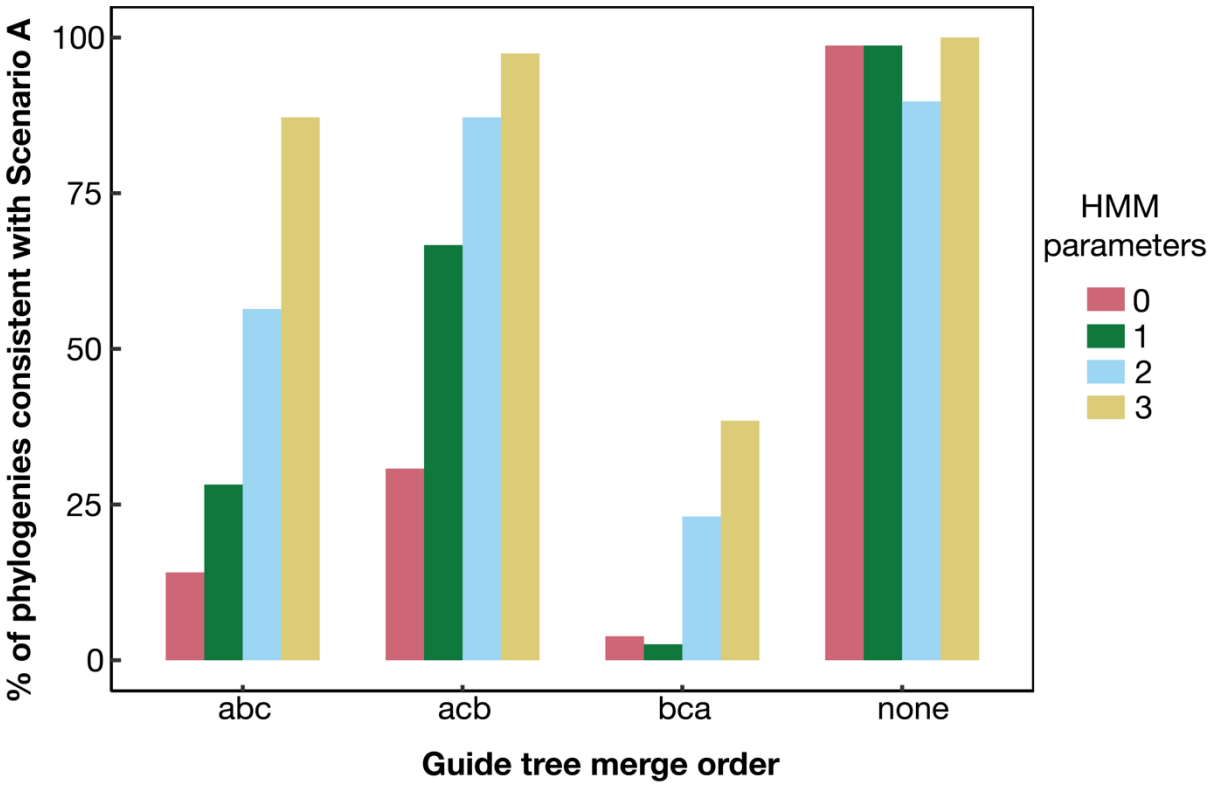



***Supplementary Figure 6.5. Two-dimensional MDS plot of the NS5b stratified MUSCLE replicate phylogenies.*** *a, Coloured by whether they were generated in the initial alignment analysis or from the stratified MUSCLE replicates b, Excluding trees generated from Clustal*

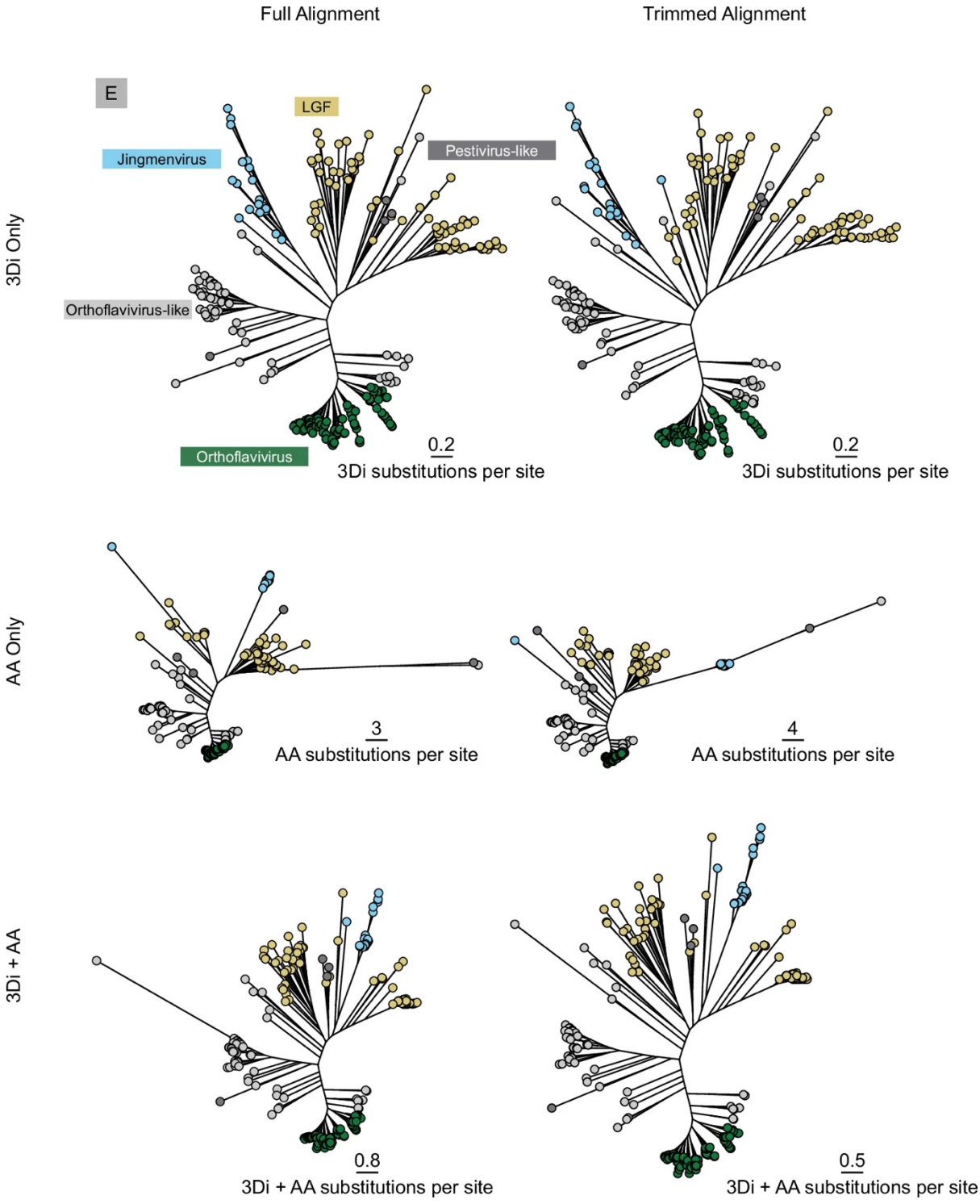
*Omega alignments.none.1_cons17.5_gt0.7_20240420_LG.treefile was excluded from the scatterplot as it was a clear outlier.*




***Supplementary Figure 6.6. Illustrative cladogram showing the major Flaviviridae clades topologies observed in trees estimated from the stratified MUSCLE alignment replicates.*** *Each clade is individually coloured and the pestiviruses bolded to highlight the differences in topology between the scenarios. Under each scenario the percentage and number of phylogenies it was observed in are shown.*

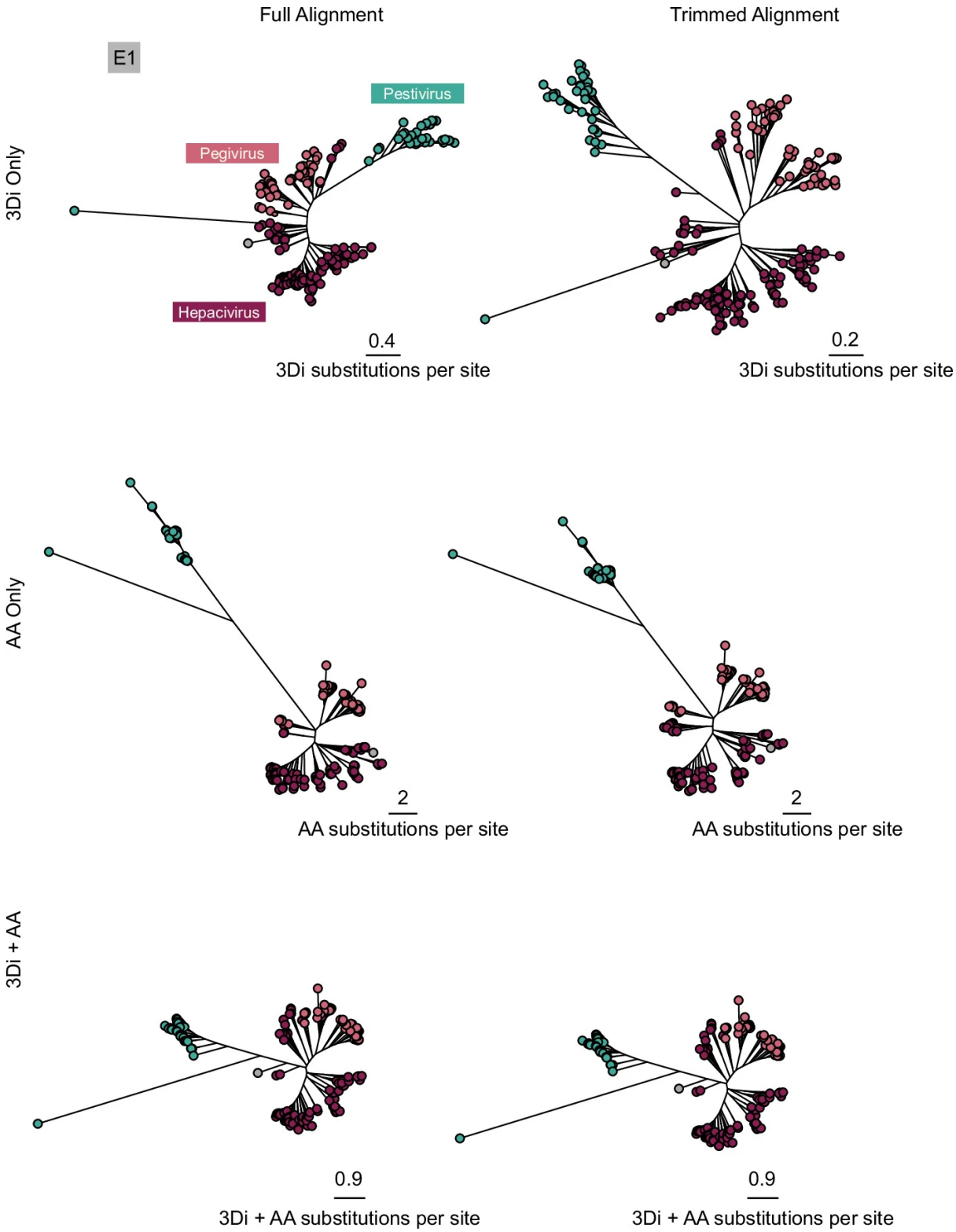



***Supplementary Figure 6.7. Barplot showing the percentage of NS5b stratified MUSCLE replicate phylogenies that are consistent with Scenario A.*** *Consistent refers to the deep branching clades (see Methods). Tree are compared with Tree 18 (Scenario A, Supplementary Figure 6.5) and compared across guide tree merge order levels and coloured by HMM parameter choice.*



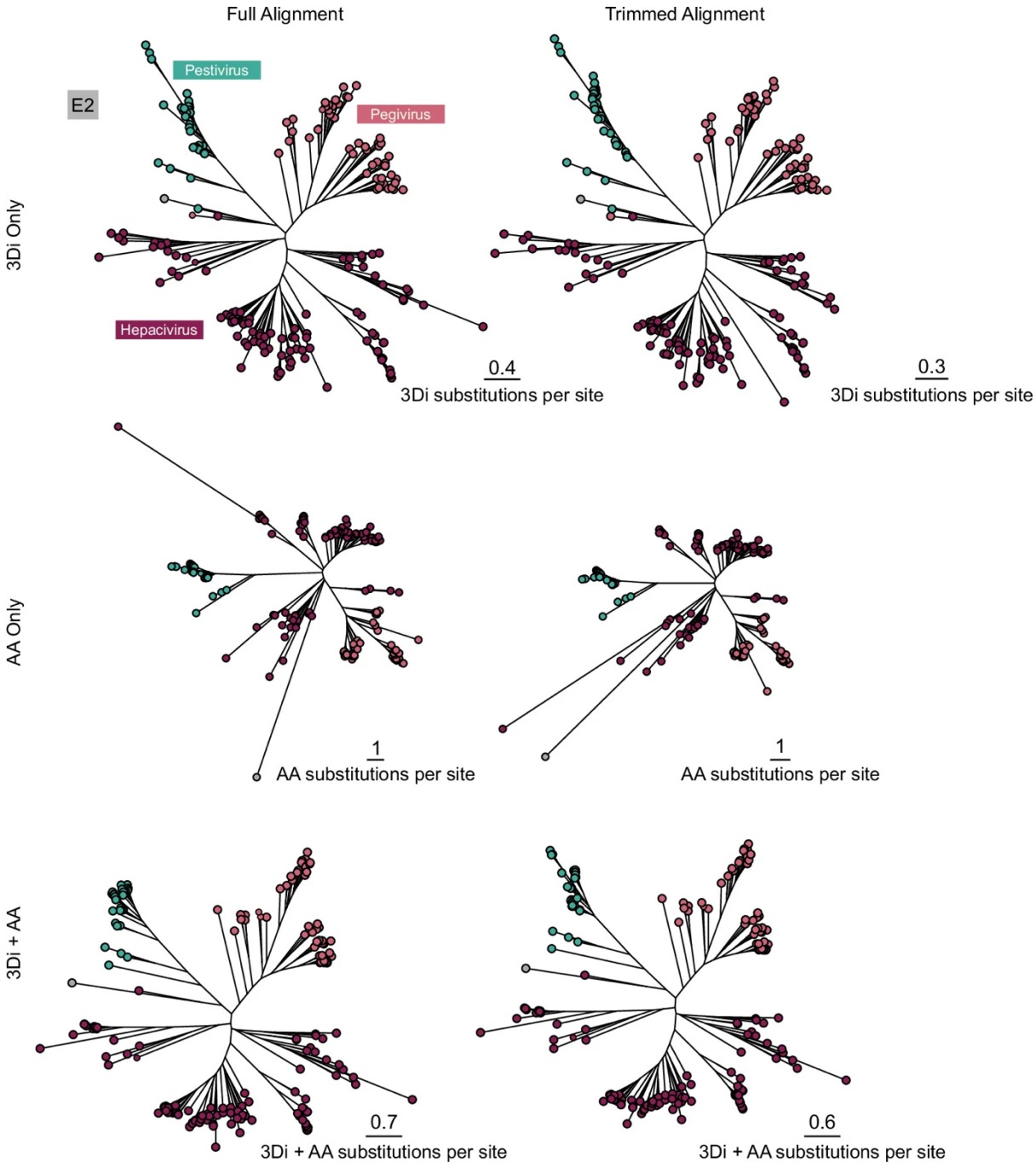

***Supplementary Figure 6.8. Structurally aligned E glycoprotein phylogenies.*** *Protein sequences were aligned using their 3Di structural representation (see Methods for details). Phylogenies were reconstructed using 3Di sequence alone (top), amino acid (AA) sequence alone (middle) or combined 3Di and AA sequence. Right hand trees are derived from alignments trimmed with a gap threshold of 35%. Scale bars indicate substitutions per site for either the 3Di, AA or combined sequences, respectively. Tip shapes are colour-coded by*

*genus/subclade as in Figure 6.1a All phylogenetic trees are provided in the associated Zenodo repository.*



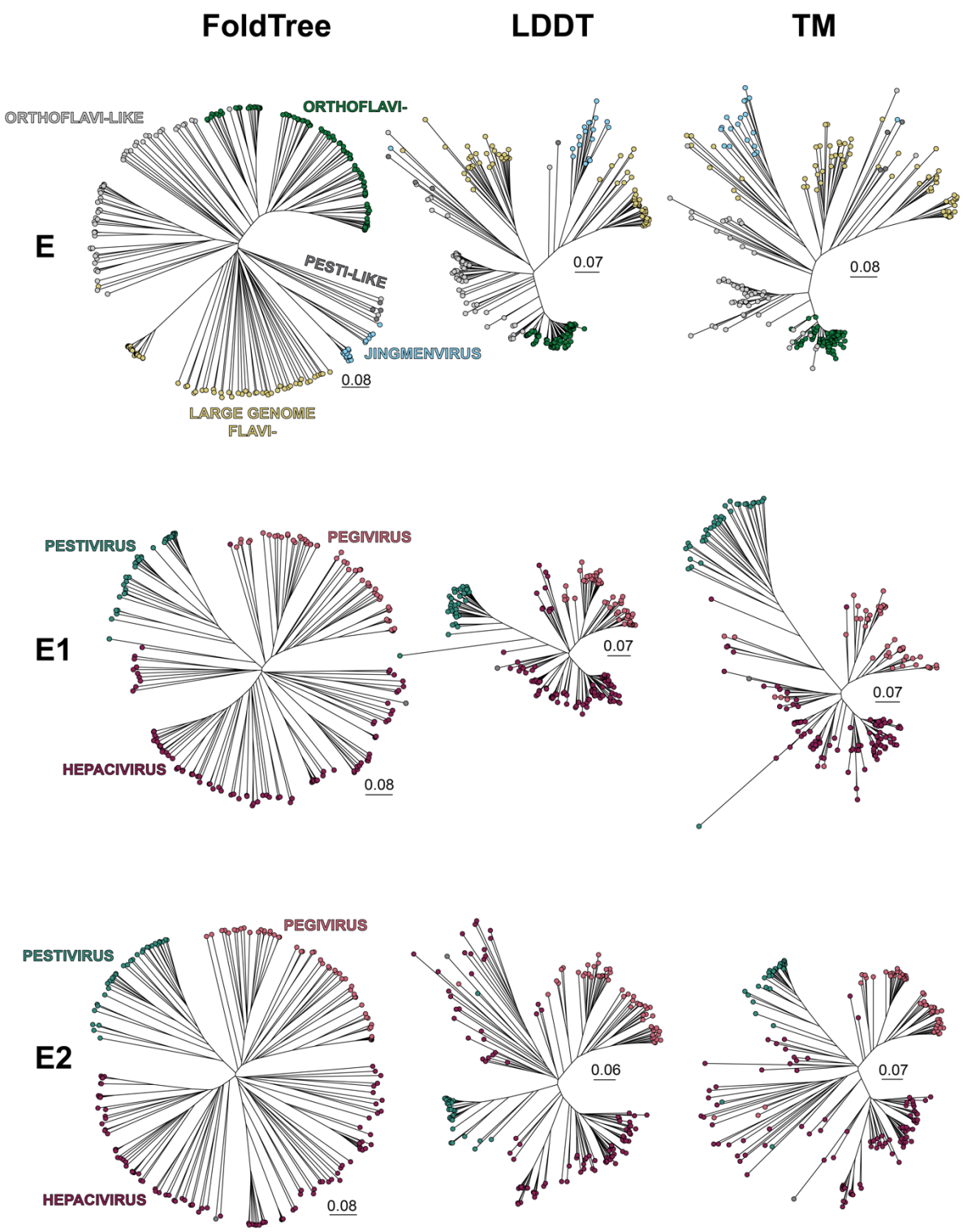

***Supplementary Figure 6.9. Structurally aligned E1 glycoprotein phylogenies.*** *Protein sequences were aligned using their 3Di structural representation (see Methods for details).*

*Phylogenies were reconstructed using 3Di sequence alone (top), amino acid (AA) sequence alone (middle) or combined 3Di and AA sequence. Right hand trees are derived from alignments trimmed with a gap threshold of 35%. Scale bars indicate substitutions per site for either the 3Di, AA or combined sequences, respectively. Tip shapes are colour-coded by genus/subclade as in Figure 6.1a. All phylogenetic trees are provided in the associated Zenodo repository.*



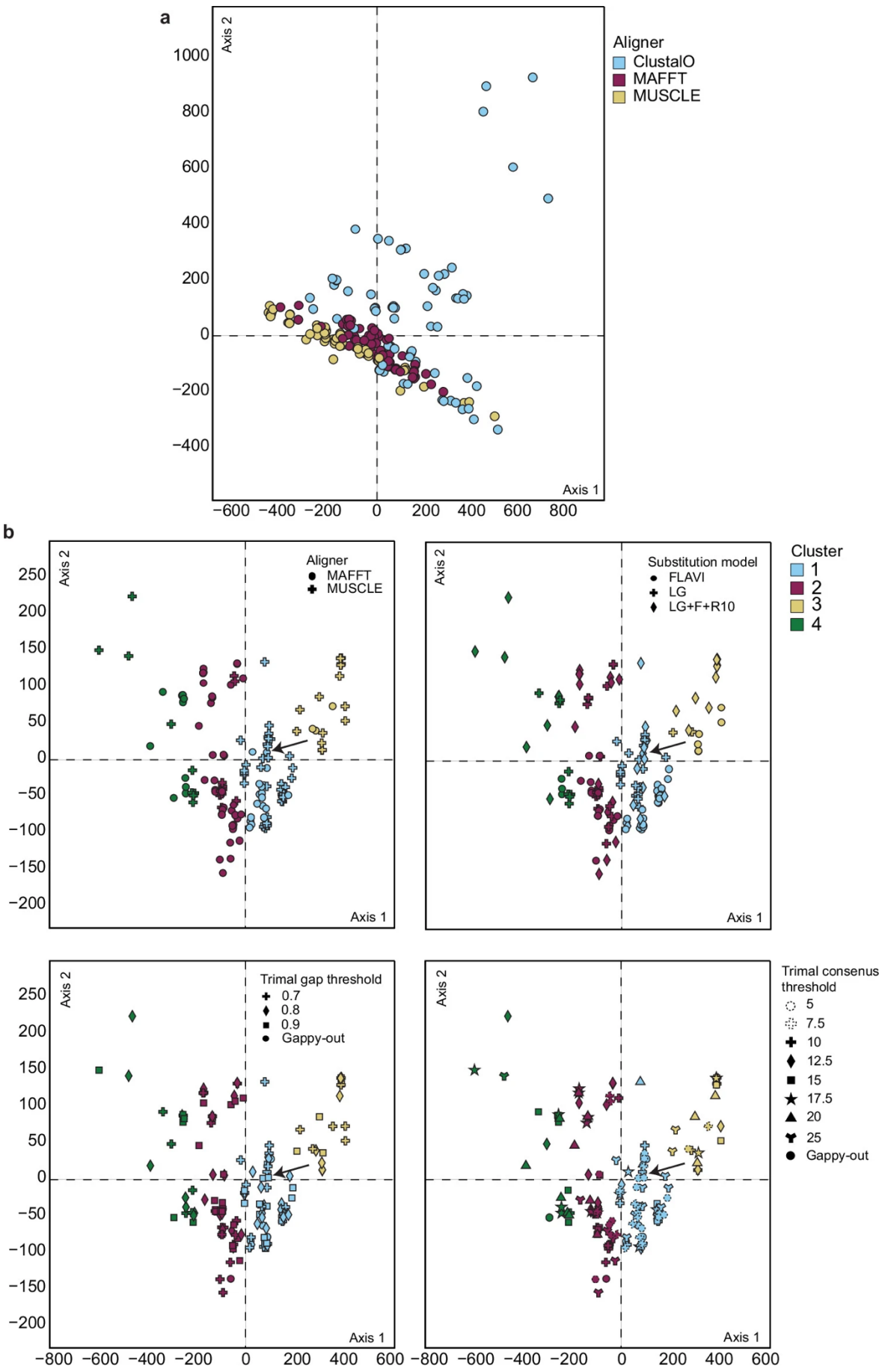

***Supplementary Figure 6.10. Structurally aligned E2 glycoprotein phylogenies.*** *Protein sequences were aligned using their 3Di structural representation (see Methods for details).*

*Phylogenies were reconstructed using 3Di sequence alone (top), amino acid (AA) sequence alone (middle) or combined 3Di and AA sequence. Right hand trees are derived from alignments trimmed with a gap threshold of 35%. Scale bars indicate substitutions per site for either the 3Di, AA or combined sequences, respectively. Tip shapes are colour-coded by genus/subclade as in Figure 6.1a. All phylogenetic trees are provided in the associated Zenodo repository.*

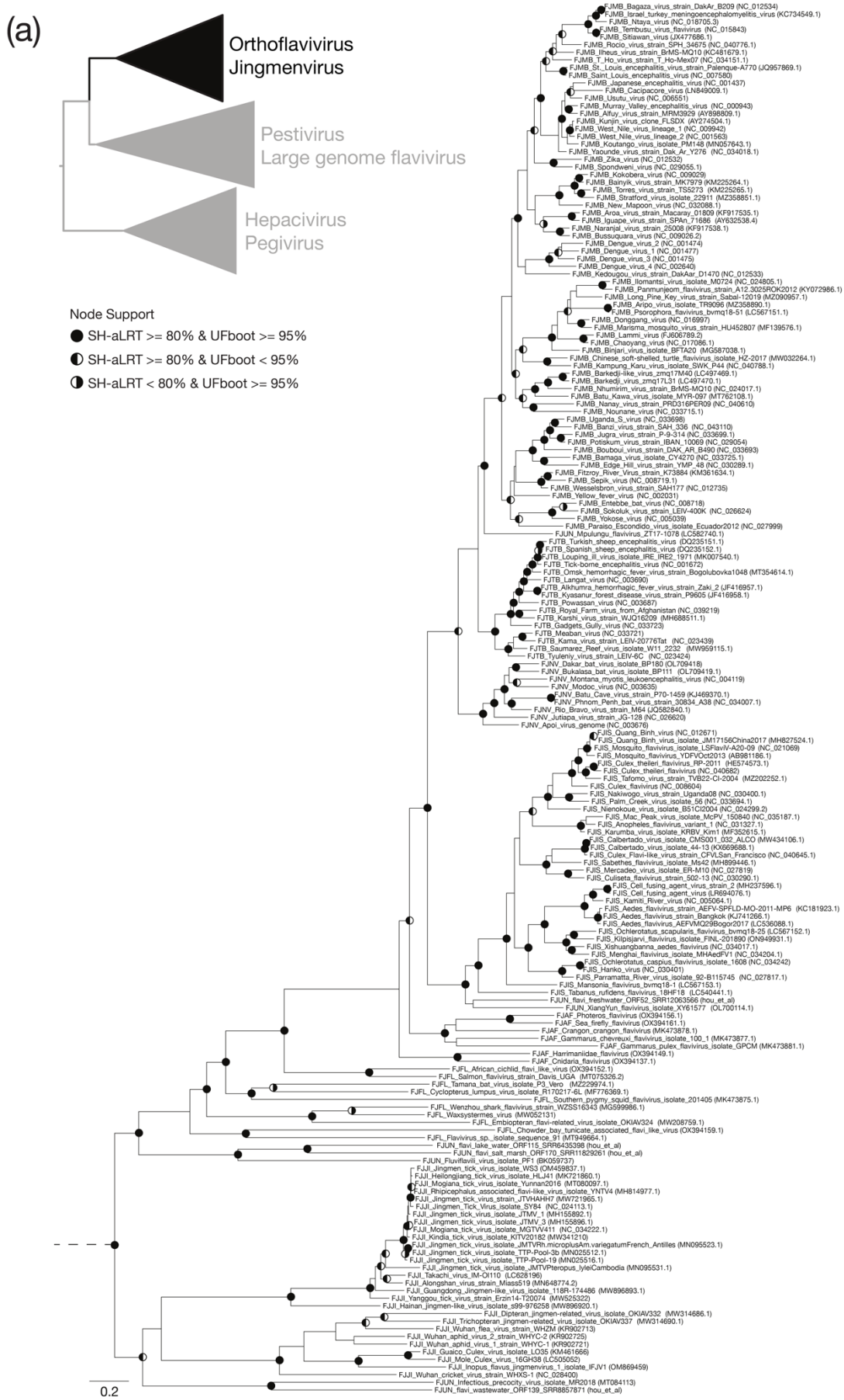



***Supplementary Figure 6.11. Structurally aligned glycoprotein phylogenies using FoldTree.*** *For each glycoprotein (E, E1 and E2) phylogenies generated using three different*

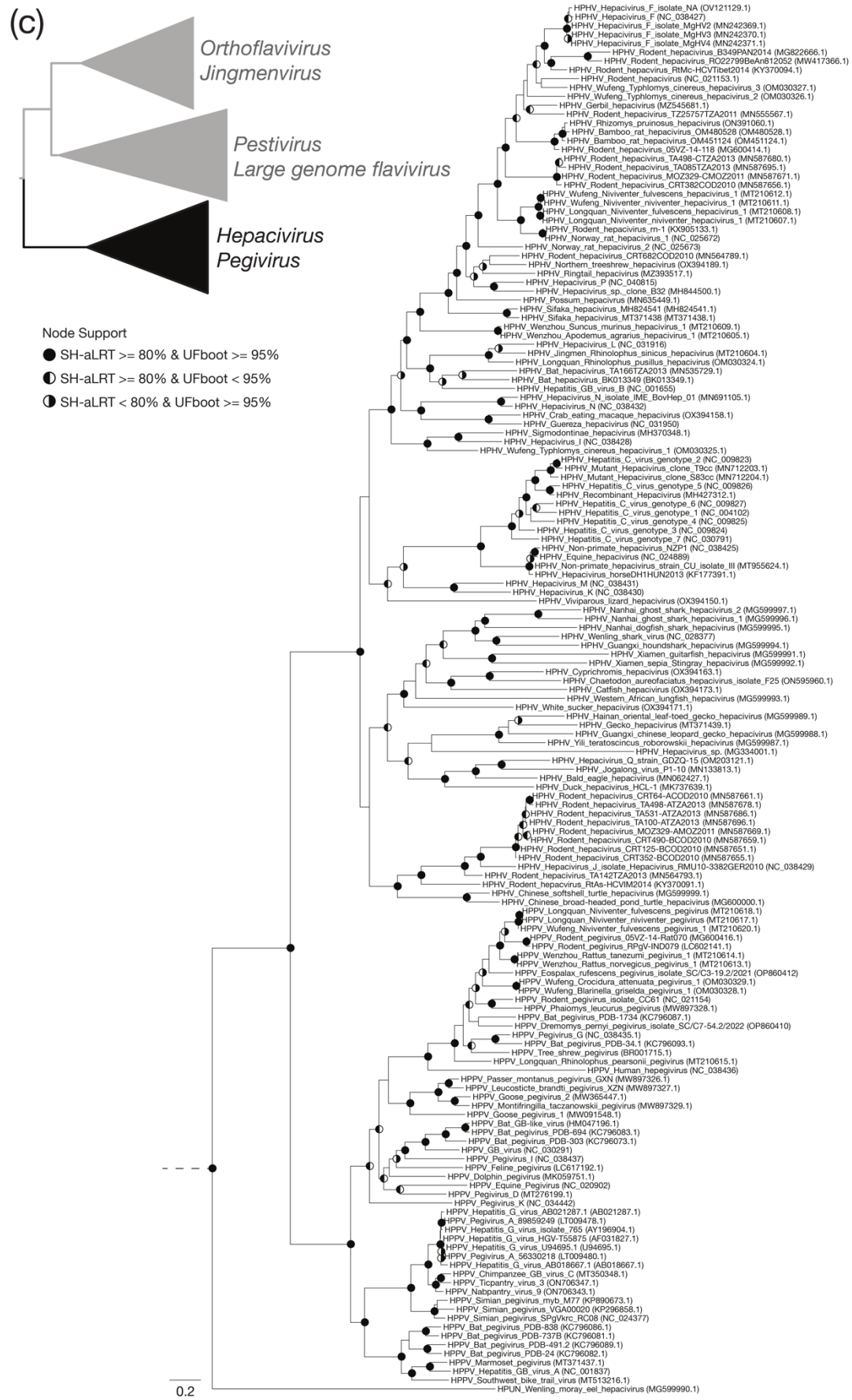
*measures of structural distance, Foldtree metric, LDDT and TM score are shown. For visualisation the unrooted phylogenies with statistical correction are shown. Tip circles are coloured by Flaviviridae clade. PP unrooted trees chosen for visualisation. All resulting phylogenetic trees are provided in the associated Zenodo repository.*


a
Axis 2
Axis 1
Aligner
ClustalO
MAFFT
MUSCLE
b
Aligner
MAFFT
MUSCLE
Substitution model
FLAVI
LG
LG+F+R10
Cluster
1
2
3
4
Trimal gap threshold
0.7
0.8
0.9
Gappy-out
Trimal consenus threshold
5
7.5
10
12.5
15
17.5
20
25
Gappy-out

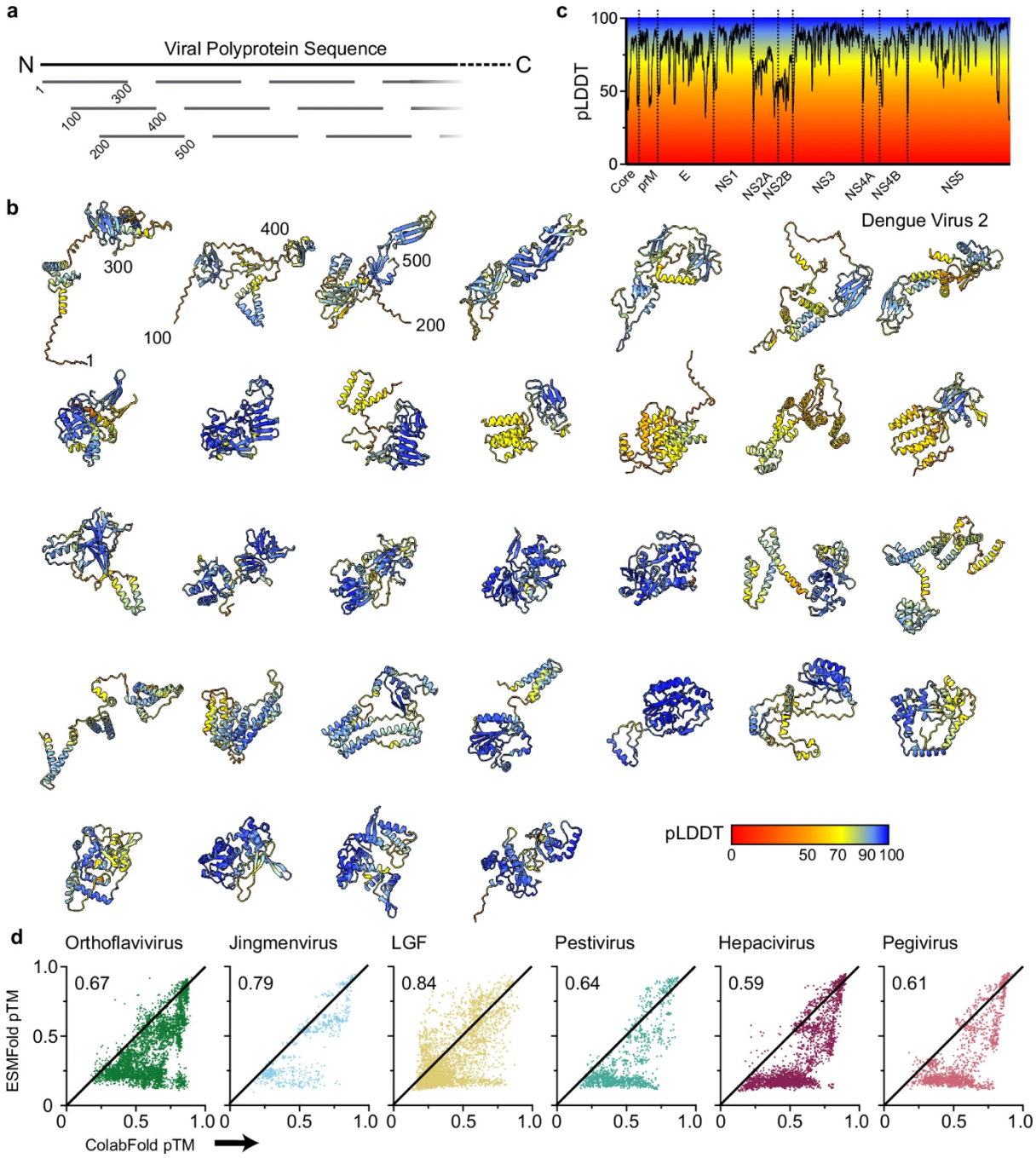
***Supplementary Figure 6.12. Two-dimensional MDS plot of the NS5b phylogeny variations.*** *(a), Coloured by alignment method (b), Excluding trees generated from Clustal Omega alignments. Points, which represent individual phylogenies, are colour-coded based on the clusters identified using the "findGroves" function. Point shapes indicate the alignment software, the trimAl gap and consensus thresholds, and the substitution model applied. An arrow signifies the master phylogeny (Tree 18) chosen for further analysis.*


(a)
Orthoflavivirus
Jingmenvirus
Pestivirus
Large genome flavivirus
Hepacivirus
Pegivirus
Node Support
SH-aLRT >= 80% & UFboot >= 95%
SH-aLRT >= 80% & UFboot < 95%
SH-aLRT < 80% & UFboot >= 95%
0.2

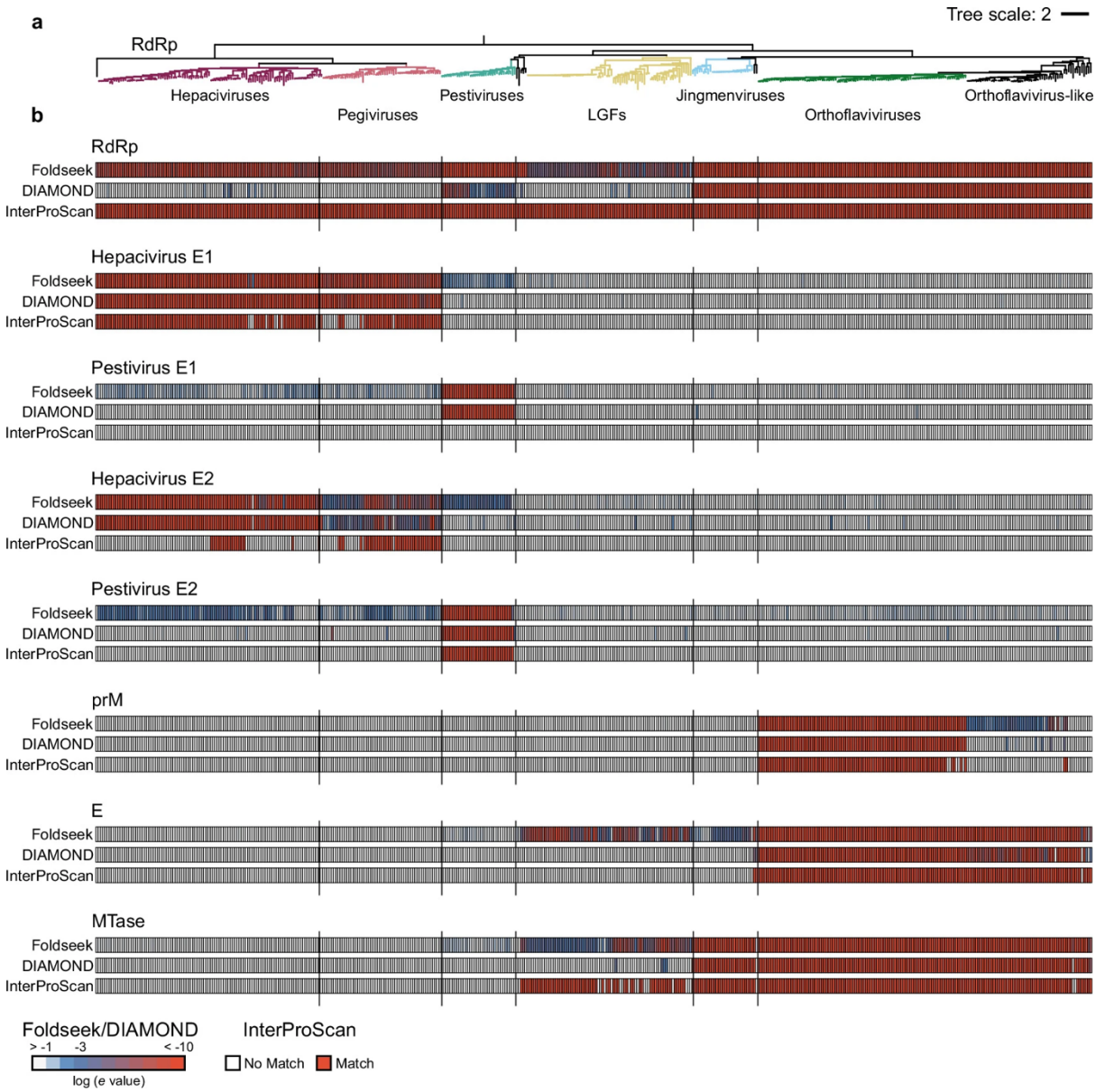
(b)
Orthoflavivirus
Jingmenvirus
Pestivirus
Large genome flavivirus
Hepacivirus
Pegivirus
Node Support
SH-aLRT >= 80% & UFboot >= 95%
SH-aLRT >= 80% & UFboot < 95%
SH-aLRT < 80% & UFboot >= 95%
PLLG_Flaviviridae_sp._isolate_TIGMIC_24 (ON811725.1)
PLLG_Bole_tick_virus_4_strain_NE-DH2 (ON408074.1)
PLLG_Bole_tick_virus_4_strain_NE-ShL2 (ON408075.1)
PLLG_Flaviviridae_sp._isolate_TIGMIC_28 (ON811729.1)
PLLG_Flaviviridae_sp._isolate_TIGMIC_12 (ON811713.1)
PLLG_Flaviviridae_sp._isolate_TIGMIC_13 (ON811714.1)
PLLG_Flaviviridae_sp._isolate_TIGMIC_30 (ON811731.1)
PLLG_Flaviviridae_sp._isolate_TIGMIC_36 (ON811737.1)
PLLG_Flaviviridae_sp._isolate_TIGMIC_27 (ON811728.1)
PLLG_Flaviviridae_sp._isolate_TIGMIC_25 (ON811726.1)
PLLG_Flaviviridae_sp._isolate_TIGMIC_20 (ON811721.1)
PLLG_Flaviviridae_sp._isolate_TIGMIC_5 (ON811706.1)
PLLG_Flaviviridae_sp._isolate_TIGMIC_8 (ON811709.1)
PLLG_Flaviviridae_sp._isolate_TIGMIC_1 (ON811700.1)
PLLG_Flaviviridae_sp._isolate_TIGMIC_33 (ON811734.1)
PLLG_Flaviviridae_sp._isolate_TIGMIC_32 (ON811733.1)
PLLG_Zhangjiakou_Flavi_tick_virus_1 (N746389.1)
PLLG_Flaviviridae_sp._isolate_TIGMIC_38 (ON811741.1)
PLLG_Flaviviridae_sp._isolate_TIGMIC_39 (ON811742.1)
PLLG_Flaviviridae_sp._strain_FVHAHH8 (MZ964986.1)
PLLG_Dermacentor_reticulatus_pestivirus-like_virus_1_isolate_CT2 (ON684360.1)
PLLG_Haseki_tick_virus_strain_Novosibirsk-2020 (MW808978.1)
PLLG_Rhipicephalus_sanguineus_associated_flavi-like_virus (BK067815)
PLLG_Trinbago_virus_isolate_TTP-Pool-4 (MN025505.1)
PLLG_Bole_tick_virus_4_isolate_TIGMIC_1 (ON811701.1)
PLLG_Bole_tick_virus_4_isolate_Thailand_tick_flavivirus (MN095535.1)
PLLG_Bole_tick_virus_4_strain_BangaliH.truncatum2018 (MW561974.1)
PLLG_Bole_tick_virus_4_strain_IftinH.dromedarii2018 (MW561976.1)
PLLG_Bole_tick_virus_4_isolate_TIGMIC_2 (ON811702.1)
PLLG_Camel_tick_virus_isolate_Alain (ON262323.1)
PLLG_Bole_tick_virus_4_strain_BLP-1 (KR902736)
PLLG_Bole_tick_virus_4_isolate_GSC346flaviV (MZ244286.1)
PLLG_Bemisia_tabaci_pesti-like_virus_1_isolate_ZJU-Q2 (MW256672.1)
PLLG_Bemisia_tabaci_pesti-like_virus_2 (BK067808)
PLLG_Artemia_sinica_associated_flavi-like_virus (BK067806)
PLLG_Hangzhou_ochthera_mantis_flavivirus_1_isolate_SYFY8 (MZ209722.1)
PLLG_Semibalanus_balanoides_associated_flavi-like_virus (BK067816)
PLLG_Beihai_barnacle_viurs_1_strain_BHGZ (NC_028376)
PLLG_flavi_estuary_ORF249_SRR7087980 (hou_et_al)
PLLG_Xiangshan_flavi-like_virus_1 (OK491477.1)
PLLG_Xiangshan_flavi-like_virus_2 (OK491478.1)
PLLG_Brachymeria_lasus_associated_flavi-like_virus (BK067807)
PLLG_Fushun_laodelphax_striatellus_flavivirus_1_isolate_HFSFS147 (MZ210009.1)
PLLG_Shayang_fly_virus_4_strain_SYCY (KR902732)
PLLG_Locusta_migratoria_associated_flavi-like_virus (BK067811)
PLLG_Shuangao_lacewing_virus_2_strain_SCL (KR902734)
PLLG_Plodia_interpunctella_associated_flavi-like_virus (BK067812)
PLLG_Macrosiphum_euphorbiae_virus_1_isolate_K01 (NC_028137.1)
PLLG_Sitobion_miscanthi_flavi-like_virus_1_07BaA1 (LC516843.1)
PLLG_Gamboa_mosquito_virus_strain_GW (KR902733)
PLLG_Sanxia_water_strider_virus_6_strain_SYSZZ-2 (KR902740)
PLLG_Fushun_acrida_cinerea_flavivirus_1_isolate_ZHJJHFS91 (MZ210052.1)
PLLG_Hangzhou_nephotettix_cincticeps_flavivirus_1_isolate_HWYCFY169513 (MZ209632.1)
PLLG_Hangzhou_flavivirus_1_isolate_FSFY137086 (MZ209617.1)
PLLG_Eriocheir_sinensis_cholera_like_virus (OP019096.1)
PLLG_Gecarcinus_lateralis_associated_flavi-like_virus (BK067809)
PLLG_Gentian_Kobu-sho-associated_virus (AB698917.1)
PLLG_Carrot_flavi-like_virus_1_isolate_744 (OM681407.1)
PLLG_Sonchus_asper_flavi-like_virus (BK062903)
PLLG_Coptis_teeta_flavi-like_virus (BK062902)
PLLG_Phenacoccus_solenopsis_associated_flavi-like_virus (BK067813)
PLLG_Drosophila-associated_flavivirus_7_isolate_Con0316 (MZ852365.1)
PLLG_Lampyris_noctiluca_flavivirus_1_isolate_17FIN1 (MH620810.1)
PLLG_Insect_associated_flavi-like_virus (BK067810)
PLLG_Apis_flavivirus_isolate_RI-A (NC_035071.1)
PLLG_Hangzhou_frankliniella_intonsa_flavivirus_2_isolate_JM1FY91827 (MZ209659.1)
PLLG_Frankliniella_occidentalis_flavi-like_virus_1 (MN714664.1)
PLLG_Hangzhou_frankliniella_intonsa_flavivirus_1_isolate_JM1FY93485 (MZ209656.1)
PLLG_Diaphorina_citri_flavi-like_virus (NC_030453.1)
PLLG_Atrato_Flavi-like_virus_1_strain_Cx_1773-1 (MN661082.1)
PLLG_Soybean_cyst_nematode_virus_5_strain_ScV5 (NC_024077.1)
PLLG_Xinzhou_spider_virus_2_strain_XZZZ-7 (KR902735)
PLLG_Tacheng_tick_virus_8_strain_TCP-2 (KR902741)
PLLG_Wuhan_centipede_virus_strain_WHWG-11 (KR902737)
PLLG_flavi_soil_ORF358_SRR7976360 (hou_et_al)
PLPV_Classical_swine_fever_virus (NC_002657)
PLPV_Pestivirus_sp._isolate_ovinelt338710-32017 (MK618726.1)
PLPV_Tunisian_small_ruminant_pestivirus_strain_920192007AG (MZ664274.1)
PLPV_Pestivirus_Burdur05-TR (KM408491.1)
PLPV_Pestivirus_strain_Aydin04-TR (NC_018713.1)
PLPV_Border_disease_virus_X818 (NC_003679)
PLPV_Pestivirus_reindeer-1_V60-Krefeld (AF144618.2)
PLPV_Bovine_viral_diarrhea_virus_1 (NC_001461)
PLPV_Bovine_viral_diarrhea_virus_3_Th04_KhonKaen (NC_012812)
PLPV_Bovine_viral_diarrhea_virus_2 (NC_039237)
PLPV_Pestivirus_giraffe-1_H138 (NC_003678)
PLPV_Pestivirus_PG-2 (KJ660072.1)
PLPV_Pronghorn_antelope_pestivirus (NC_024018)
PLPV_Pangolin_pestivirus_strain_PPeVBIME1P019Tpangolin2018 (ON024090.1)
PLPV_Dongyang_pangolin_virus_isolate_DYCS (MK636875.1)
PLPV_Pangolin_pestivirus_strain_PPeVBIME3P019Tpangolin2018 (ON024097.1)
PLPV_Pangolin_pestivirus_strain_PPeVBIME5P141Tpangolin2018 (ON024101.1)
PLPV_Pangolin_pestivirus_strain_PPeVBIME7P022Tpangolin2018 (ON024104.1)
PLPV_Pangolin_pestivirus_strain_PPeVBIME6P226Tpangolin2018 (ON024103.1)
PLPV_Pangolin_pestivirus_strain_PPeVBIME6P141Tpangolin2018 (ON024102.1)
PLPV_Pangolin_pestivirus_1_isolate_ZJ-F1 (OM451127.1)
PLPV_Wenzhou_Pipistrellus_abramus_pestivirus_1_isolate_YJB_Pabr (OM030320.1)
PLPV_Linda_virus_strain_Austria1 (NC_035432)
PLPV_Phocoena_pestivirus_isolate_NS170386 (MK910229.1)
PLPV_Porcine_pestivirus_isolate_Bungowannah (NC_023176)
PLPV_Jingmen_Crocidura_shantungensis_pestivirus_1 (OM030319.1)
PLPV_Norway_rat_pestivirus_isolate_NrPVNYC-D23 (NC_025677.1)
PLPV_Rodent_pestivirus_isolate_RtNn-PestVHuB2014 (KY370101.1)
PLPV_Rodent_pestivirus_isolate_RtAp-PestVJL2014 (KY370100.1)
PLPV_Zikole_virus (OU592965.1)
PLPV_Bat_pestivirus_BtSk-PestV-1GX2017 (MH282908.1)
PLPV_Atypical_porcine_pestivirus (MH307700.1)
PLPV_Frog_pestivirus (OX394182.1)
PLPV_Glass_knifefish_pestivirus (OX394178.1)
PLPV_Wenzhou_pesti-like_virus_1_strain_WZSS31331 (MG599982.1)
PLPV_Xiamen_fanray_pesti-like_virus_strain_XMTSYG10372 (MG599985.1)
PLUN_Xinzhou_spider_virus_3_strain_XZZZ-3 (KR902730)
PLUN_Parasteatoda_tepidariorum_associated_flavi-like_virus (BK067814)
PLUN_Shayang_spider_virus_4_strain_SYZZ-1 (KR902739)
PLUN_Diatom_colony_associated_ssRNA_virus_1 (NC_040589.1)
PLUN_flavi_freshwater_sediment_ORF251_SRR7080628 (hou_et_al)
PLUN_Inner_Mongolia_sediment_flavi-like_virus_3 (MW896903.1)
PLUN_flavi_lake_water_ORF238_SRR7883614 (hou_et_al)
PLUN_flavi_soil_ORF389_SRR11565438 (hou_et_al)
0.3


(c)
Orthoflavivirus
Jingmenvirus
Pestivirus
Large genome flavivirus
Hepacivirus
Pegivirus
Node Support
SH-aLRT >= 80% & UFboot >= 95%
SH-aLRT >= 80% & UFboot < 95%
SH-aLRT < 80% & UFboot >= 95%
HPHV_Hepacivirus_F_isolate_NA (OV121129.1)
HPHV_Hepacivirus_F (NC_038427)
HPHV_Hepacivirus_F_isolate_MgHV2 (MN242369.1)
HPHV_Hepacivirus_F_isolate_MgHV3 (MN242370.1)
HPHV_Hepacivirus_F_isolate_MgHV4 (MN242371.1)
HPHV_Rodent_hepacivirus_B349PAN2014 (MG822666.1)
HPHV_Rodent_hepacivirus_RO22799BeAn812052 (MW417366.1)
HPHV_Rodent_hepacvirus_RtMc-HCVTibet2014 (KY370094.1)
HPHV_Rodent_hepacivirus (NC_021153.1)
HPHV_Wufeng_Typhlomys_cinereus_hepacivirus_3 (OM030327.1)
HPHV_Wufeng_Typhlomys_cinereus_hepacivirus_2 (OM030326.1)
HPHV_Gerbil_hepacivirus (MZ545681.1)
HPHV_Rodent_hepacivirus_TZ25757TZA2011 (MN555567.1)
HPHV_Rhizomys_pruinosus_hepacivirus (ON391060.1)
HPHV_Bamboo_rat_hepacivirus_OM480528 (OM480528.1)
HPHV_Bamboo_rat_hepacivirus_OM451124 (OM451124.1)
HPHV_Rodent_hepacivirus_05VZ-14-118 (MG600414.1)
HPHV_Rodent_hepacivirus_TA498-CTZA2013 (MN587680.1)
HPHV_Rodent_hepacivirus_TA085TZA2013 (MN587695.1)
HPHV_Rodent_hepacivirus_MOZ329-CMOZ2011 (MN587671.1)
HPHV_Rodent_hepacivirus_CRT382COD2010 (MN587656.1)
HPHV_Wufeng_Niviventer_fulvescens_hepacivirus_1 (MT210612.1)
HPHV_Wufeng_Niviventer_niviventer_hepacivirus_1 (MT210611.1)
HPHV_Longquan_Niviventer_fulvescens_hepacivirus_1 (MT210608.1)
HPHV_Longquan_Niviventer_niviventer_hepacivirus_1 (MT210607.1)
HPHV_Rodent_hepacivirus_rn-1 (KX905133.1)
HPHV_Norway_rat_hepacivirus_1 (NC_025672)
HPHV_Norway_rat_hepacivirus_2 (NC_025673)
HPHV_Rodent_hepacivirus_CRT682COD2010 (MN564789.1)
HPHV_Northern_treeshrew_hepacivirus (OX394189.1)
HPHV_Ringtail_hepacivirus (MZ393517.1)
HPHV_Hepacivirus_P (NC_040815)
HPHV_Hepacivirus_sp._clone_B32 (MH844500.1)
HPHV_Possum_hepacivirus (MN635449.1)
HPHV_Sifaka_hepacivirus_MH824541 (MH824541.1)
HPHV_Sifaka_hepacivirus_MT371438 (MT371438.1)
HPHV_Wenzhou_Suncus_murinus_hepacivirus_1 (MT210609.1)
HPHV_Wenzhou_Apodemus_agrarius_hepacivirus_1 (MT210605.1)
HPHV_Hepacivirus_L (NC_031916)
HPHV_Jingmen_Rhinolophus_sinicus_hepacivirus (MT210604.1)
HPHV_Longquan_Rhinolophus_pusillus_hepacivirus (OM030324.1)
HPHV_Bat_hepacivirus_TA166TZA2013 (MN535729.1)
HPHV_Bat_hepacivirus_BK013349 (BK013349.1)
HPHV_Hepatitis_GB_virus_B (NC_001655)
HPHV_Hepacivirus_N_isolate_IME_BovHep_01 (MN691105.1)
HPHV_Hepacivirus_N (NC_038432)
HPHV_Crab_eating_macaque_hepacivirus (OX394158.1)
HPHV_Guereza_hepacivirus (NC_031950)
HPHV_Sigmodontinae_hepacivirus (MH370348.1)
HPHV_Hepacivirus_I (NC_038428)
HPHV_Wufeng_Typhlomys_cinereus_hepacivirus_1 (OM030325.1)
HPHV_Hepatitis_C_virus_genotype_2 (NC_009823)
HPHV_Mutant_Hepacivirus_clone_T9cc (MN712203.1)
HPHV_Mutant_Hepacivirus_clone_S83cc (MN712204.1)
HPHV_Hepatitis_C_virus_genotype_5 (NC_009826)
HPHV_Recombinant_Hepacivirus (MH427312.1)
HPHV_Hepatitis_C_virus_genotype_6 (NC_009827)
HPHV_Hepatitis_C_virus_genotype_1 (NC_004102)
HPHV_Hepatitis_C_virus_genotype_4 (NC_009825)
HPHV_Hepatitis_C_virus_genotype_3 (NC_009824)
HPHV_Hepatitis_C_virus_genotype_7 (NC_030791)
HPHV_Non-primate_hepacivirus_NZP1 (NC_038425)
HPHV_Equine_hepacivirus (NC_024889)
HPHV_Non-primate_hepacivirus_strain_CU_isolate_III (MT955624.1)
HPHV_Hepacivirus_horseDH1HUN2013 (KF177391.1)
HPHV_Hepacivirus_M (NC_038431)
HPHV_Hepacivirus_K (NC_038430)
HPHV_Viviparous_lizard_hepacivirus (OX394150.1)
HPHV_Nanhai_ghost_shark_hepacivirus_2 (MG599997.1)
HPHV_Nanhai_ghost_shark_hepacivirus_1 (MG599996.1)
HPHV_Nanhai_dogfish_shark_hepacivirus (MG599995.1)
HPHV_Wenling_shark_virus (NC_028377)
HPHV_Guangxi_houndshark_hepacivirus (MG599994.1)
HPHV_Xiamen_guitarfish_hepacivirus (MG599991.1)
HPHV_Xiamen_sepia_Stingray_hepacivirus (MG599992.1)
HPHV_Cyprichromis_hepacivirus (OX394163.1)
HPHV_Chaetodon_aureofaciatus_hepacivirus_isolate_F25 (ON595960.1)
HPHV_Catfish_hepacivirus (OX394173.1)
HPHV_Western_African_lungfish_hepacivirus (MG599993.1)
HPHV_White_sucker_hepacivirus (OX394171.1)
HPHV_Hainan_oriental_leaf-toed_gecko_hepacivirus (MG599989.1)
HPHV_Gecko_hepacivirus (MT371439.1)
HPHV_Guangxi_chinese_leopard_gecko_hepacivirus (MG599988.1)
HPHV_Yili_teratoscincus_roborowskii_hepacivirus (MG599987.1)
HPHV_Hepacivirus_sp. (MG334001.1)
HPHV_Hepacivirus_Q_strain_GDZQ-15 (OM203121.1)
HPHV_Jogalong_virus_P1-10 (MN133813.1)
HPHV_Bald_eagle_hepacivirus (MN062427.1)
HPHV_Duck_hepacivirus_HCL-1 (MK737639.1)
HPHV_Rodent_hepacivirus_CRT64-ACOD2010 (MN587661.1)
HPHV_Rodent_hepacivirus_TA498-ATZA2013 (MN587678.1)
HPHV_Rodent_hepacivirus_TA531-ATZA2013 (MN587686.1)
HPHV_Rodent_hepacivirus_TA100-ATZA2013 (MN587696.1)
HPHV_Rodent_hepacivirus_MOZ329-AMOZ2011 (MN587669.1)
HPHV_Rodent_hepacivirus_CRT490-BCOD2010 (MN587659.1)
HPHV_Rodent_hepacivirus_CRT125-BCOD2010 (MN587651.1)
HPHV_Rodent_hepacivirus_CRT352-BCOD2010 (MN587655.1)
HPHV_Hepacivirus_J_isolate_Hepacivirus_RMU10-3382GER2010 (NC_038429)
HPHV_Rodent_hepacivirus_TA142TZA2013 (MN564793.1)
HPHV_Rodent_hepacvirus_RtAs-HCVIM2014 (KY370091.1)
HPHV_Chinese_softshell_turtle_hepacivirus (MG599999.1)
HPHV_Chinese_broad-headed_pond_turtle_hepacivirus (MG600000.1)
HPPV_Longquan_Niviventer_fulvescens_pegivirus (MT210618.1)
HPPV_Longquan_Niviventer_niviventer_pegivirus (MT210617.1)
HPPV_Wufeng_Niviventer_fulvescens_pegivirus_1 (MT210620.1)
HPPV_Rodent_pegivirus_05VZ-14-Rat070 (MG600416.1)
HPPV_Rodent_pegivirus_RPgV-IND079 (LC602141.1)
HPPV_Wenzhou_Rattus_tanezumi_pegivirus_1 (MT210614.1)
HPPV_Wenzhou_Rattus_norvegicus_pegivirus_1 (MT210613.1)
HPPV_Eospalax_rufescens_pegivirus_isolate_SC/C3-19.2/2021 (OP860412)
HPPV_Wufeng_Crocidura_attenuata_pegivirus_1 (OM030329.1)
HPPV_Wufeng_Blarinella_griselda_pegivirus_1 (OM030328.1)
HPPV_Rodent_pegivirus_isolate_CC61 (NC_021154)
HPPV_Phaiomys_leucurus_pegivirus (MW897328.1)
HPPV_Bat_pegivirus_PDB-1734 (KC796087.1)
HPPV_Dremomys_pernyi_pegivirus_isolate_SC/C7-54.2/2022 (OP860410)
HPPV_Pegivirus_G (NC_038435.1)
HPPV_Bat_pegivirus_PDB-34.1 (KC796093.1)
HPPV_Tree_shrew_pegivirus (BR001715.1)
HPPV_Longquan_Rhinolophus_pearsonii_pegivirus (MT210615.1)
HPPV_Human_hepegivirus (NC_038436)
HPPV_Passer_montanus_pegivirus_GXN (MW897326.1)
HPPV_Leucosticte_brandti_pegivirus_XZN (MW897327.1)
HPPV_Goose_pegivirus_2 (MW365447.1)
HPPV_Montifringilla_taczanowskii_pegivirus (MW897329.1)
HPPV_Goose_pegivirus_1 (MW091548.1)
HPPV_Bat_GB-like_virus (HM047196.1)
HPPV_Bat_pegivirus_PDB-694 (KC796083.1)
HPPV_Bat_pegivirus_PDB-303 (KC796073.1)
HPPV_GB_virus (NC_030291)
HPPV_Pegivirus_I (NC_038437)
HPPV_Feline_pegivirus (LC617192.1)
HPPV_Dolphin_pegivirus (MK059751.1)
HPPV_Equine_Pegivirus (NC_020902)
HPPV_Pegivirus_D (MT276199.1)
HPPV_Pegivirus_K (NC_034442)
HPPV_Hepatitis_G_virus_AB021287.1 (AB021287.1)
HPPV_Pegivirus_A_89859249 (LT009478.1)
HPPV_Hepatitis_G_virus_isolate_765 (AY196904.1)
HPPV_Hepatitis_G_virus_HGV-T55875 (AF031827.1)
HPPV_Hepatitis_G_virus_U94695.1 (U94695.1)
HPPV_Pegivirus_A_56330218 (LT009480.1)
HPPV_Hepatitis_G_virus_AB018667.1 (AB018667.1)
HPPV_Chimpanzee_GB_virus_C (MT350348.1)
HPPV_Ticpantry_virus_3 (ON706347.1)
HPPV_Nabpantry_virus_9 (ON706343.1)
HPPV_Simian_pegivirus_myb_M77 (KP890673.1)
HPPV_Simian_pegivirus_VGA00020 (KP296858.1)
HPPV_Simian_pegivirus_SPgVkrc_RC08 (NC_024377)
HPPV_Bat_pegivirus_PDB-838 (KC796086.1)
HPPV_Bat_pegivirus_PDB-737B (KC796081.1)
HPPV_Bat_pegivirus_PDB-491.2 (KC796089.1)
HPPV_Bat_pegivirus_PDB-24 (KC796082.1)
HPPV_Marmoset_pegivirus (MT371437.1)
HPPV_Hepatitis_GB_virus_A (NC_001837)
HPPV_Southwest_bike_trail_virus (MT513216.1)
HPUN_Wenling_moray_eel_hepacivirus (MG599990.1)
0.2

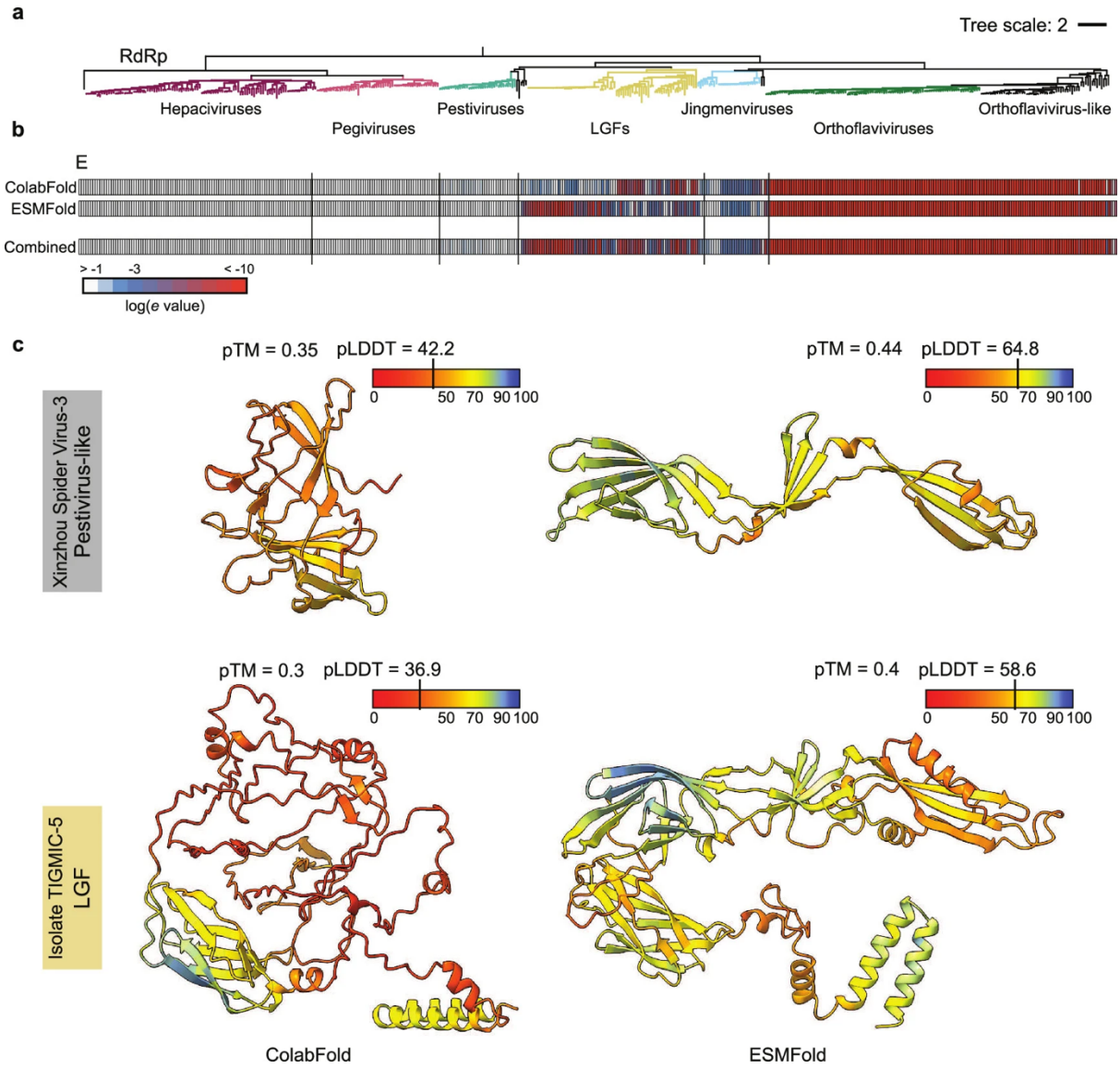
***Supplementary Figure 6.13.*** *NS5b consensus phylogeny (Tree 18) with tip labels. The complete NS5b phylogeny (underlying Figure 6.1a and 2a) split by the major Flaviviridae lineages (a), Orthoflavivirus/Jingmenvirus, (b), Large genome flavivirus/Pestivirus and (c), Hepacivirus/Pegivirus. A scale bar denotes the number of amino acid substitutions per site. Node support (SH-aLRT >=80% and UFboot >=95%) is indicated by a black circle while intermediate support (SH-aLRT <80% and UFboot >=95% or SH-aLRT >80% and UFboot <95%) is indicated by a half filled circle corresponding to the indice above the support threshold.*



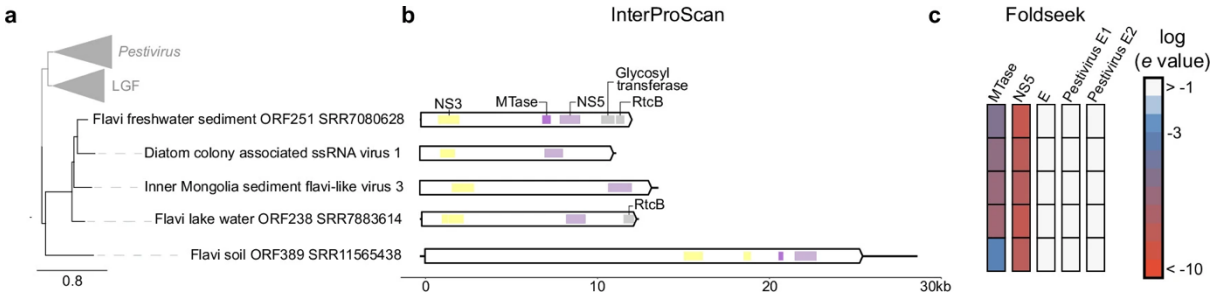
***Supplementary Figure 6.14.*** *Structure prediction performance. (a), For structural inference, all Flaviviridae polyprotein sequences were split into blocks of 300 residues, each overlapping by 100 residues (461 species, 16,463 blocks in total). Residue numbers are provided for the first three blocks. (b), Representative ColabFold protein structure predictions spanning the entire Dengue Virus 2 (DENV-2) polyprotein. Residue numbers are provided as in a. Structures are colour-coded by prediction confidence scores (pLDDT), as denoted in the key. (c), pLDDT confidence scores along the length of the DENV-2 polyprotein. Dotted lines delineate the mature proteins, which are labelled on the x-axis. d, Scatter plots representing predicted TM-score (pTM) confidence metric for ColabFold and ESMFold for each sequence block in each genus/subclade. Numerical values provide the performance ratio between either protein structure prediction method; values below 1 indicates better performance by ColabFold.*



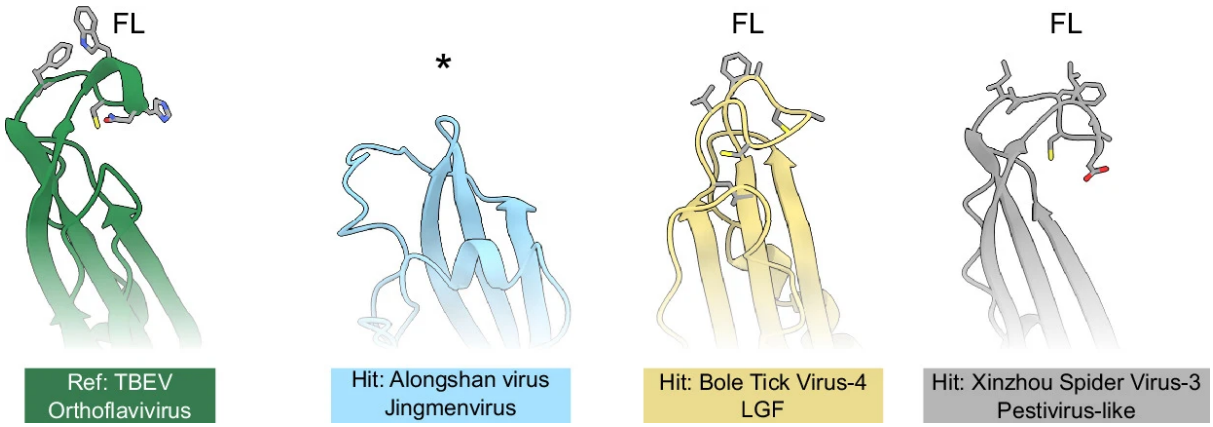

***Supplementary Figure 6.15. Benchmarking with sequence-based homology searches.*** *(a), RdRp phylogeny, as in Figure 6.2a. (b), Heatmap comparison of homology detection using Foldseek (as in Figure 6.2), DIAMOND and InterProScan for the stated reference proteins. DIAMOND results represent a pure sequence-based recapitulation of the Foldseek search (i.e., all query and reference structures used in Foldseek were represented by their cognate protein sequences for DIAMOND analysis). Foldseek and DIAMOND data are log e-values and are colour-coded as in the key. InterProScan results provide a simple binary score: match (red) or no match (white). Vertical lines demark divisions between major clades.*



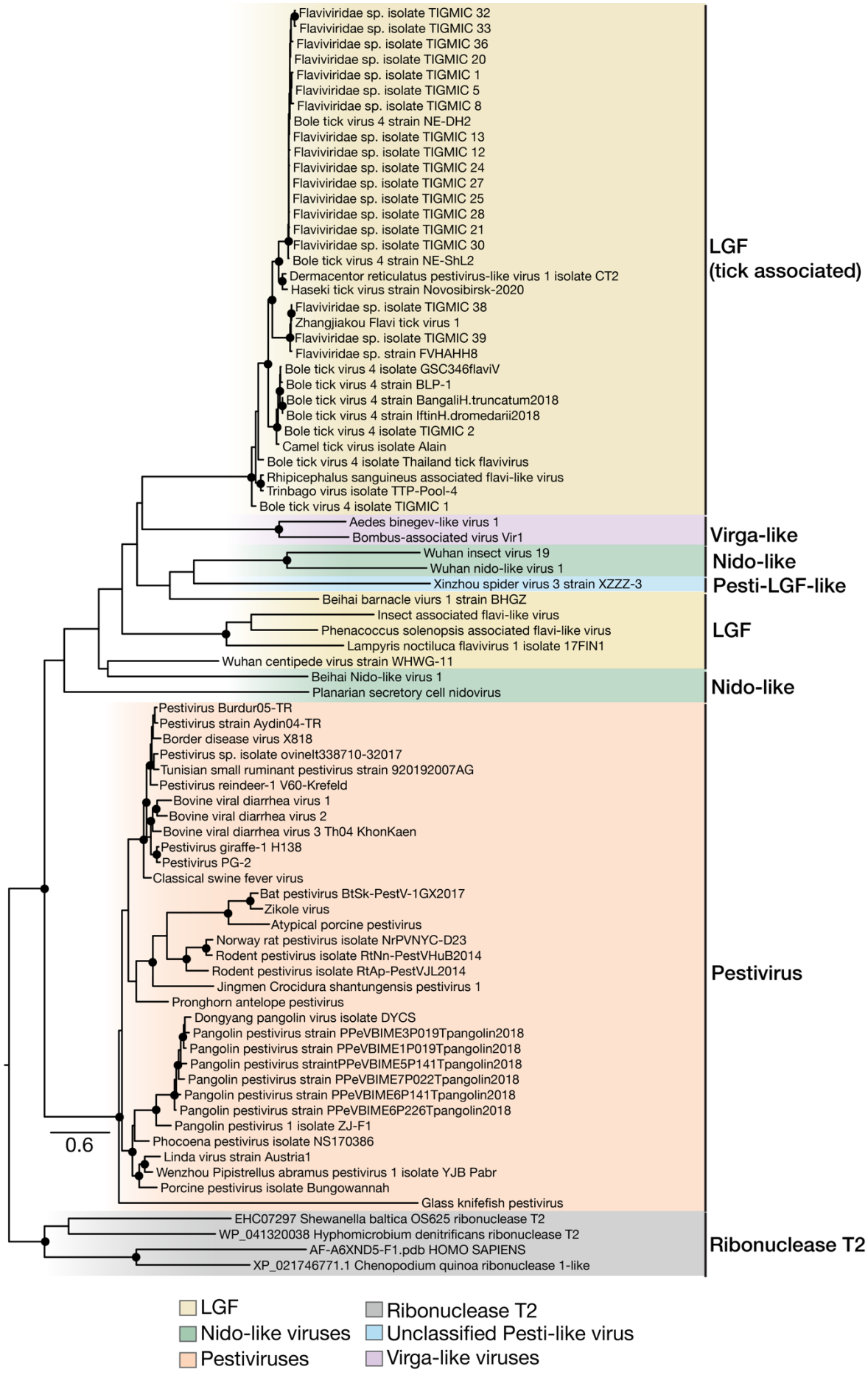

***Supplementary Figure 6.16. ESMFold permits unambiguous detection of glycoproteins in divergent species.*** *(a), RdRp phylogeny, as in Figure 6.2a. (b), E glycoprotein Foldseek e-value heatmaps for Flaviviridae structures predicted only by ColabFold (top), only by ESMFold (middle) or when combined (bottom). (c), Representative examples of targets that are predicted well by ESMFold, but poorly by ColabFold. Structures are colour-coded by pLDDT confidence scores, as shown in the key. pLDDT and pTM metrics are provided for each model.*



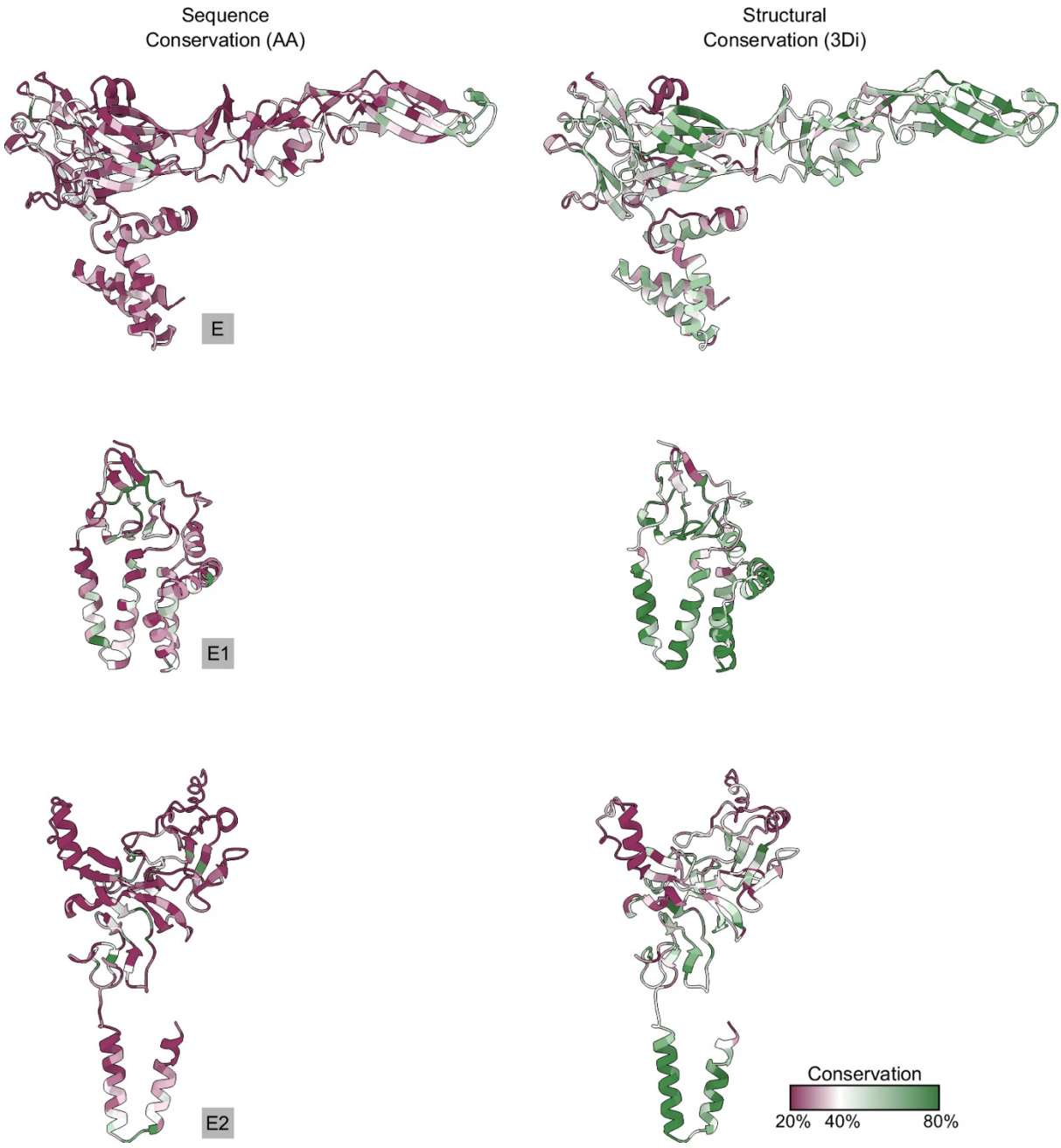

***Supplementary Figure 6.17. Analysis of environmental pesti-like viruses.*** *(a), Large genome flavivirus/Pestivirus subset of the Flaviviridae phylogeny (Tree 18) collapsed to highlight the environmental pesti-like viruses for which no glycoproteins were identified. The scale bar denotes the number of amino acid substitutions per site. (b), Genome organisation is provided for each species, with annotations based on conserved domain sequence searches. (c), Foldseek e-value heatmaps for the indicated reference proteins, values are log transformed and colour-coded as shown in the key. For E, E1 and E2 the values represent summary e-values after comparison with a range of relevant reference structures, as described in the methods.*




***Supplementary Figure 6.18. Absence of a fusion loop in E glycoprotein homologues of the jingmenviruses.*** *Structurally conserved E protein fusion loops (FL) were found in orthoflavi-, LGF-, and in pesti-like viruses. The FL is absent from the E protein homologue of the jingmenviruses (its expected location is marked by an asterisk). Amino acid side chains are shown for the FL only.*

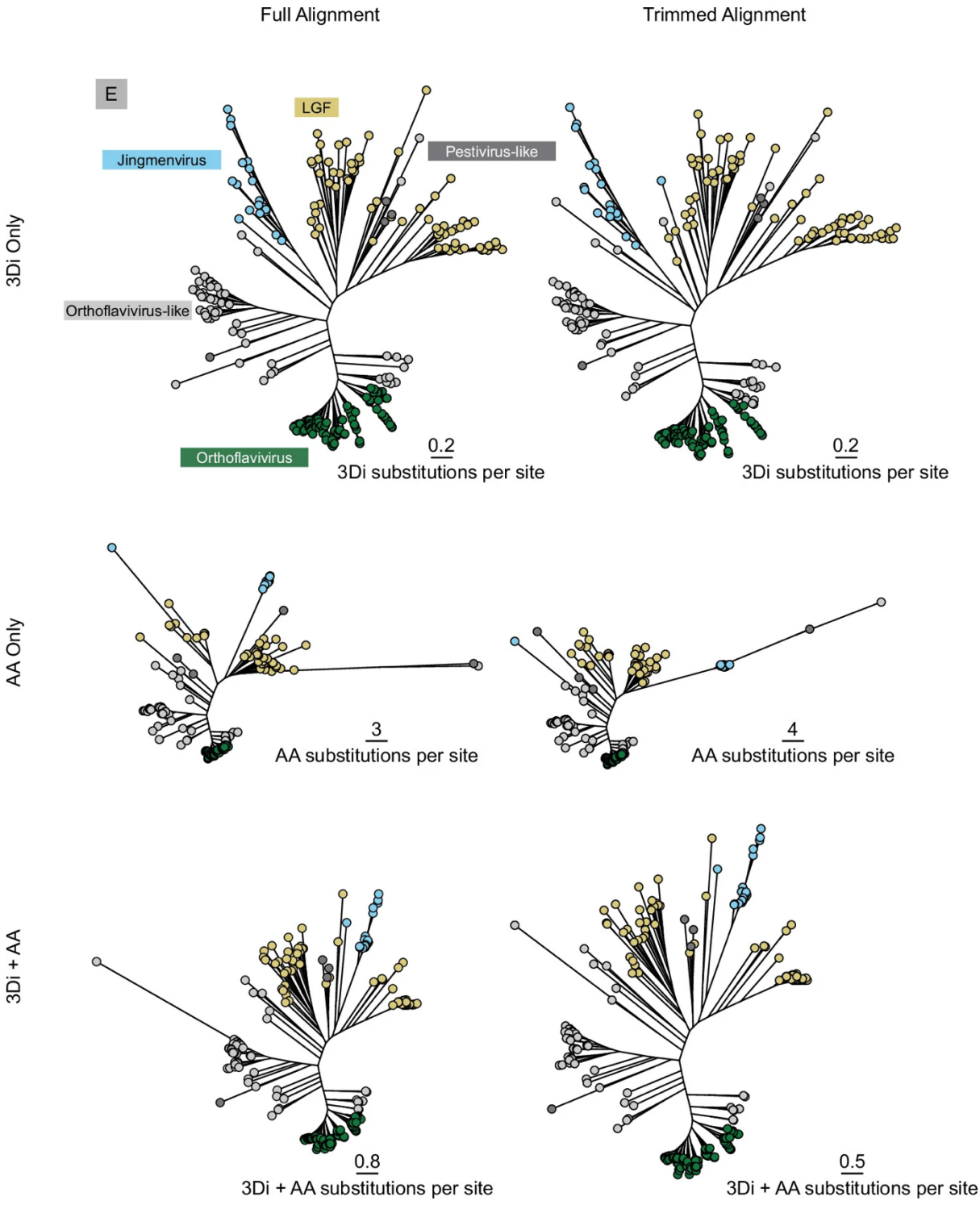

Flaviviridae sp. isolate TIGMIC 32
Flaviviridae sp. isolate TIGMIC 33
Flaviviridae sp. isolate TIGMIC 36
Flaviviridae sp. isolate TIGMIC 20
Flaviviridae sp. isolate TIGMIC 1
Flaviviridae sp. isolate TIGMIC 5
Flaviviridae sp. isolate TIGMIC 8
Bole tick virus 4 strain NE-DH2
Flaviviridae sp. isolate TIGMIC 13
Flaviviridae sp. isolate TIGMIC 12
Flaviviridae sp. isolate TIGMIC 24
Flaviviridae sp. isolate TIGMIC 27
Flaviviridae sp. isolate TIGMIC 25
Flaviviridae sp. isolate TIGMIC 28
Flaviviridae sp. isolate TIGMIC 21
Flaviviridae sp. isolate TIGMIC 30
Bole tick virus 4 strain NE-ShL2
Dermacentor reticulatus pestivirus-like virus 1 isolate CT2
Haseki tick virus strain Novosibirsk-2020
Flaviviridae sp. isolate TIGMIC 38
Zhangjiakou Flavi tick virus 1
Flaviviridae sp. isolate TIGMIC 39
Flaviviridae sp. strain FVHAHH8
Bole tick virus 4 isolate GSC346flaviV
Bole tick virus 4 strain BLP-1
Bole tick virus 4 strain BangaliH.truncatum2018
Bole tick virus 4 strain IftinH.dromedarii2018
Bole tick virus 4 isolate TIGMIC 2
Camel tick virus isolate Alain
Bole tick virus 4 isolate Thailand tick flavivirus
Rhipicephalus sanguineus associated flavi-like virus
Trinbago virus isolate TTP-Pool-4
Bole tick virus 4 isolate TIGMIC 1
LGF
(tick associated)
Aedes binegev-like virus 1
Bombus-associated virus Vir1
Virga-like
Wuhan insect virus 19
Wuhan nido-like virus 1
Nido-like
Xinzhou spider virus 3 strain XZZZ-3
Pesti-LGF-like
Beihai barnacle viurs 1 strain BHGZ
Insect associated flavi-like virus
Phenacoccus solenopsis associated flavi-like virus
Lampyris noctiluca flavivirus 1 isolate 17FIN1
Wuhan centipede virus strain WHWG-11
LGF
Beihai Nido-like virus 1
Planarian secretory cell nidovirus
Nido-like
Pestivirus Burdur05-TR
Pestivirus strain Aydin04-TR
Border disease virus X818
Pestivirus sp. isolate ovineIt338710-32017
Tunisian small ruminant pestivirus strain 920192007AG
Pestivirus reindeer-1 V60-Krefeld
Bovine viral diarrhea virus 1
Bovine viral diarrhea virus 2
Bovine viral diarrhea virus 3 Th04 KhonKaen
Pestivirus giraffe-1 H138
Pestivirus PG-2
Classical swine fever virus
Bat pestivirus BtSk-PestV-1GX2017
Zikole virus
Atypical porcine pestivirus
Norway rat pestivirus isolate NrPVNYC-D23
Rodent pestivirus isolate RtNn-PestVHuB2014
Rodent pestivirus isolate RtAp-PestVJL2014
Jingmen Crocidura shantungensis pestivirus 1
Pronghorn antelope pestivirus
Dongyang pangolin virus isolate DYCS
Pangolin pestivirus strain PPeVBIME3P019Tpangolin2018
Pangolin pestivirus strain PPeVBIME1P019Tpangolin2018
Pangolin pestivirus straintPPeVBIME5P141Tpangolin2018
Pangolin pestivirus strain PPeVBIME7P022Tpangolin2018
Pangolin pestivirus strain PPeVBIME6P141Tpangolin2018
Pangolin pestivirus strain PPeVBIME6P226Tpangolin2018
Pangolin pestivirus 1 isolate ZJ-F1
Phocoena pestivirus isolate NS170386
Linda virus strain Austria1
Wenzhou Pipistrellus abramus pestivirus 1 isolate YJB Pabr
Porcine pestivirus isolate Bungowannah
Glass knifefish pestivirus
Pestivirus
0.6
EHC07297 Shewanella baltica OS625 ribonuclease T2
WP_041320038 Hyphomicrobium denitrificans ribonuclease T2
AF-A6XND5-F1.pdb HOMO SAPIENS
XP_021746771.1 Chenopodium quinoa ribonuclease 1-like
Ribonuclease T2
LGF
Nido-like viruses
Pestiviruses
Ribonuclease T2
Unclassified Pesti-like virus
Virga-like viruses

***Supplementary Figure 6.19. Phylogeny of the RNA virus RNase T2/Erns clade rooted on non-viral sequences.*** *Viral clades colour-coded as shown in the key. A scale bar denotes the number of amino acid substitutions per site. Node support (SH-aLRT >=80% and UFboot >=95%) is indicated by a black circle.*



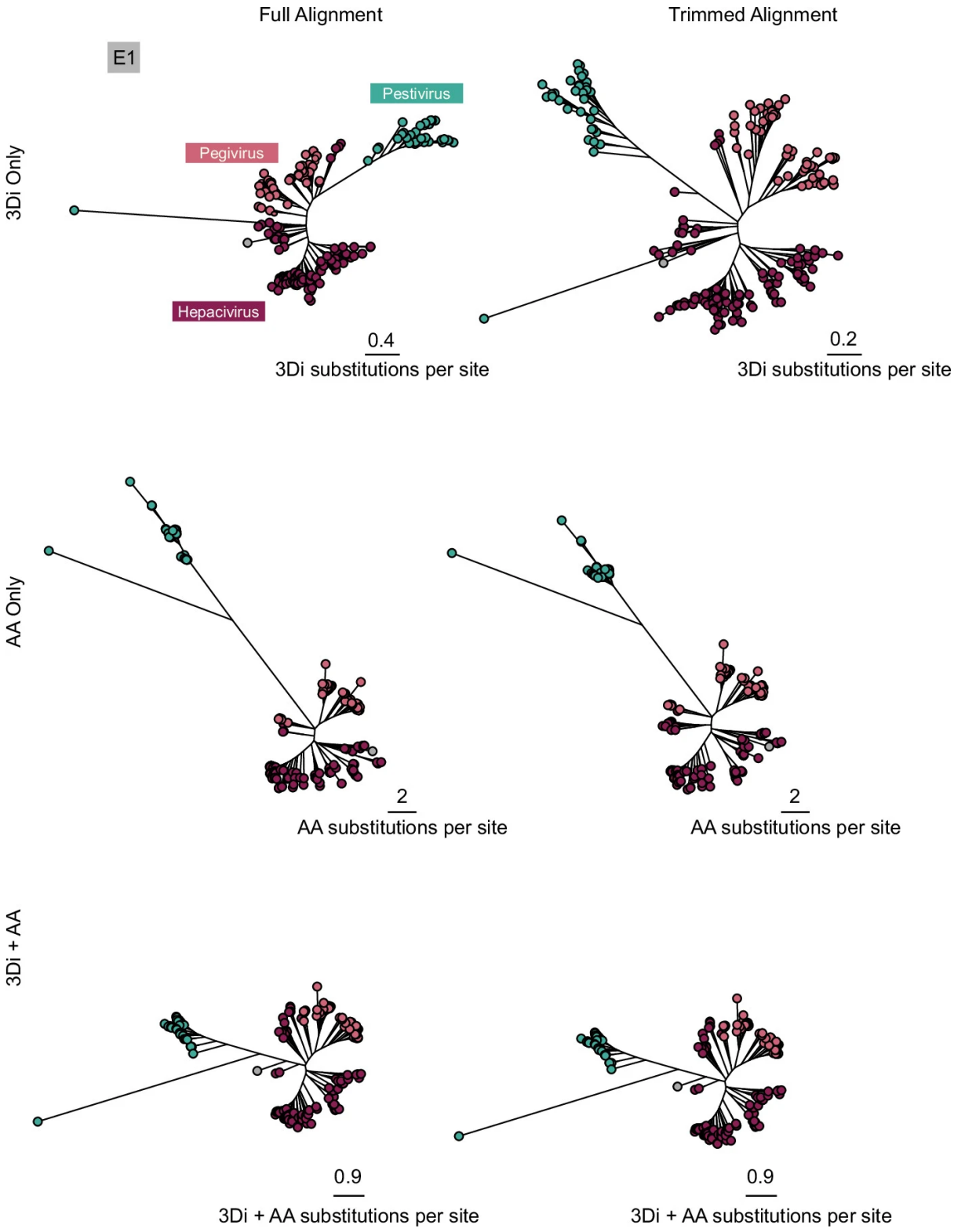

***Supplementary Figure 6.20. Conservation of structure revealed through the Foldseek 3Di alphabet.*** *Representative structures of E (West Nile virus) or E1 and E2 (Hepacivirus F) colour-coded by either sequence or structural conservation, as denoted in the key. In both cases values represent percentage conservation of the consensus for each structurally*

*aligned protein (see Methods for details), with amino acid residues representing protein sequence and the 3Di structural alphabet representing protein structure.*

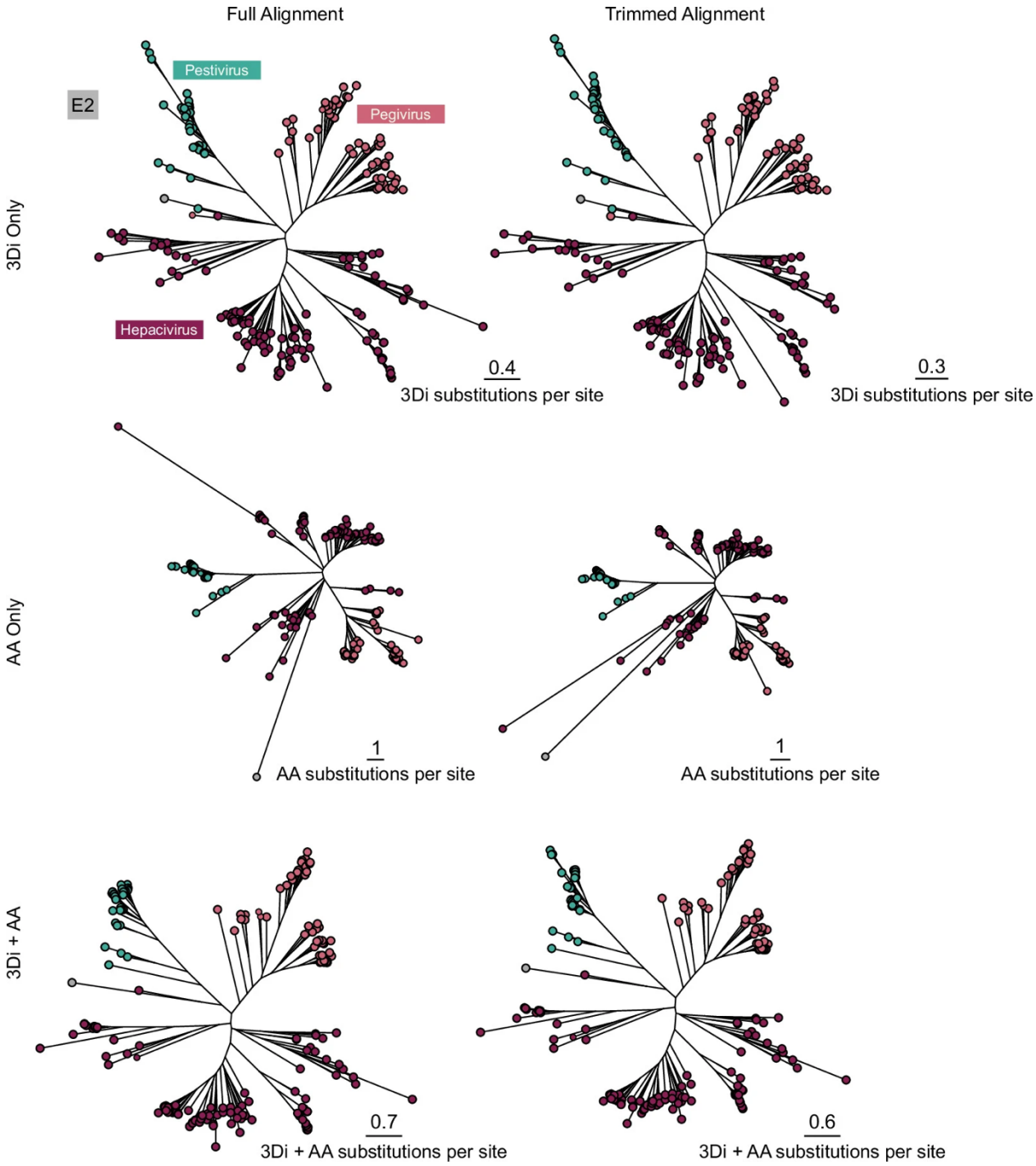



***Supplementary Figure 6.21. Structurally aligned E glycoprotein phylogenies.*** *Protein sequences were aligned using their 3Di structural representation (see Methods for details). Phylogenies were reconstructed using 3Di sequence alone (top), amino acid (AA) sequence alone (middle) or combined 3Di and AA sequence. Right hand trees are derived from*

*alignments trimmed with a gap threshold of 35%. Scale bars indicate substitutions per site for either the 3Di, AA or combined sequences, respectively. Tip shapes are colour-coded by genus/subclade as in Figure 6.1a. All phylogenetic trees are provided in the associated Zenodo repository.*



***Supplementary Figure 6.22. Structurally aligned E1 glycoprotein phylogenies.*** *Protein sequences were aligned using their 3Di structural representation (see Methods for details). Phylogenies were reconstructed using 3Di sequence alone (top), amino acid (AA) sequence alone (middle) or combined 3Di and AA sequence. Right hand trees are derived from alignments trimmed with a gap threshold of 35%. Scale bars indicate substitutions per site for either the 3Di, AA or combined sequences, respectively. Tip shapes are colour-coded by genus/subclade as in Figure 6.1a. All phylogenetic trees are provided in the associated Zenodo repository.*




***Supplementary Figure 6.23. Structurally aligned E2 glycoprotein phylogenies.*** *Protein sequences were aligned using their 3Di structural representation (see Methods for details). Phylogenies were reconstructed using 3Di sequence alone (top), amino acid (AA) sequence alone (middle) or combined 3Di and AA sequence. Right hand trees are derived from alignments trimmed with a gap threshold of 35%. Scale bars indicate substitutions per site for either the 3Di, AA or combined sequences, respectively. Tip shapes are colour-coded by genus/subclade as in Figure 6.1a. All phylogenetic trees are provided in the associated Zenodo repository.*

**Supplementary Note 6.1: MUSCLE alignment analysis**

To demonstrate the robustness of the NS5b alignment and the resulting phylogenies we created a stratified MUSCLE alignment ensemble with the same NS5b sequences used to build the representative *Flaviviridae* phylogeny (Tree 18). The stratified ensemble comprised four replicates for different HMM parameters and four for guide tree merge order permutations, resulting in 16 total alignments (see Edgar (2022)). We measured the dispersion (D), the average distance between these replicate alignments, and found it to be 0.1320, thereby reflecting the presence of regions of low sequence conservation that are vulnerable to misalignment (D = 0 signifies that replicates are identical). To evaluate whether D > 0 affects the robustness of the MSA and tree topology, we first evaluated the consistency of the core *Flaviviridae* RdRp motif columns across the stratified MUSCLE replicate alignments and compared these to Alignment 18. Using the Palmscan software (v1.0) (Babaian and Edgar 2022), we extracted RdRp palm motifs from 448 of the 461 NS5b sequences in the *Flaviviridae* amino acid sequence set. Sequences without detectable palm motifs were confided to the LGF and PLUN groups, and Motif C was manually detected in these. The extracted RdRp palm motifs were then aligned using MUSCLE and used to identify sequence logos, which were compared to those identified in the Kitrinoviricota (Babaian and Edgar 2022; Charon et al. 2022) to obtain boundaries for each motif (Supplementary Figure 6.2).

We calculated letter confidence—a measure of how consistently a letter at a specific position is aligned—using the stratified MUSCLE replicates as input with the untrimmed Alignment 18 as the reference. Averaging across sequences, all positions within the palm motifs achieved the maximum possible letter confidence score, with the exception for position 2016 within motif B, which scored 8/9. Hence, high level of letter confidence suggests that despite the observed dispersion in replicates, the core motifs remain unaffected (Supplementary Figure 6.2). We next evaluated how dispersion (D) affects deep tree topology. Notably, the majority (56%, 9/16) of phylogenies estimated from the untrimmed MUSCLE replicates produced similar topologies to Tree 18, with the *Flaviviridae* placed into three distinct clades: (i) large genome flaviviruses and Pestivirus, (ii) Orthoflavivirus-Jingmenvirus, and (iii) Pegivirus-Hepacivirus. Two other topologies were also observed: (1) Scenario B in Supplementary Figure 6.6, occurred in 38% (6/16) of phylogenies and is similar to that of Tree 18 with the exception that the pestiviruses fall basal to the Orthoflavivirus, Jingmenvirus and LGF: (2) Scenario C in Supplementary Figure 6.6 was only observed once. This phylogeny placed the pestiviruses basal to the hepaci-

and pegiviruses. These alternative topologies/scenarios for *Flaviviridae* evolution were considered in further detail below. There was no pattern according to alignment length, pairwise identity, percentage identical sights and topology. Phylogenetic analysis of divergent families such as the *Flaviviridae* typically employ quality trimming to remove poorly aligned regions from alignments before estimating phylogenies. To replicate this and the initial analysis used to select Tree 18, we trimmed each of the 16 MUSCLE replicates using trimAl with three gap thresholds (0.7, 0.8, and 0.9), eight consensus values (5, 7.5, 10, 12.5, 15, 17.5, 20, and 25), and one automated parameter selection mode (gappyout), resulting in 416 alignments (including the 16 untrimmed). Phylogenies were then estimated using three different substitution models resulting in a total of 1248 phylogenies. Neither a treespace PCoA analysis, nor mapping the distances between phylogenies in an MDS plot revealed clear clustering of the various guide tree merge order variations and HMM parameters, suggesting that the effects of dispersion from these perturbations is not easily predicted (Supplementary Figure 6.4). We then repeated this in combination with the original phylogeny set constructed with MUSCLE and MAFFT finding that the original tree set (without ClustalO alignments) sits within the core diversity of the trees estimated from the stratified MUSCLE replicates (Supplementary Figure 6.5a & b). This suggests our original analysis set covers much of the tree space explored by the MUSCLE stratified perturbations, and that both have a consistent core grouping.

Our estimation of the *Flaviviridae* phylogeny is particularly concerned with the deep evolutionary history of the family. We then sought to assess whether these phylogenies were consistent with the deep tree topology observed in our original analysis. After removing unclassified tips (i.e., those with clade labels PLUN, FJUN or HPUN [n = 19] in Supplementary Table 3) as these were typically on long branches and had highly inconsistent topologies, we assessed whether monophyly was present across various clade groupings consistent with our original phylogeny using the R package MonoPhy (v1.3) (Schwery and O'Meara 2016). We found that 58% of the phylogenies matched Scenario A in the deep branching order (Supplementary Figure 6.6). Notably, the LGF-Pestivirus clade was responsible for 99.8% (527/528) of deviations from Scenario A, with Scenario B the most commonly observed (38%) (Supplementary Figure 6.6). Geometric median phylogenies created for the levels of HMM and guide tree perturbations showed similar topological ratios (Scenario A 56%, Scenario B 44%).

The varying frequencies of each topology across the guide tree and HMM parameters suggest that both contribute bias to the downstream analysis. For example, different topologies were consistently observed across the levels of HMM parameters when guide tree selection was held consistent (Supplementary Figure 6.3). There is an exception to this; in phylogenies estimated from untrimmed MUSCLE replicates HMM level 3 and Guide tree variation level “none”, all reflect the topology of Scenario A. This was also noted in the trimmed variations, where Scenario A was more consistently observed when these parameters were selected. Overall, the ratios of topologies across the parameter levels were similar to those observed in the untrimmed trees, suggesting that while trimming can reduce noise it appeared to have a limited effect on the deep branching order compared to the effects of parameters perturbed by MUSCLE.

Critically, therefore, despite the challenges of creating alignments and phylogenies over such extensive evolutionary spans, that our alignments and phylogenies almost exclusively reflect one of two deep topologies that only differ in the placement of the pestiviruses by a single node.

We argue that there are several key pieces of evidence that suggest Scenario A is the most parsimonious topology for the *Flaviviridae*:

**1. Scenario A was the most common topology observed across aligners MAFFT and MUSCLE In our original analysis we used multiple aligners to construct MSA.**
We note that Scenario A was the tree topology most commonly observed in both phylogenies estimated from MUSCLE (97%) and MAFFT (88%) alignments. Of note, 60% of ClustalO phylogenies were consistent with Scenario B. In our initial analysis we found that several of the ClustalO topologies were discordant both in manual topological comparison and distance-based metrics: for example, the genus Pegivirus appeared in the middle of the hepaciviruses, while the Hepacivirus and Pegivirus clade would occasionally appear sister to the orthoflavi- and jingmenviruses. It is interesting to note that in benchmarks testing sequence aligners against simulated viral sequence data ClustalO performed poorly on sequences modelled on Hepatitis C virus compared to MUSCLE and MAFFT (Young et al. 2022). Due to these inconsistencies, we made the decision to exclude these alignments from our final tree selection in the original analysis.

**2. Scenario A is consistent with the previously published *Flaviviridae* phylogenies despite methodological variations between studies.**

While studies presenting family level NS5b phylogenies that include representatives from the jingmenviruses and LGF are rare, where they have been constructed they all follow the same deep branching topology as Scenario A despite methodological differences (Bamford et al. 2022; Debat and Bejerman 2023; Dong et al. 2021; Paraskevopoulou et al. 2021; Petrone et al. 2024; Shi et al. 2016a)

**3. Scenario A is the most parsimonious biologically.**

Scenario B assumes that the LGF, jingmenviruses and orthoflavivirues (including orthoflavi-like) diverged from a Pestivirus. Notably, orthoflavi-like viruses and endogenous viral elements (EVEs) have been identified in basal metazoans, including Cnidaria, suggesting a divergence from a flavi-like ancestor between 652-973 mya (Bamford et al. 2022). In contrast, pestiviruses which exclusively infect vertebrates, exhibit a clear pattern of virus-host co-divergence (Mifsud et al. 2023) and have no identified EVEs outside of mammals appear to have diverged more recently, 450-497 million years ago (Bamford et al. 2022). Given the ancient association of the orthoflavi-like viruses and that most vertebrate virus groups fall within larger invertebrate virus clades (Shi et al. 2016b) it is unlikely that the LGF, orthoflaviviruses and jingmenviruses emerged from an ancestral vertebrate infecting Pestivirus. However, a scenario where this ancestral virus infected invertebrates cannot be entirely ruled out, despite the absence of evidence for a vast, unsampled viral diversity or EVEs along this branch. It is important to note that the Spider pesti-like viruses don't fulfil this role. In Scenario B, where the pestiviruses fall basal to the LGF, the Spider pesti-like viruses typically cluster with the LGF rather than the pestiviruses. Lastly, the presence of a homologous RNase T2 in the pestiviruses and some members LGF but not in the jingmen- and orthoflaviviruses is provides further evidence for Scenario A. Taken together, it is more likely that the pestiviruses came from an invertebrate associated LGF-like ancestor (Scenario A) given the enormous diversity of these viruses and large gene pool available.

*The following Supplementary Tables are available online as .xlsx files:*

***Supplementary Table 1.*** *Flaviviridae host association and vector status metadata related to Figure 6.2c.*
https://static-content.springer.com/esm/art%3A10.1038%2Fs41586-024-07899-8/MediaObjects/41586_2024_7899_MOESM5_ESM.xlsx

***Supplementary Table 2.*** *Contribution of 3Di vs amino acid in the glycoprotein phylogeny partition models.* https://static-content.springer.com/esm/art%3A10.1038%2Fs41586-024-07899-8/MediaObjects/41586_2024_7899_MOESM6_ESM.xlsx

***Supplementary Table 3.*** *Flaviviridae sequence metadata including clade designations and GenBank nucleotide accession numbers.* https://static-content.springer.com/esm/art%3A10.1038%2Fs41586-024-07899-8/MediaObjects/41586_2024_7899_MOESM3_ESM.xlsx

***Supplementary Table 4.*** *Combination of sequence alignment, quality trimming methods, and amino acid substitution models used to infer the NS5b phylogenies.* https://static-content.springer.com/esm/art%3A10.1038%2Fs41586-024-07899-8/MediaObjects/41586_2024_7899_MOESM4_ESM.xlsx

# Chapter 7. General discussion

Metatranscriptomics (i.e., total RNA sequencing) has proven a powerful tool for discovering and characterising viral diversity. Even over the time-course of my thesis other researchers have conducted large scale virus discovery projects spanning diverse environments and planetary metatranscriptomes which have collectively contributed to at least a 10-fold expansion of the known RNA virosphere (Chen et al. 2022; Edgar et al. 2022; Hou et al. 2024; Olendraite et al. 2022; Zayed et al. 2022). Despite these advances, major gaps remain in our knowledge of the virosphere, in large part reflecting uneven sampling effort across the tree of life (Harvey and Holmes 2022). Currently, the number of viruses present in the NCBI taxonomy database stands at 258,034, approximately 200,000 of which are RNA viruses. The International Committee on Taxonomy of Viruses has formally classified 6,712 of these. However, the estimated number of eukaryotic virus species far exceeds this number, with a conservative estimate of 87 million species (Geoghegan and Holmes 2017). As a result, ~99.7% of eukaryotic viruses remain undiscovered alongside a likely equally vast diversity of bacteriophage. This thesis aimed to fill some of these gaps, at the same attempting to reveal more about the fundamental mechanisms, patterns, and processes of virus evolution.

My approach combined metatranscriptomic sequencing and the secondary analysis of transcriptome data with state-of-the-art phylogenetic and structural analysis, including the use of methods that utilise artificial intelligence-based analysis of protein structures. In Chapters 2 through 5, I focused on describing virus diversity across a wide array of species, including animals ranging from domesticated and wild mammals, to marine non-bilaterians, as well as those spanning the entire plant kingdom. Through this work, I identified 150 novel viruses and 16 known canine virus species. In Chapter 6, I demonstrated how the large diversity of metagenomically identified viruses, particularly those in hosts of limited economic significance, provide key insights into the evolutionary history of virus lineages as well as the evolution of viral proteins that influence host range, tissue tropism, and pathogenesis.

Below, I discuss the core themes of this thesis, the implications of the work, ongoing challenges in the field, and outline questions for future research.

As discussed in Chapter 1, metatranscriptomics (has fundamentally changed how we identify potential aetiological agents in animals. Despite this, there are challenges and limitations associated with this approach. Metatranscriptomics generates large amounts of sequencing data from all organisms present in a sample. For example, in a mammalian faecal sample, sequencing reads may originate from the host, its diet, or gut flora such as bacteria, fungi, and invertebrates. All of these organisms in turn harbour their own viruses. Faced with this deluge of information, virus-host assignment can be challenging particularly when the virus is highly divergent. One approach to help solve this involves combining metatranscriptomic sequencing with traditional veterinary diagnostic techniques such as microscopy and histology to provide further evidence for host assignment.

In Chapter 2, I demonstrate the effectiveness of combining metatranscriptomic with traditional veterinary diagnostic techniques through the examination of New Zealand fur seals, both those with and without clinical or pathological signs of viral infection. Here, metatranscriptomics proved valuable as it provided information on the presence of non-viral organisms and their relative abundance which aided virus-host assignment of potential aetiological agents. Using this method, I identified a novel papillomavirus and established a likely aetiology for an individual presenting with clinically and histologically identified oral papilloma-like lesions. This represented the first papillomavirus described in fur seals, and the first report of oral papilloma-like lesions in this host group. Although these findings were of limited clinical significance in this case, since the lesions self-resolved, it is possible that plaques of this nature could progress into invasive squamous cell carcinomas in other individuals (Munday et al. 2016). Indeed, aided by transcriptomics I was able to provide evidence that this papillomavirus may have existed in both episomal and integrated forms, a factor relevant to papillomavirus-induced oncogenesis (Münger et al. 2004; Schmitz et al. 2012). Further research is needed to determine the prevalence of this virus in New Zealand fur seals and its clinical relevance, as this observation was based on a single individual. With the ongoing outbreaks of highly pathogenic avian influenza viruses (e.g., H5N1) in seal populations (Mirolo et al. 2023), there is a pressing need for continued viral surveillance in seals.

A reoccurring approach for virus discovery I used in this thesis was the secondary analysis of publicly available sequencing data. The Sequence Read Archive (SRA) hosts over 5 million individual sequencing runs from more than 48,461 unique organisms, and this number continues to grow rapidly (Leinonen et al. 2010). This resource provides a snapshot

of the diversity of life on Earth and is a valuable resource for virus discovery. It maximises prior monetary investments in sampling and sequencing while enabling access to a broader diversity of host species and the expertise required to sample and extract nucleic acids from them. In Chapter 4, I analysed data from the 1000 Plant Transcriptomes Initiative (1KP), which generated transcriptomes for plants across all major taxonomic groups and broad geographical regions, many of which hold little economic importance (Leebens-Mack et al. 2019). At the time of this study, aside from a few viruses previously identified in ferns (Nibert et al. 2018; Valverde and Sabanadzovic 2009), little was known about viruses in earlier-diverging plant lineages such as lycophytes, bryophytes, and charophytes (i.e., "lower plants"). Through the analysis of these transcriptomes and the identification of 104 novel viruses, I expanded the known host range of 13 virus families to include lower plants and four virus families to include algal hosts. My findings revealed that numerous RNA virus families were linked to older evolutionary plant lineages than previously thought and suggested that the apparent scarcity of RNA viruses in lower plants likely reflected a lack of investigation rather than their true absence.

Plant virus genome structure and composition appeared to vary between the major plant clades. For example, the recent discovery of an unsegmented *Varicosavirus* (family *Rhabdoviridae*) in gymnosperms (Bejerman et al. 2021; Walker et al. 2018), differs from the typical bi-segmented viruses associated with flowering plants and raised the question as to whether this represented an adaptation specific to gymnosperms or perhaps the ancestral form of the varicosaviruses (Bejerman and Debat 2024). I identified bi-segmented varicosavirus genomes for the first time in ferns and liverworts which predate the gymnosperms. This provides evidence that the ancestral state of this lineage was likely bi-segmented for much of plant evolution and further evidence for a gymnosperm-specific clade of unsegmented varicosaviruses. To the best of my knowledge, there is no evidence of mono-segmented viruses evolving from bi-segmented ancestors. A question that arises from this observation is what evolutionary factors underpin this transition. Could having an unsegmented genome simplify replication and require a lower multiplicity of infection (Dall'Ara et al. 2016)? Perhaps the versatility provided by segmentation—such as differential accumulation of genome segments in the host, finer control of viral gene expression, and adaptability through segment reassortment (Bonnamy et al. 2024; Dall'Ara et al. 2016) – is not necessary for vertical and persistent infections in gymnosperm hosts which are typically longer-lived and have slower reproductive cycles than flowering plants, ferns, and liverworts (Mauseth 2014).

Despite the viruses identified in this study there are still substantial gaps remaining in our sampling of plants, and in general far less attention has been paid to sampling plant viruses compared to animal viruses. To the best of my knowledge, viruses are yet to be described in the Glaucophytes and only a handful have been documented in the Charophytes. Viruses from these plants will undoubtedly inform us about the acquisition of viruses in plants and how the extant plant virus families have evolved.

In this thesis I focused on exploring the viral diversity of organisms of evolutionary significance, particularly those representing key transitions such as the terrestrialisation of plants or the development of a bilaterally symmetric body plan. In Chapter 5, prompted by the finding of an orthoflavi-like derived endogenous viral element (EVE) in the peach blossom jellyfish genome (Bamford et al. 2022), I conducted a broad survey of flavivirus-like sequences across publicly available transcriptomes. This survey spanned the diversity of the metazoans. This led to the identification of 32 novel flavivirus sequences including an orthoflavivirus-like sequence in a species of Cnidaria. This virus occupied a basal lineage to the genus *Orthoflavivirus* and provided further evidence for an ancient origin of the orthoflaviviruses. The identification of a virus in a non-bilaterian (organism lacking a bilaterally symmetric body plan) raises important questions regarding the ancestral modes of flavivirus transmission and their tissue tropisms. The absence of a circulatory system in non-bilaterians suggests the possibility of alternative modes of cell-to-cell virus transmission in these organisms. This could include exosome-like vesicles as previously reported in other orthoflaviviruses, although these are not well documented in Cnidaria (Bamford et al. 2022; Zhou et al. 2018). Other possible routes for cell-to-cell transmission of viruses in marine non-bilaterians could possibly include extracellular diffusion of virus particles or interactions with the Cnidaria's symbiotic bacteria or protists.

Given the hepatotropism observed in human, rodent, and bovine hepaciviruses (Baron 2018; Jiang et al. 2022) and the identification of these viruses across all major vertebrate groups (Shi et al. 2018), it is possible that hepatotropism is a universal and ancient feature of the hepaciviruses, which potentially dates back to the origin of the vertebrate liver. Indeed, the fish, reptile, and bird hepaciviruses identified in Chapter 5, as well as in subsequent studies (Costa et al. 2024; Mahar et al. 2024), were predominately associated with liver tissue. Evidence of hepacivirus-like ancestors in invertebrates is limited—apart from the identification of a distant hepaci-pegi-like EVE basal to the hepaci- and

pegiviruses (Bamford et al. 2022). Assuming hepatotropism is conserved, lancelets (family Branchiostomatidae) would be a logical target for such a search as they contain homologs of organ systems found in fish, and vertebrates more broadly. This includes the hepatic caecum which appears to be a homologue or ancestral form of the vertebrate liver (Zhang et al. 2023).

While the SRA is a widely used resource for virus discovery there are several limitations that must be acknowledged during study design. Between sequencing projects, libraries are often prepared using various approaches, employing different positive or negative selection methods, and are sequenced on different platforms at varying depths. For example, the libraries of the 1KP were prepared with enrichment for polyA+ RNA, which limited the detection of viruses with non-polyadenylated RNA such as those in the *Tospoviridae*, *Arenaviridae* and *Orthomyxoviridae* (Visser et al. 2016). Although these variations are unlikely to impact the conclusions of a pure virus discovery study such as Chapter 4, they introduce biases that limit the ability to conduct cross-study statistical comparisons of virus diversity and abundance (Visser et al. 2016). Additionally, the use of public data limits control over sampling methodology and contamination, which poses challenges to data reliability and, more specifically, the accurate assignment of virus host species. Consequently, SRA host assignments cannot be assumed to reflect the actual host of a virus without additional investigation. As discussed in Chapter 1, this could include examining library composition using read taxonomic classification tools such as KMA or CCMetagen (Clausen et al. 2018; Marcelino et al. 2020), employing phylogenetic analyses to infer the host relationships, and reviewing the sampling and sequencing methodologies associated with the libraries. Lastly, it is often difficult to obtain the original sample for further virological analyses and PCR confirmation. Nonetheless, with proper caution and the right application, the secondary analysis of transcriptome data will continue to lead to impactful discoveries.

Expanding our understanding of viruses does not always necessitate sampling new host species. Instead, new insights into virus evolution and disease ecology can be gained from examining well-studied species in unique geographical and ecological contexts. For example, Australia's geographic isolation, strict biosecurity measures, and native dingo populations present a unique model for understanding the spread and evolution of canine viruses. In Australia, available virus sequence data from domestic dogs (*Canis lupus familiaris*) is largely limited to DNA viruses such as canine parvovirus (CPV) (Kelman et al.

2020). Where RNA viruses have been examined, this has been largely limited to clinical and epidemiological investigations (Marshall et al. 1984; Wyllie et al. 2016). This gap in research is even more pronounced for dingoes (*Canis lupus dingo*), with only CPV reported (Van Arkel et al. 2019), limiting our understanding of canine virus evolution and disease ecology. In Chapter 3, using a metatranscriptomic approach, I examined the viruses in Australia's domestic dog and dingo revealing the presence of 86 vertebrate-associated viruses belonging to 16 distinct species including important dog pathogens and the first RNA virus – rotavirus A (RVA) – in a dingo.

One of the aims of this study was to compare the viromes of domestic dogs and dingoes to assess the extent of cross-species virus transmission. It has been established that CPV can be transmitted between domestic dogs and dingoes (Van Arkel et al. 2019). Other canine pathogens (e.g., canine distemper virus) have broad host ranges including non-canids such as rodents (Wyllie et al. 2016). Given this, and the close genetic relatedness between dingoes and domestic dogs, it is likely that there is a large overlap in their virome, at least in terms of viral susceptibility. However, behavioural differences between dingoes and domestic dogs—such as dingoes living in highly structured packs, occupying larger home ranges and exhibiting distinct temporal activity patterns (Ward et al. 2021)—may lead to varying levels of exposure to viruses, particularly those circulating in wildlife. While interactions between dingoes and domestic dogs have been observed (Ward et al. 2021) this is likely to be infrequent for the majority of domestic dogs. Of particular note in this context, the dingo RVA discovered in Chapter 3 was phylogenetically distinct from dog-associated RVA sequences, and more closely related to those found in humans and bats. As well as representing the first dingo RNA virus, this observation also shows that dingo viruses do not necessarily come from domestic dogs. This finding also indicates that a past cross-species transmission event took place and highlights that, despite their close genetic relationship, domestic dogs and dingoes do not necessarily share identical viromes. Additional research is required to recover the complete genome of this virus and confirm its association with dingoes.

My analysis revealed that Australian canine viruses occupied multiple distinct phylogenetic clades and exhibited minimal geographic structure suggesting multiple virus introductions into Australia and high connectivity between dogs despite separation of over 2000 km in some cases. This raises an additional question: how are these viruses entering Australia? Several of the enteric viruses identified in this study are environmentally stable for extended

periods and can be transmitted via fomites (Boone and Gerba 2007). As such, it is plausible that these viruses could be transmitted to Australian dogs through contaminated clothing or footwear. Other routes of viral entry could potentially include the illegal entry of dogs or other carnivores on boats and the importation of dog semen (Charles 2022). Although Australia's strict quarantine measures for importing dogs likely greatly reduce this risk, they do not completely eliminate it. Since all imported dogs are sent to a single quarantine facility in Mickleham, Victoria, it is possible that viral transmission could occur at this location. Once a virus is introduced into Australia, it is likely spread through human-mediated interstate movement of dogs. A broader sampling of domestic dogs, particularly those recently imported could provide the virus sequences necessary to reconstruct these introduction events.

Major advancements in protein structure prediction have taken place over the duration of this thesis, most notably the development and public release of AlphaFold2 in 2021 (Jumper et al. 2021). This software and the abundance of available structures (Varadi et al. 2024) have opened countless avenues of research, including the identification of distant homology between virus proteins in the twilight zone" (20-35% amino acid sequence identity) and below. Over the past decade, our understanding of flavivirus evolution has undergone a fundamental transformation, and with the continued identification of viruses in diatoms, plants, fungi, and diverse invertebrate taxa this progress is certain to continue. Making use of the recent advancements in protein structure prediction, Chapter 6 demonstrates the power of structural based homology searches to map the distribution of glycoproteins across the *Flaviviridae*. When mapped onto the underlying RdRp phylogeny the distribution of glycoproteins uncovered a complex evolutionary history characterised by bacterial gene capture and potential inter-genus recombination. This work furthers our understanding of the diversity of viral fusion mechanisms in flaviviruses and the events that have shaped the virology and ecology of this diverse family.

Several notable questions remain: (i) what was the entry system of the common ancestor of the *Flaviviridae*? Given the diverse host range of the class II envelope system and its presence in early diverging metazoan lineages it is tempting to hypothesise that this was the ancestral entry system. However, when examining other members of the phylum Kitrinoviricota, all families (n = 21) with the exception of the *Flaviviridae*, *Togaviridae*, *Matonaviridae*, are non-enveloped. A comprehensive structural similarity-based search across the virus kingdom is likely required to identify the extent of the class II entry system,

followed by structural phylogenetic analysis. The diversity of these structures would likely pose challenges for current methodologies. (ii) What are the evolutionary origins of the E1/E2 system. No homologs of the E1/E2 system appear to exist outside of the hepaci-, pegi- and pestviruses. Likewise, no intermediate form between the class II and E1/E2 systems is apparent. Uncovering greater viral diversity and a larger diversity of predicted structures will likely be critical to addressing this question. One striking finding across flavivirus virus discovery projects (Mifsud et al. 2023; Paraskevopoulou et al. 2021; Shi et al. 2016) is the absence of pegi and hepaci-like viruses in invertebrates. However, as mentioned previously, an EVE with distant homology to pegi- and hepaciviruses was identified in the genome of a mollusc (Bamford et al. 2022) suggesting that this group may have its origins in marine invertebrates. (iii) If an invertebrate pegi-/hepacivirus ancestor were found, does this virus possess an E1/E2 envelope, marking the first detection of this feature outside vertebrates? (iv) If so, does it also contain a methyltransferase or internal ribosomal entry site and (v) are there shared characteristics between this envelope protein and class II envelope proteins, representing an intermediate form between the two systems? Flavivirus genomes are typically 9-13 kb in length, although members of the large genome flaviviruses (LGF) have been found to extend up to ~40 kb (Petrone et al. 2024). (vi) Was the ancestor of the *Flaviviridae* large in genome size, with subsequent genome reduction events occurring in the other genera, or have the LGF increased in size due to the acquisition of a yet unknown proofreading mechanism? (Petrone et al. 2024). An invertebrate pegi-/hepacivirus sequence would likely be informative regarding how the genome size of the *Flaviviridae* has fluctuated over its evolutionary history.

Due to the resilience of structural features to saturation over deep evolutionary timescales, phylogeneticists and structural biologists have long explored the use of protein structures to infer phylogenetic trees (Bujnicki 2000; Eventoff et al. 1975). The growing abundance of available protein structures and the ease with which they can now be generated are likely to make structural phylogenetics a rapidly advancing area of research in the coming years. One exciting approach is the use of structural alphabets, such as 3Di tokens, to generate distance matrices (Moi et al. 2023) or directly as phylogenetic characters (Puente-Lelievre et al. 2024). Incorporating 3Di tokens directly as phylogenetic characters in traditional bioinformatics and phylogenetic programs allows for the use of methods such as maximum likelihood (ML), Bayesian inference, and modern ultrafast statistical measures of phylogenetic uncertainty. The key advantage of 3Di tokens is the ease at which they can be used in current alignment and phylogenetic programs and their potential to overcome the

site saturation problem associated with amino acids. However, the transition from amino acids to 3Di may result in the loss of information about evolutionary events occurring over shorter timescales. As demonstrated by Puente-Lelievre et al. (2023) and replicated in Chapter 6, 3Di tokens can be integrated in partition models alongside amino acids *to* simultaneously resolve both deep evolutionary and finer-scale relationships. Despite its promise, the use of 3Di tokens as phylogenetic characters remains in its infancy and requires further development and testing. As discussed further in Garg and Hochberg (2024) there are several limitations with the use of 3Di, including that it explicitly violates the ML assumption of site independence, as each character reflects the interaction of six or more amino acid positions in three-dimensional space. An exciting extension of this is the training of protein language models, such as ProstT5, to predict 3Di characters directly from amino acid sequence (Heinzinger et al. 2024). This method would eliminate the need for structural prediction drastically increasing computational speed, and could be useful for long polyprotein such as the extreme 12,236 amino acids of Maximus pestivirus, a large genome flavivirus identified in a *Haliclona* sponge (Petrone et al. 2024). Polyproteins of this size far exceed the current limit of AlphaFold (2000 amino acids for ColabFold). Although it is currently unknown how well this approach will perform with polyproteins, if successful, it could eliminate the need to determine cleavage sites, which is often challenging for highly divergent viruses, or, as we did in Chapter 6, to segment and fold polyproteins into small overlapping building blocks, which proved a computationally intensive procedure. Taken together, these advancements highlight a clear use case for structural phylogenetics in resolving challenging rooting problems, such as reconstructing the deep evolutionary relationships within the broad *Riboviria*.

Metatranscriptomic sequencing for virus discovery has proven a powerful tool for discovering and characterising viral diversity. Despite the exponential increase in the number of known viruses over the past decade, much of the tree of life remains unexplored. In this thesis, I integrated metatranscriptomic sequencing with phylogenetic and structural analyses to characterise viruses in species critical to understanding the ecological and evolutionary processes that shape viral diversity. This approach resulted in the identification of aetiological agents in wildlife, the discovery of highly divergent viruses in early diverging taxa, an expansion of the known host ranges for several virus families, and the distribution of glycoproteins in the *Flaviviridae*. Through these findings I highlighted how novel and highly divergent viruses—often discovered in hosts of limited economic importance—can be key to reconstructing virus origins and evolutionary histories. With recent advancements

in AI technology and the rapid generation of vast sequencing datasets, we are entering an exciting period for viral discovery and evolutionary research.

### 6.1 References


Bamford, C. G., et al. (2022), 'Comparative analysis of genome-encoded viral sequences reveals the evolutionary history of flavivirids (family Flaviviridae)', *Virus Evolution,* 8 (2).

Baron, A. L. (2018), 'Biological characterization of Bovine Hepacivirus', (Tierärztliche Hochschule Hannover).

Bejerman, N. and Debat, H. (2024), 'Alpha-and Betagymnorhavirus: two new genera of gymnosperm-infecting viruses in the family Rhabdoviridae, subfamily Betarhabdovirinae', *Archives of Virology,* 169 (10), 193.

Bejerman, N., Dietzgen, R. G., and Debat, H. (2021), 'Illuminating the Plant Rhabdovirus Landscape through Metatranscriptomics Data', *Viruses,* 13 (7), 1304.

Bonnamy, M., et al. (2024), 'The genome formula of a multipartite virus is regulated both at the individual segment and the segment group levels', *PLOS Pathogens,* 20 (1), e1011973.

Boone, S. A. and Gerba, C. P. (2007), 'Significance of fomites in the spread of respiratory and enteric viral disease', *Appl Environ Microbiol,* 73 (6), 1687-96.

Bujnicki, J. M. (2000), 'Phylogeny of the restriction endonuclease-like superfamily inferred from comparison of protein structures', *Journal of Molecular Evolution,* 50 (1), 39-44.

Charles, C. (2022), 'First detection of canine herpesvirus (CHV) in dog semen: a contribution to the study of venereal transmission risk by real-time PCR and complementary serological analysis', *Sciences du Vivant*, 4021592.

Chen, Y. M., et al. (2022), 'RNA viromes from terrestrial sites across China expand environmental viral diversity', *Nature Microbiology,* 7 (8), 1312.

Clausen, P. T., Aarestrup, F. M., and Lund, O. (2018), 'Rapid and precise alignment of raw reads against redundant databases with KMA', *BMC Bioinformatics,* 19 (1), 1-8.

Costa, V. A., et al. (2024), 'Host adaptive radiation is associated with rapid virus diversification and cross-species transmission in African cichlid fishes', *Current Biology,* 34 (6), 1247-57. e3.

Dall'Ara, M., et al. (2016), 'Ins and Outs of Multipartite Positive-Strand RNA Plant Viruses: Packaging versus Systemic Spread', *Viruses,* 8 (8), 228.

Edgar, R. C., et al. (2022), 'Petabase-scale sequence alignment catalyses viral discovery', *Nature,* 602 (7895), 142-47.

Eventoff, W., Rossmann, M. G., and Brändén, C.-I. (1975), 'The Evolution of Dehydrogenases and Kinase', *CRC Critical Reviews in Biochemistry,* 3 (2), 111-40.

Garg, S. G. and Hochberg, G. K. (2024), 'A general substitution matrix for structural phylogenetics', *bioRxiv*, 2024.09.19.613819.

Geoghegan, J. L. and Holmes, E. C. (2017), 'Predicting virus emergence amid evolutionary noise', *Open Biology,* 7 (10), 170189.

Harvey, E. and Holmes, E. C. (2022), 'Diversity and evolution of the animal virome', *Nature Reviews Microbiology,* 20 (6), 321-34.

Heinzinger, M., et al. (2024), 'Bilingual language model for protein sequence and structure', *NAR Genomics and Bioinformatics,* 6 (4).

Hou, X., et al. (2024), 'Using artificial intelligence to document the hidden RNA virosphere', *Cell,* 187 (24), 6929-42.e16.

Jiang, J., et al. (2022), 'Severe Acute Hepatitis Outbreaks Associated with a Novel Hepacivirus in Rhizomys pruinosus in Hainan, China', *Journal of Virology,* 96 (17), e00782-22.

Jumper, J., et al. (2021), 'Highly accurate protein structure prediction with AlphaFold', *Nature,* 596 (7873), 583-89.

Kelman, M., et al. (2020), 'A history of canine parvovirus in Australia: what can we learn?', *Australian Veterinary Journal,* 98 (10), 504-10.

Leebens-Mack, J. H., et al. (2019), 'One thousand plant transcriptomes and the phylogenomics of green plants', *Nature,* 574 (7780), 679-85.

Leinonen, R., et al. (2010), 'The sequence read archive', *Nucleic Acids Research,* 39, 19-21.

Mahar, J. E., et al. (2024), 'The diverse liver viromes of Australian geckos and skinks are dominated by hepaciviruses and picornaviruses and reflect host taxonomy and habitat', *Virus Evolution,* 10 (1).

Marcelino, V. R., et al. (2020), 'CCMetagen: comprehensive and accurate identification of eukaryotes and prokaryotes in metagenomic data', *Genome Biology,* 21 (1), 103.

Marshall, J., et al. (1984), 'Viruses and virus-like particles in the faeces of dogs with and without diarrhoea', *Australian Veterinary Journal,* 61 (2), 33-38.

Mauseth, J. D. (2014), *Botany: an introduction to plant biology* (Jones & Bartlett Publishers).

Mifsud, J. C. O., et al. (2023), 'Transcriptome mining extends the host range of the Flaviviridae to non-bilaterians', *Virus Evolution,* 9 (1).

Mirolo, M., et al. (2023), 'Highly pathogenic avian influenza A virus (HPAIV) H5N1 infection in two European grey seals (Halichoerus grypus) with encephalitis', *Emerging Microbes & Infections,* 12 (2), 2257810.

Moi, D., et al. (2023), 'Structural phylogenetics unravels the evolutionary diversification of communication systems in gram-positive bacteria and their viruses', *bioRxiv*.

Munday, J. S., et al. (2016), 'Genomic characterisation of canine papillomavirus type 17, a possible rare cause of canine oral squamous cell carcinoma', *Veterinary Microbiology,* 182, 135-40.

Münger, K., et al. (2004), 'Mechanisms of human papillomavirus-induced oncogenesis', *Journal of Virology,* 78 (21), 11451-60.

Nibert, M. L., et al. (2018), 'Evidence for contemporary plant mitoviruses', *Virology,* 518, 14-24.

Olendraite, I., Brown, K., and Firth, A. E. (2022), 'Identification of RNA virus-derived RdRp sequences in publicly available transcriptomic datasets', *bioRxiv*, 2022.10.18.512700.

Paraskevopoulou, S., et al. (2021), 'Viromics of extant insect orders unveil the evolution of the flavi-like superfamily', *Virus Evolution,* 7 (1).

Petrone, M. E., et al. (2024), 'A ~40-kb flavi-like virus does not encode a known error-correcting mechanism', *Proceedings of the National Academy of Sciences,* 121 (30), e2403805121.

Puente-Lelievre, C., et al. (2023), 'Tertiary-interaction characters enable fast, model-based structural phylogenetics beyond the twilight zone', *bioRxiv*.

Puente-Lelievre, C., et al. (2024), 'Tertiary-interaction characters enable fast, model-based structural phylogenetics beyond the twilight zone', *bioRxiv*, 2023.12.12.571181.

Schmitz, M., et al. (2012), 'Loss of gene function as a consequence of human papillomavirus DNA integration', *International Journal of Cancer,* 131 (5), E593-E602.

Shi, M., et al. (2018), 'The evolutionary history of vertebrate RNA viruses', *Nature,* 556 (7700), 197-202.

Shi, M., et al. (2016), 'Redefining the invertebrate RNA virosphere', *Nature,* 540 (7634), 539.

Valverde, R. A. and Sabanadzovic, S. (2009), 'A novel plant virus with unique properties infecting Japanese holly fern', *Journal of General Virology,* 90 (10), 2542-49.

Van Arkel, A., et al. (2019), 'The relationship between reported domestic canine parvovirus cases and wild canid distribution', *Heliyon,* 5 (9).

Varadi, M., et al. (2024), 'AlphaFold Protein Structure Database in 2024: providing structure coverage for over 214 million protein sequences', *Nucleic Acids Research,* 52 (D1), D368-D75.

Visser, M., et al. (2016), 'Next-generation sequencing for virus detection: covering all the bases', *Virology Journal,* 13 (1), 85.

Walker, P. J., et al. (2018), 'ICTV virus taxonomy profile: Rhabdoviridae', *Journal of General Virology,* 99 (4), 447-48.

Ward, M. P., Sivieng, K. Q.-L., and Brookes, V. (2021), 'The dingo-domestic dog interface: Implications for disease spread', *Australian zoologist,* 41 (3), 388-407.

Wyllie, S., Kelman, M., and Ward, M. (2016), 'Epidemiology and clinical presentation of canine distemper disease in dogs and ferrets in Australia, 2006–2014', *Australian Veterinary Journal,* 94 (7), 215-22.

Zayed, A. A., et al. (2022), 'Cryptic and abundant marine viruses at the evolutionary origins of Earth's RNA virome', *Science,* 376 (6589), 156-62.

Zhang, S., Shen, Z., and Li, H. (2023), 'Hepatic caecum of amphioxus and origin of vertebrate liver', *Acta Oceanologica Sinica,* 42 (11), 1-8.

Zhou, W., et al. (2018), 'Exosomes serve as novel modes of tick-borne flavivirus transmission from arthropod to human cells and facilitates dissemination of viral RNA and proteins to the vertebrate neuronal cells', *PLoS pathogens,* 14 (1), e1006764.